UNIVERSITY OF THE AEGEAN

PHD THESIS

Bayesian Methods & Estimation of Nonlinear Dynamical Systems

Author:
Konstantinos Kaloudis

Supervisor:
Spyridon J. Hatjispyros

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

in the

Department of Statistics and Actuarial-Financial Mathematics

Karlovassi, Greece
April 2019

Bayesian Methods & Estimation of Nonlinear Dynamical Systems

Konstantinos Kaloudis

PHD THESIS

COMMITTEE

Spyridon J. Hatjispyros (SUPERVISOR–Participant of the 3 and 7 member committee)
Assoc. Professor, Dept. of Statistics and Actuarial-Financial Mathematics,
University of the Aegean.

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

Athanasios N. Yannacopoulos (Participant of the 3 and 7 member committee)
Professor, Dept. of Statistics,
Athens University of Economics and Business.

Nikolaos Halidias (Participant of the 7 member committee)
Assoc. Professor, Dept. of Statistics and Actuarial-Financial Mathematics,
University of the Aegean.

Theodoros Karakasidis (Participant of the 7 member committee)
Professor, Dept. of Civil Engineering,
University of Thessaly.

John Tsimikas (Participant of the 7 member committee)
Assoc. Professor, Dept. of Statistics and Actuarial-Financial Mathematics,
University of the Aegean.

Stavros Vakeroudis (Participant of the 7 member committee)
Assistant Professor, Dept. of Statistics and Actuarial-Financial Mathematics,
University of the Aegean.

April 2019

I, KONSTANTINOS KALOUDIS, 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 TeX LIVE DISTRIBUTION. WRITING WAS MADE USING THE PROGRAM "TeXSTUDIO" ON THE LINUX FEDORA 26 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. Christos Merkatas, Konstantinos Kaloudis, and Spyridon J Hatjispyros. A Bayesian nonparametric approach to reconstruction and prediction of random dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(6):063116, 2017.

2. Konstantinos Kaloudis, Spyridon J. Hatjispyros. A Bayesian nonparametric approach to dynamical noise reduction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):063110, 2018.

3. Spyridon J. Hatjispyros, Konstantinos Kaloudis. On the stochastic approximation of the global stable manifold. *Submitted for publication*, 2019.

This thesis has been funded by the "YPATIA" Scholarship Program for Doctoral Studies, of the University of the Aegean.

Acknowledgements

I want to express my warmest thanks to my supervisor Spyridon Hatjispyros, for his motivation, guidance and support. Spyros has not only been my supervisor, but also a mentor and a friend. Thank you for everything.

I express my gratitude to the University of the Aegean, for providing me with the "YPATIA" Scholarship Program for Doctoral Studies.

I would also like to thank all the members of my committee, for their valuable comments on this work.

It is my pleasure to thank Ass. Prof. Eleftherios Tachtsis and Ass. Prof. Stelios Xanthopoulos for their support and all the discussions we had at the "foyer" of the Provatari building.

I am also grateful to my (academic) brother Christos M. and my beloved friends Alice S., Artemis B., Antonis M., Dimitris P., Dimitris Gk., Elena K., Kostas T., Mihalis S., Panos D., Pavlos D., Tassos Gk. for their encouragement and for always being there for me.

Special thanks to Sofia for her patience, love and support.

Finally, words are not enough to express my gratitude to my family, for their constant and unconditional support. This thesis is dedicated to them.

Abstract

This thesis is concerned with the interplay between Bayesian Statistics and Nonlinear Dynamical Systems. Specifically, the main goal of the thesis is the development of new Markov Chain Monte Carlo (MCMC) methods that have applications in the general field of nonlinear dynamics. The motivation for this approach is the decomposition of the modeling procedure into two interacting parts: the deterministic part and the stochastic noise process. Using this kind of modeling, we are able to capture a wide variety of phenomena, utilizing the complex behavior of the nonlinear part and the new characteristics emerging from the interaction with the noise process. The proposed methods are nonparametric, based on the Geometric Stick-Breaking process as a prior over the space of probability measures. An important aspect of this work is the relaxation of a very common assumption in the literature: the normality of the noise distribution.

In Chapter 1 we present the basic notions and results of the Bayesian parametric and nonparametric framework, that are essential for the understanding of the methods developed in this thesis. We discuss the concept of exchangeability and the Theorem of de Finetti. Following, we introduce the notion of a Bayesian nonparametric model and define the most popular Bayesian nonparametric prior, the Dirichlet Process. We continue with the definition of the main Bayesian nonparametric prior for the scope of this thesis, the Geometric Stick-Breaking process and discuss the properties of Bayesian nonparametric Mixture models. Finally, the main MCMC sampling methods are presented and discussed.

In Chapter 2, which is also introductory, we give the definitions and basic properties of dynamical systems, both in the deterministic and the stochastic case. The concept of chaos is rigorously analyzed, along with the properties of chaotic systems, homoclinic tangencies and invariant manifolds. We continue with the presentation of the most popular in the literature noise-induced effects, as well as with the effects of dynamical noise on non-hyperbolic nonlinear maps. We conclude with an explanation of the inefficiency of parametric reconstruction models under a non-Gaussian noise process.

In Chapter 3 we propose a Bayesian nonparametric framework, the Geometric Stick Breaking Reconstruction (GSBR) model, suitable for the full **reconstruction and prediction** of dynamically-noisy corrupted time series, when the additive noise may exhibit significant departures from normality. In particular, any 0-mean symmetric density can be recovered, even in cases where

the size of the observed time series is small, hence statistical inference for the system is improved and reliable. With the proposed model, we have also shown that the associated quasi-invariant measure of the underlying random dynamical system, appears naturally as a prediction barrier, similarly as the invariant measure appears as the prediction barrier in the deterministic case. We have used the Geometric Stick Breaking process as a prior over the unknown noise density, showing that it yields almost indistinguishable results from the more commonly used, but computationally more expensive, Dirichlet Process prior. The GSBR model is also generalized in order to include arbitrary number of finite lag terms and finally extended in the multivariate case, where the noise process is modeled as an infinite mixture of multivariate Normal kernels with unknown precision matrices, using Wishart distributions.

In Chapter 4, the thesis proceeds with the proposal of a new Bayesian nonparametric method, the Dynamic Noise Reduction Replicator (DNRR) model, suitable for **noise reduction** over a given chaotic time series, subjected to the effects of (the perhaps non-Gaussian) additive dynamical noise. The DNRR model provides a highly accurate reconstruction of the dynamical equations, while in parallel it replicates the dynamics under reduced noise level dynamical perturbations. The advantages of this method are that the estimated noise-reduced orbit has approximately the same estimated deterministic part, while it evolves in a neighborhood of the original noisy orbit. The two orbits remain close even in the regions of the noise-induced prolongations of the attractor, or in cases of perturbed multistable maps exhibiting noise-induced jumps. We were also able to relate the regions of primary homoclinic tangencies of the associated deterministic system, with regions of persistent high determinism deviations.

Further, in relating the random dynamical systems with their associated deterministic parts, in Chapter 5 we present an extention of the GSBR sampler, in order to provide a MCMC-based **stochastic approximation of the global stable manifold**. Specifically, we have introduced the Backward GSBR (BGSBR) model, in order to estimate past unobserved observations, namely performing prediction in reversed time. We have emphasized on the support of the posterior marginals of the unknown initial conditions, demonstrating that the union of the supports contain the associated deterministic stable manifold of the attractor. The BGSBR sampler can be applied multiple times over proper subsets of the noisy observations, each time generating posterior samples for the various initial conditions. Then the global stable manifold of the associated deterministic map can be stochastically approximated as the union of the supports of the posterior marginal distributions. The method is parsimonious and efficient both in invertible and non-invertible maps.

Finally, in Chapter 6 we present the main conclusions and address some relevant topics for further research, based on the results obtained during this thesis. We discuss the generalization of the proposed methods in continuous-time systems, as well as the possibility of dropping the assumption of a known functional form of the underlying deterministic part. Furthermore, we discuss the development of a RJMCMC-based extension of the GSBR model, in order to impute a -possibly unknown- finite number of missing observations of an observed time-series. Finally, we discuss the problem of application of Bayesian nonparametric methods in the context of Coupled Map Lattices.

We conclude with the bibliography and 3 Appendices. In Appendix A we present analytical results regarding the MCMC sampling schemes appearing in the Chapters 3-5. In Appendix B we analyze the dynamical behavior of the polynomial maps appearing in Chapters 3-4. Finally, Appendix C provides information regarding the development of the algorithms using the Julia language, as well as a URL in order to have access to the programs.

Περίληψη

Η παρούσα διατριβή αφορά τη διάδραση μεταξύ Μπεϋζιανής στατιστικής και μη γραμμικών δυναμικών συστημάτων. Ειδικότερα, ο βασικός στόχος της διατριβής είναι η ανάπτυξη νέων μεθόδων Markov Chain Monte Carlo (MCMC) με εφαρμογές στο ευρύτερο πεδίο της μη γραμμικής δυναμικής. Το κίνητρο για την ανάπτυξη τέτοιων μεθόδων, αφορά την διάκριση της διαδικασίας μοντελοποίησης σε δύο βασικά διαδραστικά μέρη: το αιτιοκρατικό (ντετερμινιστικό) μέρος και τη στοχαστική διαδικασία θορύβου. Μέσω μιας τέτοιου είδους μοντελοποίησης, επιτυγχάνεται η σύλληψη μιας ευρείας συλλογής φαινομένων, αξιοποιώντας την πολυπλοκότητα της δυναμικής συμπεριφοράς λόγω του μη γραμμικού μέρους και τα νέα χαρακτηριστικά που αναδεικνύονται λόγω της εμπλοκής των στοχαστικών διαταραχών. Οι προτεινόμενες στατιστικές μέθοδοι είναι μη παραμετρικές και βασίζονται στη χρήση τυχαίων μέτρων πιθανότητας με γεωμετρικά βάρη (Geometric stick breaking process (GSB)) ως εκ των προτέρων κατανομές στο χώρο των μέτρων πιθανότητας. Μια σημαντική πτυχή των προτεινόμενων μεθόδων είναι η επίτευξη της χαλάρωσης μιας πολύ συχνής υπόθεσης στη βιβλιογραφία: της κανονικότητας της διαδικασίας θορύβου.

Στο Κεφάλαιο 1 παρουσιάζονται οι βασικές έννοιες της Μπεϋζιανής στατιστικής, στην παραμετρική και στη μη παραμετρική περίπτωση, καθώς και κάποια βασικά αποτελέσματα απαραίτητα για την κατανόηση των αποτελεσμάτων της παρούσας διατριβής. Ειδικότερα, παρουσιάζεται η έννοια της ανταλλαξιμότητας (exchangeability) και το Θεώρημα ανταλλαξιμότητας του De Finetti. Στη συνέχεια παρουσιάζεται η πιο δημοφιλής Μπεϋζιανή μη παραμετρική εκ των προτέρων κατανομή (prior), το τυχαίο μέτρο Dirichlet, καθώς και οι πιο δημοφιλείς αναπαραστάσεις του. Παρουσιάζονται, επίσης, το βασικό τυχαίο μέτρο για την ανάπτυξη των μεθόδων της διατριβής, το τυχαίο μέτρο με γεωμετρικά βάρη GSB και τα Μπεϋζιανά μη παραμετρικά μοντέλα μίξεων. Τέλος, αναλύονται βασικά χαρακτηριστικά μεθόδων MCMC, οι αλγόριθμοι Metropolis, Gibbs, ο αλγόριθμος χρήσης βοηθητικών μεταβλητών (slice sampler) και MCMC μέθοδοι δειγματοληψίας από μη παραμετρικά μοντέλα μίξεων με τυχαία μέτρα Dirichlet και GSB.

Στο -επίσης εισαγωγικό- Κεφάλαιο 2 παρατίθενται οι βασικές έννοιες και ορισμοί σχετικά με το πεδίο των δυναμικών συστημάτων, στην ντετερμινιστική και την στοχαστική περίπτωση. Γίνεται αναλυτική επισκόπηση της έννοιας του χάους σε δυναμικά συστήματα, καθώς και των αναλλοίωτων πολλαπλοτήτων (invariant manifolds) και παρουσιάζεται η έννοια της ομοκλινικής εφαπτομενικότητας (homoclinic tangency). Στη συνέχεια, παρουσιάζονται κάποια από τα πιο διαδεδομένα στη βιβλιογραφία φαινόμενα προκαλούμενα από το θόρυβο (noise-induced effects), καθώς και η επίδραση του δυναμικού θορύβου σε μη υπερβολικές (non-hyperbolic) απεικονίσεις. Το Κεφάλαιο ολοκληρώνεται με την επεξήγηση της μη αποτελεσματικότητας των παραμετρικών μεθόδων αναδόμησης στις περιπτώσεις μη-κανονικού (non-Gaussian) δυναμικού θορύβου.

Στο Κεφάλαιο 3, κατασκευάζουμε ένα μη παραμετρικό Μπεϋζιανό μοντέλο κατάλληλο για αναδόμηση των δυναμικών εξισώσεων και πρόγνωση μελλοντικών τιμών από παρατηρηθείσες χρονοσειρές μολυσμένες με προσθετικό δυναμικό θόρυβο: το μοντέλο geometric stick-breaking reconstruction (GSBR). Το GSBR μοντέλο βασίζεται στο τυχαίο μέτρο με γεωμετρικά βάρη (GSB), ενώ γίνεται επίσης παρουσίαση του αντίστοιχου μοντέλου Dirichlet process reconstruction (DPR) βασισμένου στο τυχαίο μέτρο DP, καθώς και η μεταξύ τους σύγκριση. Η βασική συνεισφορά της προτεινώμενης μεθοδολογίας αφορά τη χαλάρωση της συνήθους υπόθεσης της κανονικότητας της διαδικασίας θορύβου, καθώς μπορεί να εκτιμηθεί οποιαδήποτε συμμετρική γύρω από το μηδέν πυκνότητα, ακόμα και σε περιπτώσεις όπου το διαθέσιμο δείγμα είναι μικρό. Δείχνουμε επίσης, ότι κάνοντας χρήση των δύο μοντέλων, επιτυγχάνεται η εκτίμηση του ψευδο-αναλλοίωτου μέτρου του υποκείμενου στοχαστικού δυναμικού συστήματος, το οποίο εμφανίζεται ως φράγμα προβλεπτικής δυνατότητας μέσω των περιθώριων κατανομών των μελλοντικών παρατηρήσεων. Τα παραγόμενα αποτελέσματα μέσω του μοντέλου GSBR είναι ποιοτικά ίδια με εκείνα του αντίστοιχου DPR, απαιτώντας σημαντικά χαμηλότερο μέσο χρόνο εκτέλεσης. Τέλος, η μεθοδολογία επεκτείνεται ώστε να γίνει εφικτή η μοντελοποίηση χρησιμοποιώντας αυθαίρετο πεπερασμένο πλήθος όρων χρονικών υστερήσεων (lags), καθώς και στην πολυδιάστατη περίπτωση μέσω της άπειρης μίξης πολυδιάστατων κανονικών πυρήνων με άγνωστους πίνακες αποκρίσεων, χρησιμοποιώντας ως μέτρο μίξης το τυχαίο μέτρο GSB και μέτρο βάσης (base measure) μια κατανομή Wishart.

Η διατριβή συνεχίζει στο Κεφάλαιο 4, προτείνοντας μια μη παραμετρική Μπεϋζιανή μεθοδολογία βασιζόμενη επίσης στο τυχαίο μέτρο GSB, με σκοπό τη μείωση δυναμικού θορύβου σε διαθέσιμα δεδομένα μη γραμμικών χρονοσειρών με προσθετικό θορυβο. Το μοντέλο Dynamic Noise Reduction Replicator (DNRR) επιτυγχάνει μεγάλη ακρίβεια στην αναδόμηση των δυναμικών εξισώσεων, ώστε να αναπαράγει την υποκείμενη δυναμική σε περιβάλλον ασθενέστερου δυναμικού θορύβου. Τα πλεονεκτήματα της μεθόδου αφορούν αφενός το κοινό ντετερμινιστικό μέρος μεταξύ της διαθέσιμης και της αποθορυβοποιημένης τροχιάς και αφετέρου τη δυνατότητα της εκ των προτέρων ρύθμισης της μεταξύ τους εγγύτητας ακόμα και σε περιοχές επιμηκύνσεων του ελκυστή λόγω θορύβου (noise-induced prolongations), ή σε περιπτώσεις αλμάτων λόγω θορύβου (noise-induced jumps) σε πολλαπλά ευσταθείς (multistable) απεικονίσεις. Τέλος, μέσω της εφαρμογής του DNRR είναι δυνατή η σύνδεση των περιοχών υψηλών αποκλίσεων από τον ντετερμινισμό με τις περιοχές των πρωταρχικών ομοκλινικών εφαπτομενικοτήτων του υποκείμενου ντετερμινιστικού συστήματος.

Ακολούθως, συσχετίζοντας τα στοχαστικά δυναμικά συστήματα με τα αντίστοιχα ντετερμινιστικά τους μέρη, στο Κεφάλαιο 5 παρουσιάζεται μία επέκταση του μοντέλου GSBR, με σκοπό τη στοχαστική προσέγγιση της ολικής ευσταθούς πολλαπλότητας (global stable manifold), με χρήση μεθόδου MCMC. Ειδικότερα, γίνεται παρουσίαση του οπισθοδρομικού (backward) GSBR μοντέλου BGSBR, μέσω του οποίου επιτυγχάνεται πρόβλεψη σε αντεστραμμένο χρόνο. Ιδιαίτερη έμφαση δίνεται στο στήριγμα της περιθώριας κατανομής του διανύσματος των αρχικών συνθηκών. Συγκεκριμένα, με κατάλληλες πολλαπλές εφαρμογές του BGSBR χρησιμοποιώντας υποσύνολα των διαθέσιμων δεδομένων, δείχνουμε ότι η ένωση των στηριγμάτων των κατανομών αυτών για τις διάφορες αρχικές συνθήκες παρέχουν μια στοχαστική προσέγγιση της ευσταθούς πολλαπλότητας του υποκείμενου ντετερμινιστικού συστήματος. Η μεθοδολογία είναι εφαρμόσιμη τόσο σε αντιστρέψιμες όσο και σε μη αντιστρέψιμες απεικονίσεις.

Στο Κεφάλαιο 6 γίνεται σύνοψη των αποτελεσμάτων των προηγούμενων Κεφαλαίων και αναφορά σε θέματα για μελλοντική έρευνα, τα οποία προέκυψαν κατά τη διάρκεια εκπόνησης της παρούσας Διατριβής. Προτείνεται η γενίκευση των προτεινόμενων μεθόδων σε στοχαστικά δυναμικά συστήματα συνεχούς χρόνου, καθώς και η δυνατότητα απόρριψης της υπόθεσης γνωστής συναρτησιακής μορφής για το ντετερμινιστικό μέρος της διαδικασίας, κάνοντας χρήση της Gaussian διαδικασίας. Επιπλέον, αναλύεται η δυνατότητα κατασκευής μιας επέκτασης του GSBR μοντέλου με χρήση MCMC μεθόδων αναστρέψιμου άλματος (Reversible Jump MCMC). Τέλος, προτείνεται η κατασκευή μοντέλων βασιζόμενων στη Μπεϋζιανή μη παραμετρική στατιστική, προσανατολιζόμενων στην επίλυση προβλημάτων του πεδίου των πλεγμάτων συζευγμένων απεικονίσεων (coupled map lattices).

Η Διατριβή ολοκληρώνεται με την παράθεση της βιβλιογραφίας και τριών Παραρτημάτων. Στο Παράρτημα Α παρουσιάζονται αναλυτικά αποτελέσματα, σχετικά με τα MCMC δειγματοληπτικά σχήματα που εμφανίζονται στα Κεφάλαια 3-5. Στο Παράρτημα Β αναλύεται η δυναμική συμπεριφορά των πολυωνυμικών απεικονίσεων οι οποίες χρησιμοποιούνται στα Κεφάλαια 3-4. Τέλος, το Παράρτημα C περιέχει πληροφορίες σχετικά με την ανάπτυξη των αλγορίθμων με χρήση της γλώσσας προγραμματισμού Julia, καθώς και ένας σύνδεσμος URL μέσω του οποίου μπορεί να γίνει λήψη των προγραμμάτων τα οποία χρησιμοποιήθηκαν για την απόκτηση των παρουσιαζόμενων αποτελεσμάτων.

Contents

List of Figures

List of Tables

List of Abbreviations

Term	Description
ApEn	Approximate Entropy
BGSBR	Backward Geometric Stick–Breaking Reconstruction
BNP	Bayesian nonparametric
CRP	Chinese Restaurant Process
DNRR	Dynamic Noise Reduction Replicator
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
DPR	Dirichlet Process Reconstruction
DS	Dynamical System
ForeCa	Forecastable Component Analysis
GHM	Generalized Hénon Map
GP	Gaussian Process
GSB	Geometric stick–breaking
GSBM	Geometric stick–breaking mixture
HT	Homoclinic tangency
IRP	Informative Reconstruction–Prediction
KDE	Kernel Density Estimator
MH	Metropolis–Hastings
MAP	Maximum a–posteriori
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MET	Mean Execution Time
MPD	Marginal Posterior Density
MSE	Mean–Squared Error
NRP	Noninformative Reconstruction–Prediction
PAR	Polynomial Autoregressive Process
PARE	Percentage Absolute Relative Error
rDPR	randomized Dirichlet Process Reconstruction model
RDS	Random Dynamical System
SDIC	Sensitive Dependence on Initial Conditions
SM	Sampling Mean
WOSA	Weighted Overlapping Segment Averaging

To my parents Sakis & Roza
and to my grandmother, Rini.

Chapter 1

The Bayesian Framework

The concept of *uncertainty* is crucial in the process of mathematical modeling. We confront uncertainty in a plethora of ways, either due to inherent randomness in the system evolution, or due to a large number of active degrees of freedom. It is also known since the beginning of the development of Chaos Theory, that even in deterministic systems -if the dynamical behavior is complex enough- we may not be able to avoid uncertainty in terms of predicting the future outcomes. In order to deal with uncertainty, we need to perform statistical inference and extract information from the available data, using a proper probability model.

One of the main pillars of modern statistical inference is the Bayesian Theory, originating in the early works of Thomas Bayes and Pierre-Simone Laplace and becoming more popular through the advance of computational resources and MCMC methods. The essential feature of the Bayesian approach is the expression of uncertainty in terms of probability statements, thus the modeling of the parameters of interest (or all unknown quantities) as random variables following a *prior* distribution that reflects our knowledge prior to observing the data. Then, conditionally on the data, inference is based on the -updated- *posterior* distribution.

This thesis is devoted to the development of a new approach to nonlinear dynamic modeling, utilizing a Bayesian nonparametric framework. In this chapter we will review some of the basic definitions and results of Bayesian statistics, that are required to proceed to our main topic. We will continue with a formal introduction to Bayesian inference, both in a parametric and a nonparametric context. First, we will introduce the most widely used nonparametric prior, namely the Dirichlet Process, and discuss its properties and representations. Next, we will introduce the nonparametric prior that we will use for the construction of the models proposed in this thesis, the Geometric Stick-Breaking process, as well as the corresponding Bayesian nonparametric mixture models. We will end this Chapter with a description of the most widely used MCMC methods and their extensions suitable for the component update of random measures. For a thorough analysis of Bayesian statistics and MCMC methods the reader is referred to [Rob07], [RC04], or [BS09]. Regarding the theory of Bayesian nonparametrics, extensive works include [MQJH15], [HHMW10] and [Pha15].

1.1 Construction of the statistical model

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be an observed sample. We assume that $\boldsymbol{x}$ consists of realizations x_i of random variables X_i defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values on a complete separable metric space $(\mathbf{X}, d)$, with $\mathcal{X}$ the associated Borel σ-algebra of subsets of $\mathbf{X}$. We denote with $\mathcal{P}_{\mathbf{X}}$ the space of all probability measures over $\mathbf{X}$ and with $\mathcal{P}_{\mathbf{X}^*}$ the corresponding space of densities.

In order to define a parametric Bayesian model, we need to assume that $\vartheta \in \Theta$ is a random vector of unknown quantities $\vartheta = (\theta_1, \ldots, \theta_d)$ and $(X_1, \ldots, X_n \mid \vartheta)$ are generated independently and identically distributed (iid) from a probability measure $P_\vartheta \in \mathcal{P}_{\mathbf{X}}$. In this case the parameter space Θ is finite dimensional and we assign a prior measure $\Pi \in \mathcal{P}_{\mathbf{X}}$, that induces a prior distribution over the unknown parameters ϑ. The prior distribution incorporates our beliefs about the parameters of interest before $\boldsymbol{x}$ is observed.

The motivation behind the Bayesian approach and the assignment of the prior distribution over the random quantities of interest, is based on the concept of exchangeability. Instead of making the stronger assumption that the observations $\boldsymbol{x}$ are independent, we assume that they are exchangeable. In order to stress the importance of this distinction, suppose that the joint distribution of $\boldsymbol{x}$ is $p(x_1, \ldots, x_n)$ and we aim to predict the next future unobserved observation. Under the assumption of independence we have that $p(x_{n+1} | x_1, \ldots, x_n) = p(x_{n+1})$ and we treat individually every new data point, without taking into account the past ones.

Definition 1.1. *A sequence of random variables $X^{(n)} = (X_1, X_2, \ldots)$ is said to be exchangeable under probability measure P if the joint probability*

$$(X_1, \ldots, X_n) \overset{\mathrm{d}}{=} \left(X_{\tau(1)}, \ldots, X_{\tau(n)}\right)$$

for every permutation τ defined on $\{1, \ldots, n\}$. A sequence $(X_i)_{i \geq 1}$ is infinitely exchangeable if every finite subsequence is exchangeable.

Intuitively, there is no information contained in the specific order under which we obtain exchangeable observations, as their joint distribution remains invariant. Exchangeability is a weaker notion than independence, so we can construct sequences of exchangeable random variables that are not independent [HS76]. The famous *de Finetti's Representation Theorem* states that an infinite sequence of random variables is exchangeable if and only if it is conditionally iid given a probability measure, known as the de Finetti's measure. It was first proven for sequences of binary random variables [DF37] and later on extended for arbitrary real-valued variables [HS55]. For generalizations of the notion of exchangeability see [DF80].

Theorem 1.1 (de Finetti's Representation Theorem)**.** *A binary sequence $(X_i)_{i \geq 1}, X_i \in \{0, 1\}$ is exchangeable if and only if there exists a distribution function $F \in (0, 1)$ such that $\forall n \in \mathbb{N}$*

$$\mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n\} = \int_0^1 \theta^{t_n} (1 - \theta)^{n - t_n} F(\mathrm{d}\theta), \quad t_n := \sum_{i=1}^{n} X_i,$$

where F is the distribution function of the limiting empirical frequency, for which:

$$F(y) = \mathbb{P}\left\{Y \leq y\right\}, \quad Y = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$$

and $\mathbb{P}\left\{X_1 = x_1, \ldots, X_n = x_n \mid Y = \theta\right\} = \theta^{t_n}(1-\theta)^{n-t_n}$

Essentially, the above Theorem states that binary sequences of random variables can be expressed as mixtures of independent Bernoulli sequences with probability of success θ, if θ is distributed according to F, where F is defined as the limiting empirical frequency of the data. So, in a Bayesian context, the data generating mechanism consists of the two stages

$$\begin{aligned} X_i \mid \vartheta &\overset{\text{iid}}{\sim} f\left(\cdot \mid \vartheta\right), \quad i = 1, \ldots, n \\ \vartheta &\sim \Pi_\Theta, \end{aligned}$$

and the objective is the determination of the posterior distribution, using Bayes' theorem

$$\Pi\left(\mathrm{d}\vartheta \mid \boldsymbol{x}\right) = \frac{\prod_{i=1}^{n} f\left(x_i \mid \vartheta\right) \Pi_\Theta\left(\mathrm{d}\vartheta\right)}{\int_{\vartheta\in\Theta} \prod_{i=1}^{n} f\left(x_i \mid \vartheta\right) \Pi_\Theta\left(\mathrm{d}\vartheta\right)} \propto \mathcal{L}\left(\vartheta; \boldsymbol{x}\right) \Pi_\Theta\left(\mathrm{d}\vartheta\right).$$

This means that the posterior distribution $\Pi\left(\mathrm{d}\vartheta \mid \boldsymbol{x}\right)$, i.e. the conditional distribution of ϑ given the data $\boldsymbol{x}$, is the update of our prior beliefs $\Pi_\Theta\left(\mathrm{d}\vartheta\right)$, via the observed likelihood $\mathcal{L}\left(\vartheta; \boldsymbol{x}\right)$. Moreover, using Bayes' rule, we can derive the posterior predictive distribution of the next T future unobserved observations as the mixture of the likelihood of the future T observations, with the posterior measure serving as the mixing measure:

$$f\left(x_{n+1}, \ldots, x_{n+T} \mid x_1, \ldots, x_n\right) = \int_{\vartheta\in\Theta} \prod_{i=1}^{T} f\left(x_{n+i} \mid \vartheta\right) \Pi_\Theta\left(\mathrm{d}\vartheta \mid x_1, \ldots, x_n\right).$$

In the parametric case, the density $f(\boldsymbol{x} \mid \vartheta)$ corresponds to a probability measure

$$P \in \left\{P_\vartheta \mid \vartheta \in \Theta\right\} \subseteq \mathcal{P}_{\mathbf{X}},$$

where the parameter space Θ is finite dimensional. For example, suppose that we wish to infer the mean and variance of the Normal distribution that has generated the observed data, that is:

$$\begin{aligned} X_i \mid \mu, \sigma^2 &\overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right) \quad i = 1, \ldots, n \\ \mu, \sigma^2 &\sim \Pi_\Theta, \end{aligned}$$

with $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. After a finite amount of observations, we are still uncertain about the true value of the parameters $\vartheta = (\mu, \sigma^2)$ and this uncertainty is expressed by the posterior distribution.

There are cases, though, that this finite parametrization of the density of the data-generating process can be considered as a restriction. For example, we may have that $\mathbb{Q} \in \mathcal{P}_{\mathbf{X}}$, such that

$X_i \,|\, \mathbb{Q} \overset{\text{iid}}{\sim} \mathbb{Q}, 1 \leq i \leq n$ and we want to construct a probability model for $\mathbb{Q}$, dropping the (perhaps strong) assumption that it belongs to some parametric family of probability measures parametrized by ϑ. In this case, we aim to infer the probability measure $\mathbb{Q}$ itself. The probability measure $\mathbb{Q}$ is an element of the infinite dimensional space $\mathcal{P}_{\mathbf{X}}$ consisting of all the probability measures over $\mathbf{X}$, i.e. $\Theta = \mathcal{P}_{\mathbf{X}}$. For example, consider the problem of density estimation, where we want to infer the unknown density f, responsible for the collection of data $(x_1, \ldots, x_n)$, that is $x_i \,|\, f \overset{\text{iid}}{\sim} f, 1 \leq i \leq n$.

To this end, a *Bayesian nonparametric (BNP) model* is defined as a Bayesian model with infinite-dimensional parameter space. If we define a BNP prior over the space of probability measures, then we obtain a *random measure*, i.e. a measure drawn from some distributions over measures. Formally, we have the following definition [Cin13]:

Definition 1.2. *A random measure on $(\mathbf{X}, \mathcal{X})$ is a transition kernel from $(\Omega, \mathcal{F})$ into $(\mathbf{X}, \mathcal{X})$, i.e. a mapping $M : \Omega \times \mathcal{X} \to \mathbb{R}^+$, such that*

1. *$M(\cdot, A)$ is a random variable $\forall A \in \mathcal{X}$*
2. *$M(\omega, \cdot)$ is a measure $\forall \omega \in \Omega$.*

The meaning of the above definition is that M assigns for every possible outcome $\omega \in \Omega$, a proper measure M_ω and is essentially a stochastic process indexed by the elements (sets) of the $\sigma-$algebra $\mathcal{X}$.

In a nonparametric context, the use of a prior is motivated by the generalization of de Finetti's Theorem [HS55]. In this case the de Finetti measure is identified as a nonparametric prior distribution over an exchangeable sequence of observations.

Theorem 1.2 (Hewitt and Savage generalization of de Finetti's Representation Theorem)**.** *Let $(X_i)_{i \geq 1}$ be an exchangeable sequence of random variables over $\mathbf{X}$. Then there exists a probability measure $\Pi_F \in \mathcal{P}_{\mathbf{X}}$ such that $\forall n \in \mathbb{N}$*

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \int_{\mathbb{Q} \in \mathcal{P}_{\mathbf{X}}} \prod_{i=1}^{n} \mathbb{Q}(A_i) \Pi_F(\mathrm{d}\mathbb{Q}).$$

The random empirical measure F is distributed like Π_F, i.e. $F \sim \Pi_F$, where $\forall A \in \mathcal{X}$:

$$F(A) = \lim_{n \to \infty} F_n(A) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A).$$

$\mathbb{Q}$ is the distribution obtained by conditioning w.r.t. F , that is

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n \,|\, F = \mathbb{Q}\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \in A_i \,|\, F = \mathbb{Q}\} = \prod_{i=1}^{n} \mathbb{Q}(A_i).$$

Therefore, under the assumption of exchangeability, we can write the Bayesian model in a two-stage hierarchical form

$$X_i \,|\, \mathbb{Q} \overset{\text{iid}}{\sim} \mathbb{Q}$$
$$\mathbb{Q} \sim \Pi_F$$

and obtain the posterior distribution by applying Bayes' theorem

$$\Pi_F(\mathrm{d}\mathbb{Q}) = \frac{\prod_{i=1}^n \mathbb{Q}(x_i)\,\Pi_F(\mathrm{d}\mathbb{Q})}{\int_{\mathbb{Q}\in\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^n \mathbb{Q}(x_i)\,\Pi_F(\mathrm{d}\mathbb{Q})} \propto \mathcal{L}(\mathbb{Q};\boldsymbol{x})\,\Pi_F(\mathrm{d}\mathbb{Q}).$$

The nonparametric posterior predictive for T unobserved observations will be the mixture of the (nonparametric) likelihood for T observations with mixing measure, the (nonparametric) posterior measure:

$$\mathbb{Q}(x_{n+1},\ldots,x_{n+T} \,|\, x_1,\ldots,x_n) = \int_{\mathbb{Q}\in\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{T} \mathbb{Q}(x_{n+i} \,|\, \vartheta)\,\Pi_F(\mathrm{d}\mathbb{Q} \,|\, x_1,\ldots,x_n).$$

BNP models are in general more complex, compared to parametric models. The main reasons for their wide applicability in modern statistics are their flexibility and robustness. For example, in the process of modeling complex phenomena, there is often a challenging task: How complex should the model be, in order to be able to capture the desirable characteristics? A promising approach to address such issues is the development of BNP methods, as they allow for the model structure to be highly flexible, adapting its complexity as needed according to the observed data. Moreover, the use of parametric models under the false assumption regarding the parametric family of the data likelihood, may result in erroneous inference conclusions. An indicative inadequacy of a parametric modeling will be presented at the end of Chapter 2.

1.2 The Dirichlet Process

One of the most popular Bayesian nonparametric priors is the Dirichlet Process (DP), originally introduced by Ferguson [Fer73] as a prior over the space of probability measures. Before we give the definition of a DP, we will recall the definition of the Dirichlet distribution, which is considered a multivariate generalization of the Beta distribution, and is typically used as the conjugate prior [1] of the multinomial distribution.

Definition 1.3 (Dirichlet Distribution [NTT11]). *Let* $Z_i \overset{\text{ind}}{\sim} \mathcal{G}(a_i, 1), 1 \le i \le n$ *and* $Z := \sum_{i=1}^n Z_i$. *The random vector* $\mathbf{W} = (W_1,\ldots,W_n)$, *with* $W_i := \frac{Z_i}{Z}, 1 \le i \le n$ *has the Dirichlet distribution with parameters* $\alpha = (\alpha_1,\ldots,\alpha_n)$ *and probability density function*

$$\mathcal{D}(\mathbf{W} \,|\, \alpha) = \frac{\Gamma\left(\sum_{j=1}^n \alpha_j\right)}{\prod_{j=1}^n \Gamma(a_j)} \prod_{i=1}^n w_i^{\alpha_i - 1} \mathcal{I}(\mathbf{W} \in \Delta^n),$$

[1] For a given likelihood function, a prior is called conjugate, if it belongs to the same parametric family as the posterior distribution.

supported over the $(n-1)-$*dimensional probability simplex*

$$\Delta^n := \left\{ (w_1, \ldots, w_n) \in \mathbb{R}_n^+ : \sum_{i=1}^n w_i = 1 \right\}.$$

Now we can give the definition of a DP random measure, which can be considered as an infinite-dimensional generalization of the Dirichlet distribution. The definition is due to Ferguson [Fer73] who verified Kolmogorov's consistency conditions [KBR18] in order to formally prove the existence of this process.

Definition 1.4 (Dirichlet Process [Fer73])**.** *Let* $c > 0$ *and* $P_0 \in \mathcal{P}_{\mathbf{X}}$*. We say that a random measure* $\mathbb{P}$ *is distributed according to a Dirichlet Process* $\mathbb{P} \sim \mathcal{DP}(c, P_0)$ *with concentration parameter* c *and base measure* P_0 *if and only if for every finite partition* $\{A_1, \ldots, A_n\}$ *of* $\mathbf{X}$*, we have that the vector of the random probabilities* $(\mathbb{P}(A_1), \ldots, \mathbb{P}(A_n))$ *has the Dirichlet distribution*

$$(\mathbb{P}(A_1), \ldots, \mathbb{P}(A_n)) \sim \mathcal{D}\left(cP_0(A_1), \ldots, cP_0(A_n)\right).$$

This means that $\mathbb{P}$ is drawn from a DP if all its finite marginal distributions are Dirichlet. In order to investigate the effect of the parameters c and P_0, consider the partition (A, A'). We have that

$$\left(\mathbb{P}(A), \mathbb{P}\left(A'\right)\right) \sim \mathcal{D}\left(\left(cP_0(A), cP_0(A')\right)\right) \Rightarrow \mathbb{P}(A) \sim \mathcal{B}e\left(cP_0(A), c(1 - P_0(A))\right),$$

whence $\forall A \in \mathcal{X}$ we have that

$$\begin{aligned} \mathbb{E}\left\{\mathbb{P}(A)\right\} &= P_0(A) \\ \mathbb{V}\left\{\mathbb{P}(A)\right\} &= \frac{P_0(A)\left[1 - P_0(A)\right]}{c+1} \end{aligned}$$

Then the base measure P_0 plays the role of the location parameter of the DP and the concen-

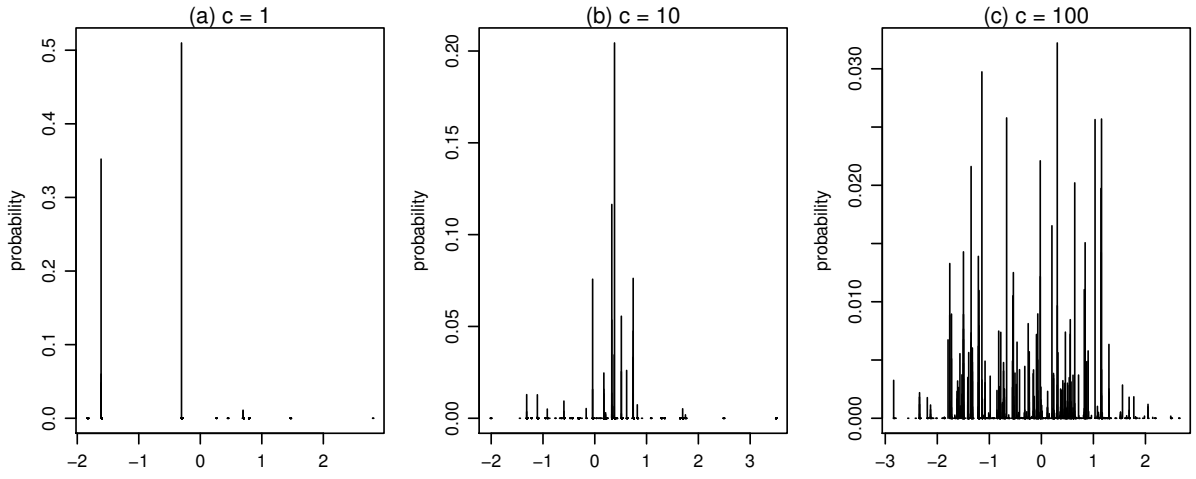

Figure 1.1: Random samples from the prior $\mathcal{DP}(c, P_0)$ for $c = 1, 10, 100$, and $P_0(\mathrm{d}x) = \mathcal{N}(x \mid 0, 1)\mathrm{d}x$.

tration parameter c is precision-like in the sense that $c \to 0$ implies $\mathbb{V}\{\mathbb{P}(A)\} \to 0$. Because $\mathbb{P}(A) = P_0(A)$ with probability 1, the draws of the DP will be concentrated tight around the base measure. In Fig. 1.2 we present draws from the DP with base measure the standard Normal distribution, for different values of the concentration parameter c. In the next section we

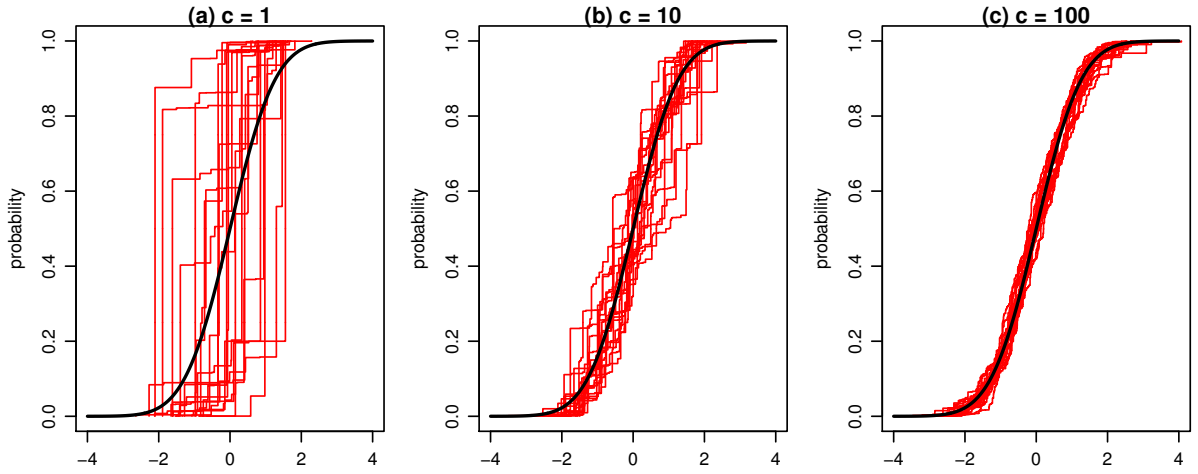

Figure 1.2: Cdf's resulting from 20 random samples from the prior $\mathcal{DP}(c, P_0)$ for $c = 1, 10, 100$, for $P_0(\mathrm{d}x) = \mathcal{N}(x \,|\, 0, 1)\mathrm{d}x$ are superimposed with the true cdf of the base measure P_0.

present the basic properties and some of the most popular representations of the DP appearing in the literature, highlighting the range of its possible applications regarding clustering and density estimation.

1.2.1 Properties and representations of the Dirichlet Process

Conjugacy of Dirichlet process

An attractive property of a DP is its conjugacy under iid sampling. This means that if we observe a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ generated directly from a DP random measure such that

$$x_i \,|\, \mathbb{P} \overset{\text{iid}}{\sim} \mathbb{P}, 1 \leq i \leq n$$
$$\mathbb{P} \sim \mathcal{DP}(c, P_0)$$

then the conditional distribution of $\mathbb{P}$ given the data, i.e. the posterior random measure, will also be a DP with parameters updated according to the observed sample, specifically

$$\mathbb{P} \,|\, \mathbf{x} \sim \mathcal{DP}\left(c + n, \frac{c}{c+n} P_0 + \frac{n}{c+n} \frac{\sum_{i=1}^{n} \delta_{x_i}}{n}\right). \tag{1.1}$$

We can see that the posterior base measure is a weighted average of the prior base measure and the empirical distribution. The weight related to the prior is proportional to the concentration parameter, while the weight related to the empirical distribution is proportional to the number

of the observed values. Moreover, due to the consistency of the empirical cdf as an estimator, if the observations x_i are iid, then we have asymptotically that $\forall A \in \mathcal{X}$

$$\mathbb{P}(A) \mid x_1, \dots x_n \xrightarrow{p} P_0(A).$$

It is also interesting the fact that the posterior base measure coincides with the posterior predictive distribution for the next future observation drawn from the DP.

Predictive distribution

Suppose again that we have a random sample $x_i \mid \mathbb{P} \overset{\text{iid}}{\sim} \mathbb{P}, 1 \leq i \leq n$, with $\mathbb{P} \sim \mathcal{DP}(c, P_0)$ and we are interested in deriving the predictive distribution of the next future unobserved observation x_{n+1}, with $\mathbb{P}$ marginalized out. Then we have that $\forall A \in \mathcal{X}$

$$\begin{aligned} P(x_{n+1} \in A \mid x_1, \dots, x_n) &= \int_{\mathbb{P} \in \mathcal{P}_{\mathbf{X}}} P(x_{n+1} \in A \mid \mathbb{P}) \, P(\mathbb{P} \mid x_1, \dots, x_n) \, \mathrm{d}\mathbb{P} \\ &= \int_{\mathbb{P} \in \mathcal{P}_{\mathbf{X}}} \mathbb{P}(A) \, P(\mathbb{P} \mid x_1, \dots, x_n) \, \mathrm{d}\mathbb{P} \\ &= \mathbb{E}\{\mathbb{P}(A) \mid x_1, \dots, x_n\} \\ &= \frac{c}{c+n} P_0(A) + \frac{n}{c+n} \frac{\sum_{i=1}^n \delta_{x_i}(A)}{n}, \end{aligned} \tag{1.2}$$

following Eq. (1.1) for the posterior base measure of $\mathbb{P}$. So we conclude that

$$x_{n+1} \mid x_1, \dots, x_n \sim \frac{cP_0 + n \sum_{i=1}^n \delta_{x_i}}{c+n} \tag{1.3}$$

and the posterior base distribution is also the predictive distribution .

Self-similarity

Another property characterizing the DP is the self-similarity property. Let $\mathbb{P}|_B$ be the restriction measure of $\mathbb{P}$ to a measurable B, $\mathbb{P}|_B(A) := \mathbb{P}(B \cap A)$, and $\mathbb{P}_B$ the conditional measure, $\mathbb{P}_B(A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, with $\mathbb{P}(B) > 0$. Similarly we obtain the restriction measure $\mathbb{P}|_{B'}$ and we have the following Theorem (for a proof we refer to [GVdV17]):

Theorem 1.3 (DP self-similarity). *If $\mathbb{P} \sim \mathcal{DP}(c, P_0)$, then $\mathbb{P}_B \sim \mathcal{DP}(cP_0(B), P_0|_B)$ and the variable $\mathbb{P}(B)$ and the processes $(\mathbb{P}_B(A), A \in \mathcal{X})$, $(\mathbb{P}_{B'}(A), A \in \mathcal{X})$ are mutually independent $\forall B \in \mathcal{X}$, such that $0 < P_0(B) < 1$.*

According to the above theorem, if we make a "localization" of the DP by conditioning on a proper set $B \in \mathcal{X}$, then the localized measure is still a (properly scaled) DP, with base measure the restriction of the original base measure to B. Moreover, the localized processes at disjoint sets are not only independent of each other, but also independent of the variable $\mathbb{P}(B)$. Essentially, we can locally obtain processes that are similar to the original process $\mathbb{P}$, so the DP exhibits self-similarity.

We now proceed to some of the most popular representations of the DP, which are indicative for the rich structure of this process and useful for the construction of Bayesian nonparametric MCMC algorithms.

Generalized Pólya Urn

The Generalized Pólya Urn scheme proposed by Blackwell and McQueen [BM+73], is a representation of the DP that has been used to provide an alternative proof for the existence of the process and is based on the integration of the underlying random measure.

Let again $\mathbf{x} = (x_1, \ldots, x_n)$ be a random sample, $x_i \,|\, \mathbb{P} \overset{\text{iid}}{\sim} \mathbb{P}, 1 \leq i \leq n$, with $\mathbb{P} \sim \mathcal{DP}(c, P_0)$ We have seen in Eq. (1.3) that by integrating out $\mathbb{P}$ we obtain the conditional predictive distribution of x_{n+1} from

$$x_{n+1} \,|\, x_1, \ldots, x_n \sim \frac{cP_0 + n\sum_{i=1}^{n} \delta_{x_i}}{c+n}.$$

Denoting by $(\tilde{x}_1, \ldots, \tilde{x}_k)$ the vector of the distinct values of $(x_1, \ldots, x_n)$ and by $n_j, 1 \leq j \leq k$, the number of times that each x_j has been observed in $\mathbf{x}$ such that $\sum_{j=1}^{k} n_j = n$ we can rewrite Eq. (1.3) as

$$x_{n+1} \,|\, x_1, \ldots, x_n \sim \frac{cP_0 + \sum_{j=1}^{k} n_j \delta_{\tilde{x}_j}}{c+n} \tag{1.4}$$

The Generalized Pólya Urn scheme provides an intuitive metaphor, describing a way to obtain a sequence of $(x_i)_{i\geq 1}$ with the above conditionals, based on the following procedure: Suppose we have an urn that is initially empty and we have an infinite set of unique colors $\mathbf{X}$. We choose at random a ball with color $x_1 \in \mathbf{X}$, $x_1 \sim P_0$ and in each subsequent $(n+1)-$step we either choose at random a ball with a new color ($x_{n+1} \sim P_0$) and place it into the urn, or we choose at random a ball from the ones already in the urn and place it back into the urn along with another ball of the same color. If we pick a new color with probability proportional to $c > 0$ and we choose a ball from the urn (that is x_{n+1} is drawn from the empirical distribution) with probability proportional to n (the number of balls in the urn), then we have the following conditional distribution:

$$x_{n+1} \,|\, x_1, \ldots, x_n \sim \begin{cases} \delta_{\tilde{x}_j}, & \text{with probability } \frac{n_j}{n+c} \text{ for j = 1,...,k.} \\ P_0, & \text{with probability } \frac{c}{c+n}. \end{cases}$$

We note that, by construction, the probability of x_{n+1} having an already observed color say $\tilde{x}_j$ is proportional to the number of times $n_j, 1 \leq n_j \leq n$ this color has appeared. These sets of identical x_i, i.e. $\mathcal{C}_j = \{i \in \{1, \ldots, n\} : x_i = \tilde{x}_j\}, 1 \leq j \leq k \leq n$, can be considered as *clusters*, as it follows that $\cup_{i=1}^{k} \mathcal{C}_j = \{1, \ldots, n\}$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \forall i \neq j$, moreover, the clusters induce a *random partition* over $[n]$[2]. Because larger clusters tend to grow faster, this property is being referred to as the "rich get richer" phenomenon.

[2]We denote with $[n]$ the set $\{1, \ldots, n\}$

So, we can construct a Generalized Pólya Urn scheme using the DP. The reverse is also possible, because a sequence of $(x_i)_{i\geq 1}$ generated by the Generalized Pólya Urn satisfies the property of infinite exchangeability, thus by de Finetti's theorem we have the existence of a random measure Π, such that

$$P(x_1 \in A_1, \ldots, x_n \in A_n) = \int_{\mathbb{P}\in\mathcal{P}_X} \prod_{i=1}^{n} \mathbb{P}(A_i)\Pi(d\mathbb{P})$$

It has been proved [BM+73] that such a random distribution, playing the role of the mixing measure is a DP. The Pólya Urn representation has applicability in marginal MCMC methods used for sampling from the DP, that result in finite-dimensional approximate samplers integrating out the random measure. For details regarding marginal samplers we refer to [Esc94, Nea00].

Chinese Restaurant Process

Another famous representation of the DP is the Chinese restaurant process (CRP), closely related to the generalized Pólya Urn. For given $c > 0$ and $n \in \mathbb{N}$, CRP (c, n) defines a distribution over the partitions of $[n]$ and is in fact, the distribution of the partitions induced by the Pólya Urn scheme, discussed above. The name of the process is due to the following intuitive interpretation.

Suppose we have a restaurant with an infinite number of tables, initially empty. When the first customer arrives, he chooses to sit at a table at random. Then, every $n+1-$customer, sits at a new table with probability proportional to c, or chooses an occupied table with probability proportional to the number of customers already sitting at the table. We can consider the different tables as clusters d_i, associated with the indices (customers) i, for $i = 1, 2, \ldots$. It has been shown [P+02] that the "rich gets richer" clustering effect of the DP is exhibited by the CRP and the growth of the number of new tables is of order $\mathcal{O}(c\log(n))$.

An important feature of CRP is exchangeability [Ald85], in the sense that the probability of any specific partition is invariant under the permutations of its elements, i.e. not affected by the ordering of the customers. Letting k the number of distinct clusters, the probability of a given partition $\pi_{[n]} \sim$ CRP (c, n) can be written as

$$P\left(\pi_{[n]}\right) = \frac{c^k}{c\,(c+1)\ldots(c+n-1)} \prod_{i\in\pi_{[n]}} (|i|-1)! \tag{1.5}$$

Equation (1.5) is called an *exchangeable partition probability function* (EPPF). For generalizations of the CRP and models of random partitions, we refer to [Pit95, P+02, IJ03].

Stick-breaking representation

Perhaps the most widely used, especially by practitioners, representation of the DP is the Stick-Breaking representation introduced by Sethuraman [Set94]. In this constructive definition, Sethuraman showed that for a random measure $\mathbb{P} \sim \mathcal{DP}(c, P_0)$ we can obtain the following representation:

$$\mathbb{P} = \sum_{j=1}^{\infty} w_j \delta_{x_j}, \tag{1.6}$$

where $x_j \overset{\text{iid}}{\sim} P_0, j \geq 1$ and $\boldsymbol{w} = (w_j)_{j \geq 1}$ is an infinite sequence of stick-breaking weights, defined by

$$w_1 = v_1, \quad w_k = v_k \prod_{l<k} (1 - v_l), \tag{1.7}$$

with $v_i \overset{\text{iid}}{\sim} \mathcal{B}e(1, c), i \geq 1$ and $\sum_{j=1}^{\infty} w_i \overset{a.s.}{=} 1$.

According to the above definition, the DP is a *discrete random measure*.

The intuition behind the construction of the infinite sequence of weights is as follows: Consider a unit-length stick. We break the stick at length v_1. Then, we normalize the remaining length $1 - v_1$ and break it again at length v_2, obtaining a piece of length $w_2 = v_2(1 - v_1)$, along with the remaining piece which is broken again in order to define w_3. The process continues ad infinitum and it is evident that $\sum_{i \geq 1} w_i \overset{a.s.}{=} 1$, inducing a countable (a.s.) random partition of $[0, 1]$. One can use the notation $\boldsymbol{w} \sim \text{GEM}(c)$[3] for the distribution obtained by Eq. (1.7). The representation of Eq. (1.6) plays a central role in the construction of *exact* MCMC sampling techniques, that -in contrast with the marginal methods discussed above- include in the inferential procedure the random measure itself and use slice sampling methods for the creation of finite dimensional Gibbs samplers.

Furthermore, an important aspect of Sethuraman's definition, is that it can be modified in order to construct a more general class of BNP priors (containing the DP), namely the class of stick-breaking priors [IJ01]. It is possible to construct this class by allowing the weights to be generated as in (1.7) but with more flexible $v_i \overset{\text{iid}}{\sim} \mathcal{B}e(a_i, b_i), a_i, b_i > 0, i \geq 1$. In this case, in order to impose $\sum_{i \geq 1} w_i \overset{a.s.}{=} 1$, it suffice that $\sum_{j \geq 1} \log\left(1 + \frac{a_j}{b_j}\right) = \infty$.

Examples of BNP priors that can be constructed using the above framework include the two-parameter Poisson-Dirichlet process [PY97], the Indian Buffet Process [TGG07] and beta two-parameter processes [IZ00]. Moreover, for a novel alternative proof of the existence of the DP based on the stick-breaking representation see [HNW19]. We will continue with the definition of a specific BNP prior that we will use extensively in the next Chapters, namely the Geometric Stick-Breaking process.

[3] from the names of Griffiths-Engen-McCloskey

1.3 The Geometric Stick-Breaking Process

The Geometric Stick-Breaking process is a random measure introduced by [FGMW10], and its weights are less complex than the DP weights. Formally, we denote by $\mathbb{G} \sim \mathcal{GSB}(a, b, P_0)$ a random measure $\mathbb{G} \in \mathcal{P}_{\mathbf{X}}$ drawn from a Geometric Stick-Breaking process with *geometric probability* λ and base measure P_0 if:

$$\mathbb{G} = \lambda \sum_{j=1}^{\infty} (1-\lambda)^{j-1} \delta_{x_j} \tag{1.8}$$

with $x_i \overset{\text{iid}}{\sim} P_0, i \geq 1$ and $\lambda \sim \mathcal{B}e(a, b)$.

