PhD THESIS DECLARATION

The undersigned

Cappello Lorenzo
PhD Registration Number 1759984

Recursive Procedures for Nonparametric Inference in Multivariate Settings

PhD in Statistics

Cycle 29^{th}

Advisor: Professor Sonia Petrone (Università Bocconi)

Co-advisor: Professor Stephen G. Walker (UT Austin)

Year of thesis defence *2018*

DECLARES

Under his responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28^{th} December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30^{th} April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio istituzionale ad accesso aperto" (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

   - thesis: Recursive Procedures for Nonparametric Inference in Multivariate Settings;
   - by Cappello Lorenzo;
   - defended at Università Commerciale "Luigi Bocconi" – Milano in 2018;
   - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo.

Date February 16th, 2018

Cappello Lorenzo

Contents

1 Introduction **1**
1.1 Outline of the Thesis 6

2 Background **9**
2.1 Mixture Models and Bayesian Inference 9
2.2 Recursive Algorithms 13
2.2.1 Recursive Algorithm of Newton et al. (1998) 14
2.2.2 Recursive Algorithm of Hahn et al. (2017) 15
2.2.3 Algorithms Implementation 18
2.2.4 Stochastic Approximation 20
2.2.5 Fixed Point Iterations 22
2.3 Asymptotic Theory for Recursive Algorithms 23
2.3.1 Convergence Results for the NQZ algorithm 24
2.3.2 Convergence Results for the HMW algorithm 26

3 Recursive Procedures for Conditional Distributions **29**
3.1 Recursive Nonparametric Predictive Regression 31
3.1.1 A parametric model 32
3.1.2 Bayesian Inference for Mixture Regression 33
3.1.3 A Recursive Predictive Regression Model 35
3.1.4 Illustrations 40
3.2 A Two-Step Recursive Predictive Algorithm 45
3.2.1 The Bivariate Dirichlet Process 47
3.2.2 A Two-step Recursive Algorithm 48
3.3 Discussion 52

4 Laplace Inversion for Multivariate Probability Distributions **55**
4.1 Motivating Application 55
4.2 Literature Review 56
4.3 A Multidimensional Inversion Method 59
4.4 Numerical Examples 62

4.5 Discussion 73

5 Asymptotic Theory for Recursive Procedures 75

5.1 One-dimensional Convergence 76
5.2 Weak Convergence of (P_n) 79
5.3 NQZ Algorithm 86
5.3.1 Inversion of the Laplace Transform. 87
5.3.2 General NQZ Algorithm 90
5.4 Algorithms based on the copula 91
5.4.1 HMW Algorithm 91
5.4.2 Recursive Regression Algorithm 92
5.4.3 Two Steps Predictive Recursion 95
5.4.4 Tightness 98
5.5 Discussion 100

6 Discussion and Future Work 103

Bibliography 106

List of Figures

3.1 RR estimates of the predictive distributions $P(y|x)$ for $x = 1$ and $x = 4$ in Example 1: P_0 (green), recursive estimate (black) and true distribution (blue). 40

3.2 RR estimates of the predictive distributions $P(y|x)$ for $x = 1$ and $x = 5$ in Example 2: P_0 (green), recursive estimate (black) and true distribution (blue). 41

3.3 Estimates of the predictive distributions $P(y|x)$ for $x = 1$ and $x = 2$ in Example 3: P_0 (green), recursive estimate given by RR (black), recursive estimate given by HMW (green), empirical CDF (red) and true distribution (blue). 41

3.4 Distributions of Kolomorov Smirnov statistics for the four groups and the three estimator: RR (Recursive), the single-p DDP of De Iorio et al. (2004) (Bayes) and the kernel regression of Aitchison and Aitken (1976) (Kernel) . 43

3.5 Distributions of Cramer von Mises statistics for the four groups and the three estimator: RR (Recursive), the single-p DDP of De Iorio et al. (2004) (Bayes) and the kernel regression of Aitchison and Aitken (1976) (Kernel) . 44

3.6 IgG distribution $P(y|x)$ for the six age groups ($x = 0.5, x = 1.5, x = 2.5, x = 3.5, x = 4.5, x = 5.5$) estimated through four procedures: recursive estimate RR (Rec, black), single-p DDP of De Iorio et al. (2004) (Bay, orange), kernel regression of Aitchison and Aitken (1976) (Kern, purple) and empirical CDF (Emp, red). 46

3.7 TSPR estimates of the predictive distributions $P^{(a)}$ (Panel (a)) and $P^{(b)}$ (Panel (b)) in Example 4: the prior guess P_0 (green), the TSPR estimate (black) and true distribution (blue). 51