It is evident that equation (1.8) can be written in the form of Eq. (1.6), if we replace the infinite sequence of the beta variables $(v_i)_{i\geq 1}$ with a single beta variable $\lambda \sim \mathcal{B}e(a, b)$, namely the geometric probability, imposing a geometric structure on the weights, by

$$w_j = \lambda(1-\lambda)^{j-1}, j = 1, 2, \ldots \tag{1.9}$$

Clearly, from the above representation, we have that $\sum_{i\geq 1} w_i \overset{a.s.}{=} 1$ and $\mathbb{G}$ is a.s. a discrete random measure. In fact, the GSB random measure removes a level from the hierarchy of the DP, by replacing the random weights with their expected values, thus having a less complex structure. Formally, if $\boldsymbol{w}$ is an infinite sequence of stick-breaking weights as in Eq. (1.7) then

$$\mathbb{E}\{w_k\} = \frac{1}{c} \prod_{i=1}^{k-1} \frac{c}{c+1} = \frac{1}{c} \left(\frac{c}{c+1}\right)^{k-1}, \tag{1.10}$$

which has the form of Eq. (1.9), under the reparametrization $\lambda = (1+c)^{-1}$.

Regarding the properties of the GSB prior, we have the following Proposition, proved in [Mer18]:

Proposition 1.4. *Let* $\mathbb{G} \sim \mathcal{GSB}(\lambda, P_0)$, *then* $\forall A \in \mathcal{X}$ *it is that*

1. $\mathbb{E}\{\mathbb{G}(A)\} = P_0(A)$
2. $\text{Var}\{\mathbb{G}(A)\} = \frac{\lambda}{2-\lambda} P_0(A)(1 - P_0(A))$.

It appears that the simple construction of the (always decreasing) weights of the GSB prior, leads to models that are at least as efficient as more complex BNP models, such as the DP or even generalizations of the DP prior. The main reason for this, is that the GSB measure - even with less complex weight structure- admits a stick-breaking representation, so it can be proven using standard results [OC04] that it has full support on the space of discrete probability measures. Intuitively, although a hierarchical level has been removed from the weights, the flexibility of the random measure persists due to the infinity of the possible location parameters that control the weights, supporting any cluster location.

In the next Chapters, we will construct methods based on the GSB prior that perform density estimation yielding qualitatively indistinguishable results from the corresponding DP-based methods, with significantly lower mean execution times (METs).

1.4 Bayesian nonparametric mixture models

As we have seen, stick breaking random probability measures, such as DP or GSB, are a.s. discrete random measures, even in the case where the base measure P_0 is continuous. This poses a significant restriction regarding possible BNP applications (e.g. density estimation), as it renders the above priors inadequate for the modeling of densities.

In order to overcome the discrete nature of the DP, Antoniak [Ant74] and Lo [L$^+$84] introduced the Dirichlet Process Mixture (DPM) model. The main idea involves the smoothing of the DP draws and making use of a parametric family of kernels.

Let $\mathcal{K}_\theta = \{K(\cdot \mid \theta), \theta \in \Theta \subseteq \mathbb{R}^d\} \subset \mathcal{P}_{\mathbb{X}^*}$, a parametric family of kernels indexed by θ, i.e. for a given $\theta \in \Theta$, $K(\cdot \mid \theta)$ is a density.

Consider $\mathbb{P} \sim \mathcal{DP}(c, P_0)$ and a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$. Using the DPM, we model $x_i, 1 \leq i \leq n$ as $x_i \mid \mathbb{P} \sim f(\cdot \mid \mathbb{P})$, where :

$$f(x \mid \mathbb{P}) = \int_{\theta \in \Theta} K(x_i \mid \theta) \mathbb{P}(\mathrm{d}\theta) \tag{1.11}$$

Eq. (1.11) describes an *infinite mixture* of kernels in $\mathcal{K}_\theta$, with mixing measure a DP random measure $\mathbb{P}$. Hierarchically, using the latent variables $(\theta_1, \ldots, \theta_n) \in \Theta^n$ we obtain:

$$\begin{aligned} x_i \mid \theta_i &\overset{\text{ind}}{\sim} K(\cdot \mid \theta_i) \\ \theta_i \mid \mathbb{P} &\overset{\text{iid}}{\sim} \mathbb{P} \\ \mathbb{P} &\sim \mathcal{DP}(c, P_0) \end{aligned} \tag{1.12}$$

where observations x_i with the same θ_j belong to the same mixture component, considered as cluster. In contrast with parametric Bayesian mixture models, the number of clusters is determined according to the observed data and this is a highlight of the flexibility characterizing BNP models.

Moreover, using the stick breaking representation of $\mathbb{P}$, we can write the mixture distribution of Eq. (1.11) as a convex combination of kernels, that is with $\boldsymbol{w} = (w_i)_{i \geq 1}$ the infinite sequence of stick breaking weights as in Eq. (1.7):

$$f(x \mid \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{i=1}^{\infty} w_i K(x \mid \theta_i) \tag{1.13}$$

Of course, the choice of the random measure as mixing measure in Eq. (1.11) does not have to be limited in the case of DP. In this thesis, we will make use of infinite mixture models based on both the DP and GSB priors. In the general case where the mixture measure is distributed according to some random measure of the class of the aforementioned stick-breaking priors, we have the following analytical representation of the corresponding mixture model, with $\boldsymbol{w} =$

$(w_i)_{i\geq 1}$ and $\boldsymbol{\theta} = (\theta_i)_{i\geq 1}$ the infinite sequences of weights and locations respectively:

$$\begin{aligned} x_i \,|\, \boldsymbol{w}, \boldsymbol{\theta} &\overset{\text{ind}}{\sim} f(x_i \,|\, \boldsymbol{w}, \boldsymbol{\theta}) \\ f(x \,|\, \boldsymbol{w}, \boldsymbol{\theta}) &= \sum_{j=1}^{\infty} w_j K(x \,|\, \theta_j) \\ w_1 = v_1, \text{and } w_k &= v_k \prod_{l<k} (1 - v_l) \\ v_i &\overset{\text{iid}}{\sim} \mathcal{B}e\left(a_i, b_i\right), i \geq 1 \\ \theta_i &\overset{\text{iid}}{\sim} \mathbb{P}_0 \end{aligned} \tag{1.14}$$

Eqs. (1.12) constitute a special case of Eqs. (1.14). The reason for the popularity of the DPM model, is mainly due to its conjugacy, in accordance with the conjugacy of the original DP, i.e. the prior and posterior are both DPMs. When DPMs are used for density estimation, a proper parametric family has to be chosen. For example, if we have evidence that the support of the underlying density is $\mathbb{R}^+$ we may use Gamma or Log-normal kernels. A strong relevant result is that any unimodal symmetric density can be represented as an infinite mixture of Normal kernels. If we choose DP mixing measure and Normal kernels, i.e.

$$\mathcal{K}_\theta = \left\{ K(\cdot \,|\, \vartheta) = \mathcal{N}\left(\cdot \,|\, \mu, \sigma^2\right), \vartheta = \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \right\},$$

then we have a Dirichlet Process Gaussian Mixture Model [GR10].

We note that there have also been proposed alternative methods, suitable for the modeling of densities using a.s. discrete random probability measures, for example using normalized continuous processes [NBPW+04].

We will now proceed with the presentation of the most popular MCMC methods used to perform Bayesian inference, in both parametric and nonparametric prior settings.

1.5 MCMC Methods

In a Bayesian framework the uncertainty about the parameters of interest is expressed in terms of probability statements, and Bayesian inference is based on the posterior distribution. Essentially, we update our prior beliefs regarding the parameters of interest via the prior distribution, using the observed data via the data likelihood. In a parametric setting, given the *conditionally* iid observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ we have that

$$\begin{aligned} x_1, \ldots, x_n \,|\, \vartheta &\overset{\text{iid}}{\sim} f(\cdot \,|\, \vartheta) \\ \vartheta &\sim \pi(\cdot) \end{aligned}$$

and the posterior distribution will be the conditional distribution of ϑ given $\boldsymbol{x}$, i.e.

$$f\left(\vartheta \mid \boldsymbol{x}\right) = \frac{\pi(\vartheta) f\left(\boldsymbol{x} \mid \vartheta\right)}{f(\boldsymbol{x})} = \frac{\pi(\vartheta) f\left(\boldsymbol{x} \mid \vartheta\right)}{\int_{\vartheta \in \Theta} \pi(\vartheta) f\left(\boldsymbol{x} \mid \vartheta\right) \mathrm{d}\vartheta}, \tag{1.15}$$

where, due to conditional independence the likelihood has the form $f\left(\boldsymbol{x} \mid \vartheta\right) = \prod_{i=1}^{n} f(x_i \mid \vartheta)$.

Now, all our posterior beliefs are expressed by $f\left(\vartheta \mid \boldsymbol{x}\right)$ and we are typically interested in finding Bayesian point estimators (e.g. the mean, the mode or a certain quantile of the posterior), as well as *credible intervals*, i.e. regions that contain ϑ with a certain -usually high- probability.

This is an easy task when we choose a prior that is conjugate under the likelihood, resulting in a posterior belonging in the same parametric family with the prior, but with updated parameters. In many cases, though, this is not possible, and the *intractability* of $\int_{\vartheta \in \Theta} \pi(\vartheta) f\left(\boldsymbol{x} \mid \vartheta\right) \mathrm{d}\vartheta$ makes the derivation of a closed form for the posterior impossible.

However, we can circumvent the problem by approximating the above integral via a probabilistic method, known as *Monte Carlo (MC) integration*. In fact, for any integral that can be expressed as an expectation with respect to a distribution $\pi(\cdot)$

$$\int_{\vartheta \in \Theta} g(\vartheta) \pi(\vartheta) \mathrm{d}\vartheta = \mathbb{E}_{\pi}\left\{g(\theta)\right\} \tag{1.16}$$

we can obtain an approximation using iid samples $x_i \overset{\text{iid}}{\sim} \pi(\cdot), 1 \leq i \leq n$. Specifically, the Strong Law of Large Numbers guarantees the a.s. convergence of the sample mean of $g(x_i)$ to the desired expected value, that is as $n \to \infty$:

$$I_n(g) = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \overset{\text{a.s.}}{\to} I(g) = \mathbb{E}_{\pi}\left\{g(\theta)\right\} \tag{1.17}$$

Furthermore, if the variance $\sigma_g^2 = \mathrm{Var}_{\pi}\left\{g(\vartheta)\right\}$ is finite, it follows from the Central Limit Theorem that the sequence of unbiased estimators $I_n(g)$ converges asymptotically to a Normal distribution:

$$\sqrt{n}\left(I_n(g) - I(g)\right) \overset{\text{d}}{\to} \mathcal{N}\left(0, \sigma_g^2\right) \tag{1.18}$$

Unfortunately, there are cases where sampling from $\pi(\cdot)$ is not possible, so Monte Carlo integration is not adequate. In order to deal with the complex (and often high-dimensional) posterior distribution, we may resort to a large class of methods called *Markov Chain Monte Carlo* (MCMC).

The main idea underlying the MCMC approach, is the construction of a Markov chain with "nice" properties, whose stationary distribution is the distribution we are interested in, called *target distribution*. Then, using iterative sampling we obtain samples that are approximately from the target distribution. Typically, the target distribution is the posterior. The crucial property of the constructed Markov Chain is *ergodicity*, so that the Chain explores the whole support of $\pi(\cdot)$ and converges to its stationary distribution -which will also be unique [Rob07]- independently of the initial point.

There exists a vast and growing literature regarding theoretical results, practical implementations and modern applications of MCMC methods. For an extensive review and more thorough analysis, we refer the reader to [RC04, BGJM11, GRS95, Tie94]. In the following, we will present two of the most commonly used MCMC methods: Metropolis-Hastings and Gibbs sampling.

1.5.1 Metropolis-Hastings Algorithm

The Metropolis–Hasting (MH) algorithm, introduced by Metropolis [MRR+53] and later extended by Hastings [Has70], is one of the most popular MCMC methods and forms the basis for an extensive class of algorithms.

Suppose we want to obtain a sample from a density $\pi(\vartheta)$, which is known up to a normalizing constant (e.g. without knowing the integral appearing in Eq. (1.15)). The basic idea of the MH algorithm involves the usage of a *proposal distribution* $q(\cdot \mid x)$ which is easy to sample from, combined with a mechanism that is responsible for the acceptance of the samples proposed by $q(\cdot \mid x)$. By construction, the MH algorithm has a stationary distribution, which has been proven [Tie94] to be the target distribution.

Specifically, the algorithm is initialized at a point $\vartheta^{(0)}$, such that $\pi(\vartheta^{(0)}) > 0$. Then, at each $t-$step, we use the proposal distribution in order to sample a new candidate point $\tilde{\vartheta} \sim q\left(\tilde{\vartheta} \mid \vartheta^{(t-1)}\right)$, possibly depending on the previous state $\vartheta^{(t-1)}$. The proposal will be accepted with probability

$$\alpha\left(\tilde{\vartheta}, \vartheta^{(t-1)}\right) = \min\left\{1, \frac{\pi\left(\tilde{\vartheta}\right) q\left(\vartheta^{(t-1)} \mid \tilde{\vartheta}\right)}{\pi\left(\vartheta^{(t-1)}\right) q\left(\tilde{\vartheta} \mid \vartheta^{(t-1)}\right)}\right\}, \tag{1.19}$$

which is called *acceptance probability* and is independent of the normalizing constant. Schematically, the MH algorithm (Algorithm 1) can be written as follows, for a $d-$dimensional parameter vector $\vartheta = (\theta_1, \ldots, \theta_d)$.

Algorithm 1 : METROPOLIS–HASTINGS sampling for multidimensional parameter.

procedure SAMPLE $\vartheta = (\theta_1, \ldots, \theta_d)$.
  Set initial point $\vartheta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_d^{(0)}\right)$.
  for $t = 1$ to N **do**
    Sample proposal $\tilde{\vartheta} \sim q\left(\tilde{\vartheta} \mid \vartheta^{(t-1)}\right)$
    Calculate acceptance probability $\alpha(\tilde{\vartheta}, \vartheta^{(t-1)})$
    Sample $u \sim \mathcal{U}(0,1)$
    Set $\vartheta^{(t)} = \begin{cases} \tilde{\vartheta}, & \text{if } u < \alpha(\tilde{\vartheta}, \vartheta^{(t-1)}) \\ \vartheta^{(t-1)}, & \text{otherwise} \end{cases}$
  end for
end procedure

The elements of the vector ϑ can be sampled simultaneously from a $d-$dimensional proposal distribution, or alternatively, we can make the sampling one variable (or sub-blocks of variables) at a time from lower dimensional proposal distributions $q_i(\cdot \mid \cdot), 1 \leq i \leq d$. The latter sampling scheme is a special case of MH, called *One-Variable-at-a-Time* MH.

The original Metropolis algorithm, required the symmetry $q(x \mid y) = q(y \mid x)$ of the proposal distribution. Hastings generalized the algorithm, in order to allow non-symmetric proposals. In the case of a symmetric proposal, the acceptance probability becomes

$$\alpha\left(\tilde{\vartheta}, \vartheta^{(t-1)}\right) = \min\left\{1, \frac{\pi\left(\tilde{\vartheta}\right)}{\pi\left(\vartheta^{(t-1)}\right)}\right\}$$

and we obtain the so-called *random walk Metropolis* (RWM) sampler. Any distribution that is simple to simulate from can be chosen as a proposal, though a minimal requirement in order to guarantee the full exploration of the state space is that the support of the target distribution is contained in the support of the proposal:

$$\operatorname{supp} \pi \subseteq \bigcup_{x \in \operatorname{supp} \pi} \operatorname{supp} q(\cdot \mid x)$$

The transition kernel of the Markov Chain generating the MH algorithm, can be written as:

$$K\left(\vartheta^{(t-1)}, \vartheta^{(t)}\right) = \alpha\left(\vartheta^{(t)}, \vartheta^{(t-1)}\right) q\left(\vartheta^{(t)} \mid \vartheta^{(t-1)}\right) + \left(1 - \alpha^{\star}\left(\vartheta^{(t-1)}\right)\right) \delta_{\vartheta^{(t-1)}}\left(\vartheta^{(t)}\right) \tag{1.20}$$

where

$$\alpha^{\star}\left(\vartheta^{(t-1)}\right) = \int \alpha\left(\vartheta^{(t)}, \vartheta^{(t-1)}\right) q\left(\vartheta^{(t)} \mid \vartheta^{(t-1)}\right) \mathrm{d}\vartheta^{(t)}.$$

The universal applicability of the MH, lies on the fact that the transition kernel of Eq. (1.20) satisfies the *detailed balance* equation

$$K\left(\vartheta^{(t-1)}, \vartheta^{(t)}\right) \pi\left(\vartheta^{(t-1)}\right) = K\left(\vartheta^{(t)}, \vartheta^{(t-1)}\right) \pi\left(\vartheta^{(t)}\right). \tag{1.21}$$

Then the generated Markov Chain is *reversible* with respect to the target density, and it has as a stationary distribution the target density $\pi(\cdot)$.

The main drawback of the MH algorithm is that it needs to be calibrated and this can be a time-consuming task, especially for complex high-dimensional target distributions. The performance of the algorithm, mainly the time to reach stationarity in order to adequately approximate the target, is highly influenced by the proposal distribution. Having chosen a parametric family for the proposal distribution (e.g. a Normal distribution centered at the last accepted point), the basic criterion in order to calibrate the parameters is the acceptance rate (computed by the empirical acceptance frequency). Very high or very low acceptance rates indicate poor mixing and convergence properties, often indicating that the variance of the proposal should increase or decrease respectively.

In any case, it is useful to perform *convergence diagnostics* after the application of MH. Apart

from the acceptance rate, the mixing properties should be examined using autocorrelation function plots and trace plots and it is often useful to run multiple chains starting from different initial conditions, checking that they give similar results. A more thorough convergence analysis may include quantitative diagnostics that exist in the literature [BR98], such as potential scale reduction factor [GR+92, BG98]. Furthermore, a *burn-in* period of discarded samples is often used in the beginning of the chain, in order to reduce the effect of the initial point, and *thinning* is performed (keeping only one point every k iterations) aiming to reduce autocorrelation of the sampled values.

The first theoretical result regarding MH scaling was proven by Roberts et. al [RGG+97] for the RWM case, under Gaussian proposals. Specifically, the authors established that the optimal acceptance rate is, under quite general conditions, 0.234 independently of the target density. Heuristically, if the proposal is Gaussian, the proposal variance should be tuned so that the acceptance rate is between 0.25 and 0.50, with higher values more appropriate for low dimensional problems. For more results regarding the problem of optimal scaling, we refer the reader to [RR+01, R+11].

In order to overcome the calibration problem, a new class of *adaptive* MCMC algorithms has been introduced [HST+01, RR09, BDH14], where the proposal distribution is "automatically" tuned during the sampling process, utilizing the information obtained by the history of the process.

Moreover, the significance of the MH algorithm is also highlighted by a plethora of powerful MCMC algorithms that are based upon it. Popular examples of such algorithms include, among others a MCMC version of Simulated Annealing [GT95, RCC10] useful for optimization, Reversible Jump MCMC [RG97] often used for model selection, and Hamiltonian Monte Carlo [DKPR87, N+11] providing an efficient alternative in order to sample from complex high-dimensional densities.

1.5.2 Gibbs sampling

The Gibbs sampler introduced by Geman & Geman [GG84] and later generalized by Gelfand & Smith [GS90] is a MCMC algorithm, that produces a sample approximately from the target density $\pi(\cdot)$ without requiring an acceptance mechanism. This is the main advantage of the algorithm, as it circumvents the often demanding task of tuning. Specifically, the Gibbs sampling scheme, described in Algorithm 2, at each step performs draws from the *full conditional* (FC) distributions. Suppose that at each step we want to sample a vector $\vartheta = (\theta_1, \ldots, \theta_d)$. The components of ϑ are sequentially updated for $j = 1, \ldots, d$ from the FCs $\pi_{\theta_j \,|\, \theta_{-j}}(\cdot \,|\, \cdot)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j-1}, \ldots, \theta_d)$.

In the simplest case, the components $\theta_j, j = 1, \ldots, d$ are one-dimensional and at each iteration are all updated in the same order. Whenever possible, it is more efficient to perform the Gibbs updates over blocks. Naturally, the higher speed compensates for the difficulty of sampling from multidimensional FCs. Furthermore, if at each iteration a cycle from θ_1 to θ_d is performed,

we have a systematic scan Gibbs sampler. It is more efficient to randomly choose an index $j \in \{1, \ldots, d\}$ at each iteration, update the single component (or block) θ_j with a draw from the corresponding FC and keep the other components at their previous values. This scheme forms a class of *random scan* Gibbs samplers, further generalized in [LYHN05, LC06].

The Gibbs sampler is preferred in cases where the FCs are of standard form, or at least easy to sample from, e.g. using variants such as the slice sampler which we will present later on. However, if some of the FCs (of a component or a block) are too complex to sample from, we could substitute the Gibbs update with a Metropolis update using a proper proposal distribution resulting in a *Metropolis-within-Gibbs* algorithm [GBT95, GT01].

Essentially, we can consider any Gibbs update as a Metropolis update with proposal density the corresponding FC and acceptance probability always equal to one.

Algorithm 2 : GIBBS sampling for multidimensional parameter.

procedure SAMPLE $\vartheta = (\theta_1, \ldots, \theta_d)$
  Set initial point $\vartheta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.
  for $i = 1$ to N **do**
    for $j = 1$ to d **do**
      Sample $\theta_j^{(i)} \sim \pi_{\theta_j \mid \theta_{-j}}(\theta_j \mid \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_d^{(i-1)})$
    end for
    Set $\vartheta^{(i)} = \left(\theta_1^{(i)}, \ldots, \theta_d^{(i)}\right)$
  end for
end procedure

The transition kernel of the Gibbs sampler is

$$K\left(\vartheta^{(t-1)}, \vartheta^{(t)}\right) = \prod_{j=1}^{d} \pi_{\theta_j \mid \theta_{-j}}(\theta_j \mid \theta_1^i, \ldots, \theta_{j-1}^i, \theta_{j+1}^{i-1}, \ldots, \theta_d^{i-1}). \tag{1.22}$$

Then it follows that the stationary distribution of the generated Markov chain is indeed the target (joint) density $\pi(\vartheta)$. For theoretical results regarding the convergence of the Gibbs sampler, we refer to [RS97, JJN+13].

1.5.3 Slice sampling

When the target density has a complex form, the derivation of standard form FC distributions is often not feasible. Although Metropolis-type updates are always a choice, a full Gibbs-type updating sampling scheme might be preferred. Auxiliary variable methods [BGJM11] provide an efficient alternative class of algorithms, by artificially augmenting the state space with auxiliary variables, leading to a set of easily sampled FCs. In [BG93], auxiliary variable methods have been used in order to remove unwanted high interactions (correlations) between the components of the sampled vector. Further applications of such methods in Bayesian image analysis are presented in [Hig98].

A general auxiliary-variables methodology with wide applicability, is known as *slice sampler* [DWW99, Nea03]. Suppose we have a target density $\pi(\cdot)$ that can be factorized as

$$\pi(\vartheta) \propto l(\vartheta) f(\vartheta), \tag{1.23}$$

where $f(\cdot)$ is a density and $l(\cdot)$ is a non-negative invertible function (not necessarily a density), so that we can obtain the sets

$$A_u = \{\vartheta : u < l(\vartheta)\} \tag{1.24}$$

The main idea of the method proposed in [DWW99] is the introduction of a strategic latent variable u defined on $(0, \infty)$, such that the marginal density $\pi(\vartheta)$ remains unchanged, while the resulting FCs $\pi(\vartheta \mid u)$ and $\pi(u \mid \vartheta)$ are now easy to sample from. Formally, with the proper introduction of u we have the joint density

$$\pi(\theta, u) \propto l(\vartheta) \mathcal{I}(u < f(\vartheta)) \tag{1.25}$$

from which we can deduce that the marginal density $\pi(\theta)$ remains unaffected by the introduction of u. The next step involves the iterative sampling from the simple forms of the FCs, that is $\mathcal{U}(0, l(\vartheta))$ for $u \mid \vartheta$ and the restriction of $f(\cdot)$ on the set A_u for $\vartheta \mid u$. The Markov chain will have as a stationary distribution the joint $\pi(\vartheta, u)$ and we are interested in the target marginal $\pi(\vartheta)$. For theoretical results concerning the convergence of auxiliary-variable MCMC methods, we refer to [MT97].

We note that, if needed, multiple latent $(u_i)_{1 \leq i \leq n}$ can be introduced, as long as we can use non-negative invertible functions $l_i(\cdot), 1 \leq i \leq n$ in order to write down the target density as

$$\pi(\vartheta) \propto \prod_{i=1}^{n} l_i(\vartheta) f(\vartheta) \tag{1.26}$$

and compute the FC distributions using the slice sets $A_u^i = \{\vartheta : u < l_i(\vartheta)\}$ in accordance with the scheme described above.

The factorization of the target density as in Eqs. (1.23-1.26) is not unique, in the sense that a variety of different choices can lead to efficient samplers , provided that the marginal density remains the target density and the FCs are substantially easier to sample from than the original ones. General methods, however, have been proposed [DW01] suitable for the sampling of truncated densities using a single latent variable. In the following Chapters, we will make extensive use of sampling procedures requiring draws from truncated Normal densities, so we illustrate the use of the slice sampler as in [DW01].

Suppose we have as target density a truncated standard Normal distribution over the interval (a, b), i.e.

$$\pi(\vartheta) \propto \exp\left\{\frac{-\vartheta^2}{2}\right\} \mathcal{I}(\vartheta \in (a, b)) \tag{1.27}$$

Introducing the latent variable u, such that the joint density can be written as

$$\pi(\vartheta, u) \propto \mathcal{I}\left(u \in \left(0, \exp\left\{\frac{-\vartheta^2}{2}\right\}\right)\right) \mathcal{I}\left(\vartheta \in (a, b)\right) \tag{1.28}$$

we have that the marginal distribution remains unaffected, as

$$\pi(\vartheta) = \int_0^{\infty} \pi(\vartheta, u) \mathrm{d}u \propto \int_0^{\exp\left\{\frac{-\vartheta^2}{2}\right\}} \mathcal{I}\left(\vartheta \in (a, b)\right) \mathrm{d}u \propto \exp\left\{\frac{-\vartheta^2}{2}\right\} \mathcal{I}\left(\vartheta \in (a, b)\right) \tag{1.29}$$

and the FC distributions are uniform, namely

$$\pi(u \mid \vartheta) = \mathcal{U}\left(0, \exp\left\{\frac{-\vartheta^2}{2}\right\}\right) \tag{1.30}$$

$$\pi(\vartheta \mid u) = \mathcal{U}\left(\max\left\{a, -\sqrt{-2\log u}\right\}, \min\left\{b, \sqrt{-2\log u}\right\}\right) \tag{1.31}$$

In Fig. 1.3 we present a sample from truncated $\mathcal{N}(0,1)$ on $(0,3)$ using slice sampling, with FCs given by Eqs. (1.31-1.31). We ran the sampler for $5,000$ iterations with a burn-in period of length $1,000$. With red vertical lines we indicate 0 and 3. Based on the above scheme it also easy to simulate draws from truncated versions of multivariate Normal, Beta or Gamma distributions [DW01].

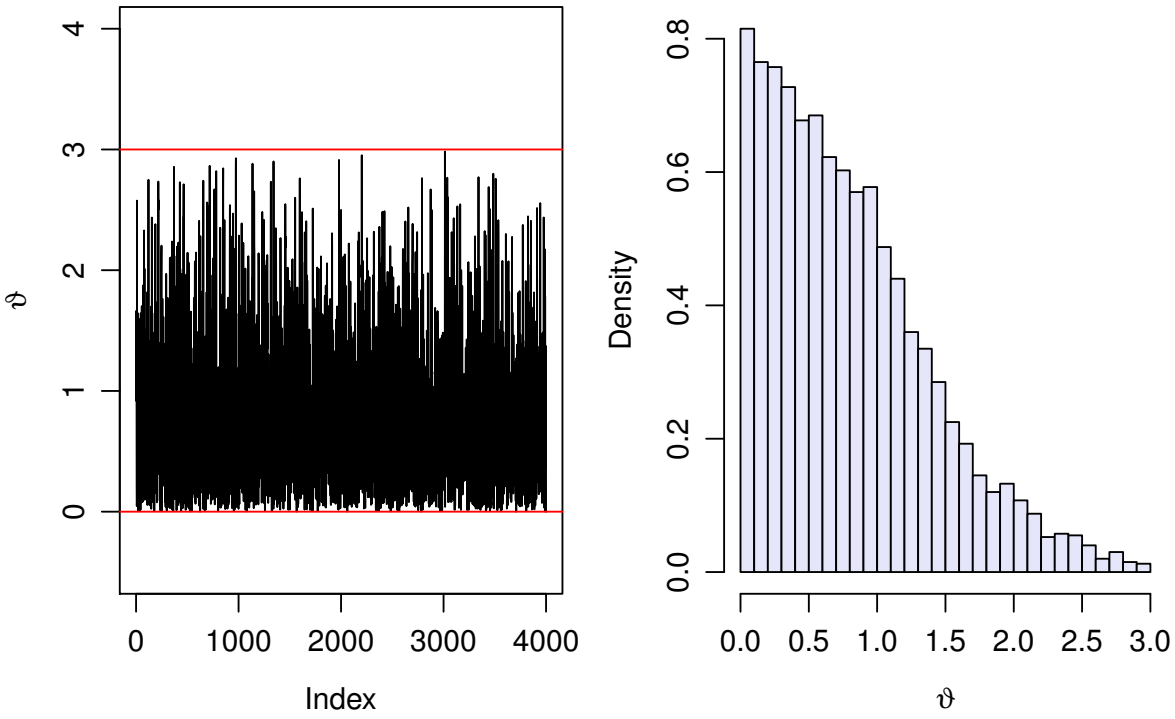

Figure 1.3: Trace (Left) and histogram (Right) of a sample from truncated $\mathcal{N}(0,1)$ on $(0,3)$, obtained using slice sampling.

In the rest of this Chapter, we will present MCMC methods, based on slice sampling techniques, suitable for the simulation of BNP mixture models, using DP and GSB priors.

1.5.4 MCMC methods for Bayesian nonparametric mixture models

Bayesian mixture models [MMR05] have been extensively used in a variety of applications in statistics [GSC+13] and machine learning [Mur12]. Different methods have been proposed in order to perform model selection [CH+08], mainly regarding the choice of the number of the mixture components (clusters), e.g. the number of Normal kernels in finite Gaussian mixture model. The importance of Bayesian nonparametric mixture models -that can also be derived as the limit of the corresponding finite mixture models when the number of components approaches infinity [Nea92, Ras00]- is highlighted by their potential to "automatically" adapt their complexity (i.e. the number of the mixture components) according to the observed data. Moreover, while in finite mixture models the number of clusters is often prespecified, the complexity of Bayesian nonparametric mixture models may grow as the number of the available data increases.

In terms of application, though, this flexibility comes with the expense of the need for more advanced MCMC methods, namely in order to sample from infinite dimensional state spaces. As it has already been mentioned, there are two general categories of MCMC algorithms for BNP models. The marginal samplers [MM98, Nea00] integrating out the infinite-dimensional random measure and use Gibbs sampling in order to sample the finite dimensional marginal distribution of the remaining variables, and the conditional methods retaining the random measure in the sampling process, keeping only a finite number of variables at each iteration, which is sufficient in the sense that the stationary distribution remains unchanged.

The main advantage of the conditional methods is that they do not suffer from the dependence between parameters, existing in the marginal methods based on the Pólya Urn representation. Initially [IZ00, IJ01, IZ02], the conditional methods were based on finite approximations of the DP prior, essentially via a proper truncation of the stick-breaking construction. Later, Papaspiliopoulos and Roberts [PR08] used retrospective sampling methods in order to construct an algorithm that generates exact DPM samples. We are particularly interested in the slice-sampling based method for BNP mixture models proposed by Walker [Wal07] and extended by Kalli et. al [KGW11].

For what follows, suppose we have a sample $\boldsymbol{x} = (x_i)_{1 \leq i \leq n}$ from a BNP mixture model, assigning a DP or GSB random measure $\Pi(\cdot)$ as a prior over the mixing measure, that is

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

So, for $i = 1, \ldots, n$, with $\boldsymbol{w} = (w_i)_{i \geq 1}, \boldsymbol{\theta} = (\theta_i)_{i \geq 1}$ the infinite sequences of stick-breaking weights and (atom) locations respectively, the conditional distribution of x_i can be written as

$$\pi\left(x_i \,|\, \boldsymbol{w}, \boldsymbol{\theta}\right) = \sum_{j=1}^{\infty} w_j K\left(x_i \,|\, \theta_j\right). \tag{1.32}$$

Th main idea of the slice sampling method [Wal07] involves the augmentation of the state space with proper auxiliary variables, so that conditionally on those variables the number of weights and locations that needs to be sampled at each iteration is finite.

In particular, we assign to each observation x_i a clustering variable d_i indicating the associated mixture component that generated x_i, along with a random set of indices $\mathbb{A}_i$ that is by construction a.s. finite. Then, if the conditional distribution $d_i \,|\, \mathbb{A}_i$ is uniform over $\mathbb{A}_i$ only a finite number of variables will be required to sample from at each iteration of the MCMC sampler.

Thus, under a suitable generic augmentation scheme, the state space effectively will become finite dimensional and no approximations are required. In the next sections we will describe in detail the slice sampling method, when the prior over the mixing measure is DP [Wal07] and GSB [FGMW10].

Sampling from the DP

Suppose we have that $\mathbb{P} \sim \mathcal{DP}(c, P_0)$. Then, by the stick-breaking representation we have that $\mathbb{P} = \sum_{j\geq 1} w_j \delta_{\theta_j}$, with weights $\boldsymbol{w} \sim \text{GEM}(c)$. We introduce the set of latent variables $\{u_i\}_{i=1}^n$ such that the joint density of (x_i, u_i) for $i = 1, \ldots, n$ is

$$\pi(x_i, u_i \,|\, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{j=1}^{\infty} \mathcal{I}(w_j > u_i) K(x_i \,|\, \theta_j) = \sum_{j \in \mathbb{A}_i} K(x_i \,|\, \theta_j) \tag{1.33}$$

with $\mathbb{A}_i$ the random slice sets of indices defined via the u_i's as

$$\mathbb{A}_i \equiv \mathbb{A}_{\boldsymbol{w}}(u_i) = \{j \in \mathbb{N} : 0 < u_i < w_j\}. \tag{1.34}$$

Furthermore, introducing the set of the latent clustering variables $\{d_i\}_{i=1}^n$, indicating the mixture components from which x_i came from, we can write

$$\pi(x_i, u_i \,|\, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{j=1}^{\infty} \pi(d_i = j)\, \pi(x_i, u_i \,|\, d_i = j, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{j=1}^{\infty} w_j \mathcal{U}(u_i \,|\, 0, w_j) K(x_i \,|\, \theta_j) \tag{1.35}$$

so that the marginal distribution of x_i remains unaffected and $\pi(u_i \,|\, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{j\geq 1} w_j \mathcal{U}(u_i \,|\, 0, w_j)$. For the FC distributions, we have that:

$$\begin{aligned} \pi(u_i \,|\, \ldots) &= \mathcal{U}(u_i \,|\, 0, w_j) \\ \pi(d_i \,|\, \ldots) &= \mathcal{DU}\{\mathbb{A}_i\} \\ \pi(x_i \,|\, \ldots) &= \frac{1}{|\mathbb{A}_i|} \sum_{j \in \mathbb{A}_i} w_j K(x_i \,|\, \theta_j) \end{aligned} \tag{1.36}$$

Eqs. (1.36) indicate the effect of the random slice sets on the the FC distribution for each observation which is an *equally weighted finite mixture*, with weights equal to the cardinality of the a.s. finite random set $\mathbb{A}_i$. Finally, the complete model for the posterior distribution will have

the form

$$\pi\left(\boldsymbol{w},\boldsymbol{\theta}\,|\,\boldsymbol{x},\{u_i\}_{i=1}^n,\{d_i\}_{i=1}^n\right)\propto\pi\left(\boldsymbol{w},\boldsymbol{\theta}\right)\prod_{i=1}^n\mathcal{I}\left(w_{d_i}>u_i\right)K\left(x_i\,|\,\theta_{d_i}\right)\tag{1.37}$$

Having derived the posterior in Eq. (1.37), it is evident that only a finite number of weights and locations needs to be sampled at every sweep of the Gibbs sampler. The sampling steps are:

1. The FC distributions for the latent u_i's are

$$\pi\left(u_i\,|\,\ldots\right)=\mathcal{U}\left(u_i\,|\,0,w_j\right)\tag{1.38}$$

2. The θ_j-locations' FCs are

$$\pi\left(\theta_j\,|\,\ldots\right)=P_0(\theta_j)\prod_{d_i=j}K\left(x_i\,|\,\theta_j\right)\tag{1.39}$$

If there are no $d_i=j$ then θ_j's are independently sampled from the prior P_0.

3. Following an efficient modification [KGW11] of the originally proposed sampling scheme, we can sample v_j's and u_i's as a block. This leads to the following FC

$$\pi\left(v_j\,|\,\ldots,\text{exclude }\{u_i\}_{i=1}^n\right)=\mathcal{B}e\left(1+\sum_{i=1}^n\mathcal{I}\left(d_i=j\right),c+\sum_{i=1}^n\mathcal{I}\left(d_i>j\right)\right)\tag{1.40}$$

so we are able to construct the stick-breaking weights following Eq. (1.7).

4. In order to sample the clustering variables from

$$\pi\left(d_i=k\,|\,\ldots\right)=\mathcal{I}\left(k\in\mathbb{A}_i\right)K\left(x_i\,|\,\theta_k\right)\tag{1.41}$$

we need to construct the random sets $\mathbb{A}_i$. Specifically, we know the $d_i\in\{k:w_k>u_i\}$, so the question is how many weights have to be sampled in order to ensure that we have the exact number of them. Noting that the weights a.s. sum up to one, it suffices to identify the smallest k^*, such that for $u^*=\min_{1\leq i\leq n}\{u_i\}$:

$$\sum_{j=1}^{k^*}w_j>1-u^*\tag{1.42}$$

There cannot exist $k'>k^*$ such that for some u_i, $w_{k'}>u_i$, as in that case the weights would not sum up to 1. So, all the weights needed for the chain to proceed are $\{w_1,\ldots,w_{k^*}\}$. Moreover, it has been proved [MT98] that k^* is distributed as $1+\mathcal{P}oi\left(-c\log u^*\right)$.

5. In order to sample the -randomized- concentration parameter c, we follow [Wes92], assigning a Gamma prior $c\sim\mathcal{G}(\alpha,\beta^{-1})$. Then, sampling an auxiliary $h\sim\mathcal{B}e\left(c+1,n\right)$ and denoting with κ the number of the distinct $d_i,1\leq i\leq n$ we can sample c from the following 2-mixture of

Gamma distributions

$$\pi\left(c \mid \kappa, h\right) = w_c \mathcal{G}\left(\alpha+\kappa, \beta-\log h\right) + \left(1-w_c\right)\mathcal{G}\left(\alpha+\kappa-1, \beta-\log h\right) \tag{1.43}$$

with weights defined by $\frac{w_c}{1-w_c} = \frac{\alpha+\kappa-1}{n(\beta-\log h)}$. It is evident that a posteriori, c depends only on the number of clusters and the sample size.

6. The draws from the predictive density $\pi\left(x_{n+1} \mid \boldsymbol{x}\right)$ will be used for density estimation purposes. Specifically, we want to sample

$$x_{n+1} \sim \sum_{j\geq 1} w_j^* K\left(\cdot \mid \theta_j^*\right) \tag{1.44}$$

where the weights and locations $\left(w_j^*, \theta_j^*\right)_{j\geq 1}$ define the posterior random measure $\mathbb{P} \mid \boldsymbol{x}$. To this end, we can use the weights in order to choose a location θ_k^* and then sample $x_{n+1} \sim K\left(\cdot \mid \theta_k^*\right)$, i.e. from the specific component of the infinite mixture of kernels. To specify k, we generate $r \sim \mathcal{U}(0,1)$ and demand

$$\sum_{j=1}^{k-1} w_j < r \leq \sum_{j=1}^{k} w_j, \tag{1.45}$$

with $w_0 = 0$. It is possible that the available weights are inadequate to satisfy Eq. (1.45). In this case we use Eq. (1.7) in order to produce more weights and sample θ_{n+1}^* from P_0.

Sampling from the GSB

When the prior measure is a GSB process, we have that $\mathbb{P} \sim \mathcal{GSB}\left(\lambda, P_0\right)$. By the stick-breaking representation we have that $\mathbb{P} = \sum_{j\geq 1} w_j \delta_{\theta_j}$, but the less complex form of the geometric weights leads to $w_j = \lambda(1-\lambda)^{j-1}, j \geq 1$. We use a GSB mixture (GSBM) modeling approach, so for data $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the infinite sequence of locations $\boldsymbol{\theta} = (\theta_i)_{i\geq 1}$, it is for $i = 1, \ldots, n$:

$$\pi\left(x_i \mid \boldsymbol{\theta}, \lambda\right) = \lambda \sum_{j=1}^{\infty} (1-\lambda)^{j-1} K\left(\cdot \mid \theta_j\right) \tag{1.46}$$

Again, we need to solve the problem of the infinite number of locations, using an appropriate augmentation scheme. In particular, we need to assign to each observation x_i a pair $(d_i, \mathbb{A}_i)$ of a clustering variable and a random set respectively, so that the conditional distribution of $d_i \mid \mathbb{A}_i$ will be discrete uniform over $\mathbb{A}_i$, $i = 1, \ldots, n$.

We note that the random sets $\mathbb{A}_i$ in the case of a DPM described above, typically are sets of integers with gaps, e.g. $\{5, 220, 530, 780, 900\}$. The difference with the GSBM slice sampler is that the simpler form of the weights allows us to construct *sequential* random sets (i.e. with no gaps), thus leading to efficient samplers with lower mean execution times.

Specifically, following [FGMW10] we introduce the random variables $N_i \overset{\text{iid}}{\sim} \pi_N\left(\cdot\right), i = 1, \ldots, n$ such that $\mathbb{A}_i = \{1, \ldots, N_i\}$ and the conditional distribution of each observation given N_i will

be an equally weighted finite mixture of kernels

$$\pi\left(x_i \,|\, N_i\right) = \frac{1}{N_i} \sum_{j=1}^{N_i} K\left(\cdot \,|\, \theta_j\right) \tag{1.47}$$

while for the clustering variables we have

$$\pi\left(d_i \,|\, N_i\right) = \mathcal{DU}\left\{d_i \,|\, \mathbb{A}_i\right\} = \frac{1}{N_i} \mathcal{I}\left(d_i \leq N_i\right) \tag{1.48}$$

and the augmented joint density of x_i with (d_i, N_i)

$$\begin{aligned} \pi\left(x_i, d_i = k, N_i = N\right) &= \pi_N(N) \pi\left(d_i = k \,|\, N_i = N\right) \pi\left(x_i \,|\, d_i = k, N_i = N\right) \\ &= \pi_N(N) \frac{1}{N} \mathcal{I}\left(k \leq N\right) K\left(x_i \,|\, \theta_k\right). \end{aligned} \tag{1.49}$$

The validity of the proposed augmentation is illustrated by the marginal distribution of x_i, which remains unaffected, as we have that

$$\pi(x_i) = \sum_{k=1}^{\infty} \underbrace{\frac{1}{N} \sum_{N=k}^{\infty} \pi_N(N)}_{w_k} K\left(\cdot \,|\, \theta_k\right). \tag{1.50}$$

It is now clear that the *ordered* (decreasing) weights $w_k = \frac{1}{N} \sum_{N=k}^{\infty} \pi_N(N)$ depend on the choice of $\pi_N(\cdot)$. If we choose a Negative Binomial distribution $\pi_N(N_i) = \mathcal{NB}\left(N_i \,|\, 2, \lambda\right)$ then we recover the familiar geometric weights, as:

$$w_k = \sum_{N=k}^{\infty} \frac{1}{N} N \lambda^2 (1-\lambda)^{N-1} = \lambda (1-\lambda)^{k-1}. \tag{1.51}$$

Then, the complete posterior distribution for the augmented GSBM model will have the following form

$$\pi\left(\boldsymbol{w}, \boldsymbol{\theta} \,|\, \boldsymbol{x}, \{N_i\}_{i=1}^n, \{d_i\}_{i=1}^n\right) \propto \pi\left(\boldsymbol{w}, \boldsymbol{\theta}\right) \prod_{i=1}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \mathcal{I}(d_i \leq N_i) K(x_i \,|\, \theta_{d_i}). \tag{1.52}$$

and only a finite number of weights, namely $N^* = \max_{1 \leq i \leq n} N_i$, will be required at each iteration of the Gibbs sampler. We complete the description of the slice sampler with the FC distributions. At each iteration we have to update the sets of latent variables $\{N_i\}_{i=1}^n, \{d_i\}_{i=1}^n$ and sample the (finite number of) weights and locations that characterize the draws from the posterior random measure.

1. The θ_j-locations are sampled from

$$\pi\left(\theta_j \,|\, \ldots\right) = P_0(\theta_j) \prod_{d_i = j} K\left(x_i \,|\, \theta_j\right) \tag{1.53}$$

2. The clustering variables are sampled from the discrete distribution

$$\pi\left(d_i \mid \ldots\right) = \mathcal{I}\left(d_i \leq N_i\right) K\left(x_i \mid \theta_{d_i}\right) \tag{1.54}$$

3. Regarding the construction of the slice sets $\mathbb{A}_i$, we only need to sample $N_i, 1 \leq i \leq n$ from the truncated Geometric FC distributions

$$\pi\left(N_i \mid \ldots\right) \propto (1-\lambda)^{N_i - 1} \mathcal{I}\left(N_i \geq d_i\right) \tag{1.55}$$

4. Assigning a conjugate Beta prior $\lambda \sim \mathcal{B}e\left(\alpha, \beta\right)$ over the geometric probability we have

$$\pi\left(\lambda \mid \ldots\right) = \mathcal{B}e\left(\lambda \mid \alpha + 2n, \beta + \sum_{i=1}^{n} N_i - n\right) \tag{1.56}$$

5. The geometric weights $w_j, j = 1, \ldots, N^*$ are constructed by

$$w_j = \lambda(1-\lambda)^{j-1} \tag{1.57}$$

6. The draws from the predictive density $\pi(x_{n+1} \mid \boldsymbol{x})$ are performed in the same way as in the DPM slice sampler, namely using the Eqs. (1.44-1.45).

The above slice sampling schemes are essential for the understanding of the methods that will be constructed in the following Chapters. We complete this Chapter with an illustrative density estimation example, using both DPM & GSBM models.

Density estimation using DPM & GSBM models

In order to illustrate the performance of the slice samplers described above, we apply them for the purpose of density estimation. Specifically we choose Gaussian Kernels $K\left(\cdot \mid \vartheta\right) = \mathcal{N}\left(\cdot \mid \vartheta\right)$, with $\vartheta = \left(\mu, \tau^{-1}\right) \in \mathbb{R} \times \mathbb{R}^+$ and use a sample $\boldsymbol{x} = (x_1, \ldots, x_{250})$ from the Normal mixture

$$\pi(x) = \frac{1}{3}\mathcal{N}(x \mid -4, 1) + \frac{1}{3}\mathcal{N}(x \mid 0, 1) + \frac{1}{3}\mathcal{N}(x \mid 6, 1). \tag{1.58}$$

For both models we have the following prior specification:

$$c \sim \mathcal{G}(\alpha_c, \beta_c), p \sim \mathcal{B}e(\alpha_\lambda, \beta_\lambda), \{(\mu_j, \tau_j) \sim \mathcal{N}(m, v^{-1})\mathcal{G}(a, b) : j \geq 1\},$$

with $(\alpha_c, \beta_c, \alpha_\lambda, \beta_\lambda, m, v, a, b) = (0.1, 0.1, 1, 1, 0, 10, 0.5, 0.5)$. For the complete slice sampler, we have to sample from the FC distribution of $\vartheta = \left(\mu, \tau^{-1}\right)$, which we set to the semi–conjugate independent Normal-Gamma prior

$$\pi\left(\mu, \tau\right) = \mathcal{N}\left(\mu \mid \frac{\tau_j s_j}{n_j \tau_j + v}, \frac{1}{n_j \tau_j + v}\right) \mathcal{G}\left(\tau \mid a + \frac{1}{2} n_j, b + \frac{1}{2} \sum_{d_i = j} \left(x_i - \mu_j\right)^2\right), \tag{1.59}$$

with $s_j = \sum_{d_i=j} x_i$ and $n_j = \sum_{d_i=j} 1$.

We ran the chains for 30,000 iterations with a burn-in period of 5,000. As a measure of distance between the true density π and the estimated densities $\hat{\pi}_{\text{DP}}, \hat{\pi}_{\text{GSB}}$ via the DPM and GSBM models, respectively, we use the Hellinger distance, defined as

$$\mathcal{H}(f,g) = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}x \tag{1.60}$$

for densities f and g.

In Fig. 1.4(a) we present the histogram of the data, superimposed with the kernel density estimators (KDEs) obtained by the posterior predictives of the DPM and GSBM models. In Fig. 1.4(b) we display the Hellinger distance over the sampling iterations after burn-in.

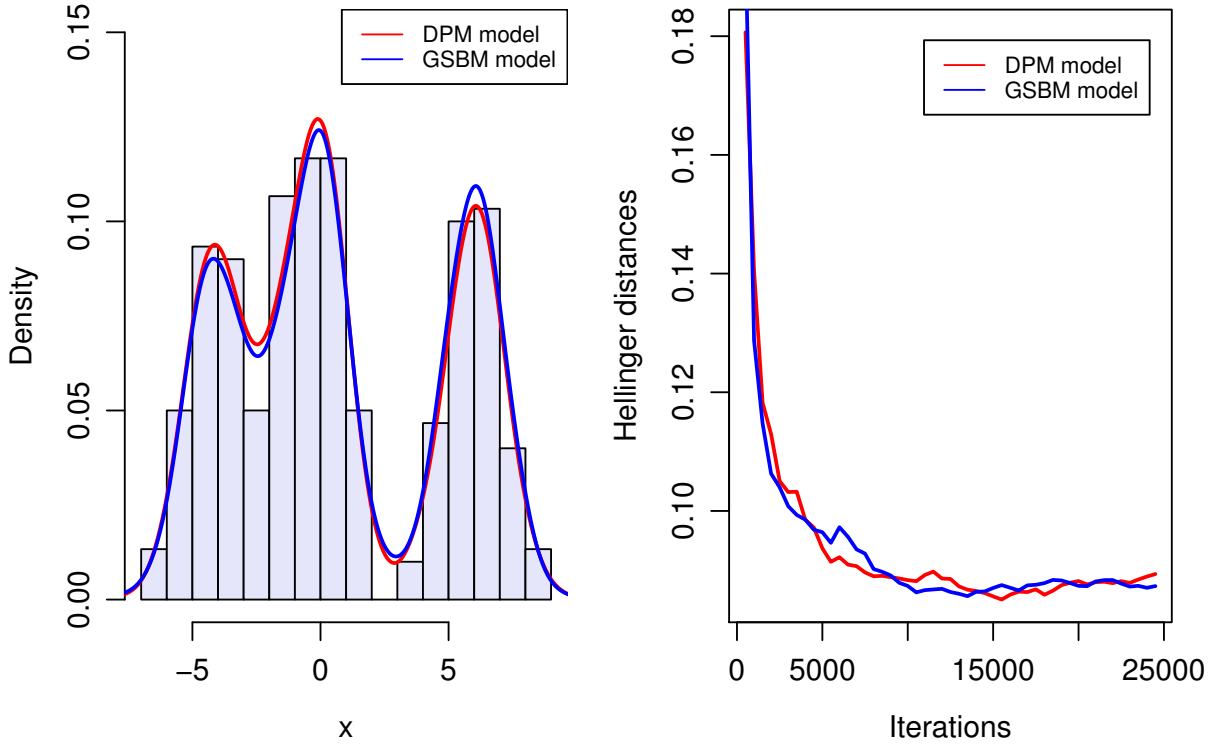

Figure 1.4: Left: Data histogram and KDE's obtained by the posterior predictive samples obtained DPM (red) GSBM (blue) models. Right: Evolution of the Hellinger distances $\mathcal{H}(\pi, \hat{\pi}_{\text{DP}})$ (red) and $\mathcal{H}(\pi, \hat{\pi}_{\text{GSB}})$ (blue) over sampling iterations after the burn-in period.

We note that the estimated densities are almost indistinguishable, with Hellinger distances $\mathcal{H}(\pi, \hat{\pi}_{\text{DP}}) = 0.089$ and $\mathcal{H}(\pi, \hat{\pi}_{\text{GSB}}) = 0.087$. Due to the less complex form of the consecutive random slice sets, the GSBM model is characterized by lower MET, compared to the corresponding DPM model. We will further analyze compare the two models on the estimation of chaotic dynamic equations in Chapter 3.

Chapter 2

Dynamical systems

In this Chapter we give the definitions and basic properties of dynamical systems, for the deterministic and the stochastic case. The concept of chaos is rigorously analyzed along with the properties of chaotic systems, homoclinic tangencies and invariant manifolds. We continue with the presentation of the most popular in the literature noise-induced effects, as well as with the effects of dynamical noise on non-hyperbolic nonlinear maps. We conclude with an explanation of the parametric reconstruction inefficiency under non-Gaussian noise processes. This gives us the main motivation for the methods included in the next Chapters. For extended and rigorous presentations of the Theory of Deterministic Dynamical Systems and Chaos we refer to [GL02, ASY96, Ott02, SJ06].

2.1 Deterministic Dynamical systems

A dynamical system can be considered as a *deterministic* mathematical description for the evolution of a system over time. We can give the following rigorous definition.

Definition 2.1 (Dynamical system)**.** *A deterministic dynamical system is a tuple* $(\mathbf{X}, \mathcal{T}, f)$ *consisting of a state space* $\mathbf{X}$*, a set* $\mathcal{T}$ *describing time and an evolution function* $f : \mathbf{X} \times \mathcal{T} \mapsto \mathbf{X}$*, such that:*

1. $f_0 = \mathbf{1}_{\mathbf{X}}$,
2. $f_{s+t} = f_s \circ f_t, \forall s, t \in \mathcal{T}$

where $f_t := f(\cdot, t)$ *and* $\mathbf{1}_{\mathbf{X}}$ *the identity function over* $\mathbf{X}$*.*

Usually the time parametrization is defined over $\mathcal{T} = \mathbb{R}$ and the dynamical system is called *continuous* with a *flow* f, or over $\mathcal{T} = \mathbb{Z}$ resulting in a discrete dynamical system with *map* f. If the system is not explicitly depended on time it is called *autonomous*, otherwise it is called non-autonomous. Moreover, the evolution function most of the times depends on some *control parameter* $\vartheta \in \Theta \subseteq \mathbb{R}^d$, controlling the asymptotic dynamical behavior. However the dynamical behavior of large classes of seemingly different dynamical systems can be essentially the same. In this case, the dynamical systems are called *conjugate*.

Definition 2.2. *The dynamical systems* $(\mathbf{X}, \mathcal{T}, f)$, $(\mathbf{Y}, \mathcal{T}, g)$ *are conjugate if there exists homeomorphism* $h : \mathbf{X} \mapsto \mathbf{Y}$ *called topological conjugacy, such that*

$$h \circ f = g \circ h \tag{2.1}$$

We are typically interested in studying the orbits of a given dynamical system, which in the case of a discrete dynamical system are the successive functional iterations of the evolution map, starting from an initial point $x_0 \in \mathbf{X}$, i.e.

$$\mathcal{O}_f(x_0) = \{f^n(x_0) : n \geq 0\} \tag{2.2}$$

with $f^n := \underbrace{f \circ f \circ \ldots \circ f}_{n\text{-times}}$ and $x_n = f^n(x_0)$. If $f(x) = x$, x is a *fixed point* of f, while if $f^k(x) = x$, such that $f^r(x) \neq x$ for $1 \leq r < k$, x is a *periodic point* of (minimal) period k, while $\mathcal{O}_f(x) = \{x, \ldots, f^{k-1}(x)\}$ is the associated *periodic orbit*. Naturally, topological conjugacy is associated with invariant objects, such as the spectrum of periodic orbits. In this thesis, we are mainly interested in discrete dynamical system. We note that any continuous dynamical system can be reduced to a discrete dynamical system using the technique of *Poincaré sections* [CE07], i.e. examining intersections of the flow with proper hyperplanes.

For the sake of completeness, we will mention some definitions which will proved useful in the sequel.

Definition 2.3. *If* p *periodic point with minimal period* k*, then if* $\frac{\mathrm{d}}{\mathrm{d}x} f^k(p) \neq 1$ p *is hyperbolic. Moreover,* p *is attracting if* $\frac{\mathrm{d}}{\mathrm{d}x} f^k(p) < 1$*, or repelling if* $\frac{\mathrm{d}}{\mathrm{d}x} f^k(p) > 1$*. Every attracting set has an associated basin of attraction, i.e. the set of the initial conditions converging to it.*

Definition 2.4. *A point* $x \in \mathbf{X}$ *is forward asymptotic to* p *if:*

1. $\lim_{n\to\infty} f^{kn}(x) = p$, $\quad p$ *is a periodic point of period* k*.*
2. $d(f^n(x), f^n(p)) \xrightarrow{n\to\infty} 0$, p *is non-periodic.*

The stable set of p *is the set* $\mathcal{W}^s(p) = \{x : x \text{ forward asymptotic to } p\}$*. If* f *is invertible the unstable set is also defined, as* $\mathcal{W}^u(p) = \{x : x \text{ backward asymptotic to } p\}$*, with backward asymptotic meaning forward asymptotic in reversed time* $n \to -\infty$*.*

We will further analyze the stable and unstable asymptotic sets for the case of higher-dimensional maps, in section 2.3.

Definition 2.5. *The set* $A \subseteq \mathbf{X}$ *is invariant under* f *if* $f(A) \subseteq A$*.*

Whenever an orbit enters an invariant set, it will remain in the set forever. Invariant sets can be further categorized in positively or negatively invariant, when the orbits are contained in the set in forward ($t > 0$) or reversed ($t < 0$) time.

2.1.1 Chaos in Dynamical Systems

If the evolution function is *nonlinear*, we may have the occurrence of complex dynamical behavior known as *chaos*. There exist various definitions [LY75, HO15] and generalizations [Li93, SS94] for the notion of Chaos in the literature. Following Devaney [Dev08] we provide some useful definitions.

Definition 2.6. *f is topologically transitive if*

$$\forall A, B \subseteq \mathbf{X}, \exists k > 0 : f^k(A) \cap B \neq \emptyset.$$

The intuition behind topological transitivity is that orbits starting at some $x_0 \in A$ will eventually visit any arbitrarily small neighborhood B of any other point $y \in \mathbf{X}$. So, there are no orbits that can be "trapped" forever in some $A \subset \mathbf{X}$. We also note that topological transitivity is equivalent to the existence of a dense orbit in $\mathbf{X}$, i.e. there is $x \in \mathbf{X} : \overline{\mathcal{O}(x)} = \mathbf{X}$, where $\overline{\mathcal{O}(x)}$ denotes the closure of the orbit $\mathcal{O}(x)$ of x.

Definition 2.7. *f exhibits sensitive dependence on initial conditions (SDIC) if*

$$(\exists \delta > 0)(\forall \epsilon > 0, x \in \mathbf{X})(\exists y \in N_\epsilon(x), n \in \mathbb{N}) : d(f^n(x), f^n(y)) > \delta$$

where $d(\cdot, \cdot)$ is a metric over $\mathbf{X}$ and $N_\epsilon(x)$ an $\epsilon-$neighborhood of x.

The property of sensitive dependence on initial conditions implies that there exist orbits of arbitrarily close initial conditions which will eventually diverge and this fact is highly important in terms of numerical computation as this will cause round–off errors to be magnified. Furthermore, sensitive dependence on initial conditions is closely related with the notion of *unpredictability*. Systems that exhibit sensitive dependence on initial conditions although they possess a deterministic evolution mechanism, are inherently unpredictable, in the sense that the limited knowledge of their initial condition due to finite arithmetic precision will lead to substantially erroneous calculation of the future states. A way to quantify the sensitive dependence on initial conditions is the *Lyapunov exponent*, defined as

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \log \left| \frac{\mathrm{d}}{\mathrm{d}x} f^k(x_0) \right| = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \log \left| f'(x_k) \right|, \tag{2.3}$$

which is the average exponential rate of divergence of nearby orbits and is independent of the initial condition [Ose68]. Positive Lyapunov exponents are associated with sensitive dependence on initial conditions as arbitrarily close initial points evolve in orbits separated by an exponentially growing distance $\Delta x_n \sim e^{\lambda n} \Delta x_0$. There exist also, *local* Lyapunov exponents [EFT91] associated with the local behavior of the nearby orbits separation. In more than one dimensions, $d > 1$, we have the existence of d-Lyapunov exponents [Str18], at least one positive of which is an indication of sensitive dependence on initial conditions. Lyapunov exponents are also related with the predictability of the dynamical system, vie the *Lyapunov time*, the inverse Lyapunov exponent, which determines the time horizon until the behavior of the dynamical system becomes unpredictable [Str18].

We can now give the definition of a chaotic dynamical system:
Definition 2.8 (Devaney [Dev08]). *$(\mathbf{X}, \mathcal{T}, f)$ is chaotic if:*

1. *f is topologically transitive,*
2. *periodic points are dense in $\mathbf{X}$,*
3. *f exhibits sensitive dependence on initial conditions.*

The above chaos definition is strong, in the sense that it implies several other definitions such as the Li-Yorke or that of topological chaos. Regarding the necessity of all three conditions it has been shown that sensitive dependence on initial conditions is implied by transitivity and the existence of a dense periodic orbit [AG92], while whenever $\mathbf{X}$ is an interval $I \subset \mathbb{R}$, then topological transitivity alone implies both sensitive dependence on initial conditions and existence of a dense set of periodic points [VB94]. For a recent survey of equivalent forms of Devaney's definition see [WH13].

Summarizing, a chaotic dynamical system is characterized by deterministic evolution of bounded, aperiodic orbits, which exhibit sensitive dependence on initial conditions. As a result, chaotic motion is unpredictable, and it seems erratic or even random while the underlying evolution mechanism is purely deterministic. Chaos Theory keeps advancing [TKR+10, BDV06] and has met wide applicability in a variety of sciences [Str18] such as physics [BR05], biology [Cof98], economics [PPP94], or medicine [Wes12].

For continuous time autonomous systems, the occurrence of chaotic behavior requires that the dimension of the state space is at least 3, while in the discrete case chaos arises even in one-dimensional systems. We give a famous example of one-dimensional chaotic map, the *logistic* map [May76] with $\mathbf{X} = \mathbb{R}$, and evolution function

$$f(\mu, x) = \mu x(1 - x). \tag{2.4}$$

The logistic map has been widely used as an ecological model of predator-prey systems. In order to qualitatively examine the dynamical behavior of the logistic map for the varying values of the control parameter $0 < \mu \leq 4$, we will use a *bifurcation diagram*. Bifurcations are qualitative changes in the dynamical behavior of a dynamical system, appearing in a variety of ways. A period-doubling bifurcation leads to the creation of a periodic orbit with double period than the original one. One of the most common routes to chaos is the *period-doubling* cascade, i.e. an infinite sequence of period doubling bifurcations, as illustrated in Fig. 2.1 for the case of logistic map. The period-doubling cascade has universal properties [Cvi17] and characterizes the whole family of *unimodal* maps. It has been proved [Fei78] that the ratio of parameter values associated with a stable 2^n periodic orbit is

$$\delta = \lim_{n \to \infty} \frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_n} = 4.669... \tag{2.5}$$

where μ_n is the point of bifurcation and δ the *Feigenbaum constant* appearing in a large class of both theoretical systems and experimental data

The map $f(\mu, x)$ as the control parameter μ varies, undergoes a sequence of period doubling bifurcations. Every $k-$period stable periodic orbit loses its stability, giving rise to a new stable $2k-$periodic orbit. Moreover, inside the chaotic region there exist infinitely many periodic windows, corresponding to finite length $\mu-$intervals.

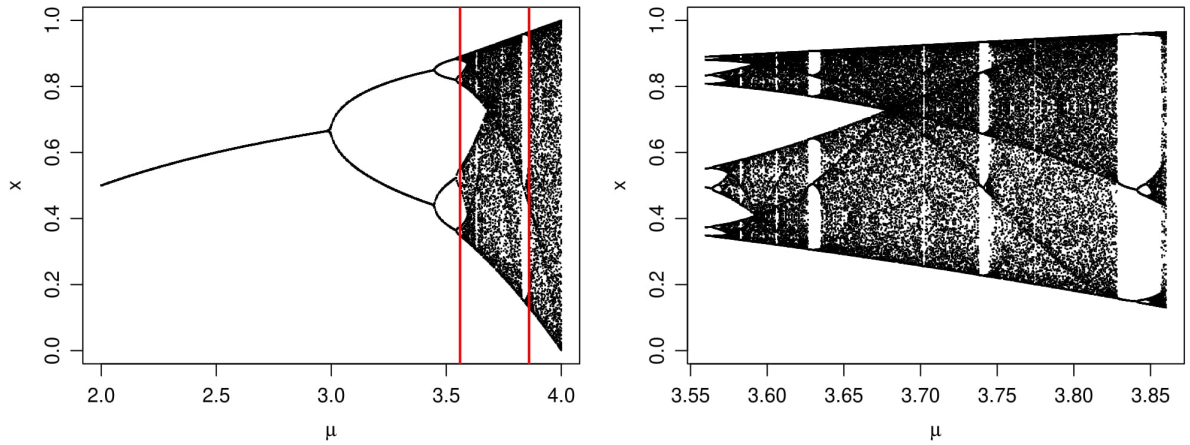

Figure 2.1: Bifurcation diagram for the logistic map. Left: $0 < \mu < 4$ Right: $3.56 < \mu < 3.86$. Red lines indicate values 3.56, 3.86.

In Fig. 2.1 we can see the period doubling cascade and the periodic windows arising as finite subintervals of the parameter range. For values of the control parameter that the dynamical system is chaotic we have the existence of an infinity of periodic orbits, which are all unstable.

2.1.2 Statistical Properties of Deterministic Systems

The irregular seemingly random behavior of chaotic orbits, led early to the connection of chaos theory with probability and statistics. The unpredictable nature of chaos is related with the existence of a certain Borel probability measure associated with the dynamic system f, such that it remains unchanged when f acts on it. This leads to the fact that although the point estimation of future states becomes more and more inaccurate as the time horizon increases we are able to make probabilistic statements regarding the long-term behavior of the orbits. This is the main goal of the Ergodic Theory [PY98] approach to the analysis of dynamical system. For an extensive analysis of the probabilistic properties characterizing deterministic dynamical system, we refer to [LM85].

In Fig. 2.2 we provide an illustration of the way that densities are associated with orbits of a dynamical system. We produce 2 orbits of the logistic map in the chaotic region, with control parameter $\mu = 3.95$, with length $n = 50,000$. The initial conditions of the two orbits are very close, namely $x_0 = 0.3$ (blue) and $y_0 = 0.30001$ (red). In Fig. 2.2(a) we present the first 100 map iterations of x_0 and y_0. The SDIC property is clearly depicted as the two orbits are initially almost identical, but after approximately 15 iterations they appear to be completely different. Of course, they are not *completely* different, as we can see in Fig. 2.2(b) where the qq-plot of the empirical densities is depicted, clearly indicating that the two orbits share the same *histogram statistics*, i.e. they visit the various subsets of the state space with the same probability. This fact

is further illustrated by the indistinguishable histograms of the two orbits in Figs. 2.2(c) (x_0) - 2.2(d) (y_0).

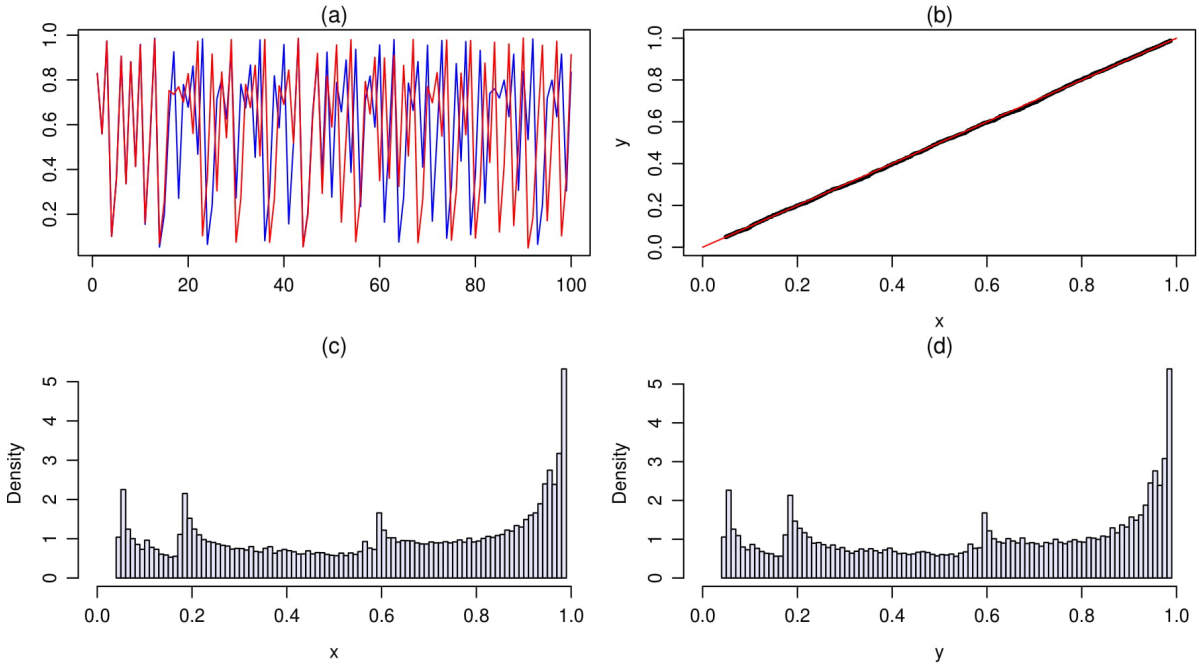

Figure 2.2: (a) First 100 iterations of the initial conditions x_0 (blue) and y_0 (red). (b) qq-plot of the empirical distributions of the 2 orbits, x and y. (c),(d) Histograms of the orbits x and y, respectively.

The above description essentially indicates the existence of a measure that remains invariant under the effect of the map. The following two definitions are required for a more rigorous statement regarding the long-term behavior of a dynamical system.

Definition 2.9. *Let $(\mathbf{X}, \mathcal{X}, \mu)$ a measure space. A map $f : \mathbf{X} \mapsto \mathbf{X}$ is called measurable if*

$$f^{-1}(A) \in \mathcal{X}, \forall A \in \mathcal{X}, \tag{2.6}$$

where $f^{-1}(A) = \{x \in \mathbf{X} : f(x) \in A\}$. Then, the measure μ is invariant for f (or $f-$ invariant) if

$$\mu\left(f^{-1}(A)\right) = \mu(A), \forall A \in \mathcal{X}. \tag{2.7}$$

The asymptotic behavior of a dynamical system is naturally related with the existence of invariant measures. We are often especially interested in invariant (probability) measures that are absolutely continuous with respect to the Lebesgue measure, provided they exist under general conditions [LY73]. Having derived an invariant probability measure μ, we can make probabilistic prediction arguments for the long term behavior of the system in the sense that $P\{x_i \in A\} = \mu(A)$, for $A \in \mathcal{X}$.

Definition 2.10. *The Frobenius–Perron operator $P_f : \mathcal{L}^1 \mapsto \mathcal{L}^1$ is defined by*

$$\int_A P_f g(x)\mu(\mathrm{d}x) = \int_{f^{-1}(A)} g(x)\mu(\mathrm{d}x), \forall A \in \mathcal{X}, g \in \mathcal{L}^1 \tag{2.8}$$

with L^1 the space of μ-integrable functions. P_f is also a Markov operator, as it satisfies $g \geq 0 \Rightarrow P_f g \geq 0$ and $\|P_f g\| = \|g\|, \forall g \in \mathcal{L}^1$, with $\|\cdot\|$ the $\mathcal{L}^1$–norm.

We may consider the Frobenius–Perron operator as the mechanism describing how densities *evolve* after applying the map f on them. Moreover, considering P as a Markov operator, a fixed point g, i.e. $Pg = g$, is a stationary density of P [LM85]. Notably, the derivation of the invariant measures is generally a difficult task, so their approximations are useful in providing intuition about the asymptotic behavior of the dynamical system. Usually, we need to provide approximations for a special invariant measure, called *natural measure*, as it assigns to each $A \in \mathcal{X}$ the probability with which typical orbits visit A. Specifically, the natural (or physical) measure ν_f is defined as

$$\nu_f(A) = \lim_{r\to 0} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(f^i(x_0) \in N(r, A)\right), A \in \mathcal{X}, \tag{2.9}$$

where $N(r, A) = \{x \in \mathbf{X} : \text{dist}(x, A) < r\}$, if the limit exists for almost all initial conditions, i.e. for all $x_0 \in \mathbf{X}$ except perhaps a set of zero measure.

For some modern applications of the theory and methods of invariant measures we refer to [DZ10].

2.1.3 Higher-Dimensional Deterministic Systems

When the state space is high-dimensional $\mathbf{X} \subseteq \mathbb{R}^d, d \geq 2$ we have d Lyapunov exponents, measuring the average exponential rate of divergence along the d orthogonal directions. In order to have the emergence of chaotic behavior, the greatest Lyapunov exponent has to satisfy $\lambda_{\max} = \max_{1\leq i\leq d} \lambda_i > 1$.

The dynamical behavior of the dynamical system is related to the Jacobian matrix of the map f, denoted by $\mathcal{J}_f^n := \left(\frac{\partial f^n}{\partial x_i}\right)_{1\leq i\leq k}$, where f^n is the $n-$th iterate of f.

Definition 2.11. *A fixed ($k-$periodic) point p is called hyperbolic, if $\mathcal{J}_f(p)$ ($\mathcal{J}_f^k(p)$) has no eigenvalues on the unit circle. Specifically:*

1. *If all the eigenvalues are smaller than 1, p is attracting.*
2. *If all the eigenvalues are greater than 1, p is repelling.*
3. *Otherwise, p is a saddle.*

If the areas of the state space are preserved under the action of the map the dynamical system is called *conservative* and $\det \mathcal{J}_f = 1$, whereas if the areas are not preserved we have a *dissipative* dynamical system with $\det \mathcal{J}_f < 1$. In this thesis we are interested in dissipative dynamical system which may have attractors in contrast with area-preserving maps.

Definition 2.12. *([ASY96]) The forward limit set of an orbit $\mathcal{O}_f(x_0)$ is the set*

$$\omega(x_0) = \{x \in \mathcal{X} : (\forall \epsilon > 0, m \in \mathbb{N})(\exists n > m) : x \in N_\epsilon(f^n(x_0))\} \tag{2.10}$$

A forward limit set with associated basin of attraction of non-zero measure is called an attractor.

In a similar fashion, but in reversed time, we can define the backward limit set. Moreover, if $\mathcal{O}_f(x_0)$ is a chaotic orbit with $x_0 \in \omega(x_0)$, then $\omega(x_0)$ is called a chaotic set. A *chaotic attractor* is an attractor which is also a a chaotic set. Attractors can have a simple form such as fixed or periodic orbits, but they can also have a highly complex geometry. This is the case of chaotic (or strange) attractors [ER85]. When more than one different attractors [1] coexist, then the DS exhibits *multistability* [Feu08]. The boundaries between the basins of attraction of multistable systems can be highly complex, especially in cases where the number of the coexisting attractors is large [FGPY98].

In the following, we assume that the evolution function is a diffeomorphism, that the partial derivatives of both f and f^{-1} exist and are continuous. Then, we introduce the tangent bundle at $x \in \mathbf{X}$, which is the collection of the tangent spaces of the state space at the given point:

$$T_{\mathbf{X}} = \bigcup_{x \in \mathbf{X}} T_x\mathbf{X} \tag{2.11}$$

For every vector $u \in T_x\mathbf{X}$, we have that $\mathcal{J}_f u \in T_{f(x)}\mathbf{X}$ with $\mathcal{J}_f u$ and we can associate any point $x \in \mathbf{X}$ with the stable and unstable vectors at x.

Definition 2.13. *([CE07])A vector $u \in T_x\mathbf{X}$ is called:*

1. *Stable, if $(\exists \lambda < 1, c < \infty)$ such that $\forall n \in \mathbb{N} : \|\mathcal{J}_{f^n} u\|_{T_{f^n(x)}\mathbf{X}} \leq c\lambda^n \|u\|_{T_x\mathbf{X}}$.*
2. *Unstable, if $(\exists \lambda < 1, c < \infty)$ such that $\forall n \in \mathbb{N} : \|\mathcal{J}_{f^n} u\|_{T_{f^{-n}(x)}\mathbf{X}} \leq c\lambda^{-n} \|u\|_{T_x\mathbf{X}}$.*

If all the tangent vectors at a fixed point p are stable with the same λ, p is a stable fixed point, whereas if there exist at least one unstable tangent vector, p is called an unstable fixed point.

Moreover, p is a *hyperbolic* fixed point, when there is a unique factorization of the tangent space as the direct sum of the stable and unstable subspaces E_p^s and E_p^u respectively, so that we can write:

$$T_p\mathbf{X} = E_p^s \oplus E_p^u \tag{2.12}$$

where E_p^s and E_p^u are spanned by the exponentially contracted and expanded vectors under the iterates of the Jacobian.

Definition 2.14. *([CE07]) A set $X^\star \subseteq \mathbf{X}$ is uniformly hyperbolic, if*

1. *$f(X^\star) \subseteq X^\star$*
2. *$T_x\mathbf{X} = E_x^s \oplus E_x^u, \forall x \in X^\star$*
3. *Vectors in E_x^s (E_x^u) are stable (unstable) as in Definition 2.13, with λ independent of x, $\forall x \in X^\star$.*

Then, the map f is hyperbolic on $X^\star$.

The property of hyperbolicity -main feature in a large class of systems called Axiom A- is very important, as it allows the derivation of general theoretical results regarding the shadowing property [PS17], structural stability [Rob77], and the existence of natural invariant measure

[1] There even exist systems with an infinity of attractors [PV94]

[Sin72]. Furthermore, non-hyperbolic systems react different to the effects of noise perturbations than the hyperbolic ones.

Definition 2.15. *Let $x \in \mathbf{X}$ be a saddle fixed point and f invertible. The local stable and unstable manifolds of x for a given neighborhood $N_\delta(x)$ of x are defined respectively as:*

$$W^s_{loc}(x) = \left\{ y \in \mathbf{X} : \lim_{n\to\infty} d(f^n(x), f^n(y)) \to 0, \text{ with } f^n(y) \in N_\delta(x), \forall n \in \mathbb{N} \right\} \tag{2.13}$$

$$W^u_{loc}(x) = \left\{ y \in \mathbf{X} : \lim_{n\to\infty} d(f^{-n}(x), f^{-n}(y)) \to 0, \text{ with } f^{-n}(y) \in N_\delta(x), \forall n \in \mathbb{N} \right\} \tag{2.14}$$

The local manifolds $W^s_{loc}(x)$ and $W^s_{loc}(x)$ are invariant under f and f^{-1} respectively and can be used in order to construct the global invariant manifolds, which contain all points of the state space that converge to x in forward (stable) or backward (unstable) time. Specifically we can write:

$$W^s(x) = \left\{ y \in \mathbf{X} : \lim_{n\to\infty} d(f^n(x), f^n(y)) \to 0 \right\} = \bigcup_{n\geq 1} f^{-n}\left(W^s_{loc}(x)\right) \tag{2.15}$$

$$W^u(x) = \left\{ y \in \mathbf{X} : \lim_{n\to\infty} d(f^{-n}(x), f^{-n}(y)) \to 0 \right\} = \bigcup_{n\geq 1} f^{n}\left(W^u_{loc}(x)\right) \tag{2.16}$$

The above definition can be extended for the case of $k-$periodic orbits or hyperbolic sets in general.

For a planar diffeomorphism $f : \mathbb{R}^2 \to \mathbb{R}^2$, the *Stable Manifold Theorem* guarantees that both the $W^s(x)$ and $W^u(x)$ are one-dimensional manifolds that contain x and moreover that at x they are tangent to the stable E^s_x and unstable E^s_x subspaces, respectively.

Let Λ be a hyperbolic set and $x, y \in \Lambda$. If there exists a point $q \in (W^s(x)\backslash\Lambda) \cap (W^u(y)\backslash\Lambda)$ then q is called *homoclinic point* and -due to the invariance of the manifolds- implies the existence of infinitely many other homoclinic points, thus the emergence of chaotic behavior. The angles between the invariant manifolds formed at homoclinic points, are closely related with the hyperbolicity of the dynamical system. Subsequently, if the map f is hyperbolic and homoclinic points exist, then the angles between the invariant manifolds are *bounded away from zero* and there exists an invariant chaotic set where the dynamics are equivalent to Smale's horseshoe [Sma67].

The *tangential* intersections of the two manifolds are called *homoclinic tangencies* (HTs), and their existence is mutually exclusive with the concept of hyperbolicity. If there exist HTs, then whenever the attractor is ergodic there exists a dense set of them [GBP88], meaning that there are infinite regions where the tangent space cannot be decomposed as a direct sum of stable and unstable directions. In fact, all HTs are images or pre-images of certain, more "important" HTs, called *primary homoclinic tangencies*. Primary HTs appear in regions where the sum of the curvatures of the invariant manifolds is minimal (along each sequence of images of HTs [GP91]), thus resulting in "wider" regions of tangential intersections. Formally, let

$$\mathcal{C}^{\delta,\epsilon}_{x^\star} = \{ x \in N_\delta(x^\star) : \sin\theta_{s,u} < \epsilon \} \tag{2.17}$$

the critical region of the HT $x^\star$, i.e. the set of points near $x^\star$ that "surround" the tangency in the sense of almost zero angle $\theta_{s,u}$ between the stable E^s_x and unstable E^u_x directions which is

less say than a predefined $\epsilon > 0$. Due to larger critical regions, primary HTs are the easier to identify, while -although there is a dense set of them- most of the other HTs have critical regions of practically negligible size. The most widely used methods for the identification of regions of (primary) HTs are based on the concept of critical regions [LGYK93] or on the curvature of the invariant manifolds [JK97b]. We will briefly describe the first method, as we will use it to compare our results coming from our method for the stochastic approximation of stable manifolds in Chapter 5.

In order to find regions of HTs, we need to identify the position of points where the stable and unstable manifolds approach tangentially. It is sufficient to determine where the angle between the stable and unstable manifolds approaches 0. It is easier to identify primary HTs, as they are characterized by large critical regions. In particular, we generate a deterministic orbit of the DS of interest. Then at each orbit point, we calculate the *stable* and *unstable directions*, using the Jacobian matrix. In order to do this, from a given point x we generate a forward orbit $\mathcal{O}^+(x) = \{x, f(x), \ldots, f^n(x)\}$ and calculate the inverse Jacobian matrix at each orbit point. Then we choose at random a unit vector and perform successive multiplications with the inverse Jacobian at each orbit point, starting from $f^n(x)$ and going backwards to x, normalizing at each step. The final vector is used to approximate the stable direction at x. The reason for not using the inverse map, is that the orbit will diverge from the original one. Regarding the unstable direction at x, we will use an orbit $\mathcal{O}^-(x_b) = \{x_b, f(x_b), \ldots, f^n(x_b) = x\}$, calculating the Jacobian at each orbit point. Similarly, we multiply the random unit vector with the Jacobian at each point of the stored trajectory, going forward from x_b and ending at x. Finally, we calculate the angle between the 2 directions at the points of the attractor and choose a small predefined angle as a threshold, in order to identify regions of HTs. The above scheme is based on the fact that Jacobian matrix of the forward map rotates a vector towards the unstable direction, while the Jacobian of the inverse map rotates a vector towards the stable direction.

In Fig. 2.3 we present the local stable and unstable directions [2], as well as points of HTs for the Hénon map. We have used an orbit of length $5,000$ for the standard parameters $(\theta_1, \theta_2) = (1.4, 0.3)$ for the Hénon map [Hén76]:

$$(x_n, y_n) = f(x_{n-1}, y_{n-1}) = \left(1 - \theta_1 x_{n-1}^2 + \theta_2 y_{n-1}, x_{n-1}\right). \tag{2.18}$$

There are critical consequences of the property of non-hyperbolicity, or equivalently the existence of HTs, leading to pathological dynamics. Such consequences include the loss of structural stability [GH13] and the shadowing property [HYG87], or the amplification of noise perturbations [JK97b]. Moreover, HTs play a crucial role in the symbolic dynamics formulation, in particular regarding the construction of generating partitions [GK85] and are also related in the deterministic case with regions of higher predictability [HF05] and interesting dynamical phenomena, such as chaotic blue sky catastrophe [PS00] in experimental data. We remark that hyperbolicity depends on the specific control parameters of the evolution function and not on its functional form per se. For example there exist regions in the parameter space of the Hénon map, for which the DS is uniformly hyperbolic [Ara07].

[2] The unstable manifold has been conjectured to be the closure of the attractor.

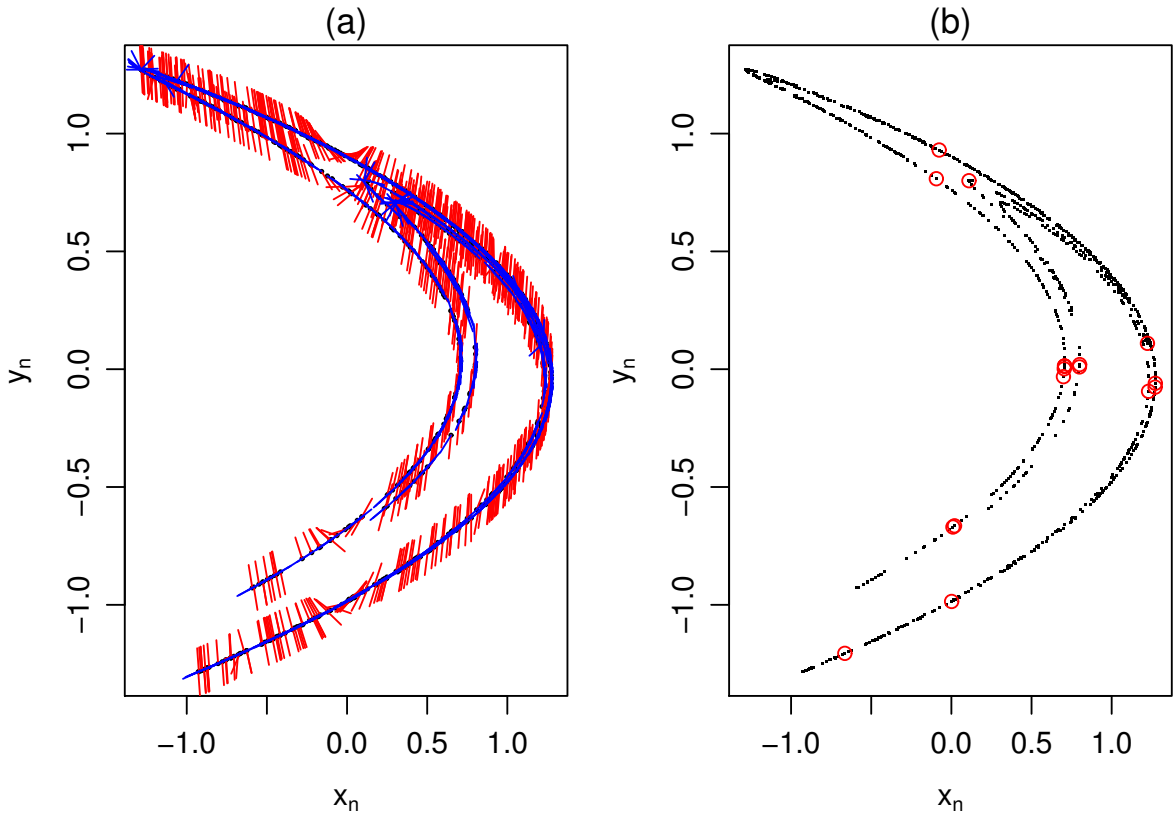

Figure 2.3: (a) Attractor of the Hénon map with the stable (red) and unstable (blue) direction on each orbit point. (b) Same set of points (black), with the primary HTs as centers of the red circles.

Delay-coordinate embedding

Suppose we have observed a scalar time series $y = \{y_i\}_{i=1}^{k}$ consisting of measurements generated from a generally unknown deterministic dynamical system. One of the main pillars of nonlinear time series analysis is the *reconstruction* of the state space, namely the construction of $m-$dimensional vectors $\mathbf{x}_n = \left(y_n, y_{n-\tau}, \ldots, y_{n-(m-1)\tau}\right) \in \mathbb{R}^m$, such that the new dynamics are topologically equivalent [PCFS80] to the true -unknown- dynamics. Of course, the main requirement of this method, called *delay-coordinate embedding* [Tak81], is the estimation of suitable embedding parameters: the embedding dimension m and the delay d. Having estimated m, such that $m > 2d_A$ [SYC91] (d_A is the capacity dimension of the true attractor) then the topological conjugacy is assured. For an extensive analysis of embedding methods and their applications we refer to [KS04].

In Fig. 2.4 we present the delay plot generated by the (x_{n+1}, x_n) embedded points from the $x-$coordinate time series of a Hénon map orbit under standard parameters $(\theta_1, \theta_2) = (1.4, 0.3)$

We will use this result in the next Chapters, in order to investigate dynamical invariants of higher-dimensional maps, having at our disposal scalar time series.

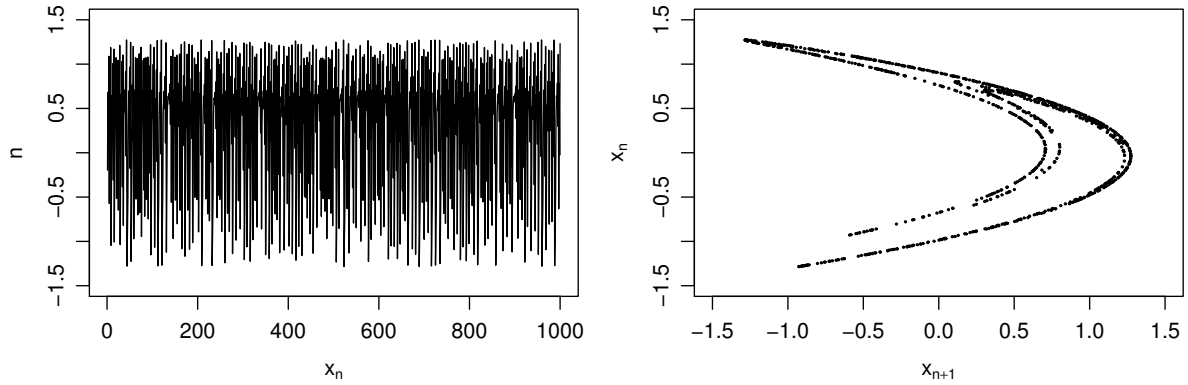

Figure 2.4: Time series of the $x-$coordinate of a Hénon map orbit (left) and the corresponding 2-dimensional (x_{n+1}, x_n) delay-plot (right).

2.2 Random Dynamical systems

In the previous sections we discussed the appearance of uncertainty in purely a deterministic setting, via the the invariant measure over the evolution function of the dynamic system. There is, though, another way that uncertainty comes in to play (thus the extensive literature on probabilistic modeling and statistics [CT13]) in the field of dynamical systems, namely the presence of random noise. In reality, it is often difficult to observe a system that is completely deterministic and even if when this is the case, the finite precision arithmetic alone, can be considered adding random fluctuations between the successive functional iterations of the map. Moreover, measurement devices are generally imperfect and the measurement errors cannot always be considered negligible.

The type of noise contaminating the data is very important, as the various noise types induce different effects on the system. Observational (or measurement noise), originating from errors in the measurement process, is independent of the dynamics and can be thought of as being added after the time evolution of the trajectories under consideration. On the other hand, *dynamical* (or interactive) noise, is added at each step of the time evolution of the trajectories, drastically modifying the underlying dynamics. Dynamical noise can also be used in order to represent the error in the assumed model, thus compensating for a small number of degrees of freedom, for example a small amplitude high dimensional deterministic part not included in the model [KS04]. Due to the different impact of the noise types, the goal of discriminating between measurement and dynamical noise, as well as estimating the associated noise density, is highly significant [HS00, SM08, SKFP04].

Noise can be further categorized as additive or multiplicative. For simplicity, let us consider a 1-d dynamical system $(\mathbf{X}, \mathcal{T}, f)$ and the noise processes $(e_i)_{i\geq 1}, (\zeta_i)_{i\geq 1}$, with $e_i \overset{\text{iid}}{\sim} \pi(\cdot)$, $\zeta_i \overset{\text{iid}}{\sim} q(\cdot)$ and let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the noise corrupted vector of observations, then we have the following categorization:

- Additive observational noise: $x_i = y_i + e_i$, where $y_i = f(y_{i-1})$. Noise is added after the orbit generation.

- Additive dynamical noise: $x_i = f(x_{i-1}) + e_i$. Noise is added at each step of the orbit evolution.
- Multiplicative dynamical noise: $x_i = e_i f(x_{i-1})$. Usually multiplicative noise is applied directly on the control parameters.
- Additive observational and dynamical noise, known as a state-space model:

$$x_i = g(y_i) + e_i$$
$$y_i = f(y_{i-1}) + \zeta_i.$$

Bayesian formulation has been of great use in the general field of noise perturbed dynamical systems. The main idea is the modeling of the states of the system and the control parameters as random variables. Thus we can model the noisy orbit as a stochastic process with a suitable transition kernel $Q(\cdot,\cdot)$ which is determined by the noise process e.g. in the simplest case $Q(x, A) = \mathbb{P}\{x_n \in A \,|\, x_{n-1} = x\}$. Early connections between chaotic systems and Bayesian statistics are discussed in [Ber92], while MCMC methods were used for nonlinear noise reduction by Davies [Dav98]. In [MC00] and [MC01], MCMC methods were applied for the parameter estimation of state-space nonlinear models, extending maximum likelihood-based existing methods [MS99a]. Later, in [SLTB05] a path integral representation was proposed for the likelihood function in order to make inference in stochastic nonlinear dynamics, extended for nonstationary systems in [LSDM08]. In [MNS+01] and [NMKY05] Bayesian methods were suggested for reconstruction and prediction of nonlinear dynamical systems. More recently in [MLMF12], a Bayesian technique was proposed for the prognosis of the qualitative behavior of random dynamical systems under different forms of dynamical noise.

In this Thesis, we are mainly interested in the case of *additive dynamical noise*, using a flexible modeling based on Bayesian nonparametric random measures. In general, dynamical systems subjected to the effects of dynamical noise are called random dynamical systems (RDS). RDS have been rigorously defined and studied, initially by Arnold [Arn13]. We will now proceed with some useful definitions in order to formally define a RDS, as we will use this notion in subsequent Chapters. For an extensive review of the Theory of RDS we refer to [Arn13, Dua15, AJMR06], while [MMP15] provides a recent review of statistical methods concerning RDS.

Essentially, a RDS consists of two parts: the first part is the model for the deterministic DS that is perturbed by a general noise process, while the second part is the modeling of the noise process itself - emphasizing in particular on ergodic noise processes. Regarding the modeling of the noise process, the notion of a *metric DS* is required.

Definition 2.16 ([AJMR06]). *A metric dynamical system* $\left((\Omega, \mathcal{F}, \mathbb{P}), (\theta_t)_{t\in\mathcal{T}}\right)$ *consists of a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and a family of maps* $\{\theta_t : \Omega \to \Omega, t \in \mathcal{T}\}$ *indexed by a set of times* $\mathcal{T}$*, such that:*

1. $(t, \omega) \mapsto \theta_t\omega$ *is* $(\mathcal{B}(\mathcal{T}) \otimes \mathcal{F}, \mathcal{F})$*-measurable, where* $\mathcal{B}(\mathcal{T})$ *the Borel* σ*-algebra generated by* $\mathcal{T}$*.*

2. θ_t *is measure-preserving, i.e.* $\theta_t \mathbb{P} = \mathbb{P}$, *where*

$$\theta_t \mathbb{P}(A) = \mathbb{P}\{\omega : \theta_t \omega \in A\}, \forall A \in \mathcal{F}.$$

Moreover, a measurable set $A \in \mathcal{F}$ *is called* θ*-invariant if* $\theta_t A = A, \forall t \in \mathcal{T}$ *and the metric DS is ergodic under* $\mathbb{P}$ *if*

$$\mathbb{P}(A) \in \{0,1\}, \forall\ \theta\text{-invariant } A \in \mathcal{F}.$$

The measure-preserving property is essential in terms of defining a noise process that is stationary. For example [SH98], let $(\xi_t)_{t\in\mathcal{T}}$ be a càdlàg continuous time stochastic process, i.e. with $\mathcal{T} = \mathbb{R}$ and right-continuous trajectories having limits from the left. We can define a metric DS, using the shifts $\xi(t) \mapsto \theta_\tau \xi(t) = \xi(t+\tau)$. In fact, every stationary stochastic process on $(\mathbf{X}, \mathcal{X})$ can be used to define a measure $\mathbb{P}$ on the product space $(\mathbf{X}^{\mathcal{T}}, \mathcal{X}^{\mathcal{T}})$, so that the associated quadruple $\left((\mathbf{X}^{\mathcal{T}}, \mathcal{X}^{\mathcal{T}}, \mathbb{P}), (\theta_t)\right)$ is a metric DS, with (θ_t) the shift operator. We could also consider a DS as measure-preserving with respect to its natural measure, since it remains invariant under the action of the associated map.

Having defined the metric DS, thus modeling the noise process, we can now proceed with the formal definition of a random dynamical system, covering a large class of DS subjected to stationary stochastic perturbations, both in discrete (stochastic difference equations) or continuous (stochastic differential equations) time.

Definition 2.17. *[Arn13] A random dynamical system with state space* $(\mathbf{X}, \mathcal{X})$ *over a metric DS* $\theta := \left((\Omega, \mathcal{F}, \mathbb{P}), (\theta_t)_{t\in\mathcal{T}}\right)$ *is a* $(\mathcal{B}(\mathcal{T}) \otimes \mathcal{F} \otimes \mathbf{X}, \mathbf{X})$*-measurable mapping*

$$\phi : \mathcal{T} \times \Omega \times \mathbf{X} \mapsto \mathbf{X}, \quad (t, \omega, x) \mapsto \phi(t, \omega, x)$$

forming a cocycle over θ*, i.e.* $\phi(t, \omega) := \phi(t, \omega, \cdot)$*, when* $\forall \omega \in \Omega$ *it is that:*

1. $\phi(0, \omega) = \mathbf{1}_{\mathbf{X}}$, *and*
2. $\phi(t+s, \omega) = \phi(t, \theta_s \omega) \circ \phi(s, \omega), \forall s, t \in \mathcal{T}$.

In the spirit of the above definition, a DS is a special case of a RDS, when ϕ is independent of ω. Moreover, based on the definition of a RDS, the notions of attractor and bifurcation take the generalized forms of random attractors and stochastic bifurcations [SH98]. For an extensive analysis of the analogy between DS and RDS, as well as a survey of the RDS applications, we refer to [Chu04].

As mentioned above, we are interested in the case of additive dynamical noise, meaning that in the simplest case, using the notation of Definition 2.17, we will model our observations $\mathbf{x} = (x_1, \ldots, x_n)$ as

$$x_n = \phi(n, \omega, x) = g_{\tau^n \omega}(x_{n-1}) = f(\vartheta, x_{n-1}) + e_n, \text{ with } x_0 = x \tag{2.19}$$

where $f : \Theta \times \mathbf{X} \mapsto \mathbf{X}$ is a continuous map, depending on a vector of control parameters $\vartheta \in \Theta$ with $e_i \overset{\text{iid}}{\sim} \pi(\cdot)$, and $\pi(\cdot)$ the unknown zero-mean symmetric density, with $e(\tau^n \omega) =$

$e_n(\omega) \equiv e_n$. Subsequently [Arn13], the states of the RDS are iid random variables forming a homogeneous Markov chain $\{X_n = \phi(n, \omega, x_0), n \in \mathbb{N}, \omega \in \Omega\}$ with state space $\mathbf{X}$ and transition probability kernels

$$Q(x, A) = \mathbb{P}\{X_{n+1} \in A \mid X_n = x\} = \mathbb{P}\{\omega : g_{\tau^n\omega}(x) \in A\} \tag{2.20}$$

It is straightforward that the above transition kernels can be generalized for higher order Markovian random processes. In the context of the full reconstruction of a dynamical equation, which it will be discussed in Chapter 3, we aim to estimate the vector of the control parameters, the initial condition x_0, as well as the unknown density of the additive dynamical perturbations which are perhaps non-Gaussian.

2.2.1 Quasi-invariant measures

In analogy with the deterministic case, the long-term dynamical behavior of a RDS is dominated in some sense by its deterministic part. In particular, the inherent unpredictability persists. If the associated deterministic part exhibits chaotic behavior, in terms of prediction we are limited to to a Lyapunov time horizon and the making of probabilistic statements regarding the asymptotic frequency of the system orbits passage for the various subsets of the state space. To this end, we want to obtain the *quasi-invariant measure* of the RDS, formally defined as

$$\tilde{\mu}_f(A) = \lim_{n\to\infty} \mathbb{P}\{x_n \in A | n < \tau_{\mathbf{X}'}\} \tag{2.21}$$

where $\tau_{\mathbf{X}'}$ is the random time denoting the first time the system enters the trapping set $\mathbf{X}'$, the complement of $\mathbf{X}$. For example, we are interested in how often the system visits subsets of the state space *before* the noise drives away the orbit e.g. enters the basin of attraction of ∞. The quasi-invariant measure is a deformation of the associated invariant measure of the DS, as it is affected by the smoothing effect of the noise process, up to a length scale. Notably, when the noise is observational, the invariant measure of the DS can be deconvoluted (isolated) from the stationary measure of the noise process. This is not possible in the case of dynamical noise. Moreover, the estimation of the quasi-invariant measure is a difficult task, especially in cases where a small number of data is available. We present a method for the approximation of the quasi-invariant measure, based on BNP modeling, in Chapter 3.

In Fig. 2.5 we present two noisy orbits (different realizations), generated by the same initial conditions as in Fig. 2.2, perturbed with Gaussian additive noise with variance $\sigma^2 = 10^{-6}$, along with the associated qq-plot of the two orbits and the corresponding smoothed-out histograms. The stochastic system behavior is dominated by the underlying quasi-invariant measure.

We note that in Chapter 3 we will use the RDS framework, in the sense that we will provide a stochastic approximation of the quasi-invariant measure of the system, while in Chapters 4-5 we will essentially consider the zero-noise limit case, i.e. we will provide methods for estimating dynamical invariants (HTs and global stable invariant manifolds) for the *unknown underlying* deterministic system.

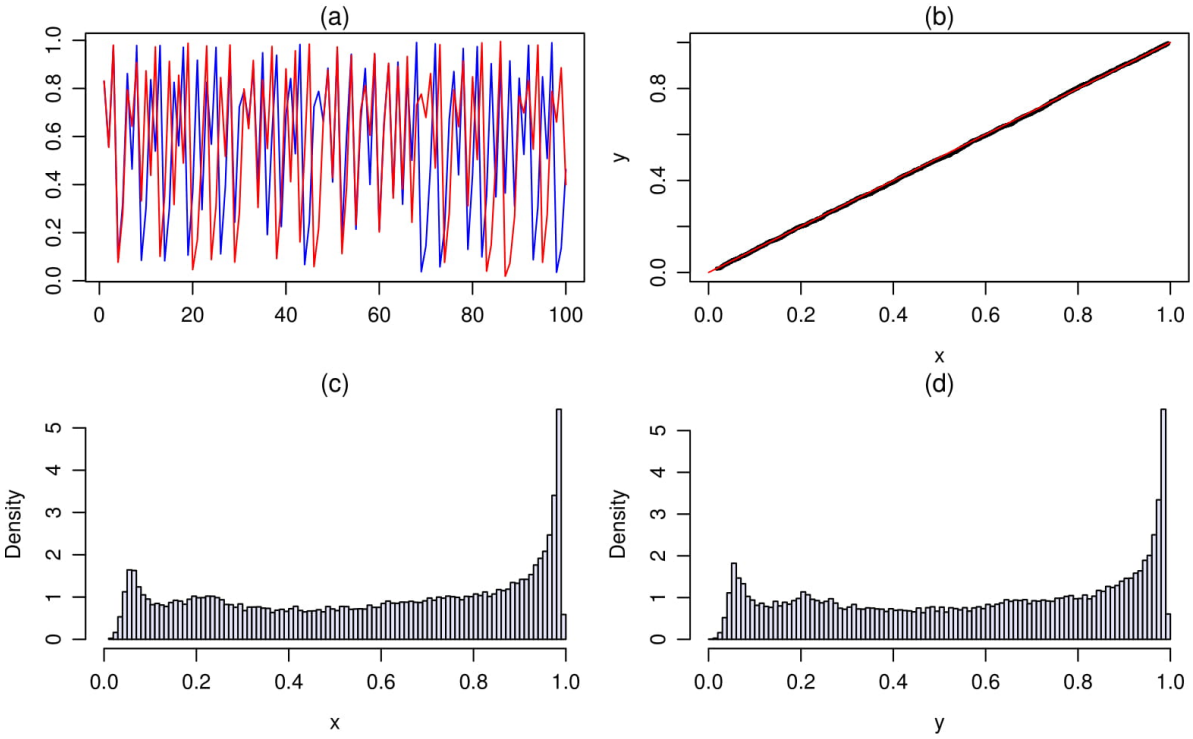

Figure 2.5: (a) First 100 noisy iterations of the initial conditions x_0 (blue) and y_0 (red). (b) qq-plot of the empirical distributions of the 2 noisy orbits, x and y. (c),(d) Histograms of the noisy orbits x and y, respectively.

2.2.2 Noise induced effects

The most generic effect of the noise is that it smears out the fine and rich structure of chaotic dynamics, e.g. fractal structures such as Cantor sets or self-similarity. This is caused by the dominance of the noise (and not of the map) up to a length scale depending on the noise magnitude (affected by the variance and the support of the noise distribution). In Figure 2.6 we present an illustrating example regarding the smearing-out effect on the structure of the deterministic attractor. We use $10,000$ points of the reduced Lorentz map [EEZ14]

$$f : \begin{pmatrix} x_i \\ y_i \end{pmatrix} \mapsto \begin{pmatrix} (1+\theta_1\theta_2)x_{i-1} - \theta_2 x_{i-1}y_{i-1} \\ (1-\theta_2)y_{i-1} + \theta_2 x_{i-1}^2 \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \tag{2.22}$$

with the control parameters set to $(\theta_1, \theta_2) = (0.85, 1)$. The noisy attractors were obtained using additive diagonal dynamical noise, $e_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, with covariance matrix $\Sigma = \sigma^2 \mathbb{I}_2$, i.e. $e_{1i}, e_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with variance $\sigma^2 \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ for Figs. 2.6 (b)-(c)-(d), respectively.

Moreover, dynamical noise has a severe effect on non-hyperbolic maps due to the amplification of the noise perturbations in the neighborhoods of HTs [JK97b]. Specifically, the noisy orbits are driven away from the attractor, due to the geometrical structure in regions of HTs, leading to deformations of the invariant measure called *noise-induced prolongations*. We will refer to this effect and its relation with the loss of the shadowing property in Chapter 4. Another significant impact of noise perturbations, is the conversion of the attractor into a repellor,

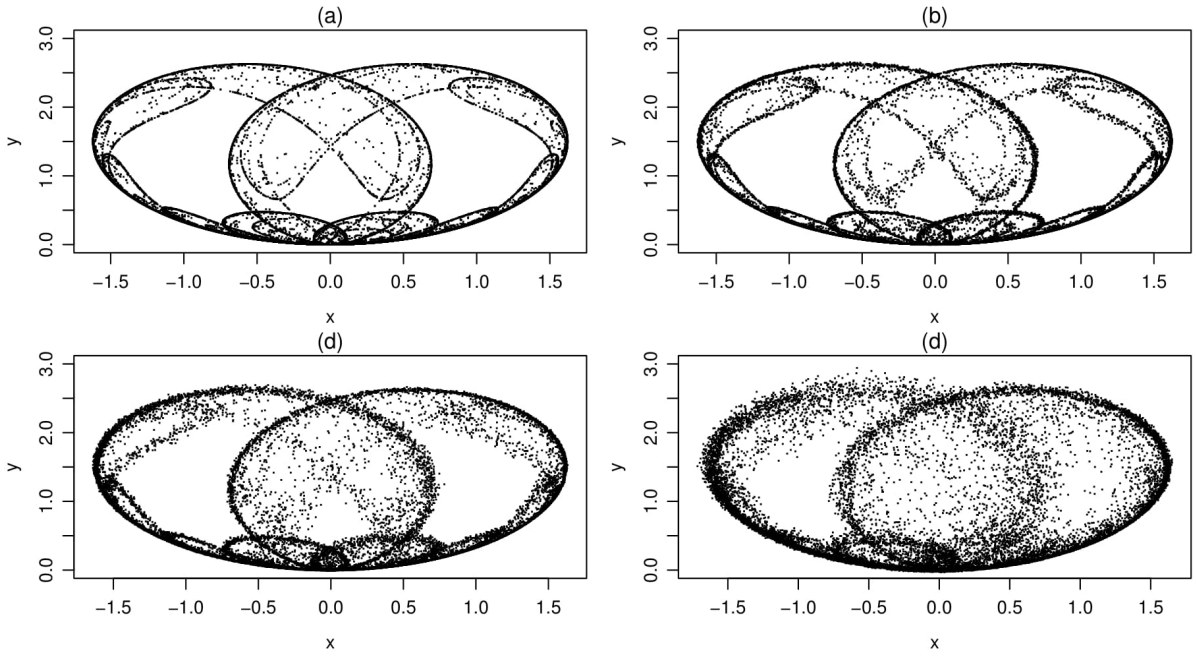

Figure 2.6: Deterministic (a) and noisy attractors of the reduced Lorentz map for additive diagonal noise with variance: (b) $\sigma^2 = 10^{-6}$, (c) $\sigma^2 = 10^{-5}$, (d) $\sigma^2 = 10^{-4}$.

called *noise-induced crisis*[3] [SDG+91]. This conversion is possible, if e.g. the noise magnitude is large enough, in order to force a noisy trajectory to visit the basin of attraction of infinity, leading to escape. Often, this happens in regions of HTs, as they are close to the basin boundaries.Estimation of relevant statistical features such as escape times has been studied in [Som93, HTG94].

Another interesting effect, is the so called *noise-induced chaos* [BSJ90, HES87]. We use this term to refer to the phenomenon where for a certain value of the control parameters, we have the emergence of a noisy attractor with SDIC, thus positive Lyapunov exponents over bounded noisy orbits, while in the absence of noise such attractors disappear. The emergence of noise-induced chaos has been connected [TLG08] with the presence of transient chaos [KG85], i.e. arbitrarily long orbit segments (transients) with chaotic behavior, eventually reaching non-chaotic steady states. Notably, the opposite effect, called *noise-induced order* [MT83], also exists. In this case, the presence of noise results in the transition from a chaotic to a non-chaotic behavior, in terms of e.g. Lyapunov exponents or entropy. The importance of the noise effects lies mainly on the fact that although noise does destruct fractal structures, it does not necessarily result to a more "simple" or to a more complex dynamical behavior. We can observe transitions from order to chaos and vice versa, depending on the specific system subjected to the stochastic perturbations.

In the next Chapters we will occasionally refer to a noise induced effect called *noise-induced jumps* [ABP85]. This is the case when the underlying dynamic system $(\mathbf{X}, \mathcal{T}, f)$ exhibits more than one coexisting attractors. Now, suppose that there exist simultaneously k attractors with with corresponding basins of attraction $X_j \subset \mathbf{X}$, with $X_i \cap X_j = \emptyset, \forall i \neq j$, such that for initial

[3]Generally, the effect of crisis refers to the sudden disappearance of a chaotic attractor.

conditions $x_0^{(i)} \in X_i, x_0^{(j)} \in X_j$, the corresponding forward limit sets are isolated, in the sense that $\omega(x_0^{(i)}) \cap \omega(x_0^{(j)}) = \emptyset$. The term noise-induced jumps refers to the existence of orbits that traverse, under the effect of dynamical perturbations, some of the previously *isolated* basins of attraction.

As an example, we consider the case of the following dynamically perturbed cubic map:

$$x_i = f(\theta, x_{i-1}) + z_i, \quad z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{2.23}$$

with corresponding deterministic map $f(\theta, x) = 0.05 + \theta x - 0.99x^3$. The system becomes bistable for values of the control parameter θ in the interval $\Theta_{\text{bi}} = [1.27, 2.54]$, exhibiting two mutually exclusive period-doubling routes to chaos. The two attractors collapse into a single attractor for for values of θ slightly larger that 2.54.