3.8 TSPR estimates of the predictive distributions $P^{(a)}$ (Panel (a)) and $P^{(b)}$ (Panel (b)) in Example 5: the prior guess P_0 (green), the TSPR estimate (black) and true distribution (blue). 52

4.1 Contour plots of the *prior* $G_0(s,t)$ (Panel 4.1a), *estimated* $G_n(s,t)$ (Panel 4.1b), *Gaver–Stehfest* (Panel 4.1c) and the *true* $G(s,t)$ (Panel 4.1d): Example 1 66
4.2 Surface plots of the densities *recursive* $g_n(s,t)$ (Panel 4.2a) and *Gaver–Stehfest* $g_N(s,t)$(Panel 4.2b): Example 2 68

5.1 Plots of $\widehat{P}_n(B)$ and $\widehat{P}_n(M)$ 99

List of Tables

4.1 Cumulative distribution function of a bivariate gamma for selected values. 65
4.2 Laplace transform of a bivariate positive stable for selected values (x, y) . . 67
4.3 Cumulative distribution function of a Downton Exponential for selected values. 69
4.4 Cumulative distribution function of a trivariate gamma for selected values. 71
4.5 Cumulative distribution function of a Lognormal for selected values. . . . 73

Acknowledgements

The four years spent as a PhD student have been exciting and rewarding.

First of all, I would like to thank Stephen G. Walker. His support to my research with ideas, insights and experience, have only be a fraction of his role as an advisor in the past years. His passion for this profession and friendly approach are contagious and has been a great motivation for me. His advices on life in academia and views on how research in science should be carried are very valuable. I believe that the best I can do to repay him is to try to live up to these precious advices.

I am also grateful to Sonia Petrone at Bocconi University, for the time spent discussing my research: her valuable insights has helped me greatly to improve this thesis, broaden its focus and clarify the presentation. Also for having given valuable inputs for future research, with the help of Sandra Fortini.

I would like to acknowledge the support of the professors in the Department of Decision Sciences at Bocconi University. In particular, Pietro Muliere, for his genuine support, encouragement and teachings to challenge current beliefs, Isadora Antoniano Villalobos for the many helpful technical pointers and friendship, and Igor Prünster and Antonio Lijoi for the support and opportunities they have offered me since they have joined the faculty.

Moving back and forth between Austin and Milan, and the numerous conferences I have had the opportunity to attend, I must have caused quite some troubles to the administration and the secretarial staff of both Bocconi University and UT Austin. I want to acknowledge the work of Angela Baldassarre, Silvia Acquati and Gualtiero Valsecchi at Bocconi, and Vicki Keller at UT Austin.

A special acknowledgement goes to the other graduate students at Bocconi and UT Austin. Many of them have helped me greatly in my research through helpful questions and remarks. In particular I am thinking about Marta and Irina, I sincerely value their friendship.

Living in two countries during the PhD has been enriching but also very intense and from time to time emotionally draining. Luckily, I could count on the support of good friends, some of whom were already in my life, some I have met along the way. They have been a wonderful support, helped me to take off my mind from the research when

needed, challenged my opinions, stimulated by curiosity. I do not want to make a list of names, but many of whom know they are a second family.

Lastly I want to thank my parents, Maurizia and Maurizio, for their unconditional support. My brother Stefano for being such a positive influence in my life. I would like to thank my sister in law Eleonora, and my niece and nephew Nina and Umberto, which are my family in Milan.

Abstract

Within the literature over the past two decades we have seen the emergence of recursive algorithms that target probability distributions and are motivated by Bayesian heuristics. These algorithms are fast, accurate, easy to implement, and theoretically sound. As such, they appear capable of addressing several central issues within the field, such as low computational time, sequential learning, as well as possibility to paralyze.

Despite this, there remains many unanswered questions about recursive algorithms within the scholarship, and still yet little work being done to find these answers. This thesis aims to begin bridging this gap by addressing a few key issues, in hopes to advance the understanding around these recursive procedures and their applicability. This thesis focuses on three main research areas: estimation of covariate-dependent distributions, application to inverse problem in a dimension higher than one and asymptotic theory.

The first contribution is an extension of the range of application of these recursive estimators. We propose two new algorithms to estimate covariate-dependent distributions in a regression setting with regressors taking finitely many values. Again, these methods retain a Bayesian interpretation: the functional form of each algorithm is suggested by the first update of a Bayesian nonparametric model. The information loss that follows from the departure from a fully-fledge Bayesian procedure is compensated by a substantial gain in computational speed. In a simulation studies and real data illustrations, we show that the numerical distance between our estimates and the data generating distributions is comparable to that calculated for the estimates given by a kernel and a Bayesian nonparametric regressions. We also study the convergence properties of these algorithms.