In Fig. 2.7 we present cases of coexisting isolated attractors. The coexisting, but mutually exclusive, deterministic attracting orbits are depicted in blue and green colors. Noisy orbits, under the $f_{2,4}$ noise process are depicted in red color. Such noisy orbits are able to visit the deterministically isolated attractors. Approximated deterministic basins of attraction, are given in matching colors, on the y-axis.

1. In Fig. 2.7(a), for $\vartheta_1 = 1.90$ there are 2 isolated stable 1-cycles at approximately $\mathcal{O}^+_{\text{st},1} = \{0.9801\}$ (blue) and $\mathcal{O}^-_{\text{st},1} = \{-0.9244\}$ (green) which are all roots of the cubic equation $g(\vartheta_1, x) = x$.

2. In Fig. 2.7(b), for $\vartheta_1 = 1.95$ there are in coexistence a stable 2-cycle and a stable 1-cycle at approximately $\mathcal{O}^+_{st,2} = \{0.9156, 1.0755\}$ and $\mathcal{O}^-_{\text{st},1} = \{-0.9521\}$ respectively, which are all roots of the 9th degree equation $x = g^{(2)}(\vartheta_2, x)$.

3. In Fig. 2.7(c), for $\vartheta_3 = 2.15$ there are 2 isolated stable 2-cycles $\mathcal{O}^+_{\text{st},2} = \{0.7831, 1.2582\}$ (blue) and $\mathcal{O}^-_{\text{st},2} = \{-1.1661, -0.8873\}$ (green) which are all roots of the 9th degree equation $x = g^{(2)}(\vartheta_3, x)$.

4. In Fig. 2.7(d), for $\vartheta_4 = 2.25$ we have the coexistence of a stable strange set (chaotic attractor) (blue) $\mathcal{O}^+_{st,\infty}$ and an isolated 2-cycle $\mathcal{O}^-_{st,2} = \{-1.2507, -0.8269\}$ (green). For example, orbits starting from $x_0 = 1$ are chaotic and orbits starting from $x_0 = -1$ will eventually stabilize on a 2-cycle.

The noise process we have used for the results presented in Fig. 2.7 is

$$f(z) = \frac{9}{10}\mathcal{N}(z \,|\, 0, \sigma^2) + \frac{1}{10}\mathcal{N}(z \,|\, 0, (200\sigma)^2), \quad \sigma^2 = 4 \cdot 10^{-6}. \tag{2.24}$$

If the noise is Gaussian, perhaps the noise variance has to be large in order to have jumps in orbits of small length. However, if the noise is not Gaussian we may have noise-induced jumps due to rare perturbations of high amplitude. Such case could be the noise process of Eq. (2.24). The Bayesian nonparametric framework will allow us to perform accurate inference on RDS subjected to such non-Gaussian noise processes, used for a variety of different tasks.

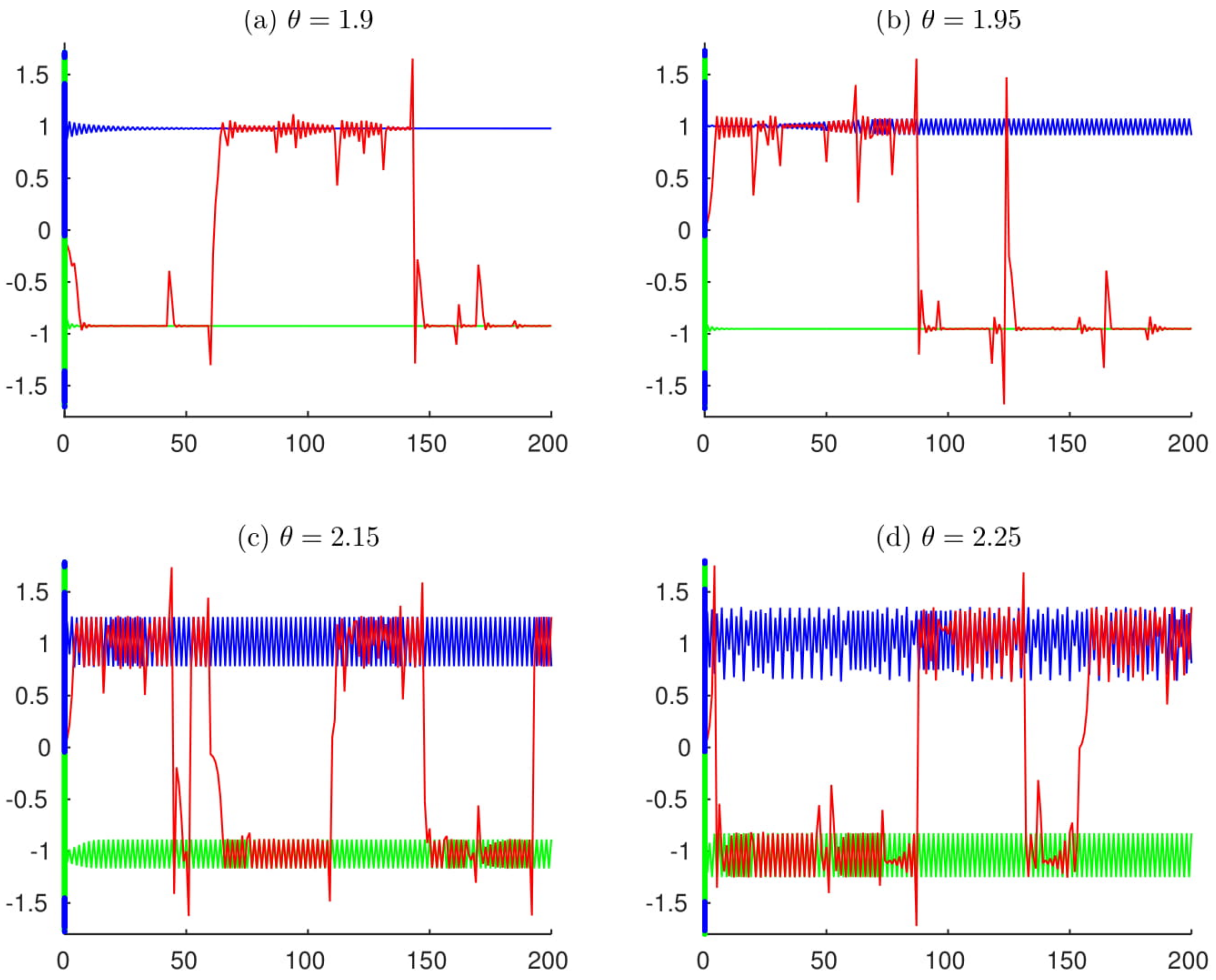

Figure 2.7: Orbits of $g(\vartheta, x) = 0.05 + \vartheta x - 0.99x^3$. Blue and green show deterministic orbits, red shows noisy orbit.

We will further analyze the cubic random map in Chapters 3-4 and we will give an example of noise-induced jumps in 2-dimensional maps in Chapter 5, where the boundaries of the disjoint basins of attraction will be identified using a stochastic approximation of the global stable manifold.

In the next section, we will give an example of Bayesian parametric reconstruction of a simple RDS, illustrating the inadequacy of parametric methods when the underlying noise process departs from normality.

2.3 Bayesian Parametric Reconstruction

We can use the Bayesian framework in order to perform full reconstruction of the dynamical equation of a random map, i.e. to estimate the control parameters, the initial condition and the noise process. Choosing a parametric model, we assume that the observed data are generated by a modeling random polynomial map, perturbed by additive Gaussian noise of unknown precision, which we aim to recover.

Following Hatjispyros et. al [HNW07a] we can use slice sampling techniques [DWW99] in order to be able to perform Gibbs sampling steps with known distributions. Alternatively, we refer to [MC00] for the description of adaptive rejection Metropolis (ARMS) sampling scheme, in similar settings of state space models.

However, if the true noise density departs from normality, such as the noise process appearing in Eq. (2.24), the parametric models are not efficient. The main reason for the inefficiency of the parametric model when the noise is not Gaussian, is the *erroneous estimation* of the variance of the noise process. For example the less frequent, but high magnitude noise perturbations of $f(\cdot)$ in Eq. (2.24), force the variance of the (falsely) assumed Gaussian noise distribution to grow artificially, thus resulting in non-reliable estimation of all the rest parameters. Alternatively, the false assumption of normality could also lead to underestimation of the true variance. In any case, the true underlying noise density is impossible to be estimated, in cases where it is not Gaussian.

In the next Chapters, we will develop novel BNP methods based on the GSB random measure, in order to perform reconstruction, prediction, noise reduction and approximation of the stable manifold, using nonlinear time series of random maps with non Gaussian perturbations.

Chapter 3

Bayesian Nonparametric Reconstruction Models

3.1 Introduction

This chapter is devoted to the presentation of a novel Bayesian approach to the reconstruction and the prediction of random dynamical systems. A common assumption in the literature is 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, as we mentioned in Section 2.3, the estimated variance of the normal errors is artificially enlarged causing poor inference for the system parameters of interest. So for example we could have two sources of random perturbations. An environmental source caused by spatiotemporal inhomogeneities [SM08] producing weak and frequent perturbations, and, a high dimensional deterministic component interpreted in our model as stronger but less frequent perturbations in the form of outlying errors. Other cases include systems containing impulsive noise [SG74, Mid77], where the noise probability density function does not decay in the tails like Gaussian. Also, in situations where the system under consideration is coupled to multiple stochastic environments, the driving noise term may exhibit non-Gaussian behavior, see for example Refs. [KSSH15b] and [KSSH15a]. It is our intention therefore to model the dynamical noise using a highly flexible family of density functions, providing a Bayesian nonparametric formulation [Fer73, FGMW10]. The BNP 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 [HNW09] and joint state-measurement noise density estimation with non-Gaussian and Gaussian observational and dynamical noise components respectively [JDVS13].

The layout of the chapter is as follows. In section 3.2, we derive the two competing nonparametric inferential models. The first model is based on DP mixtures, and we develop its randomized-efficient version. It is based on the model that has been used for the reconstruction of random quadratic maps in [HNW09]. It involves two infinite dimensional parameters

in the form of random probability weights and locations. The second one, being our main contribution, is simpler and is based on GSB mixtures leading to a faster estimation algorithm as it involves only one infinite dimensional parameter in the form of locations. The DP and GSB based Gibbs samplers are described in detail in sections 3.3 and 3.4 respectively. In section 3.5 we specialize, for simplicity, to dynamical equations with polynomial nonlinearities to an arbitrary degree, resorting to simulation. We use simulated time series produced by a cubic map that is dynamically perturbed by outlying errors of varying intensity. We compare the performance of the proposed GSB based Gibbs sampler against its randomized DP based and plain parametric counterparts in the quality of reconstruction, out-of-sample forecasting and quasi-invariant measure estimation. In section 3.6 we extend our model in cases of arbitrary lags and higher-dimensional data, while in section 3.7 we conclude with a summary and future work.

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 control parameters ϑ, and the particular realization of the noise process.

We model the errors in recurrence relation (3.1) as an infinite mixture of zero-mean normal kernels of the form $\mathcal{N}(x \,|\, 0, \tau^{-1})$ with 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

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

As it has been pointed out in [HNW09], a straightforward application of Gibbs sampling ideas, for sampling from the posterior distribution $f(\vartheta,x_0 \mid x_1,\dots,x_n)$, is not possible. In particular, we have the following two difficulties:

1. The Gibbs sampler is infinite-dimensional, due to the infinite mixture of kernels.
2. The FCs are of non-standard form, due to the nonlinearities of the associated deterministic map.

For example, if we assign to the initial condition x_0 a uniform prior over the compact set $\mathbb{X}$, the FC for x_0 will be

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

namely an infinite mixture of truncated kernels.

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 techniques for infinite mixtures. For each observation x_i, we introduce the pair $(d_i,\mathbb{A}_i)$, consisting of the d_i clustering variable indicating the component of the infinite mixture the observation x_i came from and the associated random slice set $\mathbb{A}_i$, an almost surely finite set of indices. Notably, the marginal distribution of the clustering variable is $d_i \mid \pi \sim \sum_{j\geq 1}\pi_j\delta_j$, so the variables d_i have an infinite state space.

Our aim is to have $x_i \mid \tau,\mathbb{A}_i$ 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

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

we obtain

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

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 A Dirichlet process reconstruction model

The DPR model is obtained as a special case of the general reconstruction model if we use non–sequential slice sets. 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 [Wal07] and

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

For the auxiliary variable u_i we obtain

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

Moreover, the joint distribution of u_i, d_i is $f_\pi(u_i, d_i = j) = \pi_j \mathcal{U}(u_i \,|\, 0, \pi_j)$, while for a given $d_i = j$ we have $f(x_i \,|\, d_i = j) = \mathcal{N}(x_i \,|\, 0, \tau_j^{-1})$, so we can write the augmented random densities as

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

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 DP, that is $w_1 = z_1$ and for $j > 1$:

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

with $z_j \overset{\text{iid}}{\sim} \mathcal{B}e(1, c)$ for some concentration parameter $c > 0$, 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 $j \geq 1$:

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

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

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

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

From Eq. (eq. (3.7)) we can see that the usage of the auxiliary variables resulted in finite mixtures of Normal kernels appearing in the likelihood, hence the problem of infinite-dimensionality has been solved. We note that in [HNW09] the authors used a different model for sampling, with two main differences. The first difference regards the concentration parameter, which in their case is fixed, while we let it to be random in contrast with [HNW09]. The second difference is the way we deal with the nonstandard FCs appearing in eq. (3.7).

In [HNW09], the authors introduce an auxiliary variable v_i for each observation x_i for $1 \leq i \leq n$ defined as

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

Using the representation of the normal distribution as a gamma mixture of uniforms, the obtained likelihood for the DPR model can be written as:

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

This approach, also used in [HNW07a], has the advantage that all the variables are sampled from mixtures of uniform distributions. Nevertheless, this comes at the cost of larger execution times -when the sample is large- as the number of the auxiliary variables is equal to the sample size. We will circumvent this problem by using an embedded Gibbs sampler of strategic auxiliary variables, whose size is equal to the length of the vector of the control parameter, hence resulting in standard FCs with lower execution times.

3.3.1 Extending the DPR model for prediction

The DPR model can be extended for prediction purposes, for a given prediction horizon $T > n$, given an observed time series $x^{(n)} = (x_1, \dots, x_n)$,. In this case our aim is the estimation of the future unobserved observations $(x_{n+1}, \dots, x_{n+T})$.

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

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

In the next section we describe an MCMC based algorithm for the randomized DPR (rDPR) model. The same procedure will be followed for the model based on the GSB random measure, as we are interested in their comparison in terms of reconstruction and prediction efficiency.

3.3.2 Slice sampler for the rDPR model

In this section we describe an MCMC slice sampling algorithm for estimating the model. Specifically we are interested in sampling (x_0, ϑ) and the future unobserved variables $x_{n+1}, \dots, x_{n+T}$. Before the description of the algorithm, we propose a certain prior specification scheme.

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, following [Wes92], the concentration parameter c attains a Gamma prior $\mathcal{G}(\alpha, \beta)$, and will be updated as described in Eq. (1.43).

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

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

and

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

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

1. We begin with the update of the stick–breaking weights, given the clustering variables. Namely, we update the z_j–s using

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

for $1 \leq j \leq N$. Then, we use the stick–breaking representation in order to construct the updated weights $(w_j)_{j\geq 1}$.

2. We can now sample 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

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

If $j > N$ we sample the additional τ_j's from the prior $\mathcal{G}(a, b)$.

4. In order to determine the number of the additional weights and precision we let $u^* = \min_{1\leq i\leq n_T}\{u_i\}$ and find the smallest integer N^* for which

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

This step is crucial, as it provides an exact number of variables that we need to sample without using any approximation, as discussed in Section 1.5.

5. We proceed with the sampling of the allocation variables d_i for $i = 1, \ldots, n_T$. It is that

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

6. Then, we can sample the randomized concentration parameter c of the DP. Following [Wes92], we let $\kappa \in \{1, \ldots, n_T\}$ denote the number of unique values of the clustering variables. Then c can be sampled from the following two-step scheme:

1. Sample $s \sim \mathcal{B}e(n_T + 1, c)$
2. $c \,|\, s, \kappa \sim \rho_c\, \mathcal{G}(\alpha + \kappa, \beta - \log c) + (1 - \rho_c)\, \mathcal{G}(\alpha + \kappa - 1, \beta - \log c)$,

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

7. We can now 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 choose τ_j with $1 \leq j \leq N^*$ from

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

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

8. For the vector of parameters ϑ, the FC becomes

$$f(\vartheta \mid \cdots) \propto \mathcal{I}(\vartheta \in \tilde{\Theta}) \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_T} \tau_{d_i} h_\vartheta(x_i, x_{i-1})\right\}. \tag{3.14}$$

9. The FC for x_0, will also be a truncated distribution

$$f(x_0 \mid \cdots) \propto \mathcal{I}(x_0 \in \tilde{\mathbb{X}}) \exp\left\{-\frac{\tau_{d_1}}{2} h_\vartheta(x_1, x_0)\right\}. \tag{3.15}$$

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

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

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

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

3.4 A Geometric stick–breaking reconstruction model

For the construction of the GSBR model, we will use the generic reconstruction model with sequential slice sets of the form $\mathbb{A}_i = \{1, \ldots, N_i\}$, as proposed in [FGMW10]. In particular, letting $\mathbb{A}_i = \{1, \ldots, N_i\}$, with N_i being an almost surely finite discrete random variable of mass $f_N(\cdot \mid \lambda)$, and letting

$$f(d_i = j \mid N_i) \equiv f(d_i = j \mid \mathbb{A}_i) = \frac{\mathcal{I}(j \in \mathbb{A}_i)}{\sum_{k=1}^{\infty} \mathcal{I}(k \in \mathbb{A}_i)} = N_i^{-1} \mathcal{I}(j \leq N_i), \quad 1 \leq i \leq n$$

a discrete uniform distribution on the set $\mathbb{A}_i$, we obtain a GSB mixture based augmented random density

$$f_\tau(x_i, N_i = l, d_i = j) = f_N(l \mid \lambda)\, l^{-1} \times \mathcal{I}(j \leq l) \mathcal{N}(x_i \mid 0, \tau_j^{-1}). \tag{3.18}$$

Marginalizing (3.18) with respect to (N_i, d_i), it is that

$$f_\tau(x_i) = \sum_{j=1}^{\infty} \pi_j \mathcal{N}(x_i \mid 0, \tau_j^{-1}) \text{ with } \pi_j = \sum_{l=j}^{\infty} l^{-1} f_N(l \mid \lambda).$$

When N_i comes from the negative binomial distribution

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

the weights π_j, $j \geq 1$ are geometric, that is

$$\pi_j = \lambda\,(1-\lambda)^{j-1} \mathcal{I}(j \geq 1). \tag{3.19}$$

The geometric weights can be thought of as a reparametrization of the expectation of the stick-breaking weights given in Eq. (3.5), in the sense that $\pi_j = \mathbb{E}\{w_j\}$ with $\lambda = (1+c)^{-1}$. This is important in terms of the comparison between the two models, as the main difference between the DPR and the GSBR regards the nature of the weights. Specifically, the nonconsecutive slice sets of the DPR lead to the necessity of a complete search of the vector where the weights are stored, in order to identify which ones should be chosen. The ordered (decreasing) weights of the GSBR model lead to lower mean execution times, as they circumvent the need for a complete search.

From relations (3.2) and (3.18), and letting $f_N(\cdot \,|\, \lambda) = \mathcal{NB}(\cdot \,|\, 2, \lambda)$, we have

$$f_\tau(x_i, N_i = l, d_i = j |\, x_{i-1}, \theta) = \mathcal{NB}(l|\, 2, \lambda)\, l^{-1}\, \mathcal{I}(j \leq l) \times \mathcal{N}\left(x_i \,|\, g(\theta, x_{i-1}), \tau_j^{-1}\right) \tag{3.20}$$

In a hierarchical fashion using the slice variables N_i we have that for $i = 1, \ldots, n$ and $j \geq 1$:

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

Therefore, the obtained likelihood based on a sample of size n can be written as

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

3.4.1 Extending the GSBR model for prediction

Similarly with the DPR model, we can extend the GSBR model to include the random variables $(x_{n+1}, \ldots, x_{n+T})$, for prediction purposes, at a given horizon $T > n$. In this case the augmented likelihood of the GSBR model becomes

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

In the special case where there is additional information on the path of the future unobserved orbit, for example there are intervals A_{n+i} such that $x_{n+i} \in A_{n+i},\ i = 1, \ldots, T$ we have:

$$f(x_0, x_{n+1}, \ldots, x_{n+T}, \vartheta \,|\, x_1, \ldots, x_n, \{x_{n+1} \in A_{n+1}\}, \ldots, \{x_{n+T} \in A_{n+T}\}) \propto$$
$$f(x_0, \vartheta) f(x_1, \ldots, x_n \,|\, x_0, \vartheta) f(x_{n+1}, \ldots, x_{n+T} \,|\, x_1, \ldots, x_n, x_0, \vartheta) \times$$
$$\underbrace{\mathrm{P}(x_{n+1} \in A_{n+1}, \ldots, x_{n+T} \in A_{n+T})}_{\prod_{i=1}^{T} \mathcal{I}(x_{n+i} \in A_{n+i})}.$$

3.4.2 The slice sampler for the GSBR model

To choose the fittest between the rDPR and GSBR models, we adapt to a "synchronized" prior specification. More specifically, due to the choice of a fully stochastic version of the DPR algorithm, which involves imposing a $\mathcal{G}(\alpha, \beta)$ prior over the concentration parameter c, we introduce "synchronized" prior specifications, assigning a transformed gamma prior over the geometric probability λ via $\lambda = (1 + c)^{-1}$. So as a prior over λ we set

$$f_\tau(\lambda) = \mathcal{TG}(\lambda \,|\, \alpha, \beta) = \frac{\beta^\alpha e^\beta}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} e^{-\beta/\lambda} (1 - \lambda)^{\alpha - 1}, \tag{3.23}$$

with $\lambda \in (0, 1)$. Note that for generic applications of the GSBR model, a beta conjugate prior $f(\lambda; \alpha, \beta) = \mathcal{B}e(\lambda; \alpha, \beta)$ will be easier to implement. As a base measure for GSBR, we use $P_0(d\tau_j) = \mathcal{G}(\tau_j \,|\, a, b) d\tau_j$, $j \geq 1$ for fixed hyperparameters a and b, as in the DPR.

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 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

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

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

1. We first sample the precisions τ_j for $j = 1, \ldots, d^*$ and $d^* = \max_{1 \leq i \leq n_T} d_i$. We have that

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

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

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

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

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

$$\mathrm{P}\left\{N_i = l \,|\, d_i = j, \cdots\right\} \propto (1-\lambda)^l \mathcal{I}(l \geq j),$$

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

4. 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 relation (3.1) will be

$$f(x_0 \,|\, \cdots) \propto \mathcal{I}(x_0 \in \tilde{\mathbb{X}}) \exp\left\{-\frac{\tau_{d_1}}{2} h_\theta(x_1, x_0)\right\}. \tag{3.25}$$

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

$$f(x_{n+j} \,|\, \cdots) \propto \exp\left\{-\frac{1}{2}\left[\tau_{d_{n+j}} h_\theta(x_{n+j}, x_{n+j-1}) + \tau_{d_{n+j+1}} h_\theta(x_{n+j+1}, x_{n+j})\right]\right\}. \tag{3.26}$$

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

$$f(x_{n+T} \,|\, \cdots) = \mathcal{N}\left(x_{n+T} | \, g(\theta, x_{n+T-1}), \tau^{-1}_{d_{n+T}}\right). \tag{3.27}$$

6. 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

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

7. Taking into consideration relation (3.23), the full conditional for the geometric probability λ is

$$f(\lambda \,|\, \cdots) \propto \lambda^{2n_T - \alpha - 1} (1-\lambda)^{L_{n_T}} e^{-\beta/\lambda} \mathcal{I}(\lambda \in (0,1)), \tag{3.29}$$

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

8. Having updated λ, we construct the geometric weights π_j for $1 \leq j \leq N^*$ via Eq. (3.19). 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

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

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 $\mathcal{N}(0, \tau_j^{-1})$.

Details on sampling efficiently from the nonstandard densities arising in Eqs. (3.15) to (3.17) and Eqs. (3.25) to (3.27), via the use of embedded Gibbs samplers in order to circumvent Metropolis–within–Gibbs implementations, are provided in Appendix A.1. In Appendix A.2 we also derive the transformed posterior of the geometric probability λ given in Eq. (3.29).

We note that if we use a Beta conjugate prior for the geometric probability λ, i.e. $\lambda \sim \mathcal{B}e(\alpha, \beta)$, then the FC will be

$$f(\lambda \mid \cdots) = \mathcal{B}e\left(\lambda \mid \alpha + 2\,n_T, \beta + \sum_{i=1}^{n_T} N_i - n_T\right), \tag{3.30}$$

3.5 Simulation results

In this section we will illustrate the efficiency of the two models described above, using synthetic data sets coming from cubic polynomial maps. Moreover we will compare the DPR and the GSBR models in terms of efficiency and speed. We will see that the two models give almost indistinguishable results, while the easier to implement GSBR algorithm exhibits lower mean execution times when compared to the DPR model. For a fair comparison, we use a transformed gamma prior instead of the -otherwise more reasonable choice- conjugate beta prior for the geometric probability λ.

Furthermore, in order to emphasize on the motivation for a BNP framework, we compare the two *nonparametric* models with the associated *parametric* Bayesian reconstruction-prediction model (denoted as Param. in the Tables), i.e. a model that uses the assumption of Gaussian errors with unknown precision. We will use the Percentage Absolute Relative Error (PARE) defined as PARE $= 100 \times |x - \hat{x}|/|x|$, where x and $\hat{x}$ are the true and estimated values of the quantities of interest, respectively.

3.5.1 The experimental setup

The dynamical behavior of the cubic map: Quadratic polynomial maps can exhibit for each parameter value at most one stable attractor, while multistability -as we mentioned in Section 2.4- can be achieved with higher degree polynomial maps [KFG99]. We will generate observations from a cubic random map with a deterministic part given by

$$\tilde{g}(\vartheta^*, x) = 0.05 + \vartheta^* x - 0.99x^3. \tag{3.31}$$

For $\vartheta \in \Theta_{\text{bi}} = [\underline{\vartheta}_{\text{bi}}, \overline{\vartheta}_{\text{bi}}]$ with $\underline{\vartheta}_{\text{bi}} = 1.27$ and $\overline{\vartheta}_{\text{bi}} = 2.54$ the map becomes bistable. In the phase space of the map we can identify two mutually exclusive period-doubling cascades together with two disjoint basins of attraction. The dynamical behavior of the cubic map in Eq. 3.31 can be depicted via the bifurcation diagram given in Fig. 3.1. The two coexisting attracting sets for $\vartheta^* \in \Theta_{\text{bi}}$ 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.

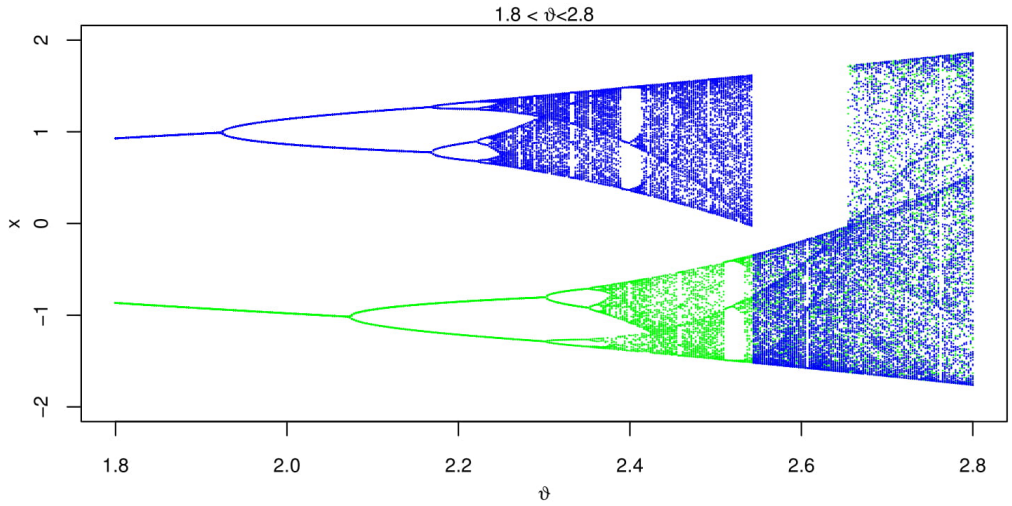

Figure 3.1: The bifurcation diagram for the deterministic map $x_i = g(\vartheta^*, x_{i-1})$.

In Fig. 3.2, we set the value of the control parameter to $\vartheta = \vartheta^\star = 2.55$ (which is the value of the control parameter we have used in our numerical experiments) and we superimpose two deterministic and one noisy 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. For values of ϑ greater than 2.65, there is only one stable attractor.

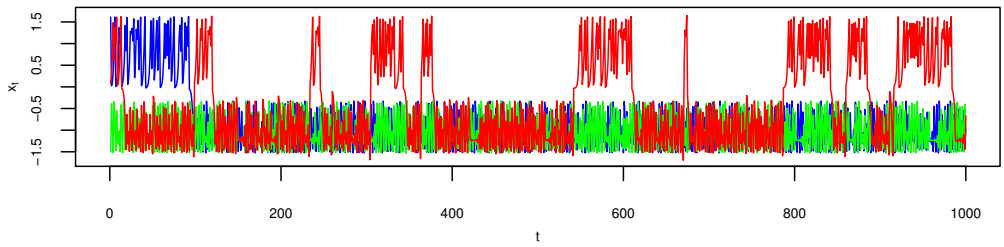

Figure 3.2: The orbits of the the deterministic map $x_i = g(\vartheta^*, x_{i-1})$, with $\vartheta^* = 2.55$, starting from $x_0 = 1$ and $x_0 = -1$ are depicted in blue and green, respectively. A dynamically-noisy orbit, starting from $x_0 = 1$, is given in red.

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

$$f_1 = \sum_{r=0}^{3} \frac{1}{4} \mathcal{N}\left(0, (5r+1)\sigma^2\right), \ \sigma = 10^{-2}. \tag{3.32}$$

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

$$f_{2,l} = \frac{5+l}{10}\mathcal{N}(0,\sigma^2) + \frac{5-l}{10}\mathcal{N}\left(0,(200\sigma)^2\right),\ \sigma = 10^{-3}. \tag{3.33}$$

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).

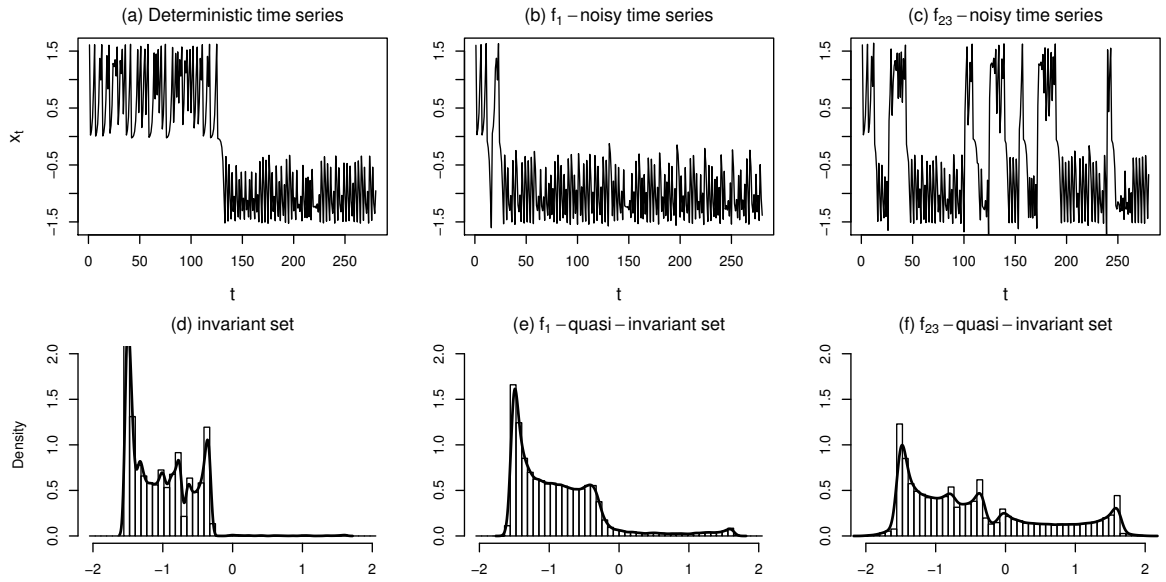

Figure 3.3: In Figure 3.3(a)-(c) we display the deterministic orbit and f_1 and $f_{2,3}$ data-realizations with initial condition $x_0 = 1$. In Figure 3.3(d)-(f) we display the deterministic and the f_1 and $f_{2,3}$ quasi-invariant set approximations respectively.

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) [Bor15, Pin91] 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) [Goe13, Goe16], 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 [Goe16]. 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

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

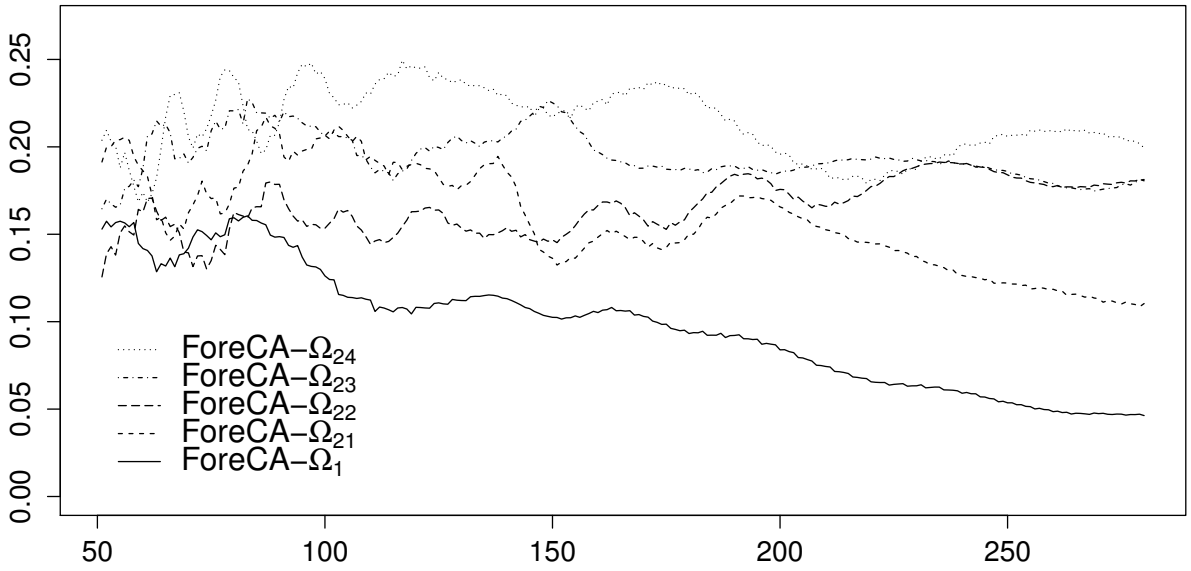

Figure 3.4: Here we display the Ω curves relating to the data sets $x_{f_1}^{(n)}$ and $\{x_{f_{2,l}}^{(n)} : 1 \leq l \leq 4\}$ for n between 50 and 280.

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) $\mathcal{PS}_{\text{IRP}}$ 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 (KDEs) 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. (3.32). 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).

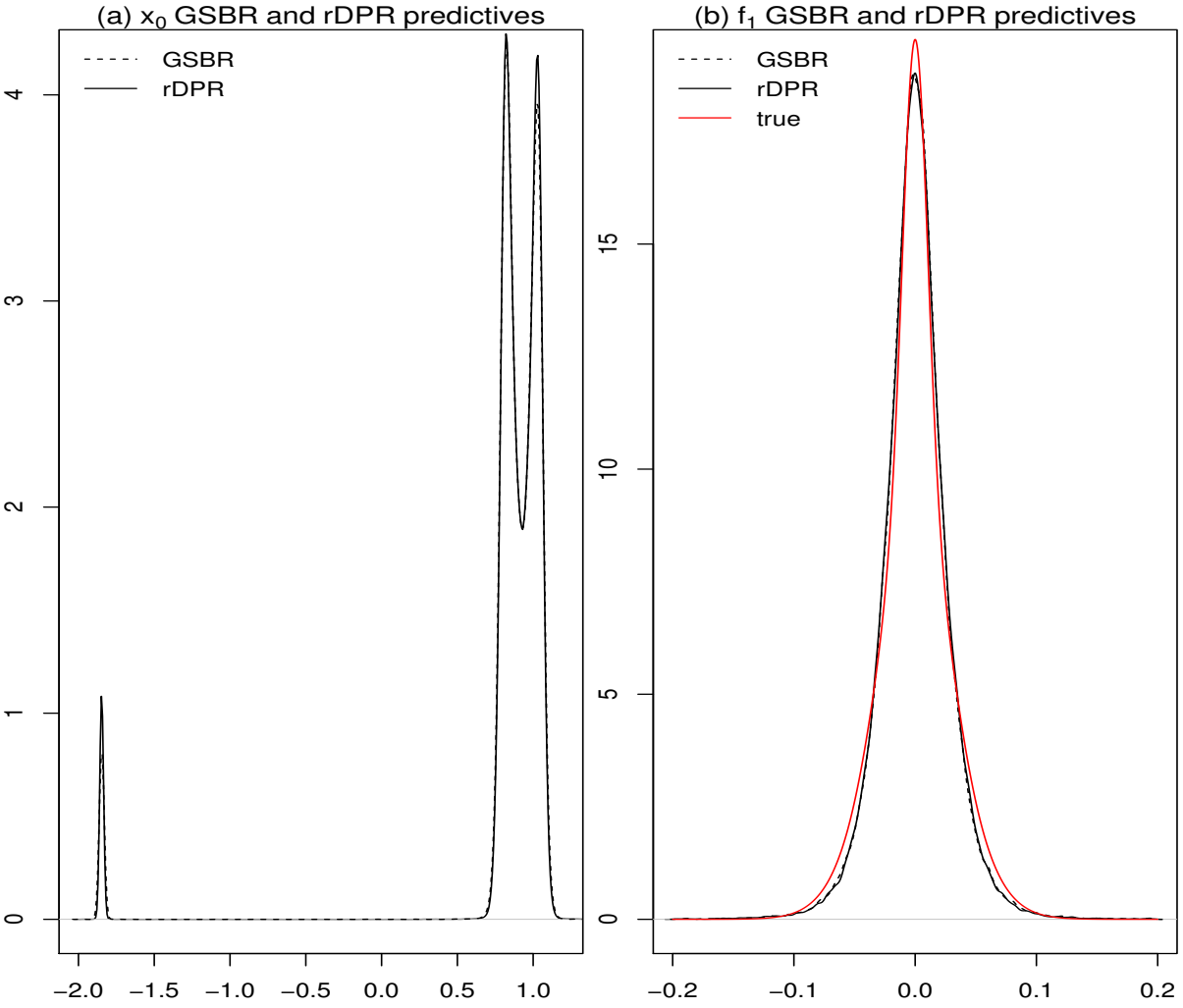

Figure 3.5: In Figure 3.5(a) we give superimposed the KDEs based on the posterior marginal predictive samples of the initial condition variable x_0. In Figure 3.5(b) we superimpose the GSBR and the rDPR noise density estimations together with the true dynamical error density.

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.

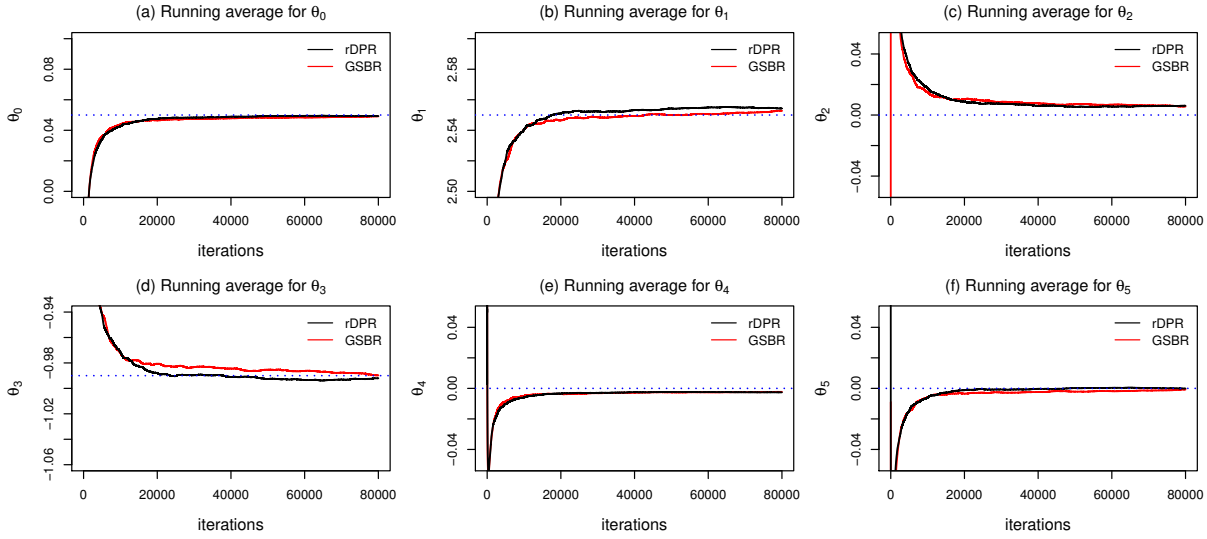

Figure 3.6: Chain ergodic averages for the θ_j variables based on the data set $x_{f_1}^{(200)}$, under prior specification $\mathcal{PS}_{\text{IR}}$, are superimposed in Figure 3.6.

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 $\mathbb{R}$.

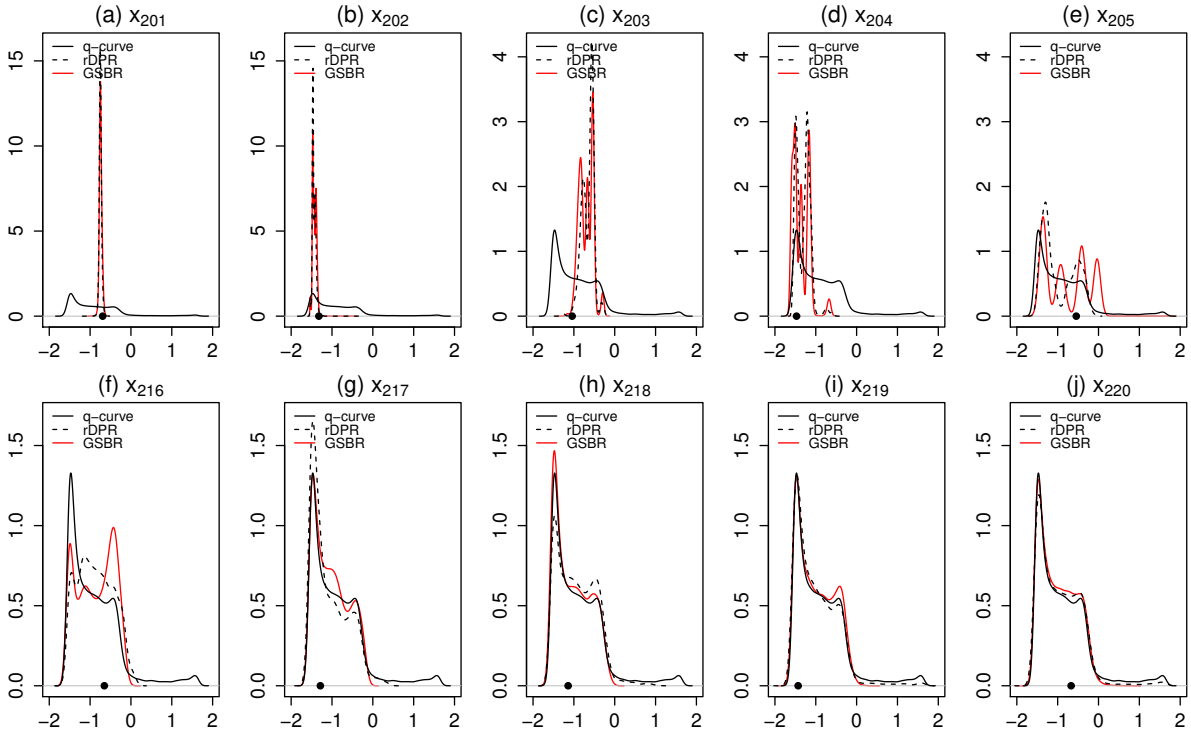

Figure 3.7: In Figure 3.7(a)-(j) we display superimposed the first five and the last five KDEs of the out–of–sample posterior marginal predictive based on data set $x_{f_1}^{(200)}$ under the informative specification $\mathcal{PS}_{\text{IRP}}$. Together we superimpose the KDE of the f_1 quasi invariant 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 KDEs 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 $x^{(200)}_{f_{2,l}} : 1 \leq l \leq 4\}$ (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).

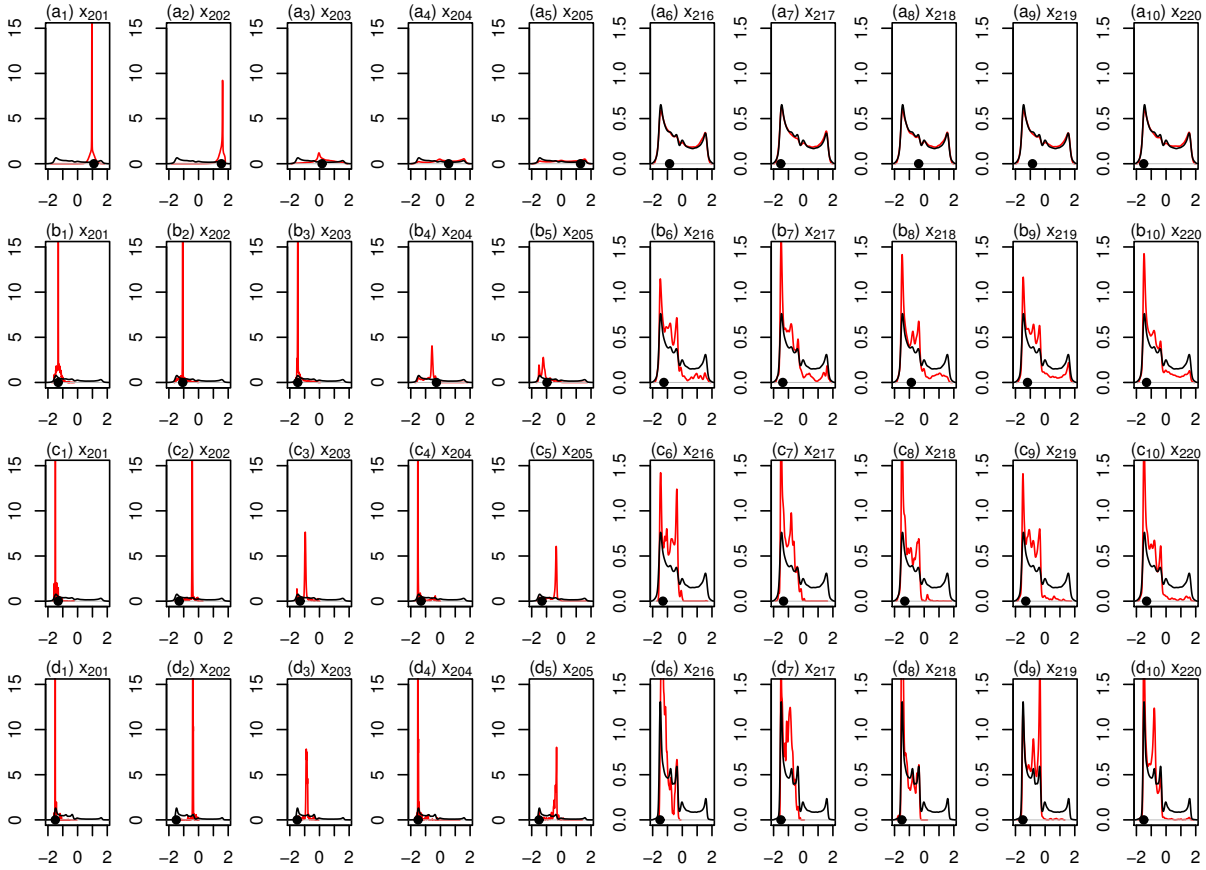

Figure 3.8: In Figure 3.8 we display the GSBR KDEs of the PPM sample of the out–of–sample variables $\{x_{201}, \ldots, x_{205}\}$ and $\{x_{216}, \ldots, x_{220}\}$(solid lines in red) based on samples $x^{(200)}_{f_{2,l}} : 1 \leq l \leq 4\}$ (rows (a) to (d)) under the noninformative prior specification. 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 $\{x^{(200)}_{f_{2,l}} : 1 \leq l \leq 4\}$.

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 Extending the GSBR sampler

In this section we make two extensions of the previously described GSBR model, in order to be able to include:

1. Deterministic counterparts with higher order lags

2. Higher-dimensional data

The above extensions are essential in terms of widening the application areas of the proposed model and being able to analyze real data sets, often characterized by high dimensionality and/or need for many autoregressive (polynomial) terms for the approximation of the deterministic part.

3.6.1 Higher order lags

We are interested in performing reconstruction and prediction of random dynamical systems, from observed time series with additive errors and deterministic part including higher order lags. Specifically, we define the random recurrence relation $T : \Theta \times \mathbb{X}^d \times \mathbb{R} \to \mathbb{R}$ given by

$$\begin{aligned} X_i &= T(\theta, X_{i-1}, \ldots, X_{i-d}, e_i) \\ &= g(\theta, X_{i-1}, \ldots, X_{i-d}) + e_i, \quad i \geq 1, \end{aligned} \tag{3.34}$$

where, $g : \Theta \times \mathbb{X}^d \to \mathbb{X}$, for some compact $\mathbb{X} \subseteq \mathbb{R}$, $(X_i)_{i \geq -d+1}$ and $(e_i)_{i \geq 1}$ are real random variables over some probability space $(\Omega, \mathcal{F}, \mathrm{P})$; we denote by $\theta \in \Theta \subseteq \mathbb{R}^m$ any dependence of the deterministic map g on the control parameters. As above, g is nonlinear, and for simplicity, continuous in $X_{i:d} := (X_{i-1}, \ldots, X_{i-d})$. The random variables e_i are assumed to be independent to each other, independent of the states X_{i-r} for $r < i + d$ and identically distributed from a $0-$mean distribution with unknown density f defined over the real line. The time series x^n is determined by the distribution of the initial vector $X_{1:d}$, the vector of control parameters θ, and the particular realization of the noise process.

Finally, notice that the lag-one stochastic process $(W_i^1, \ldots, W_i^d)$, formed out, from time-delayed values of the (X_i) process, defined by

$$W_i^k = \begin{cases} g(\theta, W_{i-1}^1, \ldots, W_{i-1}^d) + e_i & k = 1 \\ W_{i-1}^{k-1} & 1 < k \leq d\,, \end{cases}$$

is Markovian over $\mathbb{R}^d$. In other words, the time series is $d-$th order Markovian, as each observation x_i depends on the d previous states. This weakening of the Markov property makes the modeling more flexible, especially in cases of e.g. financial time series, where we have evidence of long memory. From a dynamical point of view, this is also an important assumption, as it enables us to analyze results in the embedded phase space of the one-dimensional time series in our disposal, using Taken's embedding Theorem [Tak81]. From Eq. (3.34), we obtain the data likelihood

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

Then, using sequential slice sets of the form $\mathbb{A}_i = \{1, \dots, N_i\}$ and letting $f_N(\cdot\,|p) = \mathcal{NB}(\cdot\,|2, p)$ for the distribution of the auxiliary N_i, we obtain again the geometric weights $\pi_j = \lambda(1-\lambda)^{j-1}$ and we have that for $i = 1, \dots, n$ and $j \geq 1$:

$$(x_i \,|\, x_{i:d}, \theta, d_i = j, \tau) \overset{\text{ind}}{\sim} \mathcal{N}(x_i \,|\, g(\vartheta, x_{i:d}), \tau_j^{-1})$$

For the prediction of T future unobserved observations and $n_T = n + T$ the associated likelihood becomes

$$f(x_i, d_i, N_i\,; 1 \leq i \leq n_T \,|\, \vartheta, x_{1:d}, \lambda) \propto \prod_{i=1}^{n_T} \lambda^2 (1-\lambda)^{N_i - 1} \mathcal{I}(d_i \leq N_i)\, \tau_{d_i}^{1/2} \times \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i:d}) \right\}, \tag{3.35}$$

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

The construction of the slice sampler is straightforward, following Section 2. We note that for prior distribution over the initial condition vector, we may use a Uniform distribution over a suitable subset $\tilde{X}$ of $\mathbb{R}^d$, based on our a-priori knowledge about the associated phase space of the system, instead of a constant prior.

Furthermore, there is a slight change in the full conditional distributions of the future variables, as they include more terms due to the higher lag, namely for $j = 1, \dots, T$ we can write

$$f(x_{n+j} \,|\, \dots) \propto \exp\left\{ -\frac{1}{2} \sum_{\substack{k=j \\ k \leq T}}^{j+d} \tau_{d_{n+k}} h_\vartheta(x_{n+k}, x_{(n+k):d}) \right\},$$

meaning that e.g. the FC for the last future unobserved x_{n+T} will be a Normal distribution

$$f(x_{n+T} \,|\, \dots) \propto \exp\left\{ -\frac{\tau_{d_{n+k}}}{2} h_\vartheta(x_{n+T}, x_{(n+T):d}) \right\}.$$

Simulations

We illustrate the performance of the proposed model on the reconstruction and prediction of observed time series from the Hénon map, using the one-dimensional lag-2 representation. We could also have used one-dimensional observations using the x or y coordinate of the classical 2-dimensional map. Specifically, we use data sets $x_{f_{2,1}}^{(200)}$ of length 200, generated by the Hénon map

$$x_i = 1.38 - x_{i-1}^2 + 0.27 x_{i-2} \tag{3.36}$$

perturbed by the $f_{2,l}, 1 \leq l \leq 4$ additive dynamical noise

$$f_{2,l} = \frac{5+l}{10} \mathcal{N}(0, \sigma^2) + \frac{5-l}{10} \mathcal{N}(0, 10^5 \sigma^2), \quad \sigma^2 = 10^{-7} \tag{3.37}$$

and as modeling polynomial we will use the complete quadratic map lag 2

$$g(\vartheta, x, y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 xy + \theta_4 x^2 + \theta_5 y^2. \tag{3.38}$$

The synthetic set of observations is coming from

$$x_i = g(\vartheta, x_{i-1}, x_{i-2}) + e_i, \quad e_i \overset{\text{iid}}{\sim} f_{2,l} \tag{3.39}$$

which is alternatively called a polynomial autoregressive (PAR) process of second degree with lag 2 [KKA15] PAR$(2,2)$, with an unknown vector of initial conditions $\mathbf{x}_0 = (x_0, x_{-1})$. In our case, we choose $(x_0, x_{-1}) = (-0.5, 1.2)$.

Here we simultaneously reconstruct and predict using the noninformative prior set up. More specifically for $T = 25$ we set $\alpha = \beta = 0.5, a = b = 10^{-3}, M = M_0 = 10$; we have iterated the GSBR sampler 5×10^5 after a burn-in period of $100,000$.

In Fig. 3.9 we display the data set $x^{(200)}_{f_{2,1}}$, superimposed with the union of the embedded posterior predictive marginals for $x_{201}, \ldots, x_{225}$.

We can see that the sampled values of the posterior predictives, when embedded in $\mathbb{R}^2$ lie on the noisy attractor, thus approximating the quasi-invariant measure of the system. The sampled values follow the attractor even on regions of noise-induced prolongations, which do not exist in the associated deterministic map. We will see from the results, that the choice of a BNP model is crucial when we have noise processes such as $f_{2,l}$, as the parametric model is unable to identify the true underlying model.

We note that the prediction in *forward* time is related to the propagation of the PPMs of the future unobserved observations along the *unstable invariant manifold* of the associated deterministic map of the system. In chapter 5 we will discuss the prediction in *reversed* time, associated with the *stable invariant manifold* of the underlying deterministic map, as it can be considered equivalent to prediction in forward time using the inverse map (given that the map is invertible). In this sense, the unstable manifold of the inverse map is the stable manifold of the map itself.

For the control parameters we use the posterior mean as Bayesian estimator, while for the initial conditions we use the MAP estimator. For the predictions $\hat{x}_{201} - \hat{x}_{225}$ we perform the Hartigan's statistical test [HH85] for multimodality on the posterior marginals, in order to choose the appropriate point estimator; we utilize the MAP or the sample mean estimator when the corresponding posterior marginal is multimodal or unimodal respectively.

In Figure 3.8 we display the KDEs based on the PPM samples of the out–of–sample variables $\{x_{201}, \ldots, x_{205}\}$ and $\{x_{221}, \ldots, x_{225}\}$ (solid lines in red) under data sets $x^{(200)}_{f_{2,l}} : 1 \leq l \leq 4\}$ (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).

In Tables 3.5 and 3.6 we display a summary of the results for the $(\vartheta, \mathbf{x}_0)$ estimations and out-of-sample prediction respectively, based on data sets $\{x^{(200)}_{f_{2,l}} : 1 \leq l \leq 4\}$. For (ϑ, x_0) and

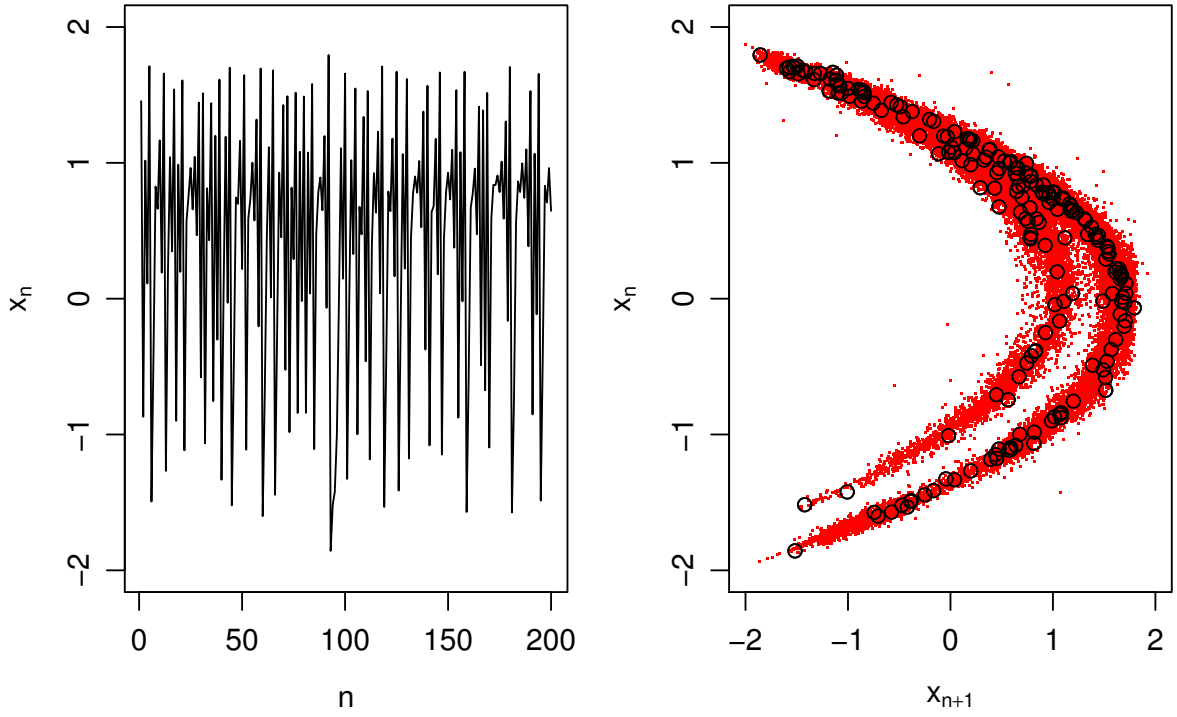

Figure 3.9: Left: Data set $x^{(200)}_{f_{2,1}}$. Right: Embedded $x^{(200)}_{f_{2,1}}$ (black) in $\mathbb{R}^2$, superimposed with the union of the embedded posterior predictive marginals for $x_{201}, \ldots, x_{225}$ (red).

$x_{201} - x_{205}$ we use the corresponding PAREs, while as measure of overall prediction quality we use the mean squared error (MSE), defined as $\text{MSE}(x, \hat{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$.

Table 3.5: Simultaneous reconstruction-prediction under the noninformative prior specification. The (ϑ, x_0) PARE's are based on the data sets $\{x^{(200)}_{f_{2,l}} : 1 \leq l \leq 4\}$ for $T = 25$.

Noise	Model	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	x_0	x_{-1}
$f_{2,1}$	Param.	4.58	5.49	4.86	0.28	0.56	4.27	24.89	32.18
	GSBR	0.04	0.05	0.03	0.01	0.01	0.03	22.70	17.43
$f_{2,2}$	Param.	0.58	0.24	5.68	0.43	0.31	0.88	20.1	19.5
	GSBR	0.10	0.14	0.08	0.03	0.03	0.11	7.40	5.44
$f_{2,3}$	Param.	1.81	1.54	2.02	0.43	0.07	2.12	9.01	12.46
	GSBR	0.01	0.03	0.05	0.00	0.01	0.02	3.63	3.22
$f_{2,4}$	Param.	1.55	1.45	0.25	0.16	0.07	1.55	0.95	2.15
	GSBR	0.00	0.00	0.01	0.00	0.00	0.00	1.03	0.73

In Table 3.5 we compare the PARE results coming from the GSBR and the parametric sampler (Param.).Again, in all cases, the accuracy of the GSBR model is considerably higher than its parametric counterpart.

In Table 3.6 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, the prediction of the GSBR model is considerably better.

In Figure 3.8 we display the KDEs based on the PPM samples of the out–of–sample variables

Table 3.6: 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 $T = 25$. 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	Model	x_{201}	x_{202}	x_{203}	x_{204}	x_{205}	Ov. MSE
$f_{2,1}$	Param.	26.83	1348.87	59.60	174.01	118.34	1.43
	GSBR	0.19	22.94	0.52	0.19	1.01	1.10
$f_{2,2}$	Param.	7.60	8.79	49.81	11.63	51.75	1.78
	GSBR	0.24	1.07	1.77	0.06	0.17	1.10
$f_{2,3}$	Param.	33.56	32.48	212.54	85.53	400.59	1.65
	GSBR	2.37	0.01	22.48	0.30	1.32	1.62
$f_{2,4}$	Param.	27.54	81.67	357.26	64.54	271.23	1.89
	GSBR	0.07	7.61	22.83	7.11	47.48	1.50

$\{x_{201}, \ldots, x_{205}\}$ 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). The predictive accuracy is high for the first five future unobserved observations. Moving ahead in time, the prediction errors become higher, until the model reaches the prediction barrier of the associated quasi-invariant measure. We note that -especially in small sizes- the quality of both reconstruction and prediction exhibits significant dependence on the particular noise realization.

In any case it is evident from the obtained results, that when the noise departs from normality, the performance of the GSBR model is significantly higher than its parametric counterpart in terms of both reconstruction and prediction.

3.6.2 Two–dimensional dynamical noise

In this subsection, we will extend the GSBR model, for the full reconstruction and prediction of nonlinear random dynamical systems, from observed high dimensional time series data, with additive dynamical noise in k dimensions. An important fact is that when the dynamical noise is added simultaneously at k dimensions, then we can no longer represent the observed system in one dimension with k lag terms, perturbed by one dimensional noise. The reason for this lies on the inadequacy of the assumed iid one dimensional noise to capture the effects of higher dimensional and perhaps correlated or non–Gaussian noise processes.

Suppose that we have observed a time series $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where for $i = 1, \ldots, n$ each observation is $k-$dimensional, i.e. $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$ generated by the random dynamical system

$$\begin{aligned} \mathbf{X}_i &= T(\theta, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_{i-d}, \mathbf{e}_i) \\ &= g(\theta, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_{i-d}) + \mathbf{e}_i, \quad i \geq 1, \end{aligned} \tag{3.40}$$

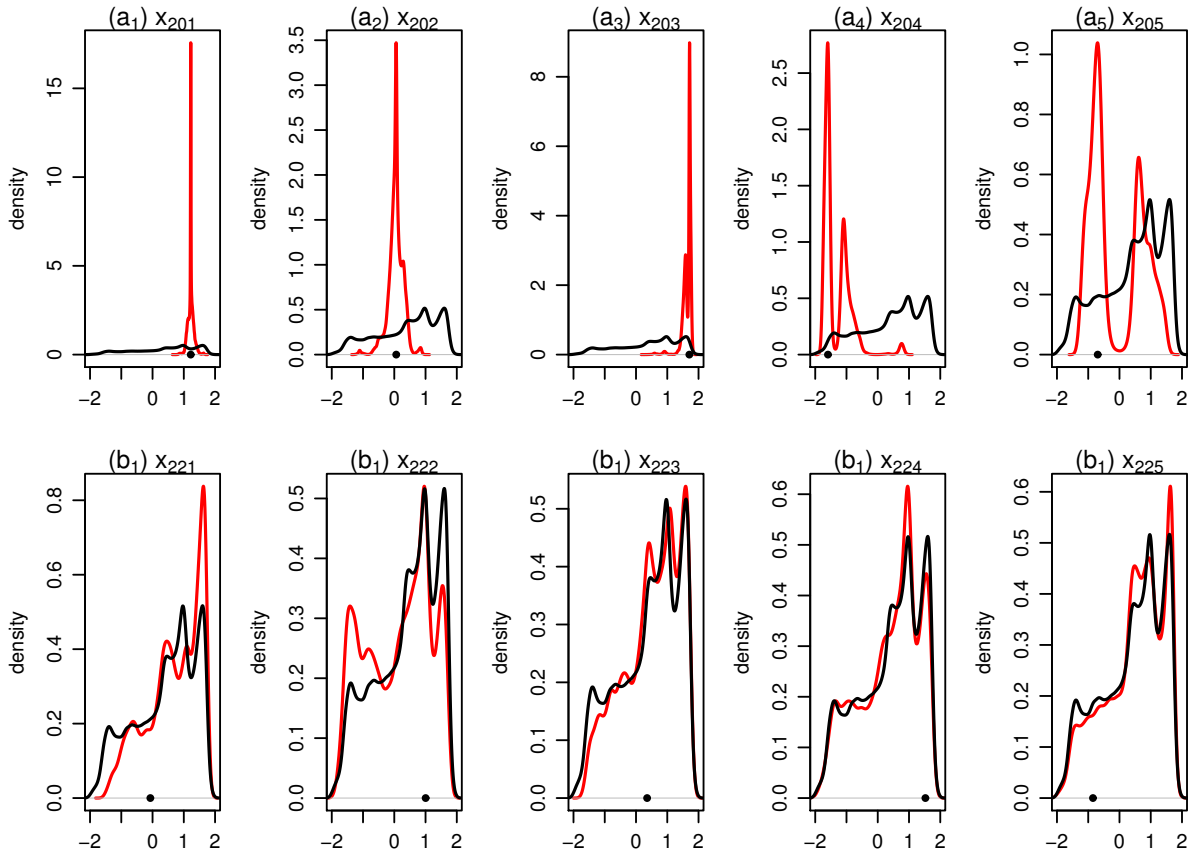

Figure 3.10: In Figure 3.10(a_1)-(a_5) we display superimposed the first five and in (b_1)-(b_5) the last five KDEs of the out–of–sample posterior marginal predictive based on data set $x_{f_1}^{(200)}$ under the noninformative specification $\mathcal{PS}_{\text{NRP}}$.Together we superimpose the KDE of the $f_{2,1}$ quasi invariant density (solid black line). In all Figures, the bullet point represents the corresponding true future value.

with initial condition vector $\mathbf{X}_{1:d} = (\mathbf{X}_0^1, \ldots, \mathbf{X}_0^d)$, where $\mathbf{X}_0^j = \left(x_{01}^j, \ldots, x_{0k}^j\right)^T$. Specifically, letting X a compact subset of $\mathbb{R}^k$ we have the nonlinear function $g : \Theta \times \mathbb{X}^d \to \mathbb{X}$ and the random recurrence $T : \Theta \times \mathbb{X}^d \times \mathbb{R}^k \to \mathbb{R}^k$, with the observations being random variables over some proper probability space.

We assume the additive perturbations $\mathbf{e}_i$ are independent of the observations, independent of each other and identically distributed from a zero mean distribution f defined over $\mathbb{R}^k$, i.e. $\mathbf{e}_i \overset{\text{iid}}{\sim} f$. In order to relax the normality assumption for the noise process, we model the unknown density f as an infinite mixture of $\mathbf{0}-$mean $k-$variate Normal kernels with precision matrices $\left(\Lambda_j\right)_{j\geq 1}$, with mixing measure a general discrete random distribution $\mathbb{G} = \sum_{j\geq 1} \pi_j \, \delta_{\Lambda_j}$. Then, we have that the noise density can be written as

$$f_{\pi,\Lambda}(z) = \sum_{j=1}^{\infty} \pi_j \mathcal{N}_k\left(\mathbf{0}, \Lambda_j^{-1}\right),$$

where $\pi = \left(\pi_j\right)_{j\geq 1}$ is an infinite sequence of random weights and $\Lambda = \left(\Lambda_j\right)_{j\geq 1}$ an infinite sequence of precision matrices $\Lambda_j \in \mathcal{S}^{+1}$.

Then, the associated data likelihood takes the form

$$f_{\pi,\Lambda}\left(\mathbf{x}_1,\ldots,\mathbf{x}_n,\,|\,\mathbf{x}_{1:d},\vartheta\right) = \prod_{i=1}^{n}\sum_{j=1}^{\infty}\pi_j\,\mathcal{N}_k\left(\mathbf{x}_i\,|\,g(\vartheta,\mathbf{x}_{i:d}),\Lambda_j^{-1}\right).$$

As a base measure we set $P_0(\mathrm{d}\Lambda) = \mathcal{W}_k\left(\Lambda\,|\,\nu,\Sigma\right)\mathrm{d}\Lambda$, the Wishart measure with ν degrees of freedom and positive semidefinite $k\times k$ covariance matrix Σ.

In order to sample from an a.s. finite dimensional space, we augment with the sequential slice sets $\mathbb{A}_i = \{1,\ldots,N_i\}$, assign $f_N(\,\cdot\,|p) = \mathcal{NB}(\,\cdot\,|2,\lambda)$ and obtain the geometric weights $\pi_j = \lambda(1-\lambda)^{j-1}$, where λ is the geometric probability of the GSB random measure. Then, taking into account the T future unobserved observations we can compute the associated augmented likelihood, with $n_T = n+T$:

$$\begin{aligned} f(\mathbf{x}_i,d_i,N_i\,;\,1\leq i\leq n_T\,|\,\vartheta,\mathbf{x}_{1:d},\lambda) \propto \prod_{i=1}^{n_T}\lambda^2(1-\lambda)^{N_i-1}\mathcal{I}(d_i\leq N_i)\,|\Lambda_{d_i}|^{1/2} \\ \times\exp\left\{-\frac{1}{2}\left(\mathbf{x}_i-g\left(\vartheta,\mathbf{x}_{i:d}\right)\right)^T\Lambda_{d_i}\left(\mathbf{x}_i-g\left(\vartheta,\mathbf{x}_{i:d}\right)\right)\right\}. \quad (3.41)\end{aligned}$$

The sampling algorithm described below, is similar to the one-dimensional sampling algorithm described in Section 2. In particular we will initialize the variables d_i, N_i for $i = 1,\ldots,n_T$, λ, $\mathbf{x}_{1:d}$ (the initial condition vector) and ϑ. At each iteration of the Gibbs sampler, we we will sample the variables:

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

and

$$(\vartheta,\ \mathbf{x}_{1:d},\ \lambda,\ \mathbf{z}_{n_T+1}),$$

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

1. The full conditional for the geometric probability $\lambda\sim\mathcal{B}e(\alpha,\beta)$ remains unchanged:

$$f(\lambda\,|\,\cdots) = \mathcal{B}e\left(\lambda\,|\,\alpha+2\,n_T,\beta+\sum_{i=1}^{n_T}N_i-n_T\right), \quad (3.42)$$

Having updated λ, we construct the geometric weights π_j for $1\leq j\leq N^*$ via Eq. (3.19). We note that in this case we will not use a transformed Gamma prior, as we are not aiming to compare the multivariate GSBR and ans DPR models. The multivariate GSBR algorithm have a faster execution time as compared with its multivariate DPM counterpart, due to the consecutive random slice sets.

[1] We denote by $\mathcal{S}^+$ the space of symmetric positive definite $k\times k$ real matrices

2. We then sample the precision matrices Λ_j for $j = 1, \ldots, N^*$ and $N^* = \max_{1 \leq i \leq n_T} N_i$. Assigning a conjugate Wishart prior over Λ_j, i.e. $\Lambda_j \sim \mathcal{W}_k(\nu_0, \Sigma_0)$ we have that

$$f(\Lambda_j \mid \cdots) = \mathcal{W}_k\left(\Lambda_j \mid \nu_0 + \sum_{i=1}^{n_T} \mathcal{I}(d_i = j), \left(\Sigma_0^{-1} + S_j\right)^{-1}\right), \tag{3.43}$$

where $S_j = \sum_{i=1}^{n_T} \mathcal{I}(d_i = j)\left(\mathbf{x}_i - g\left(\vartheta, \mathbf{x}_{i:d}\right)\right)\left(\mathbf{x}_i - g\left(\vartheta, \mathbf{x}_{i:d}\right)\right)^T$.

3. We then sample the infinite mixture allocation variables d_i for $i = 1, \ldots, n_T$ from the discrete full conditional distribution

$$\Pr(d_i = j \mid N_i, \cdots) \propto |\Lambda_j|^{1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - g\left(\vartheta, \mathbf{x}_{i:d}\right)\right)^T \Lambda_j \left(\mathbf{x}_i - g\left(\vartheta, \mathbf{x}_{i:d}\right)\right)\right\} \mathcal{I}(j \leq N_i). \tag{3.44}$$

4. The construction of the sequential slice sets $\mathbb{A}_i$ for $1 \leq i \leq n_T$ requires, as in the one–dimensional case, the sampling of the auxiliary variables N_i from

$$\Pr(N_i = r \mid d_i = j, \cdots) \propto (1 - \lambda)^r \mathcal{I}(j \leq r), \tag{3.45}$$

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

5. In this step we need to sample $\mathbf{z}_{n+1}$ from the (multivariate) noise predictive distribution $f(\mathbf{z}_{n+1} \mid x_1, \ldots, x_n)$. At each iteration of the Gibbs sampler, we have updated weights $(\pi_j)_{1 \leq j \leq N^*}$ and precision matrices $(\Lambda_j)_{1 \leq j \leq N^*}$ and we sample independently $\rho \sim \mathcal{U}(0, 1)$. Then we take the Λ_j with $1 \leq j \leq N^*$ satisfying

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

Whenever $\rho > \sum_{i=0}^{N^*} \pi_i$, we sample Λ_j from the prior $\mathcal{W}_k(\nu_0, \Sigma_0)$. In any case, we have that $\mathbf{z}_{n+1} = \mathcal{N}_k\left(\mathbf{0}, \Lambda_j^{-1}\right)$.

6. The full conditional for $\mathbf{x}_0$, with a uniform prior over the set $\tilde{\mathbb{X}} \subseteq \mathbb{R}^{kd}$ that represents our prior knowledge for the state space of the dynamical system in will be

$$f(\mathbf{x}_{1:d} \mid \cdots) \propto \mathcal{I}(\mathbf{x}_0 \in \tilde{\mathbb{X}}) \exp\left\{-\frac{1}{2} h_\vartheta\left(\mathbf{x}_1, \mathbf{x}_{1:d}, \Lambda_{d_1}\right)\right\}, \tag{3.46}$$

where $h_\vartheta\left(\mathbf{x}_i, \mathbf{x}_{i:d}, \Lambda_{d_i}\right) = \left(\mathbf{x}_i - g\left(\vartheta, \mathbf{x}_{i:d}\right)\right)^T \Lambda_{d_i} \left(\mathbf{x}_i - g\left(\vartheta, \mathbf{x}_{i:d}\right)\right)$.

7. 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

$$f(\vartheta \mid \cdots) \propto \mathcal{I}(\theta \in \tilde{\Theta}) \exp\left\{-\frac{1}{2} \sum_{i=1}^{n_T} h_\vartheta\left(\mathbf{x}_i, \mathbf{x}_{i:d}, \Lambda_{d_i}\right)\right\}. \tag{3.47}$$

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

$$f(\mathbf{x}_{n+j} \mid \ldots) \propto \exp\left\{ -\frac{1}{2} \sum_{\substack{k=j \\ k \leq T}}^{j+d} h_{\vartheta}\left(\mathbf{x}_{n+k}, \mathbf{x}_{(n+k):d}, \Lambda_{d_{n+k}}\right) \right\}. \tag{3.48}$$

Simulations

In order to illustrate the inadequacy of the parametric model in the case of higher dimension, we perform reconstruction using an observed time series $\mathbf{x}_{f_1}^{(200)}$ of length $n = 200$ from the 2-dimensional Hénon map

$$f(x, y) = (1.38 - x^2 + y, 0.27x) \tag{3.49}$$

with initial condition $\mathbf{x}_0 = (1, 0.2)$ contaminated with non Gaussian noise f_1, following the mixture

$$f_1(z) = \frac{1}{2}\mathcal{N}_2\left(z \mid \mathbf{0}, \sigma^2\mathbf{I}_2\right) + \frac{1}{2}\mathcal{N}_2\left(z \mid \mathbf{0}, (100\sigma)^2\mathbf{I}_2\right), \quad \sigma = 10^{-3},$$

where $\mathbf{I}_2$ is the 2×2 unit matrix. We present the data set $\mathbf{x}_{f_1}^{(200)}$ in Fig. 3.11, where the severe effect of the dynamical noise on the structure of the corresponding deterministic attractor is evident. We model the observations as

$$(x_i, y_i) = g\left(\vartheta, x_{i-1}, y_{i-1}\right) + e_i = (\theta_0 + \theta_1 x_{i-1} + \theta_2 x_{i-1}^2 + y_{i-1}, \theta_3 x_{i-1}) + e_i.$$

For the parametric model (Param.) we consider the noise as $e_i \overset{\text{iid}}{\sim} \mathcal{N}_2\left(\mathbf{0}, \Lambda^{-1}\right)$, where $\Lambda \sim \mathcal{W}_2\left(2, \Sigma_0\right)$ is the unknown precision matrix, with a conjugate Wishart prior.

For the extended GSBR model, the noise is represented as an infinite mixture of bivariate zero-mean Normal kernels, with unknown precision matrices $e_i \overset{\text{iid}}{\sim} \sum_{j \geq 1} w_j \mathcal{N}_2\left(\mathbf{0}, \Lambda_j^{-1}\right)$, with base measure $\mathcal{W}_2\left(2, \Sigma_0\right)$. We assign uniform priors over the initial condition vector and the control parameters $\mathbf{x}_0 \sim \mathcal{U}\left((-2, 2)^2\right)$, $\vartheta \sim \mathcal{U}\left((-10, 10)^4\right)$, while for the geometric probability we use a $\mathcal{B}e(0.5, 0.5)$ prior. We ran the chain for $85,000$ iterations with a burn-in period of length $5,000$. In Table 3.7 we present the $(\vartheta, \mathbf{x}_0)$ PAREs for the parametric and the GSBR models.

Table 3.7: $(\vartheta, \mathbf{x}_0)$ reconstruction PARE's based on the data set $\mathbf{x}_{f_1}^{(200)}$.

Model	θ_0	θ_1	θ_2	θ_3	x_0	y_0
Param.	0.21	0.74	0.63	4.50	5	80
GSBR	0.01	0.01	0.02	0.00	1	5

In Table 3.7 we see that the mean PAREs obtained from the GSBR model are significantly lower compared with those of its parametric counterpart, both for the control parameters and the initial condition vector. The main reason for the better performance of the BNP model is that it correctly identifies the true underlying noise process, even with small sample size, in contrast with the parametric model which -by construction- is unable to infer the true non-Gaussian

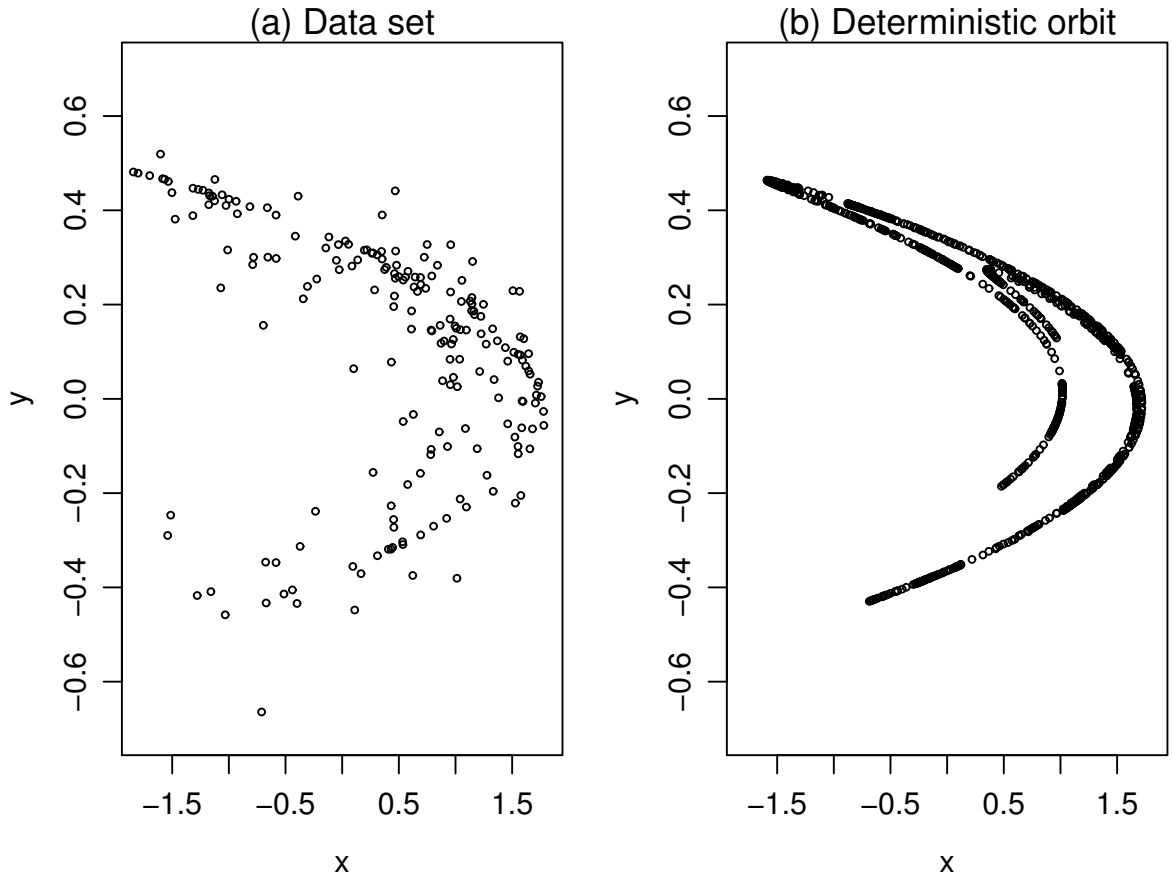

Figure 3.11: (a) Data set $\mathbf{x}_{f_1}^{(200)}$, from the 2-dimensional noisy Hénon map, (b) Deterministic orbit of length $2,000$ from the corresponding unperturbed system.

process f_1. We stress again the fact that whenever we are confident that the underlying noise is Gaussian a parametric model will be as accurate as a Bayesian nonparametric one.

Furthermore, in order to illustrate the efficiency of the GSBR model in terms of prediction, we use the data set $\mathbf{x}_{f_2}^{(500)}$ of length 500 from the noisy Hénon map, i.e.

$$(x_i, y_i) = f\left(\vartheta, x_{i-1}, y_{i-1}\right) + e_i = (1.31 - x_{i-1}^2 + y_{i-1}, 0.23 x_{i-1}) + e_i, \quad e_i \sim f_2,$$

with the noise process sampled from

$$f_2(z) = \frac{3}{4}\mathcal{N}_2\left(z \,|\, \mathbf{0}, \sigma^2 \mathbf{I}_2\right) + \frac{1}{4}\mathcal{N}_2\left(z \,|\, \mathbf{0}, 10^3 \cdot \sigma^2 \mathbf{I}_2\right), \quad \sigma^2 = 10^{-5}.$$

Again, we have used as modeling map

$$g\left(\vartheta, x, y\right) = (\theta_0 + \theta_1 x + \theta_2 x^2 + y, \theta_3 x).$$

We aim to predict the next $T = 5$ future observations. The prior specification are the same as before, with a uniform over $(-10, 10)^2$ for the initial condition vector. We ran the chain for $200,000$ iterations with a burn–in period of length $5,000$ performing thinning every 15 iterations.

In Table 3.8 we present the PAREs and the MSEs for the two coordinates x and y of the five future observations. The corresponding KDEs of $x_{501} - x_{505}$ and $y_{501} - y_{505}$ are presented in Fig. 3.12. We use the dip test as a criterion for the choice of Bayesian point estimator.

Table 3.8: Prediction PARE's and overall MSEs for the x and y coordinates based on the data set $\{\mathbf{x}_{f_2}^{(500)}\}$.

Model	x_{501}	x_{502}	x_{503}	x_{504}	x_{505}	$x-$MSE
Param.	0.23	4.39	4.97	45.84	12.91	$3.5 \cdot 10^{-5}$
GSBR	0.14	13.85	0.27	9.20	3.29	$2.6 \cdot 10^{-5}$
Model	y_{501}	y_{502}	y_{503}	y_{504}	y_{505}	$y-$MSE
Param.	2.77	2.07	356.05	7.70	34.98	$14.5 \cdot 10^{-4}$
GSBR	2.94	4.42	313.17	0.35	3.38	$4.4 \cdot 10^{-4}$

From the results in Table 3.8 we see that the GSBR model performs considerably better than its parametric counterpart, mainly due to the precise identification of the noise process. For a large prediction horizon we expect the PPMs to eventually reach the prediction barrier of the quasi-invariant set.

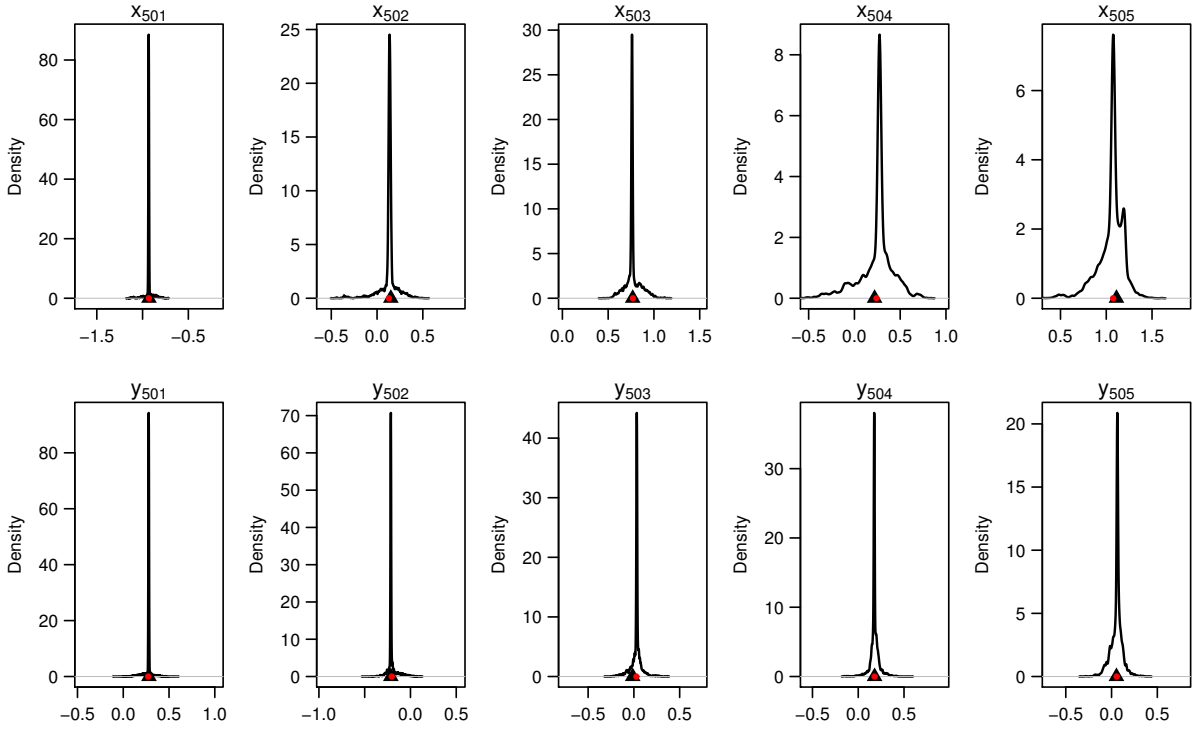

Figure 3.12: We display the five KDEs of the out–of–sample posterior marginal predictives for the x (upper) and y (lower) coordinates based on the data set $x_{f_2}^{(500)}$. In all Figures, the red bullet point represents the point estimate, while the triangle represents the true future value.

In order to further emphasize the need for a flexible nonparametric framework, we perform reconstruction using an observed time series $\mathbf{x}_{f_3}^{n}$ of length $n = 1,000$ from the 2-dimensional Hénon map:

$$f(x, y) = (1.38 - x^2 + y, 0.27x) \tag{3.50}$$

with initial condition $\mathbf{x}_0 = (1, 0.2)$, contaminated with non Gaussian noise f_3, coming from the mixture

$$f_3(z) = \frac{1}{2}\mathcal{N}_2\left(z \,|\, \mathbf{0}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) + \frac{1}{2}\mathcal{N}_2\left(z \,|\, \mathbf{0}, \sigma^2 \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\right),$$

with $\sigma^2 = 10^{-3}$ and $\rho = 0.85$. The noise process f_3 is an equally weighted mixture of two bivariate Gaussian densities with opposite correlations and is an illustrating example of a process that cannot be properly inferred if we use a simple parametric modeling. Moreover, such noise processes cannot be recovered from one dimensional observations, as the perturbations are present at every iteration in both coordinates simultaneously.

For the reconstruction we have used the same prior specification as before and have ran the chain for 150,000 iterations with a burn–in period of 30,000.

The flexibility of the GSBR model is highlighted by the accurate estimation, with a mean PARE of 0.04% for the control parameters. In Fig. 3.13 we present the noisy orbit $\mathbf{x}^n_{f_3}$ and the two dimensional KDE of the posterior marginal of the initial vector (x_0, y_0). The (x_0, y_0) posterior marginal lies on the direction of the global stable manifold, as we will analyze later on, in chapter 5.

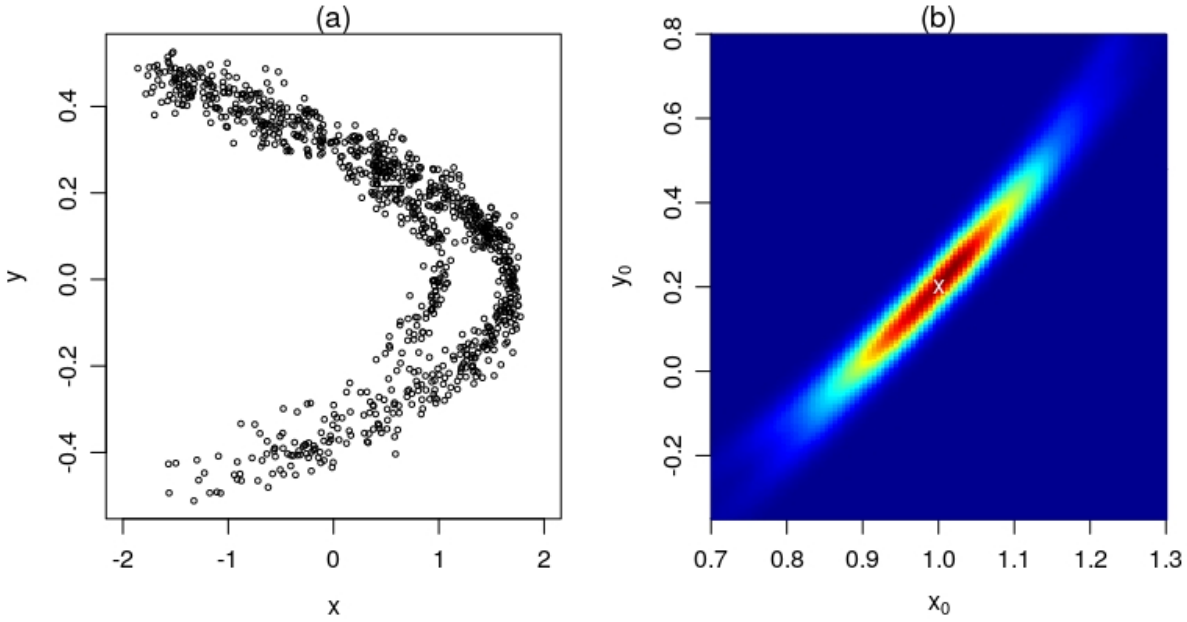

Figure 3.13: (a) Data set $\mathbf{x}^n_{f_3}$ and (b) 2-d KDE of the initial vector (x_0, y_0) posterior marginal. Position of the true (x_0, y_0) is indicated with a white "x".

In Fig. 3.14 we present the results indicating the accurate recovery of the noise process. The two components of opposite correlation resulted in a density that departs from normality, as shown in the two dimensional KDE of the noise predictive. We also present the posterior marginals for the x and y coordinates of the estimated noise process, superimposed with the true noise marginal densities.

Notably, even if a parametric model manages to achieve low errors on the control parameter estimation due to a very large sample size, it will nevertheless result to the erroneous estimation

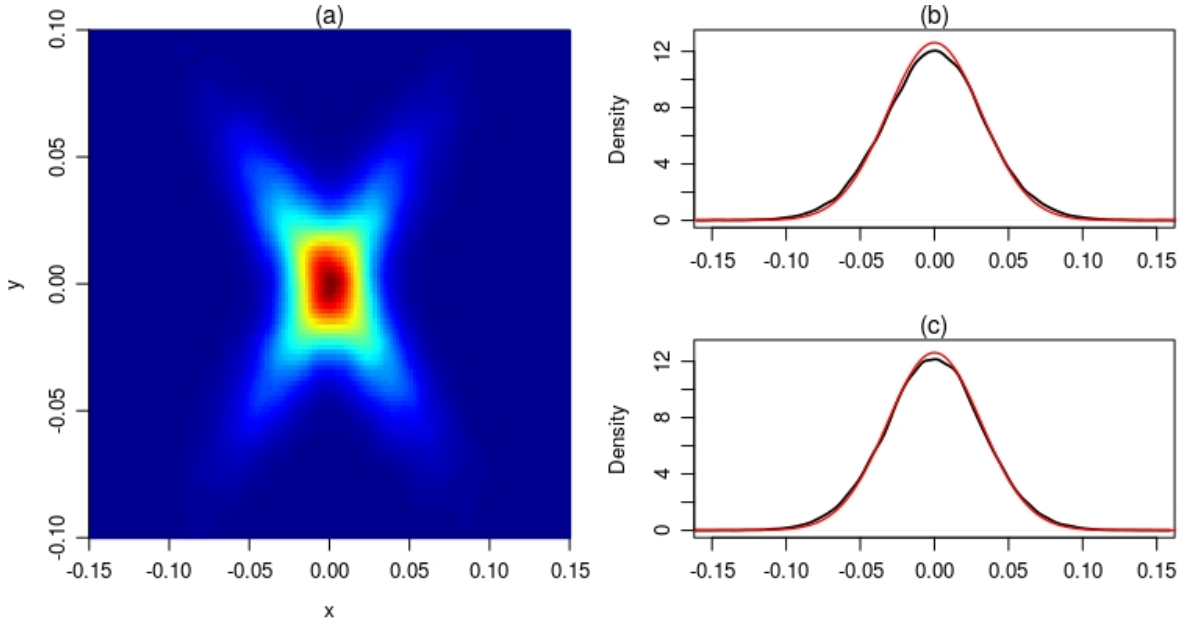

Figure 3.14: (a) 2-d KDE of the noise predictive distribution and corresponding posterior marginals of the (b) x and (c) y coordinates (black) of the estimated noise process, superimposed with the true noise marginal densities (red).

of the noise density.

3.7 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 [FGMW10], as a prior (over the space of densities) over 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, namely the locations of the atoms (τ_j) of the random measure. GSB mixture dynamical modeling is as accurate as DP based modeling but with smaller execution times, and easier implementation.

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.

Both rDPR and GSBR models are characterized by wide applicability, as infinite mixtures of zero mean Gaussians can mimic the effect of zero–mean heavy tailed behavior to an arbitrary level of accuracy.

Chapter 4

Bayesian Dynamical Noise Reduction

4.1 Introduction

This chapter is devoted to the construction of a novel BNP method, suitable for another interesting problem of the field of nonlinear dynamics, namely the problem of noise reduction. Many different approaches have been adopted to address the general issue of nonlinear noise reduction. Hammel et al. [Ham90], used techniques originated from the proof of the shadowing lemma to reduce noise in observed chaotic data. Farmer and Sidorowich [FS91], proposed the use of Lagrangian multipliers for the minimization of the distance between the observed and the denoised orbit. In order to deal with homoclinic tangencies, they used a combination of manifold decomposition and singular value decomposition techniques. Locally linear models were introduced by Schreiber and Grasssberger [SG91] for noise reduction, while Davies proposed initially gradient descent [Dav93] and later Levenberg-Marquardt [Dav94] methods for the minimization of the dynamic error. The first attempt to a Bayesian noise reduction framework [Rob07] was due to Davies [Dav98]. Other methods include the usage of shadowing methods [Jud08], wavelet transformations [Jan12], Sequential Markov Chain methods [DDFG01] in the case of state space models, and Kalman filtering techniques [WM97], while important theoretical results about the consistency of signal extraction, under measurement noise, were presented by Lalley et al. [L$^+$99, LN06].

Extensive studies on the effect of dynamical noise on the underlying deterministic system include the works of Jaeger and Kantz [JK97b], and, Strumic and Macek [SM08]. Moreover, in the presence of dynamical noise, shadowing trajectories of non-hyperbolic maps is not possible. This problem was addressed by Kantz [JK97a], introducing a noise reduction method based on "parameter shadowing". In this work, a shadowing pseudo-orbit is generated, evolving in some neighborhood of the original orbit, fulfilling the nearby rather than the exact dynamics.

We will propose a fully Bayesian nonparametric method for the reduction of the additive dynamical noise perturbing an observed noisy time series (x_i) of length n. In particular, we develop the DNRR model, whereby we introduce the n strategic hidden random variables (Y_i). Their posterior distribution describes all possible noise reduced trajectories in the neighborhood of the original trajectory, and we show that with the appropriate point estimation, we can recover a noise reduced trajectory (y_i) that for moderate noise levels is being generated

by approximately the same dynamical system, generating the observed noisy time series, yet perturbed by a weaker error process. We also show that near the homoclinic tangencies of the associated deterministic system, the posterior marginal distributions Y_i become multimodal limiting the noise reduction levels.

The novelty of our approach lies on the fact that we make no parametric assumptions for the density of the noise component. Instead, we model the additive error using a highly flexible family of density functions, which are based on a Bayesian nonparametric model, namely the Geometric Stick Breaking process [FGMW10], extending previous works regarding reconstruction and prediction of random dynamical systems [HNW07b, HNW09, MKH17].The noise reduction method proposed can be applied to cases where the noise is not assumed to be normally distributed, or even in cases where we know that the noise component has a mixture density. Such cases include, among others, scenarios where the noise is the result of multiple sources affecting the time evolution of the underlying dynamics. In this case, our method will be able to estimate the true noise density and moreover identify the number of the sources as the ergodic average of the active clusters.

The chapter is organized as follows. In Sec. 4.2 we mention some aspects of the problem and present the noise reduction algorithm steps. In Sec. 4.3 we present the MCMC procedure for the estimation of the noise-reduced orbit. In Sec. 4.4 we resort to simulation. We illustrate our method in the case of the random full quadratic and polynomial maps under non-Gaussian dynamical noise. We conclude in Sec. 4.5 giving some directions for further research.

4.2 Preliminaries

We define the random recurrence relation given by

$$\begin{aligned} X_i &= T(\theta, X_{i-1}, \ldots, X_{i-d}, e_i) \\ &= g(\theta, X_{i-1}, \ldots, X_{i-d}) + e_i, \quad i \geq 1, \end{aligned} \tag{4.1}$$

where $g : \Theta \times X^d \to X$, for some compact subset X of $\mathbb{R}$, $(X_i)_{i \geq -d+1}$ and $(e_i)_{i \geq 1}$ are real random variables over some probability space $(\Omega, \mathcal{F}, \mathrm{P})$; we denote by $\theta \in \Theta \subseteq \mathbb{R}^m$ any dependence of the deterministic map g on parameters. g is nonlinear, and for simplicity, continuous in $X_{i:d} := (X_{i-1}, \ldots, X_{i-d})$. We assume that the random variables e_i are independent to each other, and independent of the states X_{i-r} for $r < i + d$. In addition we assume that the additive perturbations e_i are identically distributed from a zero mean distribution with unknown density f defined over the real line, so that $T : \Theta \times X^d \times \mathbb{R} \to \mathbb{R}$. Finally, notice that the lag-one stochastic process $(W_i^1, \ldots, W_i^d)$, formed out, from time-delayed values of the (X_i) process, defined by

$$W_i^k = \begin{cases} g(\theta, W_{i-1}^1, \ldots, W_{i-1}^d) + e_i & k = 1 \\ W_{i-1}^{k-1} & 1 < k \leq d \,, \end{cases}$$

is Markovian over $\mathbb{R}^d$.

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 nonlinear stochastic process defined in Eq. (4.1). The time series x^n depends solely on the initial distribution of $X_{1:d}$, the vector of parameters θ, and the particular realization of the noise process.

Orbits contaminated with dynamical noise are a-pseudo-orbits of the underlying g-dynamics in the sense that for all $1 \leq i \leq n$ there is positive a for which $0 < |x_i - g(\theta, x_{i:d})| \leq a$. g-invariant measures $\mu_g(dx)$ are deformed and smoothed-out into T-quasi-invariant measures $\mu_T(dx) = \lim_{t\to\infty} \mathrm{P}\{x < X_t \leq x + dx | \tau_{X'} > t\}$, where $\tau_{X'}$ is a random time denoting the first passage time of the system to the unbounded trapping set $X' = \mathbb{R} \setminus X$. Mind that, μ_T is not a convolution of the unperturbed measure μ_g with the noise distribution, as it happens in the case of observational noise.

As a distance between the two time series x^n and y^n, we will use the average correction

$$E_0(x^n, y^n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}.$$

We will measure the overall deviation of the noisy orbit x^n from the g-determinism, with the average dynamical error

$$E_{\mathrm{dyn}}(x^n; g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - g(\theta, x_{i:d}))^2}.$$

4.2.1 Dynamical noise reduction

Dynamical noise has a severe effect on the underlying dynamics, i.e. the deterministic part of the noisy corrupted time series, especially when the system under consideration is non-hyperbolic. In the hyperbolic case, the shadowing lemma guarantees the existence of shadowing pseudo-orbits and moreover if the dynamical noise is bounded, it can be treated as measurement noise. This means that we can find a g-deterministic orbit y^n and a noise process $(\tilde{z}_i)$ such that $x_i = y_i + \tilde{z}_i$. The $\tilde{z}_i$ errors are describing the distribution of the distance between the two orbits, and the x^n-dynamical noise reduction problem can be treated as a y^n-observational noise reduction problem. This is not valid, though, when the underlying dynamics are non-hyperbolic.

In the non-hyperbolic case, the presence of homoclinic tangencies (HTs) in the phase space, points where the stable and unstable manifold of a hyperbolic orbit intersect tangentially, is responsible for the emergence of a much more complicated structure. In the vicinity of HT's, the dynamic perturbations are amplifying dynamics away from the neighborhood of the attractor. One of the effects caused by the noise amplifications due to HT's are noise-induced prolongations [JK97a]. For example, in Figure 4.1, we display the delay plots of the deterministic and a dynamically perturbed realization of the Hénon map of lengths $n = 5000$. The noisy trajectory,

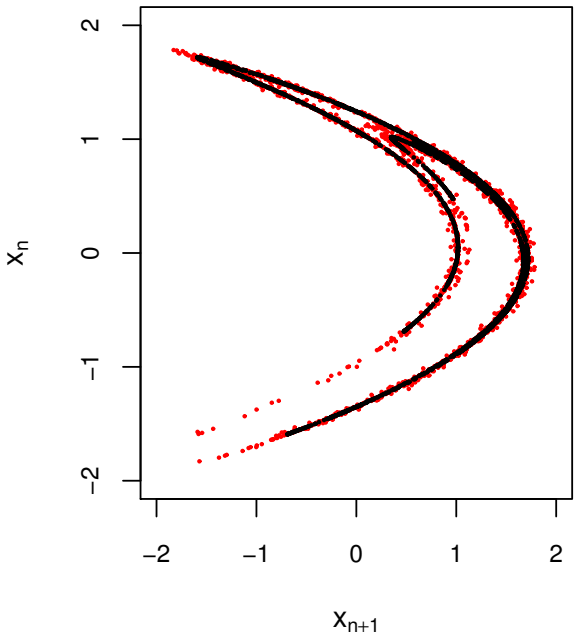

Figure 4.1: Noisy and deterministic Hénon trajectories, of length $n = 5000$, are depicted in red and black, respectively, for a 3% dynamical noise level.

has been generated via

$$x_i = g(x_{i-1}, x_{i-2}) = 1.38 - x_{i-1}^2 + 0.27x_{i-2} + e_i, \tag{4.2}$$

where $e_i \overset{\text{iid}}{\sim} 0.6\mathcal{N}(0, \sigma^2) + 0.4\mathcal{N}(0, 100\sigma^2)$, for $\sigma^2 = 0.21 \times 10^{-4}$, with initial condition $x_0 = x_{-1} = 0.5$. The time series realization has been chosen, such that, the noise level is approximately 3%. We can see the intense noise induced prolongations, as clouds of points in red, away from the neighborhood of the deterministic attractor (points in black).

The deformation of the g-invariant measure to a T-quasi-invariant measure, leads to the expansion of its support, and the perturbed map visits areas of the phase space that was not able to visit without the effect of the dynamical noise.

We aim to reconstruct the underlying deterministic dynamics in the form of a map $\hat{g}_{x^n}$, and sample a y^n trajectory, such that we will be able to control its average deviation from determinacy $E_{\text{dyn}}(y^n, \hat{g}_{x^n})$, with respect to $\hat{g}_{x^n}$, as well as its average correction $E_0(x^n, y^n)$, with respect to x^n.

4.2.2 Gaussian and non-Gaussian noise processes

We assume that the corrupting noise f, responsible for the observed time series x^n, can be represented as a countable mixture of zero mean normals $\mathcal{N}(z|0, \sigma_i^2)$ of variances σ_i^2, that is

$$f(z) := f_M(z) = \sum_{i=1}^{M} p_i \mathcal{N}(z|0, \sigma_i^2)$$

with $p_i > 0$ and $\sum_{i=1}^{M} p_i = 1$, where M can be infinite. Then, the variance associated with f_M (when it exists), is the p_i-mixture of the σ_i^2-variances i.e. $\sigma_{f_M}^2 = \sum_{i=1}^{M} p_i \sigma_i^2$. Following Jaeger and Kantz [JK97a] we define the noise level η as the percentage of the sampling standard deviation of x^n (the signal), that is, $\eta = 100\,\sigma_f / \sigma_{x^n}$.

As a measure of the departure from normality of the noise process f, we use the mean absolute deviation from the mean, normalized by the standard deviation. So for a zero mean $Z \sim f$ it is that $TF_f := \mathbb{E}|Z| / \sqrt{\mathbb{E}|Z|^2}$. The closer the quantity TF_f is to one, the thinner the tails are. We have the following lemma:

Lemma 1. *For all $M \geq 1$, it is that*

$$TF_{f_M} \leq TF_{f_1} \quad \text{with} \quad TF_{f_M} = \frac{1}{\sigma_{f_M}} \sqrt{\frac{2}{\pi}} \sum_{i=1}^{M} p_i \sigma_i. \tag{4.3}$$

Proof: We let $|Z| \sim f_+$, then it is clear that

$$f_+(z) = \frac{f_M(z)\,\mathcal{I}(z>0)}{\int_{\mathbb{R}^+} f_M(z) dz} = 2 \sum_{i=1}^{M} p_i \mathcal{N}(z|0,\sigma_i^2) \mathcal{I}(z>0),$$

where $\mathcal{I}(z > 0)$ is the characteristic function of the interval $(0, \infty)$. The equation for TF_{f_M} in (4.3) can be verified by the fact that $\int_{\mathbb{R}^+} z\,\mathcal{N}(z|0,\sigma_i^2) dz = \sigma_i \sqrt{2/\pi}$. By Jensen's concave inequality we have that $\sum_{i=1}^{M} p_i \sigma_i \leq \sigma_{f_M}$ or equivalently that $TF_{f_M} \leq TF_{f_1}$. □

We consider the noise processes f_1 and $\{f_{2,l} : 1 \leq l \leq 4\}$ given by

$$\begin{aligned} f_1(z) &= \mathcal{N}(z|0,\sigma^2) \\ f_{2,l}(z) &= \frac{5+l}{10} \mathcal{N}(z|0,\sigma^2) + \frac{5-l}{10} \mathcal{N}\left(z|0,100\sigma^2\right). \end{aligned} \tag{4.4}$$

From lemma 1, irrespective of the choice of σ^2, it is that $TF_{f_{2,l}} < TF_{f_1} = \sqrt{2/\pi}$, and the $TF_{f_{2,l}}$ sequence is decreasing, namely, it can be verified that $\{T_{f_{2,l}} : 1 \leq l \leq 4\} = \{0.58, 0.53, 0.49, 0.46\}$.

The motivation for a Bayesian *nonparametric* framework for noise reduction comes from the fact that the application of stochastic methods, under the false assumption of a normal noise process ($f = f_1$), will artificially enlarge the estimated variance of the presumed normal errors, thus, causing poor inference for the system parameters of interest, as demonstrated in Ref. [MKH17] and in Chapters 2-3.

4.3 The dynamic noise reduction replicator model

Given a noisy corrupted time series x^n, we will use a Bayesian nonparametric approach to estimate the posterior joint density of a noise reduced vector of random variables $Y^n = (Y_1, \ldots, Y_n)$. A noise reduced time series $y^n = (y_1, \ldots, y_n)$, will be formed by some central tendency statistic

applied to predictive samples of the marginal posterior densities (MPDs) for all $i = 1, \ldots, n$. We define, the estimated vectors $\hat{\theta}_{x^n}$ and $\hat{\theta}_{y^n}$, of the control parameters of g, and the associated estimated noise components $\hat{f}_{x^n}$ and $\hat{f}_{y^n}$, based on the time series x^n and y^n, respectively. Our intention is to create the y^n time series, in such a way, that it possesses an underlying estimated deterministic law, $\hat{g}_{y^n}(\,\cdot\,) := g(\hat{\theta}_{y^n}, \cdot\,)$, that is in some sense (to be made precise in the sequel) close to the estimated deterministic law, $\hat{g}_{x^n}(\,\cdot\,) := g(\hat{\theta}_{x^n}, \cdot\,)$, responsible for x^n, such that, the estimated noise component $\hat{f}_{y^n}$ influencing interactively the y^n time series, will be a weaker version of the estimated dynamic noise component $\hat{f}_{x^n}$ influencing the original x^n time series. We remark that the $\hat{g}_{y^n}$ and $\hat{f}_{y^n}$ estimations under the noise reduced trajectory y^n have been produced via the GSBR–sampler as described in Chapter 3 and in Ref. [MKH17].

4.3.1 A generic probability model

To permit a stochastic approach to the estimation of the unobserved sequence y^n, under the generic assumption of a symmetric zero mean dynamical error process, we adopt the following stochastic model:

$$
\begin{aligned}
x_i &= g(\theta, x_{i:d}) + e_i, \quad e_i \overset{\text{iid}}{\sim} f(\,\cdot\,) \\
f(\,\cdot\,) &= \sum_{k=1}^{\infty} w_k \, \mathcal{N}(\,\cdot\,|0, \lambda_j^{-1}), \; 1 \leq i \leq n \\
y_i &= g(\theta, y_{i:d}) + \zeta_i, \quad \zeta_i \overset{\text{iid}}{\sim} \mathcal{N}(\,\cdot\,|0, \delta) \\
y_{1:d} &= x_{1:d}, \; \mathrm{P}-\text{a.s.} \quad \text{and} \quad |x_i - y_i| < \gamma_i, \quad \gamma_i \overset{\text{iid}}{\sim} h(\,\cdot\,),
\end{aligned}
\tag{4.5}
$$

where we define $w^{\infty} = (w_k)_{k \geq 1}$ to be an infinite sequence of random probability weights, $\lambda^{\infty} = (\lambda_k)_{k \geq 1}$ an infinite sequence of independent and identically distributed (i.i.d) positive random variables (the precisions), with the two sequences w^{∞} and λ^{∞} independent of each other. The positive random variables γ_i are i.i.d. from some distribution h, possibly depending on parameters.

We will show numerically, that under a reasonable choice for the prior distribution of the variable $\tau = \delta^{-1}$, the posterior distribution of δ (the variance), will concentrate its mass near zero. This, will enable us, to minimize the overall deviation of the y^n trajectory from the estimated determinism. To control the similarity of the y^n trajectory, with respect to the observed x^n trajectory, we assume that both trajectories originate from the same initial point, that is, $y_{1:d} = x_{1:d}$. At the same time, a-priori, we restrict each y_i to be γ_i-close to x_i. The latter statement, conveys prior information, on the *proximity* of the variable y_i to the data point x_i. Finally, we remark that the random mixture $\omega \mapsto f(\,\cdot\,, \omega) = \sum_{k=1}^{\infty} w_k(\omega) \, \mathcal{N}(\,\cdot\,|0, \lambda_j^{-1}(\omega))$ undertakes the rôle of a non-parametric prior over the noise density assumed responsible for the time series x^n, supported over the space of densities with mean zero, which are in turn supported over $\mathbb{R}$.

We note the following lemma, which will prove useful in the sequel:

Lemma 2. *Letting* $\mathcal{P} := \cap_{i=1}^{n} \{|x_i - y_i| < \gamma_i\}$, *we have the following cases:*

1. *When $\gamma_i = \bar{\gamma}_i = const.$ a.s. for $1 \leq i \leq n$, it is that*

$$\mathrm{P}\left(\mathcal{P}|\, x^n, y^n\right) = \prod_{i=1}^{n} \mathcal{I}(x_i - \bar{\gamma}_i < y_i < x_i + \bar{\gamma}_i).$$

2. *If $\gamma_i \overset{\text{iid}}{\sim} \mathcal{W}(2, \sqrt{2/\rho})$ for $1 \leq i \leq n$, where ρ is a fixed hyperparameter, and $\mathcal{W}(a, b)$ denotes the Weibull distribution of shape a and scale b, we have that*

$$\mathrm{P}\left(\mathcal{P}|\, x^n, y^n\right) = \exp\left\{-\frac{\rho}{2}\sum_{i=1}^{n}(x_i - y_i)^2\right\}.$$

Proof: (1.) When $\gamma_i = \bar{\gamma}_i$ for all $1 \leq i \leq n$, it is that

$$\mathrm{P}\left(\mathcal{P}|\, x^n, y^n\right) = \begin{cases} 1 & y_i \in (x_i - \bar{\gamma}_i, x_i + \bar{\gamma}_i),\ 1 \leq i \leq n \\ 0 & \text{otherwise.} \end{cases}$$

(2.) Because $\gamma_i \overset{\text{iid}}{\sim} \mathcal{W}(2, \sqrt{2/\rho})$ if and only if $\gamma_i^2 \overset{\text{iid}}{\sim} \mathcal{E}(\rho/2)$, where $\mathcal{E}(a)$ denotes the exponential distribution with mean $1/a$, it is that

$$\mathrm{P}\left(\mathcal{P}|\, x^n, y^n\right) = \prod_{i=1}^{n} \mathrm{P}\{\gamma_i^2 > (x_i - y_i)^2\} = \prod_{i=1}^{n} \exp\left\{-\rho(x_i - y_i)^2/2\right\},$$

which gives the desired result. □

4.3.2 The posterior model

We consider the posterior of the stochastic quantities f, θ, $x_{1:d}$, $y_{1:d}$, τ and y^n given the data set x^n, the restriction $\mathcal{R} := \{y_{1:d} = x_{1:d}\}$, the proximity information $\mathcal{P}$, and the model space $\mathcal{M}$ for the functional representation of the deterministic part $g(\theta, x_{i:d})$; for example, the model space could be the ring $\mathbb{R}[x_{i:d}]$ of polynomial functions in the variable $x_{i:d}$, with coefficients over $\mathbb{R}$. Then, using Bayes' theorem, we have

$$\begin{aligned} &\pi\left(f, \theta, x_{1:d}, y_{1:d}, \tau, y^n |\, x^n, \mathcal{R},\, \mathcal{P}, \mathcal{M}\right) \propto \\ &\quad \pi\left(f, \theta, x_{1:d}, y_{1:d}, \tau\right)\, \pi\left(y^n, x^n, \mathcal{R}, \mathcal{P} | f, \theta, x_{1:d}, y_{1:d}, \tau, \mathcal{M}\right), \end{aligned} \tag{4.6}$$

where $\pi(f, \theta, x_{1:d}, y_{1:d}, \tau)$ is the prior density. Having in mind, that the estimation of the noise density f is equivalent to the estimation of the variables w^∞ and λ^∞, the likelihood factor on the second line of equation (4.6), becomes

$$\begin{aligned} &\pi\left(y^n, x^n, \mathcal{R}, \mathcal{P} | w^\infty, \lambda^\infty, \theta, x_{1:d}, y_{1:d}, \tau, \mathcal{M}\right) = \\ &\quad \mathrm{P}(\mathcal{R}|\, y_{1:d}, x_{1:d})\, \mathrm{P}\left(\mathcal{P}|\, x^n, y^n\right)\, \pi(x^n |\, w^\infty, \lambda^\infty, \theta, x_{1:d}, \mathcal{M})\, \pi(y^n |\, \theta, \tau, y_{1:d}). \end{aligned} \tag{4.7}$$

We believe, that it will be more efficient to control the average corrections between the two trajectories, under the assumption that $\gamma_i^2 \overset{\text{iid}}{\sim} \mathcal{E}(\rho/2)$. For this reason, we augment the conditional part of our posterior by the hyperparameter ρ. Then, taking into account the model representation for the noise components in (4.5), lemma 2, the fact that $\mathrm{P}(\mathcal{R}|\, y_{1:d}, x_{1:d}) = \mathcal{I}(y_{1:d} = x_{1:d})$ and the likelihood representation in (4.7), the posterior becomes

$$\begin{aligned}
&\pi\left(w^\infty, \lambda^\infty, \theta, x_{1:d}, \tau, y^n |\, x^n, \mathcal{R}, \mathcal{P}, \mathcal{M}, \rho\right) \propto \pi\left(w^\infty, \lambda^\infty, \theta, x_{1:d}, \tau\right) \\
&\times \mathcal{I}(y_{1:d} = x_{1:d}) \prod_{i=1}^{n} \mathcal{N}(y_i|\, x_i, \rho^{-1}) \prod_{i=1}^{n} \sum_{j=1}^{\infty} w_j\, \mathcal{N}(x_i|\, g(\theta, x_{i:d}), \lambda_j^{-1}) \prod_{i=1}^{n} \mathcal{N}(y_i|\, g(\theta, y_{i:d}), \tau^{-1}).
\end{aligned}$$

Such a likelihood will lead to a Gibbs sampler with an infinite number of full conditional distributions. To avoid that, we introduce the jointly discrete random vectors $d^n = (d_1, \ldots, d_n)$ and $N^n = (N_1, \ldots, N_n)$ (see Ref. [MKH17], and references therein). The d_i random variable, denotes the component of the random mixture f in (4.5), that the observation x_i came from. In fact, the state space of the d_i variable can be made a.s. finite, if we define the random variable $N_i \sim f_N(\,\cdot\,|\, p)$, where p is a parameter, such that, the conditional random variable $(d_i|N_i)$ attains a discrete uniform distribution over the a.s. finite set $\mathcal{S}_i = \{1, \ldots, N_i\}$. Then, it can be shown, that by letting N_i to follow the particular negative binomial distribution $f_N(N_i|\, p) = N_i\, p(1-p)^{N_i - 1} \mathcal{I}(N_i \geq 1)$, the random weights w_j in (4.5), will form the strictly decreasing geometric sequence $w_j = p(1-p)^{j-1} \mathcal{I}(j \geq 1)$. So that, in the (d^n, N^n)-augmented posterior (4.6), we can switch from the variable w^∞ to the variable p. Finally, the posterior attains the representation

$$\begin{aligned}
&\pi(p, \lambda^\infty, d^n, N^n, \theta, x_{1:d}, \tau, y^{(n)} |\, x^{(n)}, \rho, \mathcal{R}, \mathcal{P}, \mathcal{M}) \propto \pi(p, \lambda^\infty, \tau, \theta, x_{1:d}) \qquad (4.8) \\
&\quad \times \prod_{\substack{i=1 \\ d_i : d_i \leq N_i}}^{n} p^2 (1-p)^{N_i - 1} \lambda_{d_i}^{1/2} \exp\left\{ -\frac{\lambda_{d_i}}{2} (x_i - g(\theta, x_{i:d}))^2 \right\} \\
&\quad \times \quad \mathcal{I}(y_{1:d} = x_{1:d})\, \tau^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[\tau (y_i - g(\theta, y_{i:d}))^2 + \rho (y_i - x_i)^2 \right] \right\}.
\end{aligned}$$

We note that, the likelihood factor in the second line of the previous equation, is very similar to the GSBR-likelihood that appears in equation (11), of Proposition 4.1, in Ref. [MKH17].

4.3.3 Priors and full conditional distributions

To complete the model, we assign independent priors to the variables p, λ^∞, θ, $x_{1:d}$, and τ, namely:

1. We set $\pi(p) = \mathcal{B}(p|a_1, a_2)$, a beta conjugate prior, with fixed shape hyperparameters a_1 and a_2.

2. The variable λ^∞ is an infinite sequence of independent precisions (inverse variances). Nevertheless, the nonparametric MCMC will require, at each sweep, the computation of only an almost surely *finite* number, $N^* = \max_{1 \leq k \leq n} N_k$, of posterior λ_js. Standard

Bayesian modeling suggests to use gamma conjugate prior distributions over the λ_j precision parameters, so we set $\Pi(d\lambda^\infty) = \prod_{j=1}^{\infty} \mathcal{G}(\lambda_j|b_1, b_2)d\lambda_j$, where b_1 and b_2 are the fixed shape and rate hyperparameters, respectively. Similarly, because τ is a precision, we set a-priori $\pi(\tau) = \mathcal{G}(\tau|\gamma_1, \gamma_2)$.

3. For the vector of parameters $\theta = (\theta_1, \ldots, \theta_s)$ and for the the vector of initial conditions $x_{1:d} = (x_0, \ldots, x_{1-d})$, we assume the independent priors $\pi(\theta) \propto 1$ and $\pi(x_{1:d}) \propto 1$, respectively. For example, suppose that a-priori we have

$$\pi(\theta_1, \ldots, \theta_s) \propto \prod_{i=1}^{s} \exp\{-(\theta_i - \theta_{0,i})^2/2\sigma_{0,i}^2\},$$

then letting $\sigma_{0,i}^2$ tend to infinity, one obtains $\pi(\theta) \propto 1$. Such a prior is noninformative, and although improper (not a density over $\mathbb{R}^s$), leads to a proper full conditional for θ.

Note that, to reduce dynamical error, the prior expectation $\mathbb{E}(\delta)$ will have to be set close to zero. And if at the same time, we want to control the proximity between the original and the noise reduced orbit, we will have to predetermine values for the prior means of γ_is, in the interval $[2 \times 10^{-6}, 2 \times 10^{-4}]$. This is due to the fact, that the individual distances $|x_i - y_i|$ are by construction small.

We have the following proposition:

Proposition 4.1. *The full conditional distributions for the noise reduced orbit y^n, are given by $\pi(y_j|\cdots) \propto e^{-C(y_j|\cdots)/2}$, where $\pi(y_j|\cdots)$ denotes the dependence of the variable y_j to the rest of the variables. Letting $h_\theta(y_j, y_{j:d}) := (y_j - g(\theta, y_{j:d}))^2$, the function $C(y_j|\cdots)$, for $j = 1, \ldots, d$ is given by*

$$\begin{aligned} C(y_j|\cdots) = \tau \sum_{k=0}^{d} h_\theta(y_{j+k}, y_{j+k:d}) \\ \times \mathcal{I}(y_0 = x_0, \ldots, y_{-d+j} = x_{-d+j}) + \rho(y_j - x_j)^2, \end{aligned}$$

for $j = d+1, \ldots, n-d$ is given by

$$C(y_j|\cdots) = \tau \sum_{k=0}^{d} h_\theta(y_{j+k}, y_{j+k:d}) + \rho(y_j - x_j)^2,$$

and, for $j = n-d+1, \ldots, n$, by

$$C(y_j|\cdots) = \tau \sum_{k=0}^{j-n} h_\theta(y_{j+k}, y_{j+k:d}) + \rho(y_j - x_j)^2.$$

Proof. The desired result, comes from the representation of the posterior in equation (4.8). □

4.3.4 The DNRR sampler

To accelerate the convergence of the Gibbs sampler based on the posterior distribution in (4.8), we collect our variables in to the two groups:

$$G_1 = \{v, \theta, x_{1:d}\} \quad \text{and} \quad G_2 = \{\tau, y^n\}$$

with $v = \{p, \lambda^\infty, d^n, N^n\}$. We first sample, from the full conditionals of G_1 given x^n, and then, from the full conditionals of G_2 given G_1 and x^n. Then, it is not difficult to see that such a *blocked* Gibbs sampler scheme, admits the same stationary distribution as the plain Gibbs sampler scheme, coming from sampling the full conditionals of $G_1 \cup G_2$ given x^n, each one individually.

Proposition 4.2. *Given the model $\mathcal{M}$ and fixed $\rho > 0$, marginally, $(G_2|x^n)$ is distributed as*

$$(G_2|x^n) \sim \int_{\mathbb{R}^{s+d}\times\mathbb{V}} \Pi(\,\cdot\,,\,\cdot\,|\theta, x_{1:n}, x^n)\Pi(d\theta, dx_{1:n}|v, x^n)\Pi(dv|x^n), \tag{4.9}$$

where $\mathbb{V}$ denotes the support of the random vector v.

Proof. Given the model $\mathcal{M}$, and fixed $\rho > 0$, we want to sample from the variable $(\tau, y^n|x^n)$. To do so, we should first sample from the joint of θ and $x_{1:n}$ given x^n, and then from the joint of τ and y^n given θ and $x_{1:n}$, that is

$$(\theta, x_{1:n}|x^n) \sim \Pi(\,\cdot\,,\,\cdot\,|x^n)$$

and then from

$$(\tau, y^n|x^n) \sim \Pi(\,\cdot\,,\,\cdot\,|\theta, x_{1:n}, x^n)$$

whence

$$(\tau, y^n|x^n) \sim \int_{\mathbb{R}^{s+d}} \Pi(\,\cdot\,,\,\cdot\,|\theta, x_{1:n}, x^n)\Pi(d\theta, dx_{1:n}|x^n).$$

For a generic noise source, we have to sample first from $(p, \lambda^\infty|x^n)$, and then from $(\theta, x_{1:n}|p, \lambda^\infty, x^n)$. However, for the creation of an a.s. finite Gibbs sampler, the random vector (d^n, N^n) has to be introduced. Then, letting $v = (p, \lambda^\infty, d^n, N^n)$, one has

$$(\theta, x_{1:n}|x^n) \sim \int_{\mathbb{V}} \Pi(\,\cdot\,,\,\cdot\,|v, x^n)\Pi(dv|x^n),$$

which gives the desired result. □

Now, it is clear, that our model is based on the iteration of two consecutive steps, the $(\hat{g}_{x^n}, \hat{f}_{x^n})$-reconstruction step and the y^n-sampling step:

1. We have seen that the reconstruction step, stems from the GSBR-sampler introduced in Ref. [MKH17]. The differences are: the absence of the out-of-sample variables, the more general d-dimensional lag dependence, the application of a conjugate beta prior and the application of an improper prior, on the variables p and $(\theta, x_{1:d})$, respectively.

2. In the noise-reduction step, the noise reduced trajectory y^n, is sampled conditionally on the sampled values of the reconstruction step. We can think of the reconstruction stage, as providing observations from the *distributions* of the initial condition $y_{1:n}$ and the parameter θ of the estimated deterministic part $\hat{g}_{x^n}$ of the new trajectory. To replicate the $\hat{g}_{x^n}$-dynamics, under a reduced dynamical error, we use a Metropolis within Gibbs updating procedure, with a small variance random walk proposal distribution, initialized at the observed x^n trajectory.

Then, the new trajectory y^n, has the following properties:

1. We define, the relative dynamical noise reduction R_{dyn} attained by the y^n trajectory with respect to the y^n as
$$R_{\text{dyn}}(y^n, x^n; \hat{g}_{x^n}) := 1 - \frac{E_{\text{dyn}}(y^n; \hat{g}_{x^n})}{E_{\text{dyn}}(x^n; \hat{g}_{x^n})},$$
so that $R_{\text{dyn}} > r$ implies $E_{\text{dyn}}(y^{(n)}; \hat{g}_{x^n}) < (1-r)E_{\text{dyn}}(x^{(n)}; \hat{g}_{x^n})$. We will see, that in all our numerical examples, it is that with $r > 0.8$.

2. When ρ tends to infinity, the distribution of distances between the individual points of the y^n and x^n trajectories, concentrates its mass to zero.

3. The estimated underlying deterministic parts of y^n and x^n are close to each other. For suppose, that we estimate in terms of the GSBR-sampler, the g-dynamics given the x^n and the y^n trajectories. Then the distance $d(\hat{g}_{x^n}, \hat{g}_{y^n})$ between the two deterministic parts will be small; for example, when $\hat{g}_{x^n}$ and $\hat{g}_{y^n}$ are polynomials, this distance could be the l_2-norm of the polynomial $\hat{g}_{x^n} - \hat{g}_{y^n}$.

The sampling scheme: We first specify initial values for the variables $x_{1:n}$, θ, τ, and we iterate for $t = 1, \ldots, K$ the following sampling scheme:

S1: For $i = 1, \ldots, n$, generate the state space range variable $N_i^{(t)} \sim \pi(N_i| \cdots)$, of the allocation variable $d_i^{(t)}$.

S2: For $i = 1, \ldots, n$, generate the infinite mixture allocation variable $d_i^{(t)} \sim \pi(d_i| \cdots)$.

S3: For $i = 1, \ldots, N^*$, with $N^* = \max_{1 \leq k \leq n} N_k$, sample $\lambda_i^{(t)} \sim \pi(\lambda_i| \cdots)$.

S4: Generate the initial condition vector $(x_{1:n})^{(t)} \sim \pi(x_{1:n}| \cdots)$

S5: Generate $\theta^{(t)} \sim \pi(\theta| \cdots)$.

S6: Sample the geometric probability $p^{(t)} \sim \pi(p| \cdots)$.

S7: Having updated $p^{(t)}$ and $\lambda^{(t)}$ up to N^*, sample from the noise process $\hat{f}_{x^n}$

$$z_{n+1}^{(t)} \sim \sum_{j=1}^{N^*} p^{(t)}(1-p^{(t)})^{j-1} \mathcal{N}\left(z_{n+1} \,|\, 0,\, 1/\lambda_j^{(t)}\right).$$

S8: Initialize the vector of initial conditions $(y_{1:n})^{(t)}$ of the noise reduced trajectory to the previously sampled initial condition $(x_{1:n})^{(t)}$ of the x^n, and iterate for $j = 1, \ldots, n$ the following Metropolis-within-Gibbs sampling scheme:

(a) Generate proposal

$$y_j^* \sim y_j^{(t-1)} + \nu\,\mathcal{N}(0,1). \tag{4.10}$$

(b) Calculate the acceptance probability $\alpha(y_j^{(t-1)}, y_j^*)$ given by

$$\min\left\{1, \exp\left\{-\frac{1}{2}\left[C(y_j^*|\cdots) - C\left(y_j^{(t-1)}|\cdots\right)\right]\right\}\right\}.$$

(c) Accept $y_j^{(t)} = y_j^*$ with probability $\alpha(y_j^{(t-1)}, y_j^*)$.

S9: Generate $\tau^{(t)} \sim \pi(\tau|\cdots)$.

4.4 Simulation Results

In this section, we will provide numerical illustrations of the DNRR algorithm for the random Hénon map, and the random bistable cubic map introduced in Ref. [MKH17]. In all cases, except in the case for the variable τ, the prior specifications are completely noninformative.

As a prior for the geometric probability variable, we take the arcsine density $p \sim \mathcal{B}(0.5, 0.5)$, which coincides with the Jeffrey's prior for p. On the precisions $(\lambda_j)_{j\geq 1}$ of the random density f, we place the vague gamma prior $\lambda_j \sim \mathcal{G}(10^{-3}, 10^{-3})$, which is very close to a scale invariant prior. On the control variable θ, and the initial condition variable $x_{1:d}$, we assign the translation invariant priors $\pi(\theta) \propto 1$ and $\pi(x_{1:d}) \propto 1$, respectively. Because we want a-posteriori to force the variance $\delta = \tau^{-1}$ to concentrate its mass near zero, we have to set its prior mean and variance close to zero. We can achieve this by setting $\tau \sim \mathcal{G}(10^4, 10^{-2})$. Finally, to avoid mixing issues, following standard methodology, each time, we calibrate [RC04] the proposal variance ν^2 of the embedded Metropolis-within-Gibbs sampler in equation (4.10), such that, the mean acceptance probability of the sampling scheme is between 25 and 35%.

In all our numerical experiments, the DNRR Gibbs samplers have ran for 25×10^4 iterations leaving the first 5×10^4 samples as a burn-in period.

4.4.1 The Hénon map

We consider a time series realization x^n of size $n = 1000$, coming from the random recurrence relation given in (4.2) with $e_i \overset{\text{iid}}{\sim} f_{2,1}$, variance $\sigma^2 = 0.21 \times 10^{-4}$ and initial condition $x_0 = x_{-1} = 0.5$ for noise level at approximately 3%. We model the deterministic part g, with the complete quadratic polynomial in the two variables, namely

$$g(\theta, x_{i-1}, x_{i-2}) = \theta_0 + \theta_1 x_{i-1} + \theta_2 x_{i-2} + \theta_3 x_{i-1}x_{i-2} + \theta_4 x_{i-1}^2 + \theta_5 x_{i-2}^2. \tag{4.11}$$

Table 4.1: Relative dynamical noise reductions, average indeterminisms and average distances, for two different values of ρ.

ρ	$E_{\text{dyn}}(x^n, \hat{g}_{x^n})$	$E_{\text{dyn}}(y^n, \hat{g}_{x^n})$	R_{dyn}	E_0
10^2	0.02932	0.00286	0.9023	0.0428
5×10^5	0.02932	0.00710	0.7577	0.0223

1. A neutral proximity restriction: We first ran the DNRR sampler with the proximity parameter set to $\rho = 10^2$. In fact, values of ρ smaller than 10^4, due to the informative nature of τ, have a diminishing effect on the full conditional distributions of the Y_j variables of Proposition 4.1. As a result, for such 'small' ρ values, the proximity restriction $\mathcal{P}$ becomes neutral, and the DNRR sampler estimates the noise reduced orbit y^n attaining minimum average deviation E_{dyn} with respect to the estimated $\hat{g}_{x^n}$, and maximum average distance with respect to x^n.

In the first two rows of Table 4.2, we present *percentage absolute relative errors* (PAREs) of the estimated θ-coefficients, with respect to the true values, based on the noisy and the noise reduced trajectories, of the maps $\hat{g}_{x^n}$ and $\hat{g}_{y^n}$, respectively. The last two columns of the table, display average PAREs, $\bar{\theta}$, and l^2-distances. Because the y^n based quantities, $\bar{\theta}$ and l^2, are small, we consider both x^n and y^n based θ-estimations as identifying the specific Hénon map given in (4.2).

The posterior variance $\delta = \tau^{-1}$ has the interval $[1.39 \times 10^{-6}, 1.81 \times 10^{-6}]$ as a 95% highest posterior density interval. The distribution of the individual variances of the y^n trajectory, concentrates most of its mass in the interval $[0, 10^{-5}]$.

We have the following results presented from Figure 4.2 to Figure 4.6:

1.1. Noise reduction measures: In Figure 4.2(a), we present superimposed the original time series x^n (points in red), and the estimated noise reduced trajectory y^n (points in dark gray) in delay coordinates. We can see the noise reduced trajectory, shadowing the original trajectory, in the regions of noise-induced prolongations. In Figure 4.2(b), we display superimposed the individual $\log_{10}$-determinism plots of the original and the estimated time series, in red and dark gray color, respectively; for example, the individual $\log_{10}$-determinism plot of the time series (x_i) is the trace of time series $\left(\log_{10} |E_{\text{dyn}}(x_i, \hat{g})|\right)$. The red and black horizontal lines correspond to the average $\log_{10}$-determinisms of the noisy and the noise reduced times series, respectively. In the first line of Table 4.1, we exhibit the denoising measures E_{dyn}, R_{dyn} and E_0. The average noise reduction achieved by the DNRR sampler is larger than two orders of magnitude, with $R_{\text{dyn}}(y^n, x^n; \hat{g}_{x^n}) = 0.902$, $E_{\text{dyn}}(y^n; \hat{g}_{x^n}) = 0.00286$ and $E_0(x^n, y^n) = 0.0428$.

1.2. Dynamic noise estimation: In Figure 4.3, we display superimposed the true noise density $f = f_{2,1}$ (red continuous curve), the x^n based estimated noise density $\hat{f}_{x^n}$ (black continuous curve) and the y^n based estimated noise density $\hat{f}_{y^n}$ (black dashed curve). We remark the closeness of the noise densities f and $\hat{f}_{x^n}$, and the fact that the $\hat{f}_{y^n}$ density, represents a much 'weaker' error process. The latter, along with the fact that the θ-estimation based on the noise reduced trajectory identifies the specific Hénon map, validates our contention, that the noise

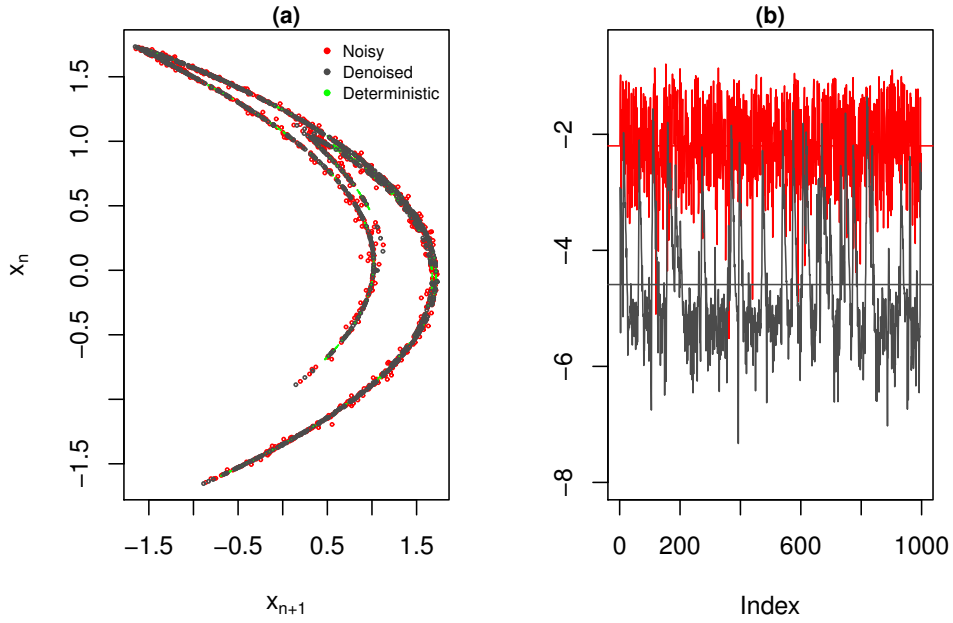

Figure 4.2: In figure (a), we present superimposed delay plots of the noisy, the noise reduced and the deterministic trajectories of the Heńon map, of length $n = 1000$. The associated $\log_{10}-$determinism plot is given in figure (b).

Table 4.2: PAREs, average PAREs and l^2-distances, for the estimated coefficients of the deterministic part of the perturbed Hénon map in (4.2), based on the noisy and the corresponding noise reduced trajectories, for two different values of ρ.

Time series	ρ	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	$\bar{\theta}$	l^2
x^n	10^2	0.089	0.096	0.046	0.044	0.011	0.070	0.059	0.00177
y^n		0.063	0.043	0.022	0.028	0.020	0.038	0.036	0.00110
x^n	5×10^5	0.079	0.071	0.041	0.031	0.002	0.059	0.047	0.00146
y^n		0.177	0.155	0.015	0.023	0.005	0.157	0.089	0.00330

reduced trajectory comes from a dynamical system very close to the original one, perturbed interactively by a 'weaker' error process.

1.3. The existence of HTs as a cause for a-posteriori multimodality: While most of the Y_i-MPDs are unimodal, a small number of them is multimodal, namely, those that their support contains the projection of a point of HT. We have used the Hartigan's statistical test [HH85] for multimodality, to choose the appropriate Y_i-point estimator; we utilize the maximum a-posteriori (MAP) estimator for the case of a Y_i-multimodal MPD, and the sample mean estimator for the unimodal case. In Figure 4.4(a) we present a delay plot of the set M_{HT} of MAP estimations (solid red circles) coming from the Y_i-posterior marginals, passing the Hartigan's test for multimodality. Alternatively, we could consider the Y_i-predictive-samples, coming from the embedded Metropolis-within-Gibbs sampler, after burn-in. For each Y_i-sample, we compute the forecastable component analysis index Ω_i [Goe13, Goe16], which is normalized in the interval $[0, 1]$. We let $\Omega = \{\Omega_i : 1 \leq i \leq n\}$, and we consider the subset of points Ω_{HT} of Ω, that are above the 99th percentile of its histogram, and thus, their predictive distribution exhibits *more structure*. In Figure 4.4(b) we depict a delay plot of Ω_{HT} (solid red circles). We

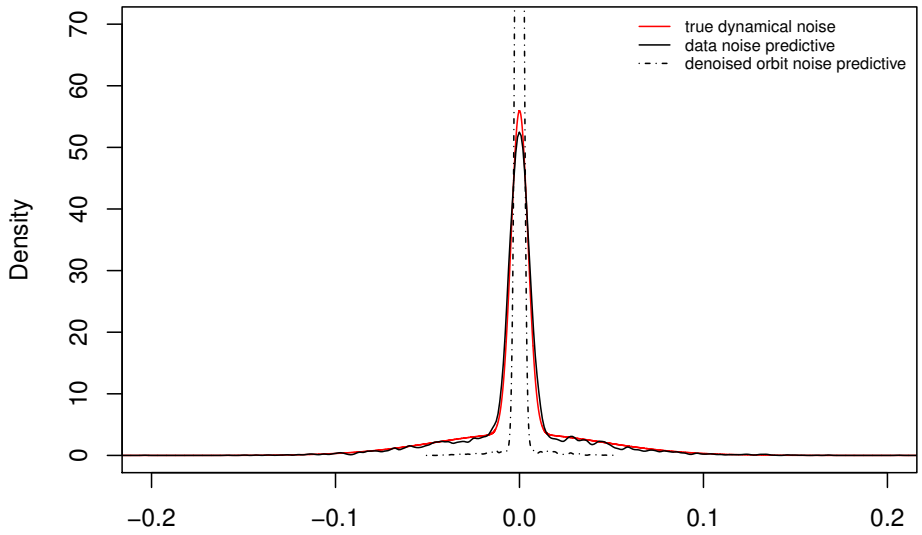

Figure 4.3: The true noise density $f = f_{2,1}$, for $\sigma^2 = 0.21 \times 10^{-4}$, is the red continuous curve. Along, we superimpose the x^n-estimated noise density $\hat{f}_{x^n}$ as a black continuous curve, and the y^n-estimated 'weaker' interactive noise density $\hat{f}_{y^n}$ as a black dashed curve.

can see that the points in the sets M_{HT} and Ω_{HT} are related to the areas of increased indeterminism depicted in Figure 4.4(c). The location of the deterministic primary HTs are given in Figure 4.4(d). We remark that the sets M_{HT} and Ω_{HT}, for fixed n, are random (point process realizations) because they depend on the particular realization of the time series $\omega \mapsto x^n(\omega)$, for example $\omega \mapsto \Omega_{\mathrm{HT}} = \Omega_{\mathrm{HT}}(y^n|x^n(\omega))$.

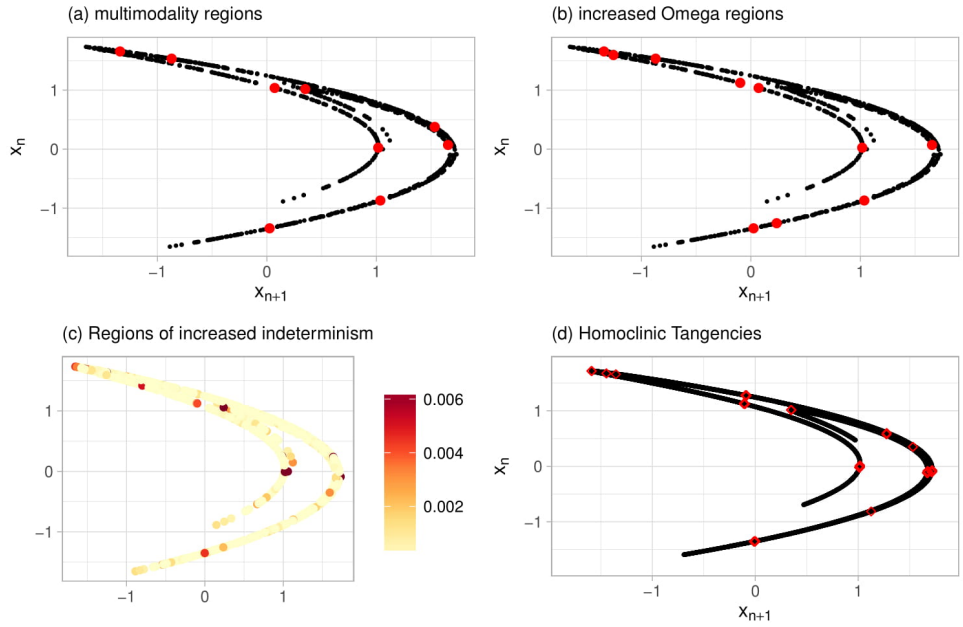

Figure 4.4: In Figure (a) we present a delay plot of the points in the set M_{HT} of the point estimators of the Y_i-posterior marginals, passing Hartigan's test for unimodality. In Figure (b) we depict the delay plot of the points in the set Ω_{HT} that are above the 99th percentile of the histogram of Ω. Regions of high E_{dyn} are depicted in Figure (c), and in Figure (d) we present the primary homoclinic tangencies of the corresponding deterministic attractor.

2. The average distance E_0 as a function of ρ: Here we perform a series of executions of the DNRR sampler with the same prior set up, and the same observed time series x^n, as in the previous subsection, for different values of the ρ parameter. We have taken $\rho \in \{\rho_j = j \times 10^4 : j = 1, \ldots, 200\}$. For example, for $\rho = 5 \times 10^5$, the effect of the proximity restriction becomes very strong. In the second line of Table 4.1, we present the noise reduction measures E_{dyn}, R_{dyn} and E_0. The average noise reduction achieved in this case decreases to$R_{\text{dyn}}(y^n, x^n; \hat{g}_{x^n}) = 0.7577$. The average indeterminism of y^n with respect to $\hat{g}_{x^n}$ escalates to $E_{\text{dyn}}(y^n; \hat{g}_{x^n}) = 0.00710$, with the average distance decreased considerably to $E_0(x^n, y^n) = 0.0223$. In Figure 4.5(a), we present superimposed, the distributions of the individual $\log_{10}$-indeterminisms of the noise reduced trajectory with respect to $\hat{g}_{x^n}$, for $\rho = 10^2$ (curve in black) and $\rho = 5 \times 10^5$ (curve in gray). We can see that for large values of ρ the density of $\log_{10}$-indeterminisms becomes more peaked and shifts to the right. In Figure 4.5(b), the density of the individual distances for the large value of ρ concentrates its mass near zero.

In Figure 4.6, we present the noise reduction measures $E_{\text{dyn}}(y^n, \hat{g}_{x^n})$ and $E_0(y^n, x^n)$ as functions of ρ. It is that as ρ increases, the average indeterminism and the average distance are increasing and decreasing, respectively.

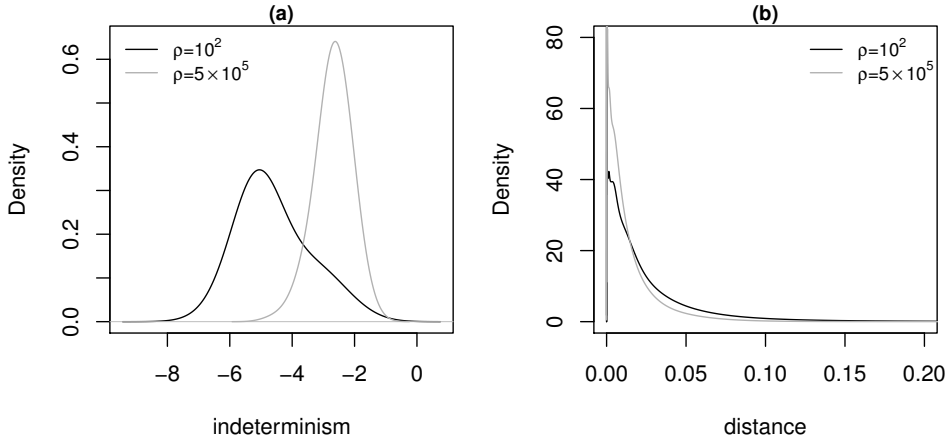

Figure 4.5: KDEs of (a) individual $\log_{10}$-indeterminism points and (b) distance between original and noise reduced orbit points, for different values of parameter ρ.

3. Fixed noise levels imply fixed relative noise reduction: In this experiment we choose the variances and the time series realizations x^n, for each $f_{2,l}$ noise process for $1 \leq l \leq 4$, such that, they give an associated noise level η of about 3%. In the fourth column of Table 4.3, we can see that the relative noise reduction measure R_{dyn}, does not undergo major changes, and it attains values between 0.871 and 0.902.

4.4.2 A bistable cubic map

Here, we consider the cubic map

$$x_i = g(\vartheta, x_{i-1}) = 0.05 + \vartheta x_{i-1} - 0.99 x_{i-1}^3. \tag{4.12}$$

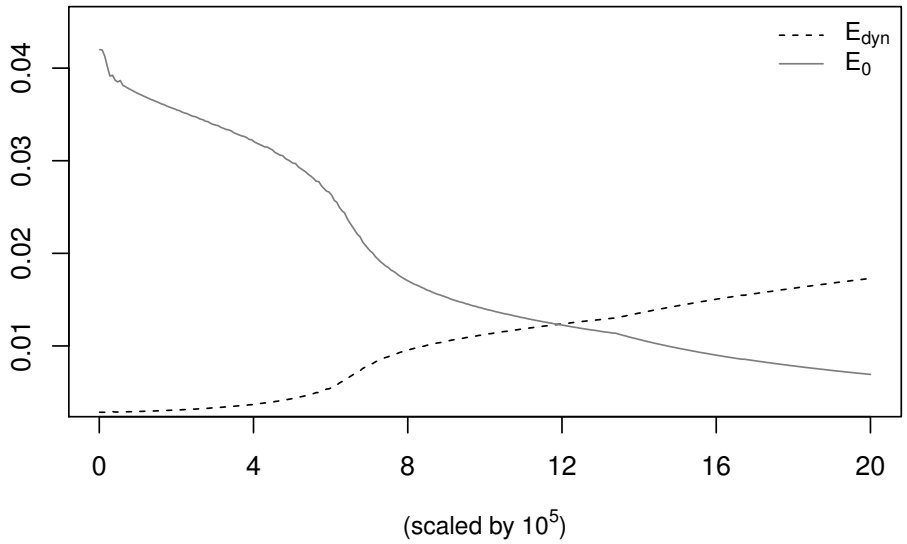

Figure 4.6: The average distance $E_0(y^n, x^n)$ and the average dynamic error $E_{\text{dyn}}(y^n, \hat{g}_{x^n})$ as functions of the parameter ρ.

Table 4.3: Measures of reconstruction and noise reduction efficiency for the $f_{2,l}$ noise processes. The variances of the noise processes, and each realization has been chosen, such that, η is fixed at about 3%, where $E_{\text{dyn}} = E_{\text{dyn}}(y^n, \hat{g}_{x^n})$.

Noise	$\sigma^2 \times 10^4$	E_0	E_{dyn}	R_{dyn}	$\bar{\theta}_{x^n}$	$\bar{\theta}_{y^n}$
$f_{2,1}$	0.21	0.0428	0.00286	0.902	0.059	0.036
$f_{2,2}$	0.29	0.0514	0.00371	0.871	0.115	0.062
$f_{2,3}$	0.40	0.0490	0.00392	0.871	0.072	0.098
$f_{2,4}$	0.77	0.0627	0.00323	0.892	0.054	0.059

For $\vartheta \in \Theta_{\text{bi}} = [1.27, 2.54]$ the map is bistable in the sense that two mutually exclusive period-doubling cascades coexist. For values of ϑ close to 2.54, we denote the two coexisting attractors by $\mathcal{O}_1 \subset I_1$ and $\mathcal{O}_2 \subset I_2$, with approximately $I_1 = [-1.60, -0.10)$ and $I_2 = [-0.10, 1.67]$. For values of ϑ slightly larger than 2.54, the set $\mathcal{O}_2$ undergoes a sudden change. It becomes repelling, and all trajectories over $I_1 \cup I_2$ are attracted by $\mathcal{O}_1$. In fact, similar behavior can be observed for all $\vartheta \in (2.54, 2.65)$.

We let $\vartheta = \vartheta^* = 2.55$ and we consider the dynamically perturbed map $x_i = g(\vartheta^*, x_{i-1}) + e_i$ with $e_i \overset{\text{iid}}{\sim} f_{2,1}$, $\sigma^2 = 0.55 \times 10^{-4}$, and $\rho = 10^2$. Then, *noise-induced jumps* are taking place between the intervals I_1 and I_2. Here we consider dynamically perturbed time series observations x^n, of small sample size $n = 200$. As a modeling polynomial, we utilize the general quintic polynomial $g(\theta, x_{i-1}) = \sum_{k=0}^{5} \theta_j x_{i-1}^k$.

Noise reduction in the neighborhood of noise induced jumps: In Figure 4.7(a), we can see the estimated y^n trajectory (in black) evolving in the neighborhood of the original trajectory x^n (in red), incorporating the weaker dynamical noise $\hat{f}_{y^n}$, given in Figure 4.9, as a black dashed density. We remark, that our method, is based on the fact that it allows only small stochastic steps around the original orbit, and thus, the noise reduced orbit follows closely the original

Table 4.4: Measures of reconstruction and noise reduction efficiency for the cubic map, for various σ^2's for the $f_{2,1}$ noise processes, where $E_{\text{dyn}} = E_{\text{dyn}}(y^n, \hat{g}_{x^n})$.

$\sigma^2 \times 10^4$	η %	E_0	E_{dyn}	R_{dyn}	$\bar{\theta}_{x^n}$	$\bar{\theta}_{y^n}$
0.33	3.5	0.0395	0.00749	0.812	0.281	0.425
0.55	4.5	0.0413	0.00695	0.842	0.605	0.804
0.59	5.5	0.0631	0.00952	0.826	0.438	0.262
0.67	6.5	0.0453	0.00847	0.848	0.872	0.958
1.00	7.5	0.0630	0.00819	0.867	0.856	0.987

orbit even to its noise-induced prolongations in the interval I_2. The corresponding $\log_{10}$ indeterminism plot is given in Figure 4.7(b). The plot of the individual distances between the original and the noise reduced trajectory is given in Figure 4.7(c). In Table 4.4 we display the noise reduction efficiency for the cubic map, for noise levels between 3.5% and 7.5%. In the last column of the table are displayed the average PAREs $\bar{\theta}_{y^n}$ of the y^n based estimation of the deterministic part of the noise reduced dynamics. We have observed, that the average PARE becomes larger than 1%, when the noise level exceeds 8%.

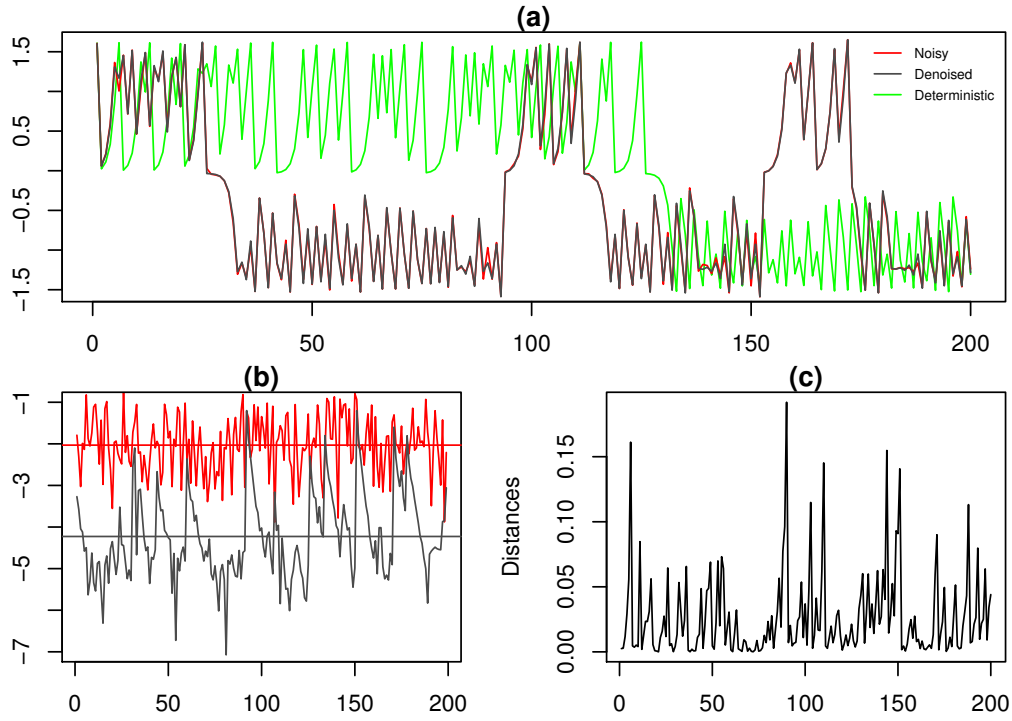

Figure 4.7: In Figure (a), we give superimposed, the deterministic trajectory, the noisy trajectory x^n and the estimated y^n trajectory. In Figure (b) we present the corresponding $\log_{10}$-indeterminism plot. The trace of the individual distances between the original and the noise reduced trajectory is given in Figure (c).

In Fig. 4.8 we presented a delay plot with the corresponding $\log_{10}$-determinism values. We note that the role of the primary HTs in the two-dimensional case, here is played by the critical points of the associated deterministic map. We have experimentally validated the above argument also in the case of the logistic map, where we have one critical point. Such analogies between the presence of HTs and the presence of critical points have also been mentioned e.g. in Refs.

[JK97b, Man85].

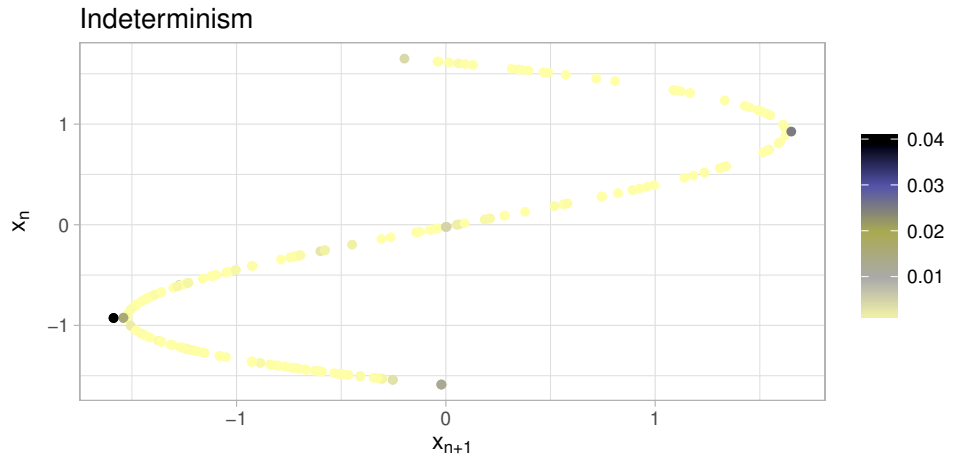

Figure 4.8: Delay plot of the noise-reduced cubic map orbit points, colored by $\log_{10}$-determinism deviations.

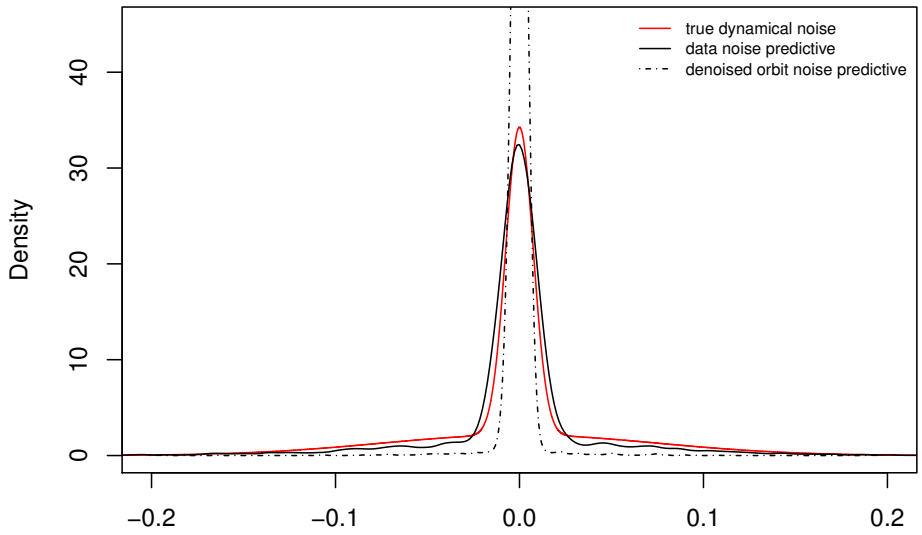

Figure 4.9: Kernel density estimations based on the predictive samples of $\hat{f}_{x^n}$ (continuous black curve), the predictive samples of $\hat{f}_{y^n}$ (dashed black curve) along with the true dynamical noise density (continuous red curve).

4.5 Conclusion

In this chapter, we have presented a novel approach to the problem of noise reduction of dynamically perturbed nonlinear maps, the DNRR sampler. Our approach is Bayesian, modeling a noise reduced trajectory y^n, that *evolves in the neighborhood* of a given noisy trajectory x^n. Our proposed DNRR algorithm, is flexible and accurate, because the assumptions for the underlying noise process f perturbing the original trajectory are relaxed. A-priori, we consider the noise as coming from a random countable mixture of zero mean Gaussians. Then, the number

of the components, the weights, and the variances of the normal mixture $\hat{f}_{x^n}$, approximating the actual noise process f, are estimated directly from the observed time series. This in turn, implies a high accuracy estimation of the deterministic part $\hat{g}_{x^n}$, which is the basic ingredient of the replication part of the DNRR sampler. Also, we have seen, that for moderate noise levels, the noise reduced trajectory y^n, has an estimated deterministic part $\hat{g}_{y^n}$ remaining close to the estimated deterministic part $\hat{g}_{x^n}$ of the original trajectory.

We could modify the proposed DNRR model, by dropping the assumption of a known functional form for the deterministic part, and instead, apply over g, a Gaussian Process prior [Ras04]. We believe, that such an approach, will be appropriate for a wide variety of real world data sets, characterized by strong nonlinearity and (or) complicated contaminating dynamic noise.

Chapter 5

On the stochastic approximation of the global stable manifold

5.1 Introduction

In the recent literature, there has been a growing interest for the construction and analysis of models that include nonlinear deterministic components and are influenced by a variety of noise processes. There exist two main distinct fields of research in relation with random perturbations of dynamical systems. The first regards the analysis of Random Dynamical Systems, by defining direct stochastic generalizations of the central notions appearing in the theory of Dynamical Systems, such as random attractors [CDF97] and random invariant manifolds [MS99b]. The second approach, which we will follow in this chapter, regards the behavior of a perturbed system in the zero-noise limit, i.e. having as reference system the underlying nonlinear deterministic part.

Invariant manifolds are essential to the theory of dynamical systems, as the behavior of any dynamical system is related to the underlying geometrical structure of the state space, more specifically the organization of the invariant stable and unstable subspaces. There exist several methods in the literature suitable for the approximation of invariant manifolds of deterministic and random dynamical systems.

The need for a development of generic methods suitably approximating the invariant manifolds of nonlinear dynamical systems is highlighted by the fact that in general it is not possible to describe the the invariant set of points via closed-form analytic expressions. A special case of a power series expansion approximation for the stable and unstable manifolds of a Hénon map saddle fixed point is presented in Ref. [FR81]. The most widely used method for the approximation of the global stable manifold of any invertible map is [YKY91], the authors also propose a method suitable for noninvertible maps in [KYY96]. Another method for the approximation of the global stable manifold of invertible maps is introduced in [KO+98] and is extended to the Search-Circle algorithm in [EKO04] for the noninvertible case. A modified version of the Search-Circle, which is faster and is based on the same concept is presented in [LFZ12]. For the numerical computation of higher-dimensional invariant manifolds, we refer to [GMJ17, KO99] and [GV04]. For a review of numerical methods for two-dimensional manifolds we refer to

[KOD+05]. A restriction of some of the above methods, can be considered the requirement of the complete knowledge of the dynamical system. A method suitable for the approximation of invariant manifolds given only by experimental data, where there is no available model, is introduced in [TBS03]. Moreover, the "parameterization method" originally introduced by Cabré et. al [CFdlL03a, CFdlL03b], forms the basis of a series of results relevant to the establishment of the existence and the computation [vdBJR16] of invariant manifolds.

In this chapter we are interested in relating the random dynamical systems with their associated deterministic parts. More specifically, we propose an extension of the GSBR sampler introduced in Chapter 3, in order to provide a MCMC based stochastic approximation of the global stable manifold based on an observed noisy time-series. We introduce the Backward GSBR model, in order to compute the support of the densities of a certain set of past unobserved observations, by performing predictions in reversed time. We emphasize on the diffuse support of the posterior marginal densities of the far-off in the past initial conditions. We demonstrate that such supports contain part of the associated noisy stable manifold. The BGSBR sampler can be applied multiple times over proper subsets of the noisy observations, each time generating posterior samples of the various parts of the stable manifold. Then the global noisy stable manifold of the associated noisy time-series observations can be approximated as the union of the supports of the posterior marginal distributions.

The special feature of the initial vector posterior marginals, is that they are the only variables with full conditional distributions depending solely on the observations that are ahead of them in time. All other variables, have full conditional distributions depending on both the future and the past values. This apparent lack of information, regarding past values, makes the support of the associated full conditional to diffuse along the direction of the local stable manifold contained on a neighborhood of the true initial condition. We will see that when we perform prediction in reversed time, the variance of the (new) initial condition vector posterior marginal is increased, spreading the sampled values along the stable manifold.

We will demonstrate the efficiency of the proposed method using different types of polynomial maps, which are of particular interest not only because of their rich dynamical behavior, but also due to their ability to approximate more complicated maps, as a finite degree Taylor approximation of non-polynomial nonlinearities. Furthermore, in essence all invertible polynomial maps, i.e. any nontrivial polynomial diffeomorphism of $\mathbb{R}^2$ with constant Jacobian, are conjugate to compositions of generalized Hénon maps [DM00].

The basic advantage of the proposed method is its ability to provide an adequate stochastic approximation of the stable manifold, under a data driven method. In particular, no knowledge regarding the parameters of the system is required, not even the location of the saddle fixed point, just a general functional representation of the deterministic part. Namely, the procedures of the system identification and of the stochastic approximation are performed in parallel, similarly with the DNRR model described in Chapter 4. Our method is parsimonious, due to the flexibility induced by the general functional form of the deterministic part and the GSB prior, while it is applicable even with small data sets corrupted by (perhaps) non-Gaussian noise. Moreover, an important feature of the BGSBR model is its wide applicability. In fact

we will demonstrate the applicability of the BGSBR algorithm for the case of invertible and non-invertible maps.

This work is related with utilizing a Bayesian non-parametric framework, for approximating dynamical invariants, based on observed time series data. As we have seen in the previous chapters one can in principle approximate

- the quasi-invariant measure of the underlying random dynamic system implying the existence of a prediction barrier.
- the positions of primary homoclinic tangencies (if they exist) revealing the non-hyperbolic nature of the underlying randomly perturbed deterministic dynamic system.

This Chapter is organized as follows. In Section 5.2, we discuss some preliminary notions. In Section 5.3, we derive the Backward Geometric Stick Breaking Reconstruction (B–GSBR) model, a Bayesian nonparametric mixture model suitable for the prediction in reversed time and for the stochastic approximation of the global stable manifold from observed time series data, by applying a GSB prior. In Section 5.4, we resort to simulation. We apply the B–GSBR model on deterministic and random polynomial maps of arbitrary degree that are dynamically perturbed by noise processes which are (perhaps) non Gaussian. Finally, we discuss the conclusions and some directions for future research.

5.2 Preliminaries

5.2.1 Invariant manifolds of deterministic systems

In what follows, we will consider a planar diffeomorphism g of $\mathbb{R}^2$ to itself. The global dynamical behavior of the map g is determined by dynamically invariant objects, such as the stable and the unstable manifolds. Let y a saddle fixed point of g, i.e. the Jacobian $D_y g$ has eigenvalues λ_s, λ_u such that $|\lambda_s| < 1$ and $|\lambda_u| > 1$. The global stable manifold [GL02, KO+98] $W^s(y)$ is the set of points whose orbit tend to y in forward time:

$$W^s(y) = \left\{ x \in \mathbb{R}^2 : g^{(n)}(x) \to y, n \to \infty \right\}.$$

Similarly, the corresponding unstable manifold is defined as the set of points whose orbits tend to y in reversed time, i.e. :

$$W^u(y) = \left\{ x \in \mathbb{R}^2 : g^{(-n)}(x) \to y, n \to \infty \right\}.$$

Since g is invertible, we note that the stable manifold of g is the unstable manifold of g^{-1} and vice versa.

Due to the stretching and folding mechanisms, the two manifolds do not only intersect at the fixed point but also in other locations called homoclinic points. If the intersection between the invariant manifolds is tangential, we have a point of homoclinic tangency and thus infinitely

many, due to invariance under both g and g^{-1}. Homoclinic tangencies (HTs) are important in the study of how noise affects chaotic systems, as they lead to noise amplification. In the symbolic dynamics literature, they have been proposed as a basis for building generating partitions. Furthermore, the existence of HTs is mutually exclusive with the property of hyperbolicity, thus making shadowing impossible. Perturbations in the vicinity of HTs are driven away from the attractor and only due to the folding effects of the nonlinear map the noisy orbits are mapped onto it again. For a detailed description of the above mechanism see Ref. [JK97b].

The Hénon map [Hén76] is defined as

$$g\left(\boldsymbol{\theta}, x, y\right) = \left(\theta_1 + \theta_2 x^2 + \theta_3 y, x\right), \tag{5.1}$$

with inverse:

$$g^{-1}\left(\boldsymbol{\theta}, x, y\right) = \left(x, -\frac{1}{\theta_3}(\theta_1 + \theta_2 y^2 - x)\right). \tag{5.2}$$

The Hénon map exhibits the same qualitative dynamical behavior with a wide variety of polynomial maps, namely quadratic maps with a constant Jacobian.

In Fig. 5.1 we present portions of the stable manifold of the saddle fixed point $(0.631, 0.631)$, for the classical parameter $\boldsymbol{\theta} = (1, -1.4, 0.3)$. We obtain points of the stable manifold by iterating a set of points in small line segments, aligned along the stable direction of the saddle fixed point [GL02]. We notice the emergence of homoclinic tangencies, as the map exhibits non-hyperbolic chaotic behavior. The attractor has been conjectured to be the closure of the unstable manifold.

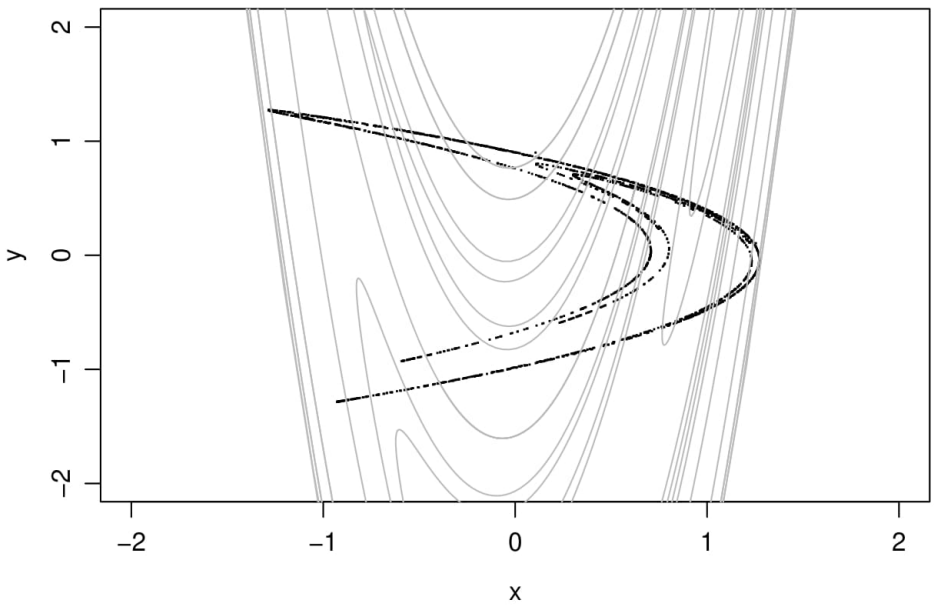

Figure 5.1: Attractor of the Hénon map (black), superimposed with portions of the global stable manifold (gray).

Another illustrating example regarding the relevance of the stable manifold with the basins of attraction in cases of multistability is the dual Hénon map [SX15], which is a special case of the

Generalized Hénon map [DM00] (GHM). The Generalized Hénon map is defined as:

$$g\left(\boldsymbol{\theta}, x, y\right) = \left(\theta_1 + \theta_2 x + \theta_3 x^3 + \theta_4 y, x\right), \tag{5.3}$$

with inverse:

$$g^{-1}\left(\boldsymbol{\theta}, x, y\right) = \left(x, -\frac{1}{\theta_4}(\theta_1 + \theta_2 y + \theta_3 y^3 - x)\right). \tag{5.4}$$

For the values of the control parameters $\boldsymbol{\theta} = (0, 2, -0.1, 0.3)$ we have the case of the dual Hénon map, exhibiting bistable behavior. In Fig. 5.2 we present two symmetric (with respect to the origin) isolated attractors along with their corresponding basins of attraction. The two basins of attraction are extend to infinity in both directions of the y axis and have a fractal boundary [SX15].

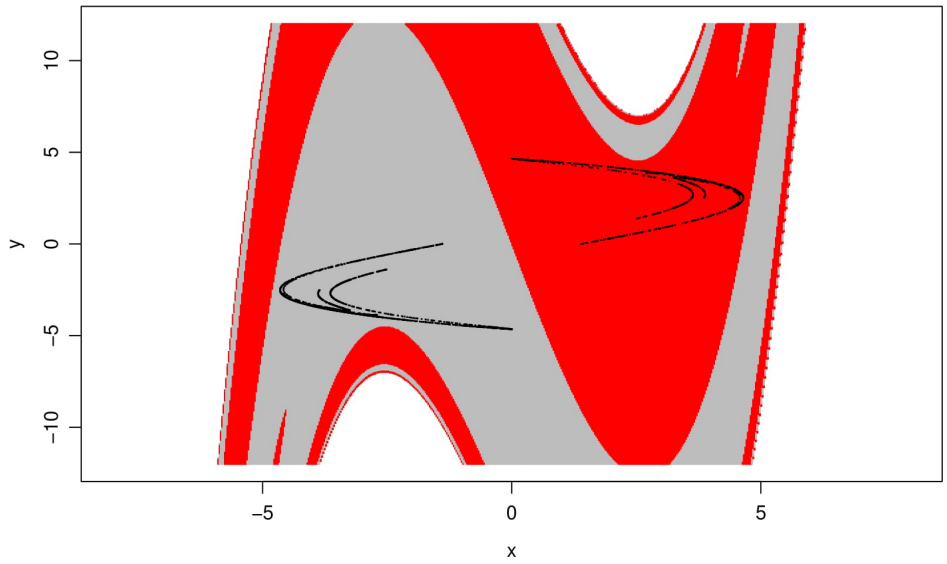

Figure 5.2: Attractors of the bistable Dual Hénon map, superimposed with the corresponding basins of attraction. We have used the red and gray colors for the basins of attraction of the coexisting strange attractors in the first and the second quadrant, respectively.

In Fig. 5.3 we plot the pair of strange attractors, along with the two global stable manifolds superimposed. The two associated symmetric saddle fixed points situated in $(3.605, 3.605)$ and $(-3.605, -3.605)$. The two interlaced global stable manifolds form the boundaries between the different basins. Points that belong in the stable manifolds they do not converge on the attractors, they will eventually converge to the associated saddle fixed points.

5.3 The BGSBR model

Suppose that we have at our disposal a time series $\mathbf{x}^n = (x_1, \ldots, x_n)$ which is a realization of the random recurrence relation given by

$$\begin{aligned} X_i &= T(\theta, X_{i-d}, \ldots, X_{i-1}, e_i) \\ &= g(\theta, X_{i-d}, \ldots, X_{i-1}) + e_i, \quad i \geq 1, \end{aligned} \tag{5.5}$$

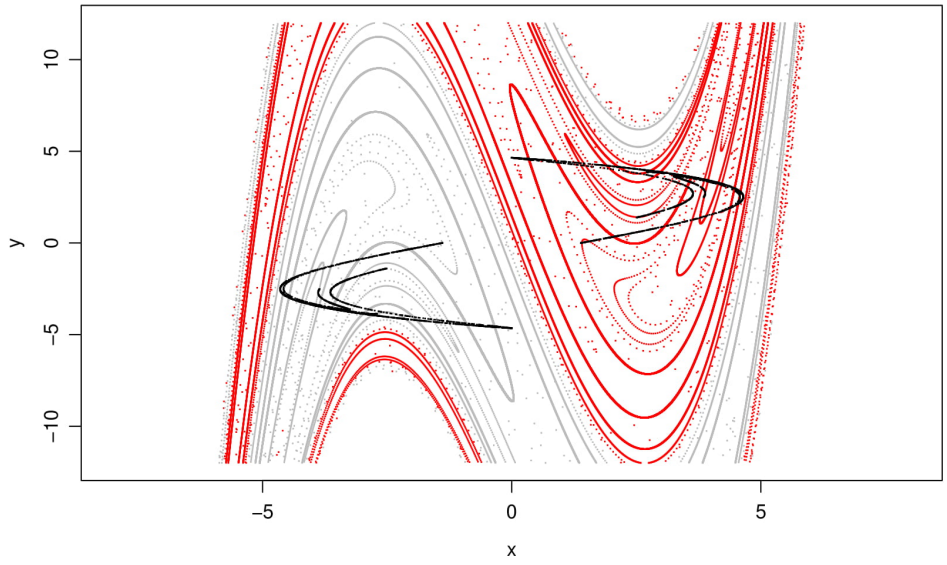

Figure 5.3: Attractors of the bistable Dual Hénon map, superimposed with the corresponding global stable manifolds. We use red and gray color for the global stable manifolds of the first and third quadrant saddle fixed points respectively.

where $g : \Theta \times X^d \to X$, for some compact subset X of $\mathbb{R}$. The states of the system $(X_i)_{i \geq -d+1}$ and the noise perturbations $(e_i)_{i \geq 1}$ are real random variables over some probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and we denote by $\theta \in \Theta \subseteq \mathbb{R}^m$ any dependence of the deterministic map g on the control parameters. g is nonlinear continuous in $\tilde{x}_{i-d-1:d} = (x_{i-d}, ..., x_{i-1})$.

As in the previous chapters, we assume that the random variables e_i are independent to each other, and independent of the states X_{i-r} for $r < i + d$. In order to relax the common assumption of normality for the noise process, we assume that the additive perturbations e_i are identically distributed from some zero mean distribution with unknown density f defined over the real line, so that $T : \Theta \times X^d \times \mathbb{R} \to \mathbb{R}$.

We do not take observational noise into consideration, so that the time series $\mathbf{x}^n$ is completely determined by the initial vector, the vector of control parameters, and the particular realization of the noise process. We note that the observed time series can also be considered as the one-dimensional measurements of the states of a dynamical system of unknown dimension, which can be embedded in a proper m–dimensional space using delay coordinates and suitably chosen time delay τ.

We propose a Bayesian nonparametric model, the backward geometric stick breaking reconstruction (B–GSBR) model, in order to jointly estimate a fixed number of consecutive past observations, namely performing prediction in reversed time. That is, given $\mathbf{x}^n = (x_1, \ldots, x_n)$ and a *reversed time* prediction horizon T, we aim to estimate the past observations vector $\tilde{x}_{-T:T} = (x_{-T+1}, \ldots, x_0)$, as well as the *shifted* to the past initial condition vector $\tilde{x}_{-T-d:d} = (x_{-T-d+1}, ..., x_{-T})$ corresponding to the augmented time-series $\tilde{x}_{-T:T+n}$.

For example, letting $d = 2$ and $T = 3$, our goal will be to estimate jointly the three past observations $\tilde{x}_{-3:3} = (x_{-2}, x_{-1}, x_0)$ and the vector of the shifted initial conditions $\tilde{x}_{-5:2} = (x_{-4}, x_{-3})$

corresponding to the augmented to the past time-series $\tilde{x}_{-3:n+3} = (x_{-2}, x_{-1}, x_0, x_1, ..., x_n)$.

To this end, we extend the GSBR model [MKH17] presented in Chapter 3, such that we will be able to jointly estimate the control parameters θ, the (perhaps) non-Gaussian density of dynamical noise process, the vector of past observations $\tilde{x}_{-T:T} = (x_{-T+1}, ..., x_0)$ together with the corresponding initial vector $\tilde{x}_{-T-d:d} = (x_{-T-d+1}, ..., x_{-T})$.

We will model the errors in recurrence (5.5) as an infinite mixture of zero-mean normal kernels of $\mathcal{N}(x \,|\, 0, \tau_j^{-1})$ for $j \geq 1$ with precision τ_j and mixing measure, the random geometric stick breaking measure

$$\mathbb{G} = \sum_{j \geq 1} \pi_j \delta_{\tau_j} \sim \mathcal{GSB}(a, b, G_0) \text{ with } \pi_j = \lambda(1-\lambda)^{j-1} \text{ and } \lambda \sim \mathcal{B}e(a, b).$$

Letting $\tau^\infty = (\tau_j)_{j \geq 1}$ and $\pi^\infty = (\pi_j)_{j \geq 1}$, it is that

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

The data likelihood of x^n conditional on the initial vector $\tilde{x}_{-d:d}$, after the augmentation with the auxiliary variables (d_i, N_i) for $1 \leq i \leq n$ takes the form

$$f(x_i, d_i, N_i : 1 \leq i \leq n \,|\, \lambda, \theta, \tilde{x}_{-d,d}) \propto \prod_{\substack{i=1 \\ d_i \leq N_i}}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \tau_{d_i}^{1/2} \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i:d}) \right\},$$

where λ the geometric probability of the GSB prior measure over the noise process and $h_\vartheta(x_i, x_{i:d}) = (x_i - g(\vartheta, x_{i:d}))^2$.

The conditional likelihood for a prediction in reversed time for $T > 0$, will be

$$f(x_i, d_i, N_i : -T+1 \leq i \leq n \,|\, \lambda, \theta, \tilde{x}_{-d:d}) \propto \prod_{\substack{i=-T+1 \\ d_i \leq N_i}}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \tau_{d_i}^{1/2} \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i:d}) \right\}. \tag{5.6}$$

We have the following proposition:

Proposition 5.1. *We let $\tau_i \overset{\text{iid}}{\sim} \mathcal{G}(\epsilon, \epsilon)$ for all $i \geq -T+1$ and we consider the cluster-slice group variable $\left(d^{n+T}, N^{n+T}\right)$ with $d^{n+T} = (d_{-T+1}, \ldots, d_n)$ and $N^{n+T} = (N_{-T+1}, \ldots, N_n)$. Then the augmented posterior $\Pi\left(\lambda, \tau^\infty, d^{n+T}, N^{n+T}, \tilde{x}_{-T-d:d}, \tilde{x}_{-T:T} \,|\, x^n, \mathcal{R}\right)$, conditionally on the event $\mathcal{R} =$*

$\{x_{-T+1} \in \mathbf{X}, \ldots, x_0 \in \mathbf{X}\}$ will be proportional to the quantity:

$$\begin{aligned}
&\mathcal{B}e(\lambda \mid a, b) \prod_{i=1}^{\infty} \mathcal{G}\left(\tau_{-T+i} \mid \epsilon, \epsilon\right) \mathcal{I}\left(\theta \in \tilde{\Theta}\right) \prod_{i=1}^{d} \mathcal{U}\left(x_{-T-d+i} \mid \mathbf{X}\right) \\
&\times \prod_{\substack{i=1 \\ d_{-T+i} : d_{-T+i} \leq N_{-T+i}}}^{T} \lambda^2 (1-\lambda)^{N_{-T+i}-1} \mathcal{U}\left(x_{-T+i} \in \mathbf{X}\right) \mathcal{N}\left(x_{-T+i} \mid g(\theta, \tilde{x}_{-T-d+i-1:d}), \tau_{d_{-T+i}}^{-1}\right) \\
&\times \prod_{\substack{i=1 \\ d_i : d_i \leq N_i}}^{n} \lambda^2 (1-\lambda)^{N_i-1} \mathcal{U}\left(x_i \in \mathbf{X}\right) \mathcal{N}\left(x_i \mid g(\theta, \tilde{x}_{-d+i-1:d}), \tau_{d_i}^{-1}\right).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\Pi\left(\lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d:d}, \tilde{x}_{-T:T} \mid x^n, \mathcal{R}\right) \propto \Pi\left(\lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d:d}, \tilde{x}_{-T:T}, x^n \mid \mathcal{R}\right) \\
&= \Pi\left(\lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d:d}\right) \Pi\left(\tilde{x}_{-T:T} \mid \lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d:d}, \mathcal{R}\right) \Pi\left(x^n \mid \lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d:d}, \tilde{x}_{-T:T}\right) \\
&= \Pi\left(\lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d:d}\right) \prod_{i=1}^{T} \Pi\left(x_{-T+i} \mid \lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d+i-1:d}, x_{-T+i} \in \mathbf{X}\right) \\
&\quad \times \prod_{i=1}^{n} \Pi\left(x_i \mid \lambda, \tau^{\infty}, \theta, \tilde{x}_{-d+i-1:d}\right) \\
&= \Pi(\lambda)\, \Pi\left(\tau^{\infty}\right)\, \Pi(\theta) \prod_{i=1}^{d} \Pi\left(x_{-T-d+i}\right) \prod_{i=1}^{T} \mathcal{U}\left(x_{-T+i} \in \mathbf{X}\right) \Pi\left(x_{-T+i} \mid \lambda, \tau^{\infty}, \theta, \tilde{x}_{-T-d+i-1:d}\right) \\
&\quad \times \prod_{i=1}^{n} \Pi\left(x_i \mid \lambda, \tau^{\infty}, \theta, \tilde{x}_{-d+i-1:d}\right).
\end{aligned}$$

Augmenting the posterior with the cluster-slice variable $\left(d^{n+T}, N^{n+T}\right)$ we obtain:

$$\begin{aligned}
&\Pi\left(\lambda, \tau^{\infty}, d^{n+T}, N^{n+T}, \tilde{x}_{-T-d:d}, \tilde{x}_{-T:T} \mid x^n, \{x_{-T+1 \in \mathbf{X}}, \ldots, x_0 \in \mathbf{X}\}\right) \\
&\propto \Pi(\lambda)\, \Pi\left(\tau^{\infty}\right)\, \Pi(\theta) \prod_{i=1}^{d} \Pi\left(x_{-T-d+i}\right) \\
&\times \prod_{\substack{i=1 \\ d_{-T+i} : d_{-T+i} \leq N_{-T+i}}}^{T} \lambda^2 (1-\lambda)^{N_{-T+i}-1} \mathcal{U}\left(x_{-T+i} \in \mathbf{X}\right) \mathcal{N}\left(x_{-T+i} \mid g(\theta, \tilde{x}_{-T-d+i-1:d}), \tau_{d_{-T+i}}^{-1}\right) \\
&\times \prod_{\substack{i=1 \\ d_i : d_i \leq N_i}}^{n} \lambda^2 (1-\lambda)^{N_i-1} \mathcal{U}\left(x_i \in \mathbf{X}\right) \mathcal{N}\left(x_i \mid g(\theta, \tilde{x}_{-d+i-1:d}), \tau_{d_i}^{-1}\right) \\
&\propto \mathcal{B}e(\lambda \mid a, b) \prod_{i=1}^{\infty} \mathcal{G}\left(\tau_{-T+i} \mid \epsilon, \epsilon\right) \mathcal{I}\left(\theta \in \tilde{\Theta}\right) \prod_{i=1}^{d} \mathcal{U}\left(x_{-T-d+i} \mid \mathbf{X}\right) \\
&\times \prod_{\substack{i=1 \\ d_{-T+i} : d_{-T+i} \leq N_{-T+i}}}^{T} \lambda^2 (1-\lambda)^{N_{-T+i}-1} \mathcal{U}\left(x_{-T+i} \in \mathbf{X}\right) \mathcal{N}\left(x_{-T+i} \mid g(\theta, \tilde{x}_{-T-d+i-1:d}), \tau_{d_{-T+i}}^{-1}\right) \\
&\times \prod_{\substack{i=1 \\ d_i : d_i \leq N_i}}^{n} \lambda^2 (1-\lambda)^{N_i-1} \mathcal{U}\left(x_i \in \mathbf{X}\right) \mathcal{N}\left(x_i \mid g(\theta, \tilde{x}_{-d+i-1:d}), \tau_{d_i}^{-1}\right).
\end{aligned}$$

□

At each iteration of the Gibbs sampler, we we will sample the variables:

$$(\tau_j),\ 1 \le j \le N^*, \quad (d_i, N_i),\ -T+1 \le i \le n,$$

$$(\tilde{x}_{-T-d:d},\ \tilde{x}_{-T:T}) \text{ and } (\vartheta,\ \lambda,\ z_{n+1}),$$

with $N^* = \max_{-T+1 \le i \le n} N_i$.

Following section 3.2, the construction of the slice sampler is straightforward. We note that for prior distribution over the initial condition vector we may use a uniform distribution over a suitable subset $\mathbf{X}$ of $\mathbb{R}^d$, based on our a-priori knowledge about the associated phase space of the system. For the geometric probability a reasonable choice will be a conjugate Beta prior.

Furthermore, we need to derive the full conditional distributions for the past variables and the new initial condition vector, namely for $j = 1, \ldots, T+d$ we can write

$$f(x_{1-j} \mid \ldots) \propto \mathcal{I}\left(x_{1-j} \in \mathbf{X}\right) \exp\left\{ -\frac{1}{2} \sum_{\substack{k=0 \\ k-j \ge -T}}^{d} \tau_{d_{1-j+k}} h_\vartheta(x_{1-j+k}, \tilde{x}_{-j+k-d:d}) \right\},$$

meaning that the FC of the last component x_{-T-d+1} of the initial condition vector $\tilde{x}_{-T-d:d}$ will be the truncated density

$$f(x_{1-(T+d)} \mid \ldots) \propto \mathcal{I}\left(x_{1-(T+d)} \in \mathbf{X}\right) \exp\left\{ -\frac{\tau_{d_{-T-d+1}}}{2} (x_{-T+1} - g(\vartheta, x_{-T-d+1}, \ldots, x_{-T}))^2 \right\}.$$

5.3.1 Stable manifold stochastic approximation

The proposed method is based on the observation that the posterior marginal distribution of the initial condition vector lies along the *stable direction*, i.e. the direction of the stable manifold [HZ98]. We use the stable direction, for example in order to find regions of homoclinic tangencies, by iterating a random (normalized) vector with the inverse Jacobian matrix at a given point, as explained in Refs. [CEP13, LGYK93]. We note that due to numerical errors, we are not able to actually multiply with the inverse Jacobian, so we use the evaluation of the inverse Jacobian at each orbit point. For more details, see subsection 2.1.3 of this thesis and Ref. [LGYK93].

On the initial condition estimation

Before we proceed with the algorithm for the stochastic approximation of the stable manifold, we will address the problem of the initial condition estimation in terms of reconstruction, i.e. with $T = 0$. Suppose we have a data set $\mathbf{x}^n = (x_1, \ldots, x_n)$ as in Eq. (5.5) and we are interested in recovering the initial condition vector $\tilde{x}_{-d:d} = (x_{-d+1}, ..., x_0)$. Our inference is based on the posterior marginal distribution of the initial vector, via the GSBR algorithm (or equivalently

the BGSBR with $T = 0$), using the proper Bayesian estimate. The quality of the estimate, is affected mainly by two factors:

1. The particular realization of the noise process.
2. The position of the initial vector in the state space.

In order to illustrate the above effects, we generate a deterministic orbit $\mathbf{y}^m$ of the Hénon map, with $y_j = g(y_{j-1}, y_{j-2}) = 1.38 - y_{j-1}^2 + 0.27 y_{j-2}, j = 1, \ldots, 250$ starting from $(y_{-1}, y_0) = (-0.61, 1.37)$ and use each vector of consecutive orbit points (y_j, y_{j+1}) as initial vector $\mathbf{x}^{(0,j)} = (x_{-1,j}, x_{0,j})$ for a random orbit $\mathbf{x}^{j,n} = (x_{1,j}, \ldots, x_{n,j})$ with the same deterministic part, i.e. for $i = 1, \ldots, 500$

$$x_{i,j} = g(x_{i-1,j}, x_{i-2,j}) + e_{i,j}, \quad e_{i,j} \overset{\text{iid}}{\sim} f \tag{5.7}$$

with noise process $f(x) = \frac{9}{10}\mathcal{N}(x \,|\, 0, 10^{-7}) + \frac{1}{10}\mathcal{N}(x \,|\, 0, 10^{-3})$. We perform reconstruction at each one of the 250 data sets and compute the absolute error (AE) of $x_{-1,j}$ and $x_{0,j}$, for $j = 1, \ldots, 250$. We model the deterministic part g, with the complete quadratic polynomial in the two variables and use as Bayesian estimate the posterior mean.

Our estimate is affected practically by the magnitude of the first noise perturbations, in the sense that stronger random perturbations are affecting the initial part of the data generating process leading to worse estimates of the initial conditions. In Fig. 5.4 we present the obtained AEs of $x_{0,j}$'s superimposed with the magnitude (absolute value) of the second noise perturbation of each data set $|e_{2,j}|, j = 1, \ldots, 250$. We can see that the locations of relatively high AEs correspond to stronger noise perturbations, with correlation between the noise magnitude and $x_{0,j}$s AE equal to 0.96. For the $x_{-1,j}$ AEs we have the same behavior, with correlation equal to 0.88. For the logistic map (and any nonlinear map with lag one), the initial condition estimation depends only on the first random noise perturbation.

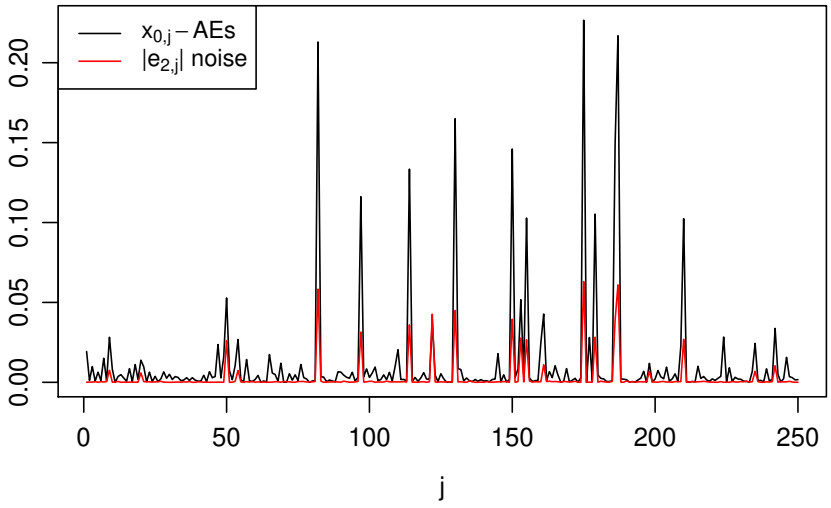

Figure 5.4: Absolute errors of the $x_{0,j}$ estimates (black) and absolute values of the second noise perturbation $|e_{2,j}|$, for $j = 1, \ldots, 250$.

The estimation of the initial condition is moreover affected by its position on the state space. More specifically, the posterior marginal of the initial vector lies along the stable direction and

depends on the curvature of the local stable manifold. In Fig. 5.5 (a) we present the deterministic orbit used to generate the initial conditions, superimposed with the sampled values of the initial condition posterior marginals for two cases with different local curvature of the stable manifold. Near the initial condition $(0.08, -1.32)$ the stable manifold exhibits locally high curvature, while near the initial condition $(1.71, -0.02)$ the curvature of the local stable manifold is very small. Note that for both cases the initial condition is in the neighborhood of a HT. In Fig. 5.5 (b)-(c) we enlarge the area around the two initial points and present the sampled values with the corresponding 90% highest posterior density regions, defined by the black polygonal curves around the initial points. The closeness of the HTs to the two initial points results in the increase of the variance of the posterior marginals (when compared to regions with no neighboring primary HTs). The supports of the posterior distributions have clearly different shapes, due to the different local curvature of the global stable manifold around the different regions of the primary HTs. Thus, apart from the intensity of the noise, the local stable direction also affects the quality of the initial vector estimation.

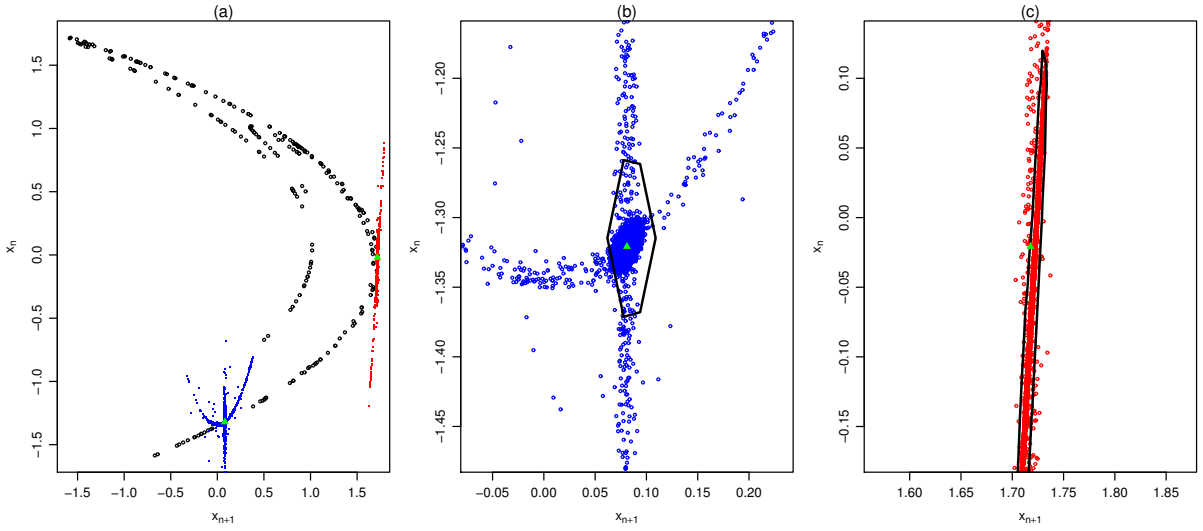

Figure 5.5: (a) Embedded deterministic orbit **y** superimposed with the sampled values for the initial conditions around $(0.08, -1.32)$ (in blue) and $(1.71, -0.02)$ (in red). In Fig. 5.5(b) and 5.5(c) we present enlargements of the regions $[-0.07, 0.23] \times [-1.47, -1.17]$ and $[1.57, 1.87] \times [-0.17, 0.13]$, respectively. The green triangles indicate the positions of the true initial conditions.

Positive reversed time prediction horizon

When we sample from the joint density of the variables $\tilde{x}_{-T-d:d} = (x_{-T-d+1}, ..., x_{-T})$, which are considered as the initial point for the time series $(\tilde{x}_{-T:T}, \tilde{x}_{0:n}) = (x_{-T+1}, ..., x_0, x_1, ..., x_n)$, for $T > 0$, there is an increase of uncertainty forcing the sampled values to diffuse along remote portions of the stable manifold, along the stable direction. So, we can utilize this *diffusion effect* caused by the increase of the reversed time prediction horizon T, in order to reveal larger and more remote portions of the global stable manifold. The approximation of the global stable manifold via sampling from the joint density of $\tilde{x}_{-T-d:d}$ for $T > 0$, can be considered as a data driven method as it is based exclusively on the observed time-series x^n.

Furthermore, in the same way that forward orbits of a map g are attracted by the unstable manifold containing the attracting set, backward orbits are attracted by the unstable manifold of the inverse map g^{-1}. However, the unstable manifold of g^{-1} coincides with the stable manifold of g, which we stochastically approximate by performing prediction in reversed time.

We demonstrate this diffusion effect along the various parts of the stable manifold by producing two versions of the last numerical experiment one for $T = 0$ and one for $T = 2$. Both versions will involve the deterministic orbit $\mathbf{y}^m$ of length $m = 250$. Each orbit point y_j is used as an initial condition for the realization of a non–Gaussian noisy orbit $\mathbf{x}^{j,n}$ of length $n = 500$.

For $T = 0$ we sample from the marginal joint density of the variable $(x_{-1,j}, x_{0,j})$, being the initial point of the $\mathbf{x}^{j,n} = (x_{1,j}, \ldots, x_{n,j})$ realization for $j = 1, \ldots, 250$ producing the sets of points $W^s(\mathbf{x}^{j,n}, T = 0)$ for $j = 1, \ldots, m$, and we denote their union by $\mathcal{W}_0^s$.

For $T = 2$ we sample from the marginal joint density of the variable $(x_{-3,j}, x_{-2,j})$ of the $T = 2$ version, being the initial point of the augmented to the past orbits $(x_{-1,j}, x_{0,j}, x^{j,n}) = (x_{-1,j}, x_{0,j}, x_{1,j}, \ldots, x_{500,j})$ producing the sets of points $W^s(\mathbf{x}^{j,n}, T = 2)$ for $j = 1, \ldots, m$, and we denote their union by $\mathcal{W}_2^s$. In Fig. 5.6(a) we display superimposed the set of points $\mathcal{W}_0^s$ in red and the the set of points $\mathcal{W}_2^s$ in gray.

The increase of the prediction horizon drives the supports of the posterior distributions along the stable direction. In a sense, the reversed time prediction at a point x, is mimicking the effect of multiplying the points of a backward orbit ending at x with the inverse Jacobian matrix, thus approximating the stable direction. In Fig. 5.6 (b)-(c) we zoom in certain regions of the state space for a more detailed inspection of this phenomenon.

Also, in Figs. 5.7(a) and 5.7(b) we present the associated density–colormaps of the set of points $\mathcal{W}_0$ and $\mathcal{W}_2$, respectively. Higher color intensity corresponds to regions of higher relative frequency of the sampled values, thus to regions associated with higher posterior probability. For $T = 0$, the posterior marginals put almost all of their mass close to the true initial conditions, along the stable direction. Fig. 5.7(a) can be compared with the local stable directions shown in Fig. 2.3(a). On the other hand, due to the increased time reversed horizon for $T = 2$ we can see in Fig. 5.7(b) that the posterior marginals are more spread across the remote parts of the stable manifold, far from the neighborhoods of the true initial conditions. We consider the union set $\mathcal{W}_2$ providing us with an approximation of the noisy global stable manifold of the of the random Hénon map given in Eq. (5.7).

Whenever we have only one available data set $(x_1, \ldots, x_n)$ generated by an unknown perturbed dynamical system, in order to reveal more portions of the stable manifold, we apply the B–GSBR algorithm multiple times, each time based on proper subsets of the observed orbit. More specifically, on the k–th consecutive application of the B–GSBR model, we will use the data set $\mathbf{x}_k^n = (x_k, \ldots, x_n)$, the last $n - k + 1$ points of the observed time–series, to sample the data set $W^s(\mathbf{x}_k^n, T = t)$. After a prespecified number of applications, say m, we use the union set $\mathcal{W}_t^s = \cup_{k=1}^m W^s(\mathbf{x}_k^n, T = t)$ of the supports of joint posterior marginals, or alternatively the associated regions of *higher posterior probability*, as an approximation to the global stable manifold. Regarding the choice of the prediction horizon T, we have observed that small values

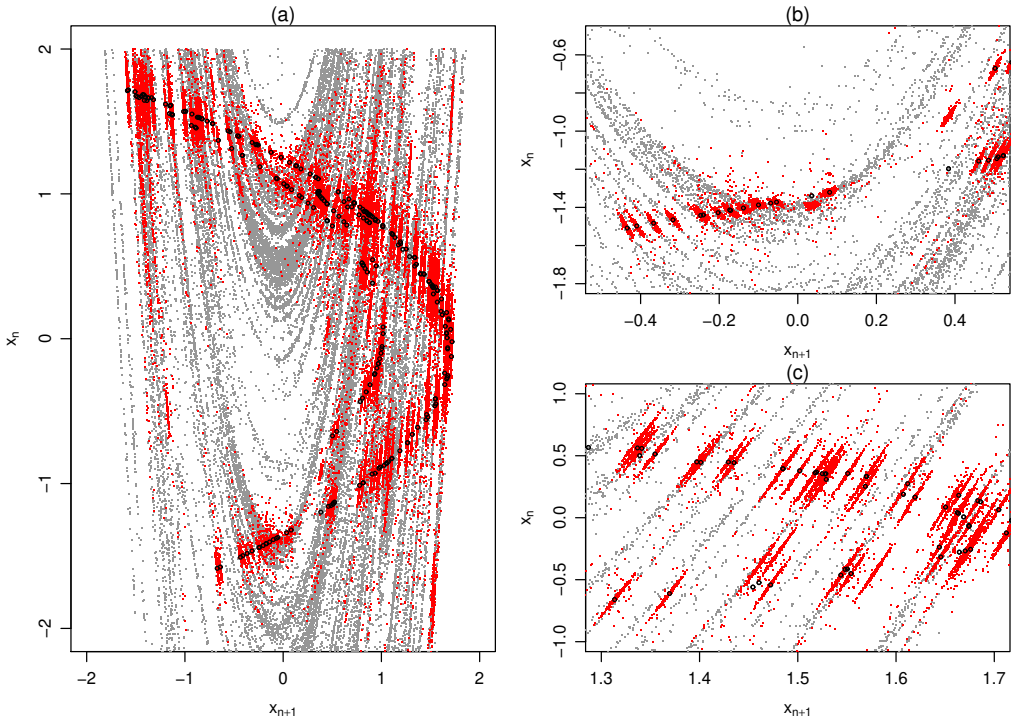

Figure 5.6: Embedded deterministic orbit **y** (black) superimposed with the sampled values of all the 250 initial conditions for $T = 0$ (red) and $T = 2$ (gray). In (b) and (c) we zoom in the regions $[-0.50, 0.55] \times [-1.8, -0.5]$ and $[1.30, 1.75] \times [-1, 1]$, respectively. At every chain we performed thinning every 50 iterations.

of T between 2 and 5 give nearly indistinguishable results. Lower values should be preferred, whenever computational resources are limited.

In Fig. 5.8 we display the propagation pattern of the support of the T–marginal densities along the stable direction, for a specific initial point and a varying time reversed prediction horizon T.

More specifically, in Fig. 5.8 we have set the initial point to $(x_{-1}, x_0) = (-2, 0)$ and we have generated a sample $(x_1, \ldots, x_n)$ of length $n = 500$, with the standard parameters of the Hénon map, using additive Gaussian noise with variance $\sigma^2 = 10^{-7}$. We have applied the B–GSBR model, for different reversed time horizons $T = 0, 1, 2, 3$.

The described method of stochastic approximation is relevant to the approximation methods described in chapters 3 and 4. We have approximated the quasi-invariant measure of the system by applying the GSBR model and the location of primary homoclinic tangencies by applying the DNRR model. The novelty of the above methods lies on the fact that we are able to approximate features of the associated random, or deterministic map, with stochastic data–driven methods.

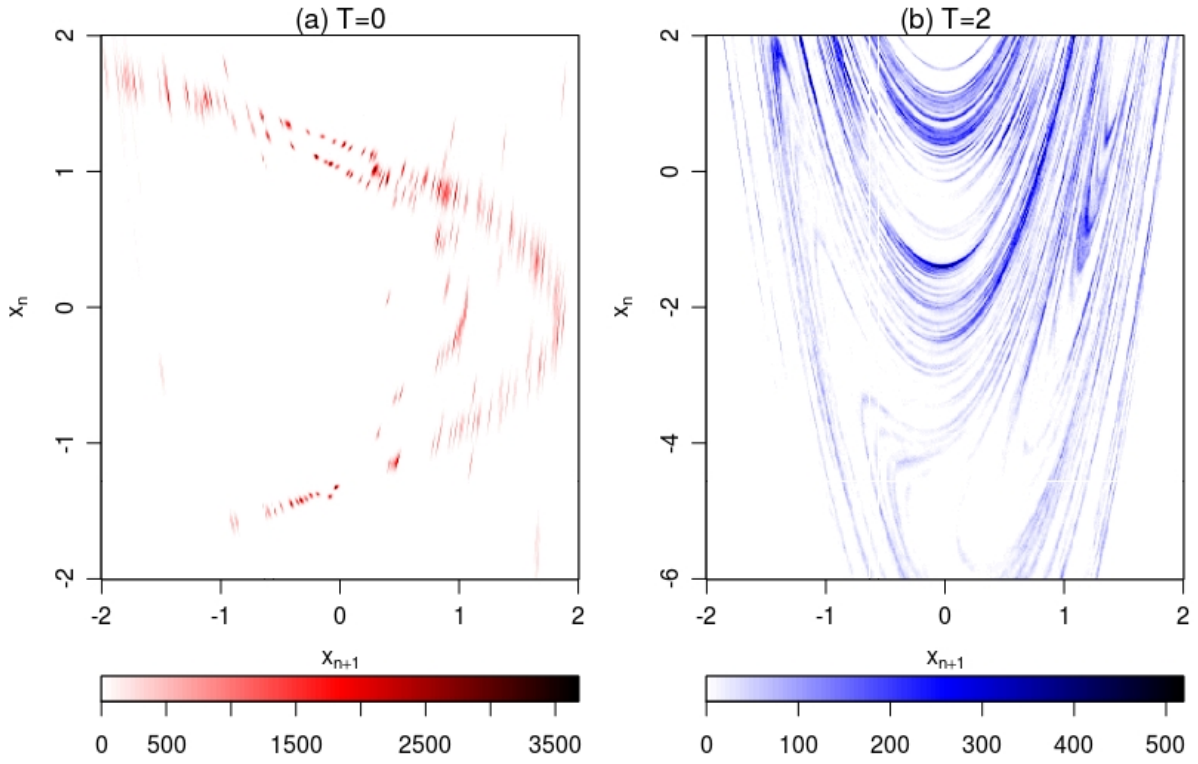

Figure 5.7: Embedded colormaps of the posterior samples for each $\mathbf{x}^{(0,j)}$, $j = 1, \ldots, 250$, when (a) $T = 0$ and (b) $T = 2$. Higher color intensity corresponds to regions of higher posterior probability.

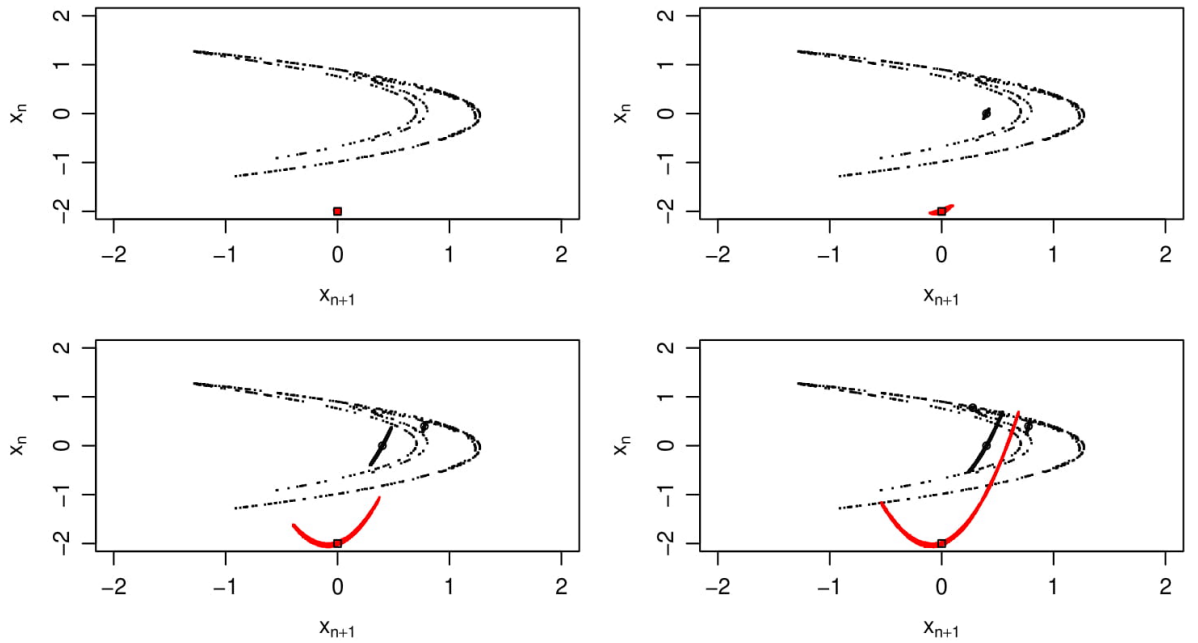

Figure 5.8: Embedded data from noisy Hénon map (black), superimposed with the posterior marginals of the initial condition vector (red) and past unobserved observations (gray). True values of initial condition and true past observations are indicated with square and circles respectively .

5.4 Simulations

In this section, we provide numerical illustrations of the proposed B–GSBR algorithm, in order to stochastically approximate the global stable manifolds of the saddle fixed points of the Hénon and Dual Hénon maps, using synthetic time series. Moreover, we will apply the B–GSBR model on orbits from deterministic and noisy noninvertible polynomial maps. We will apply the algorithms for different non–Gaussian noise processes, using mixtures of two Normal kernels with different variances.

In the following, we will use a non-informative prior specification, suitable for cases in which there is no prior information available regarding the data generating process. Specifically, as a prior for the geometric probability variable, we take the arcsine density $\lambda \sim \mathcal{B}(0.5, 0.5)$, which coincides with the Jeffrey's prior for λ. On the precisions $(\tau_j)_{j\geq 1}$ of the random density f, we place the vague gamma prior $\mathcal{G}(10^{-3}, 10^{-3})$, which is very close to a scale invariant prior. On the variables of the initial conditions we assign uniform priors based on our a priori knowledge of the map state space.

In our numerical experiments, unless otherwise stated, the B-GSBR Gibbs samplers have ran for 20×10^4 iterations for every application, leaving the first 5×10^4 samples as a burn-in period. We performed thinning every 150 iterations in order to reduce the correlation between the sampled values.

5.4.1 Hénon map

The zero noise limit case

We consider a time series realization $\mathbf{x}^n$ of size $n = 2000$, coming from the deterministic recurrence relation given in (5.1) using the parameter values $(\theta_1, \theta_2, \theta_3) = (1, -1.4, 0.3)$ with initial point $(x_{-1}, x_0) = (-1, 0.5)$. We consider the data as coming from a deterministic part g corrupted by a very low, close to zero, intensity dynamical noise process. For example we could assume the existence of such a zero limit noise process $e_i \sim f^*$ as coming from the *overall errors* produced at each iteration of the map, caused by the finite precision of the floating point computer arithmetic that is:

$$x_i = g(\boldsymbol{\theta}, x_{i-1}, x_{i-2}) + e_i, \quad e_i \sim f^* \tag{5.8}$$

For more details, see Ref. [HW11]. For greater flexibility we use $f^*(z) := \sum_{i=1}^{\infty} p_i \mathcal{N}(z|0, \sigma_i^2)$, which is an infinite mixture of Normal kernels. In this case, in principle, any symmetric $0-$mean density can be recovered.

The deterministic part g is modeled with the complete quadratic polynomial in two variables, namely

$$g(\boldsymbol{\theta}, x_{i-1}, x_{i-2}) = \theta_0 + \theta_1 x_{i-1} + \theta_2 x_{i-2} + \theta_3 x_{i-1} x_{i-2} + \theta_4 x_{i-1}^2 + \theta_5 x_{i-2}^2. \tag{5.9}$$

We used a fixed reverse time prediction horizon $T = 3$ and we have performed $m = 500$ consecutive applications of the B–GSBR algorithm.

In Fig. 5.9 we present the embedded data points over $\mathbb{R}^2$ lying on the Hénon attractor (in black), superimposed with the sampled values from the joint posterior marginals of the initial condition vector for each one of the 500 consecutive applications (in red). For the initial conditions we assigned a uniform prior over the interval $(-3, 3)$.

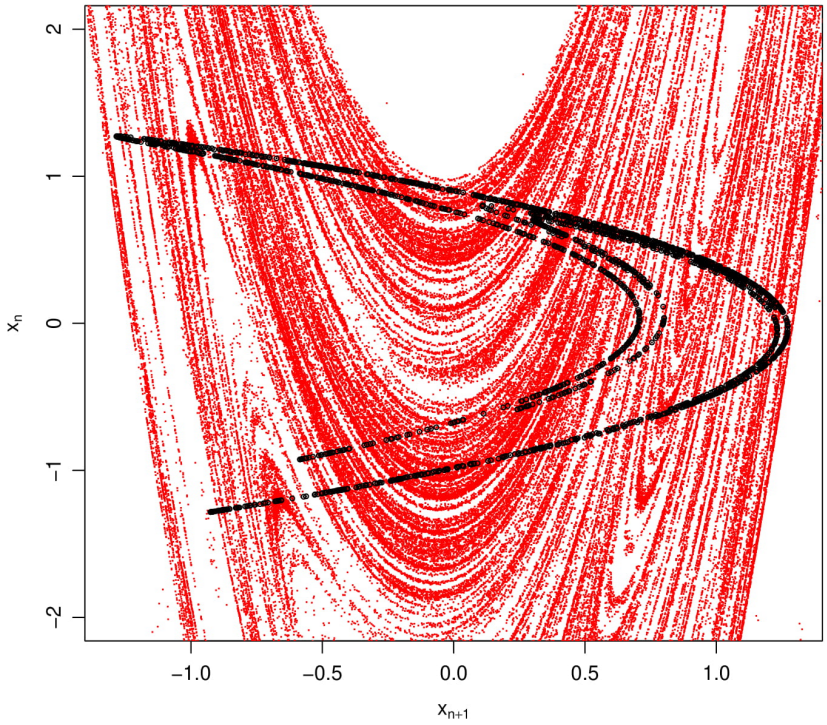

Figure 5.9: Embedded data from deterministic Hénon map (points in black), superimposed with the sampled values of the consecutive joint posterior marginals of the initial condition vector (points in red).

It is evident that the union of the supports of the joint posterior marginals forms a stochastic approximation of the associated global stable manifold of the deterministic map, as shown in Fig. 5.1.

In order to further investigate the quality of the stochastic approximation, we show in Fig. 5.10 data points and sampled values from the joint posterior marginals over the rectangle $[0.65, 1.25] \times [-0.5, 0.5]$.

We remark that whenever we are confident that we have in our disposal a deterministic data set, or a noisy data set perturbed by Gaussian noise, we can use the parametric counterpart of the B–GSBR model. In this case, we have to model the noise as $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \lambda^{-1})$ and assign a Gamma prior over the unknown precision of the noise distribution. Nevertheless, when the noise departs from normality, a parametric modeling will lead to erroneous estimations as we have analyzed in the previous chapters.

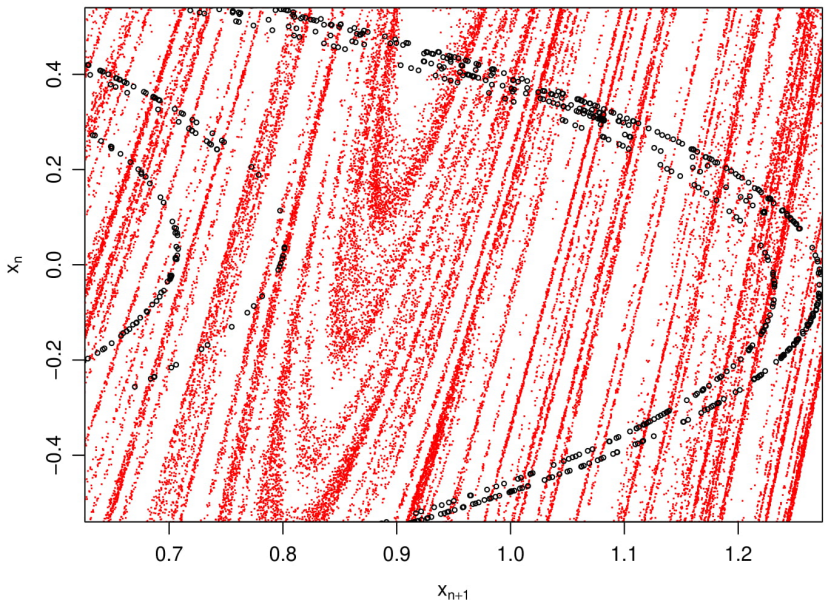

Figure 5.10: Enlargement of the rectangle $[0.65, 1.25] \times [-0.5, 0.5]$ of Fig. 5.9.

Non–Gaussian noise

We perform the same simulation as in the previous section, using a single data set from the noisy Hénon map. We consider a time–series realization $\mathbf{x}^n$ of size $n = 1000$, coming from the recurrence relation given in (5.1) using the parameters values $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (1, -1.4, 0.3)$, corrupted by the dynamical noise

$$f_1(z) = \sum_{i=1}^{2} w_i \mathcal{N}\left(z \,|\, 0, \sigma_i^2\right),$$

with $w_1 = 0.75$, $w_2 = 0.25$, $\sigma_1^2 = 10^{-6}$, $\sigma_2^2 = 10^{-3}$ and initial condition $(x_{-1}, x_0) = (-1, 0.5)$. We model again the deterministic part g, with the complete quadratic polynomial in two variables as in equation (5.4). For the initial condition we have assigned a uniform prior over the interval $(-6, 2)$.

We have used the reversed time prediction horizon $T = 3$ and we have performed $m = 250$ consecutive applications of the B–GSBR algorithm, based on the data set $\mathbf{x}^n$ with starting data set $(x_1, \ldots, x_n)$ and ending data set $(x_m, \ldots, x_n)$.

In Fig. 5.10 we present the embedded data points over $\mathbb{R}^2$ which are close to the Hénon attractor (in black), superimposed with the union of the sampled values from the joint posterior marginals of the initial condition vector coming from the 250 consecutive applications (in red).

We note that due to the presence of non–Gaussian corrupting noise, the approximation of the global stable manifold is subjected to a blurring effect. However, with the proposed method we are still able to gain insight about the qualitative characteristics of the dynamical behavior of the underlying system that generated the observed data set.

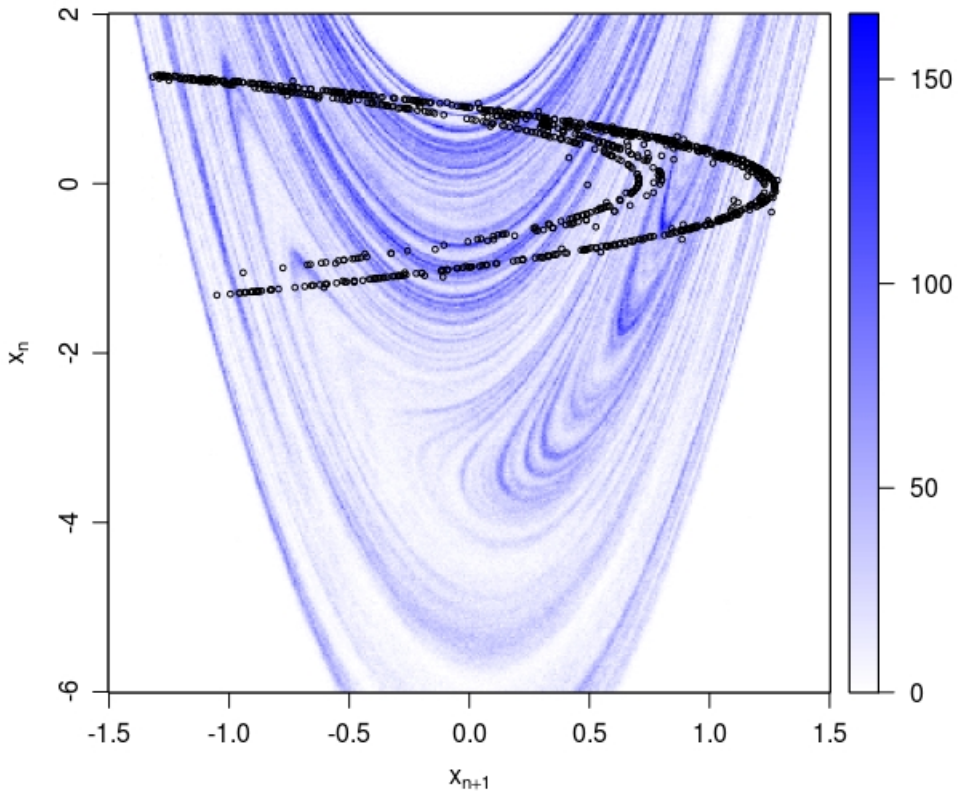

Figure 5.11: Embedded data from noisy Hénon map (points in black), superimposed with the sampled values of the consecutive joint posterior marginals of the initial condition vector. Colormap of the relative frequency, using a high analysis grid.

It is better, then, instead of approximating the stable manifold with the union of the supports of the posterior marginals to use regions of higher posterior probability. Specifically, we apply a high analysis grid over the subset of the state space that we are interested in approximating the manifold and color–code the regions according to the relative frequency of the posterior samples inside each region. Then, the regions of higher posterior probability will provide us with an increased quality stochastic approximation of the stable manifold.

5.4.2 The generalized Hénon map

The zero noise limit case

In this subsection we will consider the map

$$x_i = \theta_1 + \theta_2 x_{i-1} + \theta_3 x_{i-1}^3 + \theta_4 x_{i-2}, \tag{5.10}$$

with corresponding two-dimensional representation given in Eq. (5.3). We set the control parameters to $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 2, -0.1, 0.3)$. These are the parameter values for which two Hénon-like attractors $\mathcal{A}^+$ and $\mathcal{A}^-$ coexist (see Fig. 5.3), situated to the first and third quadrant, respectively.

We generate two deterministic orbits $\mathbf{x}_+^n = (x_1^+, ..., x_n^+)$ and $\mathbf{x}_-^n = (x_1^-, ..., x_n^-)$, with initial points $p_+ = (1, 0.5)$ and $p_- = (0.5, 1)$, lying on the basins of attraction of $\mathcal{A}^+$ and $\mathcal{A}^-$, respectively.

We use the two orbits as our observed data sets, and to each orbit we apply the B–GSBR algorithm with reversed time horizon $T = 2$. We set the restriction interval to $\mathbf{X} = (-5, 5)$. So that the prior over the initial variables (x_{-3}, x_{-2}) is uniform over $\mathbf{X}^2$ and the variables (x_{-1}, x_0) are truncated to $\mathbf{X}^2$.

In Fig. 5.12 we present the embedded data sets $\mathbf{x}_+^n$ and $\mathbf{x}_-^n$ in black over $\mathbb{R}^2$. They are lying close to the coexisting Hénon-like attractors. Together we superimpose the unions of the sampled values $W_+^s = \cup_{j=1}^m W^s(\tilde{x}_{j-1:n-j+1}^+; T = 2)$ in red and $W_-^s = \cup_{j=1}^m W^s(\tilde{x}_{j-1:n-j+1}^-; T = 2)$ in gray for $m = 500$.

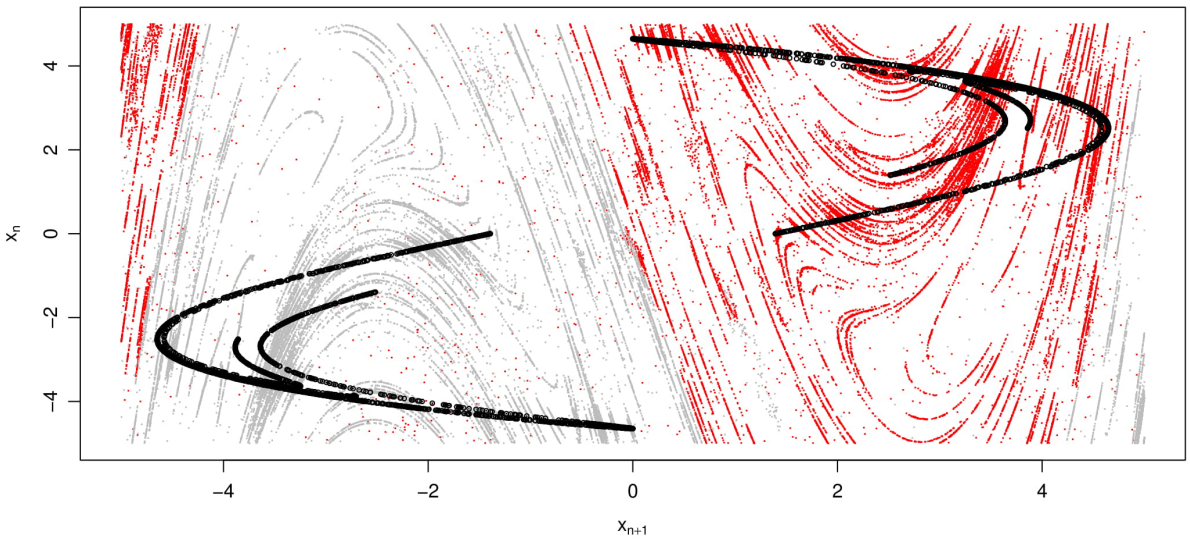

Figure 5.12: Deterministic orbits of the bistable Dual Hénon map, superimposed with the corresponding stochastically approximated global stable manifolds. We use red and gray colors for the joint posterior marginals corresponding to the first and third quadrant data sets respectively.

The union of the supports W_+^s and W_-^s of the joint posterior marginals, form a stochastic approximation of the associated global stable manifolds of the deterministic map, in Fig. 5.3. The support of the pairs of time reversed initial points is the square $\mathbf{X}^2 = (-5, 5)^2$, which we have used as a prior and a a restriction to the posterior. However, the *density* of the union of the supports is very close to zero in regions that do not belong in the associated basin of attraction.

We emphasize that we had no prior information at our disposal regarding the parameters of the map or the location of the saddle fixed points. The stochastic approximation is completely data driven, given a general functional representation for the deterministic part.

The case of an impulsive noise process

In this subsection we demonstrate the efficiency of the proposed model in the case of a sample generated by the GHM, under the influence of impulsive dynamical noise.

More specifically, we generate the realization $\mathbf{x}^n$ of length $n = 500$ using the recurrence relation given in (5.10) for the values of the control parameters $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 2, -0.1, 0.3)$, corrupted by the impulsive dynamical noise $f_2(z) = \sum_{i=1}^{2} w_i \mathcal{N}\left(z \,|\, 0, \sigma_i^2\right)$, with $w_1 = 0.9$, $w_2 = 0.1$, $\sigma_1^2 = 10^{-7}$, $\sigma_2^2 = 10^{-2}$,using the initial point $p_+ = (1, 0.5)$.

We apply the B-GSBR sampler with fixed reversed time prediction horizon $T = 2$, with the same prior specifications as above. In Fig. 5.13 we present the result of 150 consecutive applications of the model. For the initial condition (x_{-3}, x_{-2}) we have assigned the uniform prior over the square $\mathbf{X} = (-10, 10)^2$.

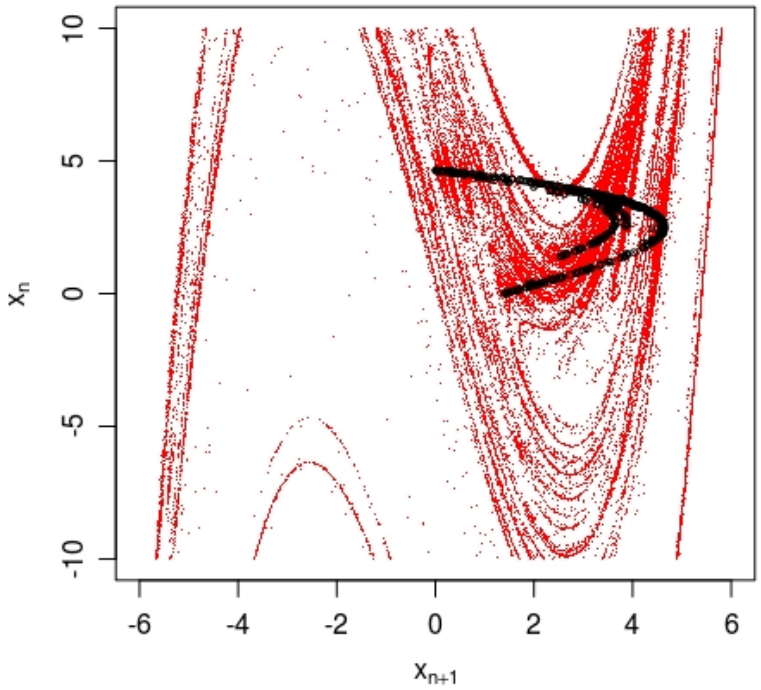

Figure 5.13: Stochastic approximation of the global stable manifold of the Dual Hénon map. Embedded data from noisy data set $\mathbf{x}^n$ (points in black), superimposed with the sampled values of the consecutive joint posterior marginals of the initial condition vector (points in red).

We remark that although the noise density is non–Gaussian and the length of the orbit small, we are able to produce an adequate approximation of the stable manifold, regarding the qualitative characteristics and the boundary of the basins of attraction. The density of the union of the supports is non–zero inside the square $(-10, 10)^2$, but (as in the zero noise limit case) it is very close to zero in regions that do not belong in the associated basin of attraction.

5.4.3 A non-invertible 2-d quadratic map

In this subsection we illustrate the performance of the B–GSBR model on nonlinear time–series data coming from noninvertible polynomial maps. As an example, we will use the noninvertible quadratic map

$$g(\boldsymbol{\theta}, x, y) = (\theta_1 - x^2 + \theta_2 y, x^2), \tag{5.11}$$

with the lag-2 representation given by

$$x_i = \theta_1 - x_{i-1}^2 + \theta_2 x_{i-2}^2, \tag{5.12}$$

for the value of the control parameters $(\theta_1, \theta_2) = (1.38, 0.211)$. For a detailed analysis of the dynamical behavior of two–dimensional non-invertible quadratic maps we refer to [ES10]. The noninvertibility of the map (5.11) is caused by the squared term in the $y-$coordinate.

Since there exist orbit points with more than one preimages, we extend the definition of the stable and unstable manifolds [San00] to

$$\begin{aligned} W^s(p) = \{x \in \mathbb{R}^2 :\ &\text{there exists an infinite forward orbit } \{x_n\} \text{ through } x \\ &\text{such that } x_k \to p, \text{ as } n \to \infty\} \end{aligned} \tag{5.13}$$

and

$$\begin{aligned} W^u(p) = \{x \in \mathbb{R}^2 :\ &\text{there exists an infinite backward orbit } \{x_n\} \text{ through } x \\ &\text{such that } x_k \to p, \text{ as } n \to -\infty\}, \end{aligned} \tag{5.14}$$

for a map g with fixed point p. Whenever the invertibility of g is not guaranteed, usually W^s and W^u are being referred to as the stable and unstable invariant *sets*, respectively. The global invariant sets are no longer guaranteed to be manifolds and due to the multiple preimages, they can have intersections, being non–smooth or being totally disconnected [FKP03]. The approximation of the stable manifold in such case is useful, because it gives us information on the complicated basins of attraction that the multiple preimages create [EKO04]. The simplest case of noninvertible quadratic maps are of the $(Z_0 - Z_2)$ type, meaning that there exist points in the state space with only two or zero preimages, defining the two mutually exclusive basins of attraction.

The prior specifications for the following two simulations, remain the same as before, while the initial point for the orbits will be $(x_{-1}, x_0) = (0.5, 1.5)$. For the initial conditions we assigned a uniform prior over the interval $(-2, 2)$.

The zero noise limit case

We generate a deterministic orbit $\mathbf{x}^n$ coming from the map in Eq. (5.12), of length $2,000$. We aim to stochastically approximate the global stable manifold of the non-invertible map in (5.12).

We set $T = 3$ and perform 500 consecutive applications of the BGSBR model, moving one point at each step. We ran all chains for $200,000$ iterations with a burn–in period of length $50,000$, performing thinning every 150 iterations.

In Fig. 5.14 we present the deterministic data $\mathbf{x}^n$ with the union of the sampled joint posterior marginals of the initial variables (x_{-4}, x_{-3}). We remark that the x–axis symmetry is caused by the y–preimages.

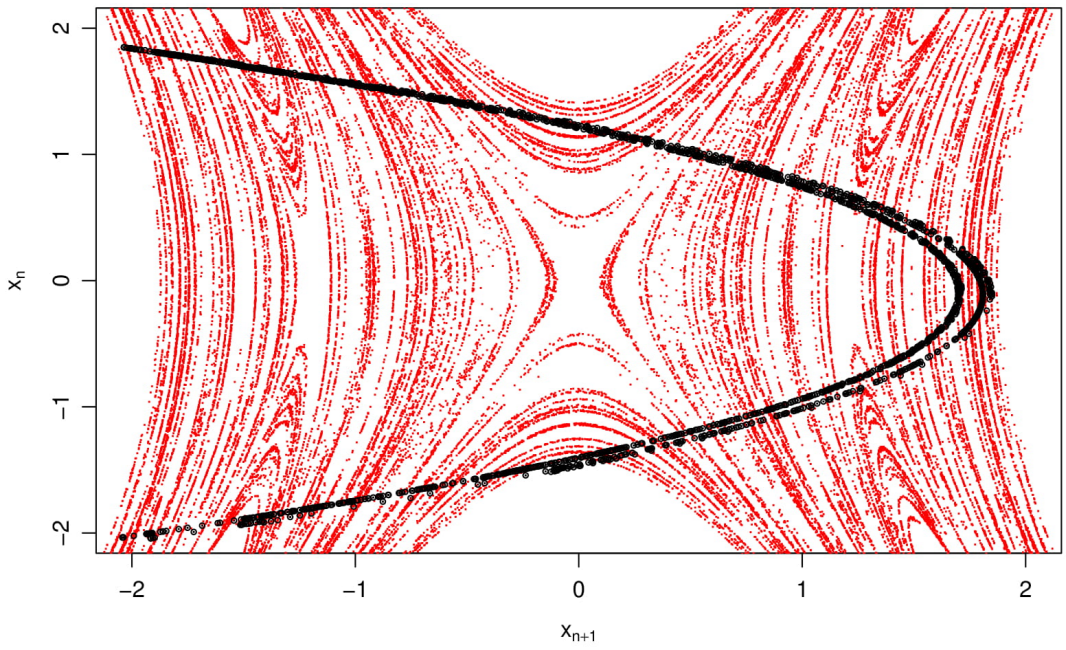

Figure 5.14: The embedded data from deterministic orbit $\mathbf{x}^n$ is depicted in black. Together we have superimposed the union of the sampled values of the consecutive joint posterior (x_{-4}, x_{-3}) in red.

The case of an impulsive noise process

We have generated a noise contaminated orbit realization $\mathbf{x}^n$, of length $n = 2000$, coming from the map

$$x_i = \theta_1 - x_{i-1}^2 + \theta_2 x_{i-2}^2 + e_i \quad e_i \overset{\text{iid}}{\sim} f_3, \tag{5.15}$$

with $f_3(z) = 0.9\mathcal{N}(z\,|\,0, 10^{-7}) + 0.1\mathcal{N}(z\,|\,0, 10^{-2})$.

In the previous chapters, we have shown that for such a noise, parametric models are not suitable for inference, so, the Bayesian nonparametric approach is essential. Nevertheless, if we are confident that the available data sets are contaminated by Gaussian noise (or by noise practically indistinguishable from Gaussian noise), then it is preferable to use a parametric counterpart to the BGSBR sampler.

Here we set $T = 2$ and perform 500 consecutive applications of the B–GSBR model, moving one point at a time. We have ran the associated chains for 200,000 iterations with a burn–in period of length 50,000, performing thinning every 150 iterations.

In Fig. 5.15 we present the stochastic approximation of the stable manifold obtained by the application of the B–GSBR algorithm to the noisy data set $\mathbf{x}^n$ and after applying a density–color–map representation. The quality of the obtained approximation is adequate, although more "noisy" than the previous illustration where the observed orbit was longer and deterministic. We remark that the qualitative characteristics of the B–GSBR algorithm persist for small and noise contaminated data sets.

Moreover, as we are interested in the recovery of the stable manifold of the underlying deterministic part, whenever we are confident that the observed data have relatively high noise

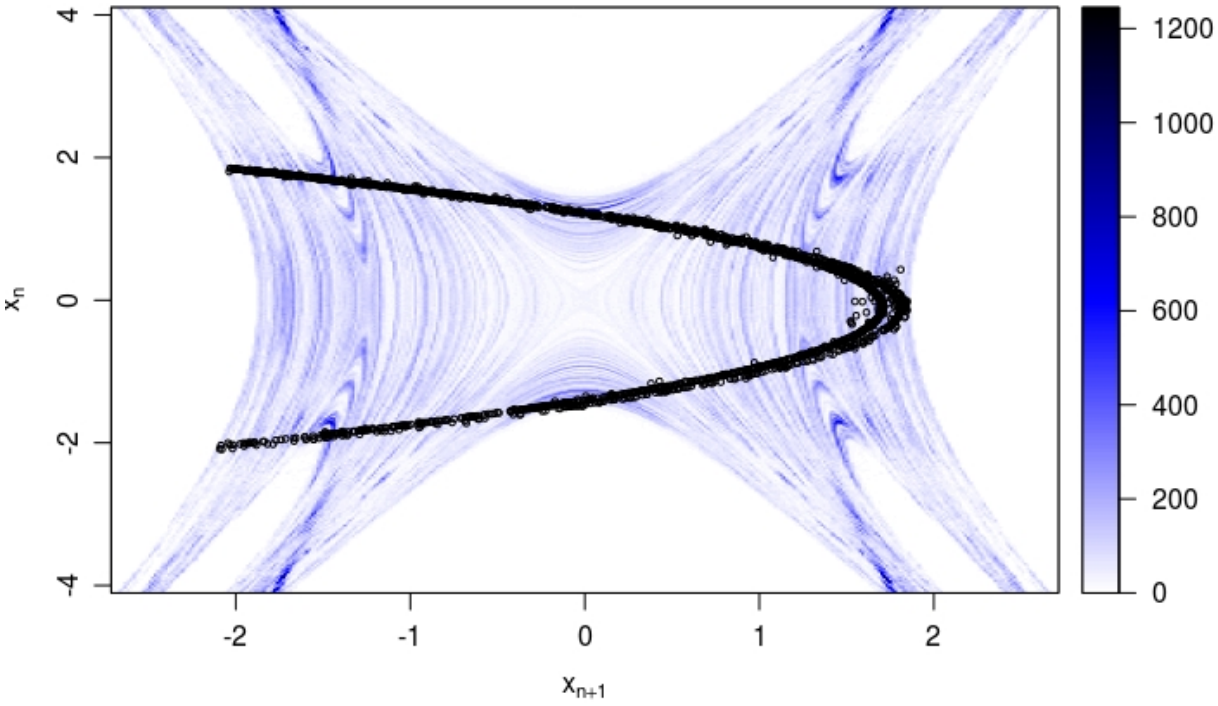

Figure 5.15: Stochastic approximation of the global stable manifold of the noninvertible map, superimposed with the noisy data set $\mathbf{x}^n$ (points in black). Colormap of the relative frequency, using a high analysis grid.

level, we can first use the DNRR model in order to obtain a noise–reduced orbit and then apply the BGSBR model on the denoised orbit.

We remark that for the noisy data sets generated by all Hénon-like maps, invertible or not, by using the complete quadratic (5.9) as modeling map, we have managed to perform identification of the underlying deterministic part with high accuracy. It is the accurate identification combined with the accurate noise density estimation that allows us to obtain adequate approximations of the associated stable manifolds. It is a straightforward task the construction of algorithms suitable for more general maps, such as for example lag 3 models, invertible or not. In fact, for the creation of Gibbs samplers for PAR processes of arbitrary degree and lag, we can apply minor modifications to the modeling polynomial equations appearing in the slice sampling step. For more details see Appendix A.

5.5 Conclusion

We have presented a new approach for the stochastic approximation of the global stable method, based on Bayesian nonparametric MCMC methods and particularly on the extension of the GSBR model, analyzed in Chapter 3. Our methods are data-driven and require no prior information about the underlying deterministic map, other than a polynomial functional representation. By a proper application of the proposed model we were able to stochastically approximate the global stable manifold and gain insight about the qualitative dynamical behavior, in

the zero noise limit (having as reference the associated deterministic system). Our method can be applied in cases of noninvertible maps.

As a future work, we propose the following directions:

1. Relaxing the assumption for a functional representation of the deterministic part, by adopting a Gaussian Process [Ras04] prior supported over the space of functions.
2. Improving manifold approximation, using manifold denoising algorithms [HM07], such as the Manifold Blurring Mean Shift algorithm [WCP10].

Chapter 6

Conclusions and future research

6.1 Conclusions

Noise effects are ubiquitous in real and experimental data sets, and are originated exogenously as measurement errors or endogenously as dynamic–interactive errors. In this thesis we have discussed the construction of novel algorithms aiming to confront the case of dynamical noise, which is vastly important mainly due to its capability to induce new dynamical phenomena.

In this thesis we have presented a new approach to nonlinear dynamic modeling, using Bayesian nonparametrics. We have relaxed the assumption of normality for the noise components, by developing a highly flexible Bayesian nonparametric framework. An important aspect of this approach is its capacity of approximating dynamical invariants which are in principle a difficult task to deal with from a computational point of view. Such dynamical invariants include quasi-invariant measures, homoclinic tangencies and global stable manifolds. We have seen that the combination of the fields of nonlinear dynamics and Bayesian nonparametrics can be useful in the direction of constructing a flexible and parsimonious framework, suitable for the analysis of real-world data and the detailed analysis of complex problems in dynamics.

More specifically, in this thesis we have applied Bayesian nonparametric methods in the problems of reconstruction, prediction [MKH17], noise reduction [KH18] and approximation of the stable manifold [HK19]. In all of the above problems we have shown that models based on the GSB random probability measures are efficient, with lower mean execution times and and at the same time are easier to implement than the corresponding Dirichlet Process based models. The main reason for this fact, is that although the two random measures have different random weights -in particular the GSB weights are simpler- they share the same support over the space of probability measures. We are confident that the Bayesian nonparametric methods in the future will assist further and successfully deal with with problems regarding complex dynamics and stochastic perturbations.

6.2 Directions for future research

There exists a plethora of research directions, which can be based on the exploitation of the connection between Bayesian nonparametrics and Random dynamical systems regarding both modeling of complex systems and novel theoretical aspects. Furthermore, there is potential for interesting applications of Random dynamical systems in Mathematical Biology and Neuroscience, where the effects of noise play crucial role and there is evidence of rich noise–induced phenomena, e.g. in neural modeling. In what follows, we address some future research projects, based on the results obtained throughout this thesis.

6.2.1 Gaussian Process & general noise scenarios

In working with real-world data sets, the assumption of a given functional family representation for the deterministic part of the unknown dynamical system, can be considered a restriction. One possible way to relax this assumption is to apply a Gaussian Process (GP) [Ras04] as a nonparametric prior over the unknown functional form. Under this approach, we could extend our models to a fully Bayesian nonparametric form.

An interesting research path also involves proper extensions of the nonparametric prior over the noise processes, in order to include more general cases such as multiplicative, multimodal or non–stationary noise processes. Furthermore, it would be worth investigating the applicability of our models in continuous-time systems, by performing proper modifications regarding essentially discretization schemes of the stochastic differential models.

6.2.2 RJMCMC Imputation & Embedding

Analysis of real data sets coming from measurements, may sometimes be challenging due to a malfunction of the measurement device. Such malfunctions could result in sets of missing observations, (perhaps) of an unknown number. A natural continuation of my research will be the development of a Bayesian nonparametric framework involving the concept of Reversible Jump Markov Chain Monte Carlo [Gre95] (RJMCMC) sampling, in order to estimate the the proper model regarding the number of consecutive missing observations and impute them using the proper Bayesian estimator.

Another interesting application of RJMCMC in the context of random dynamical systems regards the construction of a general MCMC sampling scheme, suitable for the determination of the embedding parameters (delay and embedding dimension), from one-dimensional observations. In cases where the observed time series is contaminated with dynamical noise, the traditional methods used for the determination of the embedding parameters are inefficient. An interesting alternative may include the randomization of the delay and the embedding dimension, where a RJMCMC based sampling scheme will enable transdimensional jumps between the models of different dimension, thus enabling the construction of the proper Bayesian model.

6.2.3 Perturbed Coupled–Map Lattices

Spatiotemporal phenomena arising in spatially extended systems, have been observed and analyzed in the context of many scientific disciplines, such as chemistry, neuroscience or engineering. A widely used approach for spatiotemporal modeling is the Coupled Map Lattice (CML), originally introduced by Kaneko [Kan93]. It consists of continuous states that evolve in discrete time and space and can be used to model dynamical behaviors described by partial differential equations. CMLs have been used to model and identify a wide range of complex dynamical phenomena, including spatiotemporal chaos, traveling waves, turbulence and intermittency. Essentially, in a CML we have dynamical units organized by means of a lattice that has a specific deterministic or random [Sin02] rule for spatial interaction (coupling). For example we may use a global or a local coupling scheme, depending on whether we aim to have interaction between all or only some of the lattice units, respectively.

We are interested in developing novel methods based on the Bayesian nonparametric framework, in order to perform inference based on data sets generated by stochastically perturbed CMLs. The need for a Bayesian nonparametric modeling is justified by the common assumption of Gaussian noise environment, which can be relaxed by using random measure mixtures of probability kernels as priors over the noise process. For example, suppose we have a CML at time n, with lattice sites $i \in \{1, \ldots, N\}$, that is:

$$x_{n+1}^{(i)} = (1-\epsilon) g\left(\vartheta, x_n^{(i)}\right) + \frac{\epsilon}{2}\left\{g\left(\vartheta, x_n^{(i-1)}\right) + g\left(\vartheta, x_n^{(i+1)}\right)\right\} + \zeta_n, \quad \zeta_n \overset{\text{iid}}{\sim} f(\cdot)$$

where $g(\cdot)$ is a real map, ϵ is the coupling strength of the nearest neighbor interactions, and the noise is distributed according to an unknown symmetric zero-mean density $f(\cdot)$. For example we could make the assumption that the maps are organized on a ring. Based on a data set generated by the associated perturbed CML, we are interested in estimating the control parameters ϑ, the initial point $\boldsymbol{x_0} = \left(x_0^{(1)}, \ldots, x_0^{(N)}\right)$, the coupling strength ϵ and the unknown noise density $f(\cdot)$.

Bibliography

[ABP85] FT Arecchi, R Badii, and A Politi. Generalized multistability and noise-induced jumps in a nonlinear dynamical system. *Physical Review A*, 32(1):402, 1985.

[AG92] D Assaf and Steve Gadbois. Definition of chaos. *American Mathematical Monthly*, 99(9):865–865, 1992.

[AJMR06] Ludwig Arnold, Christopher KRT Jones, Konstantin Mischaikow, and Geneviève Raugel. *Dynamical systems: lectures given at the 2nd session of the Centro internazionale matematico estivo (CIME) held in Montecatini Terme, Italy, June 13-22, 1994*. Springer, 2006.

[Ald85] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII-1983*, pages 1–198. Springer, 1985.

[Ant74] Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[Ara07] Zin Arai. On hyperbolic plateaus of the hénon map. *Experimental Mathematics*, 16(2):181–188, 2007.

[Arn13] Ludwig Arnold. *Random dynamical systems*. Springer Science & Business Media, 2013.

[ASY96] Kathleen T Alligood, Tim D Sauer, and James A Yorke. *Chaos*. Springer, 1996.

[BDH14] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *International Conference on Machine Learning (ICML)*, pages 405–413, 2014.

[BDV06] Christian Bonatti, Lorenzo J Díaz, and Marcelo Viana. *Dynamics beyond uniform hyperbolicity: A global geometric and probabilistic perspective*, volume 102. Springer Science & Business Media, 2006.

[Ber92] L Mark Berliner. Statistics, probability and chaos. *Statistical Science*, pages 69–90, 1992.

[BG93] Julian Besag and Peter J Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):25–37, 1993.

[BG98] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[BGJM11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[BKSE12] Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[BM+73] David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[Bor15] H. W. Borchers. *pracma: Practical Numerical Math Functions,*, 2015. R package version 1.8.3.

[BR98] Stephen P Brooks and Gareth O Roberts. Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing*, 8(4):319–335, 1998.

[BR05] Reinhold Blümel and William P Reinhardt. *Chaos in atomic physics*, volume 10. Cambridge University Press, 2005.

[BS09] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[BSJ90] AR Bulsara, WC Schieve, and EW Jacobs. Homoclinic chaos in systems perturbed by weak langevin noise. *Physical Review A*, 41(2):668, 1990.

[CDF97] Hans Crauel, Arnaud Debussche, and Franco Flandoli. Random attractors. *Journal of Dynamics and Differential Equations*, 9(2):307–341, 1997.

[CE07] Pierre Collet and Jean-Pierre Eckmann. *Concepts and results in chaotic dynamics: a short course*. Springer Science & Business Media, 2007.

[CEP13] Peter L Christiansen, JC Eilbeck, and Robert D Parmentier. *Future directions of nonlinear dynamics in physical and biological systems*, volume 312. Springer Science & Business Media, 2013.

[CFdlL03a] Xavier Cabré, Ernest Fontich, and Rafael de la Llave. The parameterization method for invariant manifolds i: manifolds associated to non-resonant subspaces. *Indiana University mathematics journal*, pages 283–328, 2003.

[CFdlL03b] Xavier Cabré, Ernest Fontich, and Rafael de la Llave. The parameterization method for invariant manifolds ii: regularity with respect to parameters. *Indiana University mathematics journal*, pages 329–360, 2003.

[CH+08] Gerda Claeskens, Nils Lid Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.

[Chu04] Igor Chueshov. *Monotone random systems theory and applications*. Springer, 2004.

[Cin13] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[Cof98] Donald S Coffey. Self-organization, complexity and chaos: the new biology for medicine. *Nature medicine*, 4(8):882–885, 1998.

[CT13] Kung-Sik Chan and Howell Tong. *Chaos: a statistical perspective*. Springer Science & Business Media, 2013.

[Cvi17] Predrag Cvitanovic. *Universality in chaos*. Routledge, 2017.

[Dav93] Mike Davies. Noise reduction by gradient descent. *International Journal of Bifurcation and Chaos*, 3(01):113–118, 1993.

[Dav94] Mike Davies. Noise reduction schemes for chaotic time series. *Physica D: Nonlinear Phenomena*, 79(2-4):174–192, 1994.

[Dav98] ME Davies. Nonlinear noise reduction through Monte Carlo sampling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 8(4):775–781, 1998.

[DDFG01] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[Dev08] Robert Devaney. *An introduction to chaotic dynamical systems*. Westview press, 2008.

[DF37] Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

[DF80] Persi Diaconis and David Freedman. De finetti's generalizations of exchangeability. *Studies in inductive logic and probability*, 2:233–249, 1980.

[DKPR87] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Conte Carlo. *Physics letters B*, 195(2):216–222, 1987.

[DM00] Holger R Dullin and JD Meiss. Generalized hénon maps: the cubic diffeomorphisms of the plane. *Physica D: Nonlinear Phenomena*, 143(1-4):262–289, 2000.

[Dua15] Jinqiao Duan. *An introduction to stochastic dynamics*, volume 51. Cambridge University Press, 2015.

[DW01] Paul Damien and Stephen G Walker. Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10(2):206–215, 2001.

[DWW99] P Damien, John Wakefield, and Stephen Walker. 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, 1999.

[DZ10] Jiu Ding and Aihui Zhou. *Statistical properties of deterministic systems*. Springer Science & Business Media, 2010.

[EEZ14] EM Elabbasy, AA Elsadany, and Yue Zhang. Bifurcation analysis and chaos in a discrete reduced lorenz system. *Applied Mathematics and Computation*, 228:184–194, 2014.

[EFT91] A Eden, C Foias, and R Temam. Local and global lyapunov exponents. *Journal of dynamics and differential equations*, 3(1):133–177, 1991.

[EKO04] James P England, Bernd Krauskopf, and Hinke M Osinga. Computing one-dimensional stable manifolds and stable sets of planar maps without the inverse. *SIAM Journal on Applied Dynamical Systems*, 3(2):161–190, 2004.

[ER85] J-P Eckmann and David Ruelle. Ergodic theory of chaos and strange attractors. In *The Theory of Chaotic Attractors*, pages 273–312. Springer, 1985.

[ES10] Zeraoulia Elhadj and Julien Clinton Sprott. *2-D quadratic maps and 3-D ODE systems: a rigorous approach*. World Scientific, 2010.

[Esc94] Michael D Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

[Fei78] Mitchell J Feigenbaum. Quantitative universality for a class of nonlinear transformations. *Journal of statistical physics*, 19(1):25–52, 1978.

[Fer73] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[Feu08] Ulrike Feudel. Complex dynamics in multistable systems. *International Journal of Bifurcation and Chaos*, 18(06):1607–1626, 2008.

[FGMW10] R. Fuentes-García, R. H. Mena, and S. G. Walker. A new Bayesian nonparametric mixture model. *Communications in Statistics–Simulation and Computation*, 39(4):669–682, 2010.

[FGPY98] Ulrike Feudel, Celso Grebogi, Leon Poon, and James A Yorke. Dynamical properties of a simple mechanical system with a large number of coexisting periodic attractors. *Chaos, Solitons & Fractals*, 9(1-2):171–180, 1998.

[FKP03] Christos E Frouzakis, Ioannis G Kevrekidis, and Bruce B Peckham. A route to computational chaos revisited: noninvertibility and the breakup of an invariant circle. *Physica D: Nonlinear Phenomena*, 177(1-4):101–121, 2003.

[FR81] Valter Franceschini and Lucio Russo. Stable and unstable manifolds of the hénon mapping. *Journal of Statistical Physics*, 25(4):757–769, 1981.

[FS91] J Doyne Farmer and John J Sidorowich. Optimal shadowing and noise reduction. *Physica D: Nonlinear Phenomena*, 47(3):373–392, 1991.

[GBP88] Peter Grassberger, Remo Badii, and Antonio Politi. Scaling laws for invariant measures on hyperbolic and nonhyperbolic atractors. *Journal of Statistical Physics*, 51(1-2):135–178, 1988.

[GBT95] Wally R Gilks, NG Best, and KKC Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44(4):455–472, 1995.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[GH13] John Guckenheimer and Philip Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42. Springer Science & Business Media, 2013.

[GK85] Peter Grassberger and Holger Kantz. Generating partitions for the dissipative hénon map. *Physics Letters A*, 113(5):235–238, 1985.

[GL02] R Gilmore and M Lefranc. The topology of chaos: Alice in stretch and squeezeland. *Hoboken: Wiley & sons inc*, page 518, 2002.

[GMJ17] JL Gonzalez and Jason D Mireles James. High-order parameterization of stable/unstable manifolds for long periodic orbits of maps. *SIAM Journal on Applied Dynamical Systems*, 16(3):1748–1795, 2017.

[Goe13] Georg M. Goerg. Forecastable component analysis. *Proceedings of the 30th International Conference on Machine Learning*, 28:64–72, 2013.

[Goe16] Georg M. Goerg. *ForeCA: An R package for Forecastable Component Analysis*, 2016. R package version 0.2.4.

[GP91] Franco Giovannini and Antonio Politi. Homoclinic tangencies, generating partitions and curvature of invariant manifolds. *Journal of Physics A: Mathematical and General*, 24(8):1837, 1991.

[GR+92] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[GR10] Dilan Görür and Carl Edward Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

[Gre95] Peter J Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[GRS95] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

[GS90] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[GSC+13] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[GT95] Charles J Geyer and Elizabeth A Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[GT01] John Geweke and Hisashi Tanizaki. Bayesian estimation of state-space models using the metropolis–hastings algorithm within gibbs sampling. *Computational Statistics & Data Analysis*, 37(2):151–170, 2001.

[GV04] John Guckenheimer and Alexander Vladimirsky. A fast method for approximating invariant manifolds. *SIAM Journal on Applied Dynamical Systems*, 3(3):232–260, 2004.

[GVdV17] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[Ham90] Stephen M Hammel. A noise reduction method for chaotic systems. *Physics letters A*, 148(8-9):421–428, 1990.

[Has70] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Hén76] Michel Hénon. A two-dimensional mapping with a strange attractor. In *The Theory of Chaotic Attractors*, pages 94–102. Springer, 1976.

[HES87] H Herzel, W Ebeling, and Th Schulmeister. Nonuniform chaotic dynamics and effects of noise in biochemical systems. *Zeitschrift für Naturforschung A*, 42(2):136–142, 1987.

[HF05] Markus Harle and Ulrike Feudel. On the relation between predictability and homoclinic tangencies. *International Journal of Bifurcation and Chaos*, 15(08):2523–2534, 2005.

[HH85] John A Hartigan and Pamela M Hartigan. The dip test of unimodality. *The annals of Statistics*, pages 70–84, 1985.

[HHMW10] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[Hig98] David M Higdon. Auxiliary variable methods for markov chain monte carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595, 1998.

[HK19] Spyridon J Hatjispyros and Konstantinos Kaloudis. On the stochastic approximation of the global stable manifold. *Submitted*, 2019.

[HM07] Matthias Hein and Markus Maier. Manifold denoising. In *Advances in neural information processing systems*, pages 561–568, 2007.

[HNW07a] Spyridon J Hatjispyros, Theodoros Nicoleris, and Stephen G Walker. Parameter estimation for random dynamical systems using slice sampling. *Physica A: Statistical Mechanics and its Applications*, 381:71–81, 2007.

[HNW07b] Spyridon J Hatjispyros, Theodoros Nicoleris, and Stephen G Walker. Parameter estimation for random dynamical systems using slice sampling. *Physica A: Statistical Mechanics and its Applications*, 381:71–81, 2007.

[HNW09] Spyridon J Hatjispyros, Theodoros Nicoleris, and Stephen G Walker. A Bayesian nonparametric study of a dynamic nonlinear model. *Computational Statistics & Data Analysis*, 53(12):3948–3956, 2009.

[HNW19] Spyridon J Hatjispyros, Theodoros Nicoleris, and Stephen G Walker. Distributional results relating to the posterior of a dirichlet process prior. *Statistics & Probability Letters*, 2019.

[HO15] Brian R Hunt and Edward Ott. Defining chaos. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097618, 2015.

[HS55] Edwin Hewitt and Leonard J Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955.

[HS76] David Heath and William Sudderth. De finetti's theorem on exchangeable variables. *The American Statistician*, 30(4):188–189, 1976.

[HS00] JPM Heald and J Stark. Estimation of noise levels for models of chaotic dynamical systems. *Physical review letters*, 84(11):2366, 2000.

[HST+01] Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[HTG94] A Hamm, T Tél, and R Graham. Noise-induced attractor explosions near tangent bifurcations. *Physics Letters A*, 185(3):313–320, 1994.

[HW11] Spyridon J Hatjispyros and Stephen G Walker. A high accuracy stochastic estimation of a nonlinear deterministic model. *Physics Letters A*, 375(19):1954–1964, 2011.

[HYG87] Stephen M Hammel, James A Yorke, and Celso Grebogi. Do numerical orbits of chaotic dynamical processes represent true orbits? *Journal of Complexity*, 3(2):136–145, 1987.

[HZ98] Bai-Lin Hao and Wei-Mou Zheng. *Applied symbolic dynamics and chaos*, volume 7. World scientific, 1998.

[IJ01] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick–breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[IJ03] Hemant Ishwaran and Lancelot F James. Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, pages 1211–1235, 2003.

[IZ00] Hemant Ishwaran and Mahmoud Zarepour. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

[IZ02] Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

[Jan12] Maarten Jansen. *Noise reduction by wavelet thresholding*, volume 161. Springer Science & Business Media, 2012.

[JDVS13] Nouha Jaoua, Emmanuel Duflos, Philippe Vanheeghe, and François Septier. 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, 2013.

[JJN+13] Alicia A Johnson, Galin L Jones, Ronald C Neath, et al. Component-wise markov chain monte carlo: Uniform and geometric ergodicity under mixing and composition. *Statistical Science*, 28(3):360–375, 2013.

[JK97a] Lars Jaeger and Holger Kantz. Effective deterministic models for chaotic dynamics perturbed by noise. *Physical Review E*, 55(5):5234, 1997.

[JK97b] Lars Jaeger and Holger Kantz. Homoclinic tangencies and non-normal jacobians - effects of noise in nonhyperbolic chaotic systems. *Physica D: Nonlinear Phenomena*, 105(1):79–96, 1997.

[Jud08] Kevin Judd. Shadowing pseudo-orbits and gradient descent noise reduction. *Journal of Nonlinear Science*, 18(1):57–74, 2008.

[Kan93] Kunihiko Kaneko. Theory and applications of coupled map lattices. *Nonlinear science: theory and applications*, 1993.

[KBR18] Andreĭ Nikolaevich Kolmogorov and Albert T Bharucha-Reid. *Foundations of the Theory of Probability: Second English Edition*. Courier Dover Publications, 2018.

[KFG99] Suso Kraut, Ulrike Feudel, and Celso Grebogi. Preference of attractors in noisy multistable systems. *Physical Review E*, 59(5):5253–5260, 1999.

[KG85] H Kantz and Peter Grassberger. Repellers, semi-attractors, and long-lived chaotic transients. *Physica D: Nonlinear Phenomena*, 17(1):75–86, 1985.

[KGW11] Maria Kalli, Jim E. Griffin, and Stephen G. Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

[KH18] Konstantinos Kaloudis and Spyridon J Hatjispyros. A Bayesian nonparametric approach to dynamical noise reduction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):063110, 2018.

[KKA15] Oktay Karakuş, Ercan E Kuruoğlu, and Mustafa A Altinkaya. Estimation of the nonlinearity degree for polynomial autoregressive processes with RJMCMC. *In Proceedings 23rd European Signal Processing Conference (EUSIPCO)*, 953–957, 2015.

[KO+98] Bernd Krauskopf, Hinke Osinga, et al. Growing 1d and quasi-2d unstable manifolds of maps. *Journal of Computational Physics*, 146(1):404–419, 1998.

[KO99] Bernd Krauskopf and Hinke Osinga. Two-dimensional global manifolds of vector fields. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(3):768–774, 1999.

[KOD+05] Bernd Krauskopf, Hinke M Osinga, Eusebius J Doedel, Michael E Henderson, John Guckenheimer, Alexander Vladimirsky, Michael Dellnitz, and Oliver Junge. A survey of methods for computing (un) stable manifolds of vector fields. *International Journal of Bifurcation and Chaos*, 15(03):763–791, 2005.

[KS04] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[KSSH15a] Kiyoshi Kanazawa, Tomohiko G Sano, Takahiro Sagawa, and Hisao Hayakawa. Asymptotic derivation of langevin-like equation with non-gaussian noise and its analytical solution. *Journal of Statistical Physics*, 160(5):1294–1335, 2015.

[KSSH15b] Kiyoshi Kanazawa, Tomohiko G Sano, Takahiro Sagawa, and Hisao Hayakawa. Minimal model of stochastic athermal systems: Origin of non-gaussian noise. *Physical review letters*, 114(9):090601, 2015.

[KYY96] Eric J Kostelich, James A Yorke, and Zhiping You. Plotting stable manifolds: error estimates and noninvertible maps. *Physica D: Nonlinear Phenomena*, 93(3-4):210–222, 1996.

[L+84] Albert Y Lo et al. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

[L+99] Steven P Lalley et al. Beneath the noise, chaos. *The Annals of Statistics*, 27(2):461–479, 1999.

[LC06] Richard A Levine and George Casella. Optimizing random scan gibbs samplers. *Journal of Multivariate Analysis*, 97(10):2071–2100, 2006.

[LFZ12] Huimin Li, Yangyu Fan, and Jing Zhang. A new algorithm for computing one-dimensional stable and unstable manifolds of maps. *International Journal of Bifurcation and Chaos*, 22(01):1250018, 2012.

[LGYK93] Y-C Lai, C Grebogi, JA Yorke, and Ittai Kan. How often are chaotic saddles nonhyperbolic? *Nonlinearity*, 6(5):779, 1993.

[Li93] Shi Hai Li. ω-chaos and topological entropy. *Transactions of the American Mathematical Society*, 339(1):243–249, 1993.

[LM85] Andrzej Lasota and Michael C Mackey. *Probabilistic properties of deterministic systems*. Cambridge University Press, 1985.

[LN06] Steven P Lalley and Andrew B Nobel. Denoising deterministic time series. *arXiv preprint nlin/0604052*, 2006.

[LSDM08] Dmitri G. Luchinsky, Vadim N. Smelyanskiy, Andrea Duggento, and Peter V. E. McClintock. Inferential framework for nonstationary dynamics. I. Theory. *Physical Review E*, 77(6):061105, 2008.

[LY73] Andrzej Lasota and James A Yorke. On the existence of invariant measures for piecewise monotonic transformations. *Transactions of the American Mathematical Society*, pages 481–488, 1973.

[LY75] Tien-Yien Li and James A Yorke. Period three implies chaos. *The American Mathematical Monthly*, 82(10):985–992, 1975.

[LYHN05] Richard A Levine, Zhaoxia Yu, William G Hanley, and John J Nitao. Implementing random scan gibbs samplers. *Computational Statistics*, 20(1):177–196, 2005.

[Man85] Ricardo Mané. Hyperbolicity, sinks and measure in one dimensional dynamics. *Communications in mathematical physics*, 100(4):495–524, 1985.

[May76] Robert M May. Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459, 1976.

[MC00] Renate Meyer and Nelson Christensen. Bayesian reconstruction of chaotic dynamical systems. *Physical Review E*, 62(3):3535–3542, 2000.

[MC01] Renate Meyer and Nelson Christensen. Fast Bayesian reconstruction of chaotic dynamical systems via extended Kalman filtering. *Physical Review E*, 65(1):016206, 2001.

[Mer18] Christos Merkatas. *Bayesian nonparametrics and applications*. PhD thesis, University of the Aegean, School of Sciences, Dept. of Statistics and Actuarial-Financial Mathematics, 2018.

[Mid77] David Middleton. Statistical–physical models of electromagnetic interference. *IEEE Transactions on Electromagnetic Compatibility*, (3):106–127, 1977.

[MKH17] Christos Merkatas, Konstantinos Kaloudis, and Spyridon J Hatjispyros. A bayesian nonparametric approach to reconstruction and prediction of random dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(6):063116, 2017.

[MLMF12] Y. I. Molkov, E. M. Loskutov, D. N. Mukhin, and A. M. Feigin. Random dynamical models from time series. *Physical Review E*, 85(3):036216, 2012.

[MM98] Steven N MacEachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

[MMP15] Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. Statistical inference for dynamical systems: A review. *Statistics Surveys*, 9:209–252, 2015.

[MMR05] Jean-Michel Marin, Kerrie Mengersen, and Christian P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.

[MNS+01] Takashi Matsumoto, Yoshinori Nakajima, Motoki Saito, Junjiro Sugi, and Hiroaki Hamagishi. Reconstructions and predictions of nonlinear dynamical systems: A hierarchical Bayesian approach. *IEEE transactions on signal processing*, 49(9):2138–2155, 2001.

[MQJH15] Peter Müller, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. *Bayesian nonparametric data analysis*. Springer, 2015.

[MRR+53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[MS99a] Patrick E McSharry and Leonard A Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Physical Review Letters*, 83(21):4285–4288, 1999.

[MS99b] Salah-Eldin A Mohammed and Michael KR Scheutzow. The stable manifold theorem for stochastic differential equations. *Annals of probability*, pages 615–652, 1999.

[MT83] Kenji Matsumoto and Ichiro Tsuda. Noise-induced order. *Journal of Statistical Physics*, 31(1):87–106, 1983.

[MT97] Antonietta Mira and Luke Tierney. On the use of auxiliary variables in markov chain monte carlo sampling. *Scandinavian Journal of Statistics*, 1997.

[MT98] Pietro Muliere and Luca Tardella. Approximating distributions of random functionals of Ferguson–Dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297, 1998.

[Mur12] Kevin Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012.

[N+11] Radford M Neal et al. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

[NBPW+04] Luis E Nieto-Barajas, Igor Prünster, Stephen G Walker, et al. Normalized random measures driven by increasing additive processes. *The Annals of Statistics*, 32(6):2343–2360, 2004.

[Nea92] Radford M Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

[Nea00] Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[Nea03] Radford M Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003.

[NMKY05] Yohei Nakada, Takashi Matsumoto, Takayuki Kurihara, and Kuniaki Yosui. Bayesian reconstructions and predictions of nonlinear dynamical systems via the Hybrid Monte Carlo scheme. *Signal processing*, 85(1):129–145, 2005.

[NTT11] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons, 2011.

[OC04] Andrea Ongaro and Carla Cattaneo. Discrete random probability measures: a general framework for nonparametric bayesian inference. *Statistics & Probability Letters*, 67(1):33–45, 2004.

[Ose68] Valery Iustinovich Oseledec. A multiplicative ergodic theorem. liapunov characteristic number for dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231, 1968.

[Ott02] Edward Ott. *Chaos in dynamical systems*. Cambridge university press, 2002.

[P+02] Jim Pitman et al. Combinatorial stochastic processes. Technical report, Lecture notes for St. Flour course, Technical Report 621, Dept. Statistics, UC Berkeley., 2002.

[PCFS80] Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

[Pha15] Eswar G Phadia. *Prior processes and their applications*. Springer, 2015.

[Pin91] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[Pit95] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

[PPP94] Edgar E Peters, Edgar R Peters, and Donada Peters. *Fractal market analysis: applying chaos theory to investment and economics*, volume 24. John Wiley & Sons, 1994.

[PR08] Omiros Papaspiliopoulos and Gareth O Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

[PS00] Reynaldo D Pinto and José C Sartorelli. Homoclinic tangency and chaotic attractor disappearance in a dripping faucet experiment. *Physical Review E*, 61(1):342, 2000.

[PS17] Sergei Yu Pilyugin and Kazuhiro Sakai. *Shadowing and hyperbolicity*, volume 2193. Springer, 2017.

[PV94] Jacob Palis and Marcelo Viana. High dimension diffeomorphisms displaying infinitely many periodic attractors. *Annals of mathematics*, 140(1):207–250, 1994.

[PY97] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[PY98] Mark Pollicott and Michiko Yuri. *Dynamical systems and ergodic theory*, volume 40. Cambridge University Press, 1998.

[R+11] Jeffrey S Rosenthal et al. Optimal proposal distributions and adaptive mcmc. *Handbook of Markov Chain Monte Carlo*, 4(10.1201), 2011.

[Ras00] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[Ras04] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[RC04] Christian Robert and George Casella. Monte carlo statistical methods. *Springer, New York*, 2004.

[RCC10] Christian P Robert, George Casella, and George Casella. *Introducing monte carlo methods with r*, volume 18. Springer, 2010.

[RG97] Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

[RGG+97] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[Rob77] Clark Robinson. Stability theorems and hyperbolicity in dynamical systems. *The Rocky Mountain Journal of Mathematics*, pages 425–437, 1977.

[Rob07] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[RR+01] Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[RR09] Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[RS97] Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

[San00] Evelyn Sander. Homoclinic tangles for noninvertible maps. *Nonlinear Analysis: Theory, Methods & Applications*, 41(1-2):259–276, 2000.

[SDG+91] John C Sommerer, William L Ditto, Celso Grebogi, Edward Ott, and Mark L Spano. Experimental confirmation of the scaling theory for noise-induced crises. *Physical review letters*, 66(15):1947, 1991.

[Set94] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[SG74] M. Shinde and S. Gupta. Signal detection in the presence of atmospheric noise in tropics. *IEEE Transactions on Communications*, 22(8):1055–1063, 1974.

[SG91] Thomas Schreiber and Peter Grassberger. A simple noise-reduction method for real data. *Physics letters A*, 160(5):411–418, 1991.

[SH98] Klaus Reiner Schenk-Hoppé. Random attractors–general properties, existence and applications to stochastic bifurcation theory. *Discrete & Continuous Dynamical Systems-A*, 4(1):99–130, 1998.

[Sin72] Yakov G Sinai. Gibbs measures in ergodic theory. *Russian Mathematical Surveys*, 27(4):21, 1972.

[Sin02] Sudeshna Sinha. Random coupling of chaotic maps leads to spatiotemporal synchronization. *Physical Review E*, 66(1):016209, 2002.

[SJ06] Heinz Georg Schuster and Wolfram Just. *Deterministic chaos: an introduction*. John Wiley & Sons, 2006.

[SKFP04] M Siefert, M Kern, R Friedrich, and J Peinke. How to differentiate quantitatively between nonlinear dynamics, dynamical noise and measurement noise. In *EXPERIMENTAL CHAOS: 8th Experimental Chaos Conference*, volume 742, pages 337–344. AIP Publishing, 2004.

[SLTB05] V. N. Smelyanskiy, Dmitry G. Luchinsky, D. A. Timucin, and A. Bandrivskyy. Reconstruction of stochastic nonlinear dynamical models from trajectory measurements. *Physical Review E*, 72(2):026202, 2005.

[SM08] Marek Strumik and Wieslaw Macek. Influence of dynamical noise on time series generated by nonlinear maps. *Physica D: Nonlinear Phenomena*, 237(5):613–618, 2008.

[Sma67] Stephen Smale. Differentiable dynamical systems. *Bulletin of the American mathematical Society*, 73(6):747–817, 1967.

[Som93] John C Sommerer. A universal scaling for noise-induced metastability. *Physics Letters A*, 176(1-2):85–88, 1993.

[SS94] Berthold Schweizer and Jaroslav Smital. Measures of chaos and a spectral decomposition of dynamical systems on the interval. *Transactions of the American Mathematical Society*, 344(2):737–754, 1994.

[Str18] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2018.

[SX15] JC Sprott and Anda Xiong. Classifying and quantifying basins of attraction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(8):083101, 2015.

[SYC91] Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3-4):579–616, 1991.

[Tak81] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[TBS03] Ioana Triandaf, Erik M Bollt, and Ira B Schwartz. Approximating stable and unstable manifolds in experiments. *Physical Review E*, 67(3):037201, 2003.

[TGG07] Yee Whye Teh, Dilan Grür, and Zoubin Ghahramani. Stick-breaking construction for the indian buffet process. In *Artificial Intelligence and Statistics*, pages 556–563, 2007.

[Tie94] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[TKR+10] Marco Thiel, Jürgen Kurths, M Carmen Romano, György Károlyi, and Alessandro Moura. *Nonlinear dynamics and chaos: advances and perspectives*. Springer, 2010.

[TLG08] Tamás Tél, Ying-Cheng Lai, and Márton Gruiz. Noise-induced chaos: a consequence of long deterministic transients. *International Journal of Bifurcation and Chaos*, 18(02):509–520, 2008.

[VB94] Michel Vellekoop and Raoul Berglund. On intervals, transitivity = chaos. *The American Mathematical Monthly*, 101(4):353–355, 1994.

[vdBJR16] Jan Bouwe van den Berg, Jason D Mireles James, and Christian Reinhardt. Computing (un) stable manifolds with validated error bounds: non-resonant and resonant spectra. *Journal of Nonlinear Science*, 26(4):1055–1095, 2016.

[Wal07] Stephen G Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics–Simulation and Computation*, 36(1):45–54, 2007.

[WCP10] Weiran Wang and Miguel A Carreira-Perpinán. Manifold blurring mean shift algorithms for manifold denoising. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1759–1766. IEEE, 2010.

[Wes92] Mike West. *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03, 1992.

[Wes12] Bruce J West. *Fractal physiology and chaos in medicine*, volume 16. World Scientific, 2012.

[WH13] Xiaoyi Wang and Yu Huang. Devaney chaos revisited. *Topology and its Applications*, 160(3):455–460, 2013.

[WM97] David M Walker and Alistair I Mees. Noise reduction of chaotic systems by kalman filtering and by shadowing. *International Journal of Bifurcation and Chaos*, 7(03):769–779, 1997.

[YKY91] Zhiping You, Eric J Kostelich, and James A Yorke. Calculating stable and unstable manifolds. *International Journal of Bifurcation and Chaos*, 1(03):605–623, 1991.

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. (3.14) 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' [DWW99] 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 $j = 1, \dots, m$ 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. (3.15) and (3.25), 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 $r \leq m$ such that

$$\mathcal{R}_g = \bigcup_{i=1}^{r} (\rho_{2i-1}, \rho_{2i}), \tag{A.6}$$

with $\{\rho_1, \dots, \rho_{2r}\}$ 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 $s' \geq 1$ and $t' \geq 0$ 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 eq. (3.16) and **??** given in the main text are non-standard 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. (3.29) 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.* (3.31)*, 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 DNRR and the BGSBR models (as well as their parametric counterparts) have been developed in the `Julia` language [BKSE12].

The associated software is available and can be downloaded from the URL:

`Link to thesis codes`

or available upon request via e–mail:

`kkaloudis@aegean.gr` or `konst.kaloudis@gmail.com`.