Another line of investigation deals with an application: the inversion of the Laplace transform of a multivariate distribution function. We propose a new methodology which relies on an existing recursive procedure suitably adapted to this problem. This application well illustrates the attractive features of recursive algorithms: they are incredibly fast in comparison to alternatives while remaining accurate. Our proposal is also the only methodology for Laplace inversion that gives a theoretical guarantee to recover a proper distribution. Several numerical strategies needed to implement recursive procedures at dimension higher than one are discussed.

Finally, we study the asymptotic theoretical convergence of a quite general family of recursive algorithms. Each recursive procedure consists in a linear combination of the previous step and an update driven by deterministic weights. If observations are assumed i.i.d., we show that a key condition to analyze the convergence is that the update admits a unique fixed point. We introduce a novel martingale argument, which, along with fixed point theory, allows to prove weak convergence under some appropriate conditions.

We envisage three novel contributions here: first, we bridge the gap that exists within the recursive algorithms literature, for example the stochastic approximation literature, which does not cover extensively the case of infinite dimensional sequences; second, we study convergence of existing algorithms, such as the one due to Newton et al. (1998), under less stringent conditions for some given models; lastly, this proof technique allows to study the theoretical convergence of the new algorithms proposed in this thesis. Some of the ideas in this area are part of a research that it is still ongoing.

After a short introduction (Chapter 1), the thesis is structured in a way that the background (Chapter 2) and the methodological contributions (the two new algorithms in Chapter 3, the Laplace transform inversion in Chapter 4) are presented upfront. All the asymptotic arguments are included in Chapter 5. Chapter 6 includes a short discussion.

Chapter 1

Introduction

Recursive algorithms are abundant in the mathematics and statistics literature. A large amount of work has been done to characterize their theoretical and empirical properties. They are applied in a wide variety of problems, ranging from optimization - e.g. Newton-Raphson - to the tracking of dynamical systems - e.g. Kalman filter. This thesis is about stochastic recursive algorithms. In particular, we target the estimation of cumulative distribution functions and densities given samples from either unknown or known distributions.

We consider the following setting. Let (Y_n) denote a sequence of observations taking values in a sample space $\mathcal{Y}$ and let $\mathcal{P}$ denote the class of probability distributions on $(\mathcal{Y}, \mathcal{B})$, where $\mathcal{B}$ is the σ - field associated with $\mathcal{Y}$. Regardless of the stochastic law of the observables (Y_n), it is easy to see that a sequence of probability distributions (P_n) taking values in $\mathcal{P}$ can be constructed through the following scheme: fix a prior guess $P_0 \in \mathcal{P}$, a deterministic sequence $\alpha_n \in (0, 1)$, and repeat recursively

$$P_n(\cdot) = (1 - \alpha_n)P_{n-1}(\cdot) + \alpha_n Z(\cdot, P_{n-1}, Y_n), \tag{1.1}$$

where Z needs to be a cumulative distribution function (CDF) belonging to $\mathcal{P}$. A fixed sample of observations $y_{1:N} := (y_1, \ldots, y_N)$ will produce an estimate $P_N(\cdot)$ which is obtained by repeating the scheme described in (1.1) for $n = 1, \ldots, N$. In the following we will often refer to Z as the *update*.

The definition and the properties of Z are of utmost importance. Newton et al. (1998) proposed a recursive estimator in the context of mixture model. The algorithm is not exactly of the type (1.1) because it defines a sequence of probability distributions on a latent space, but through some suitable assumptions the study of Newton et al. (1998) can be reconciled with (1.1). The algorithm became popular because it allows for an online

estimation of the mixing distribution in mixture models, which is fast and numerically accurate. Its striking feature is that the function Z is suggested by the first step of a Bayesian update for a well-studied model: the Dirichlet Process mixture model (Lo 1984). Specifically

$$Z(\cdot, P, y) = \frac{\int_{-\infty}^{(\cdot)} k(y|\theta)\mathrm{d}P(\theta)}{\int_{\Theta} k(y|\theta)\mathrm{d}P(\theta)},$$

where y is a random sample from $f(y) = \int_{\Theta} k(y|\theta)\mathrm{d}P^*(\theta)$, and the algorithm searches for P^*.

Another recent application is in Hahn et al. (2017). The authors first show an insightful characterisation of predictive densities in terms of copula. Let $Y_1 \sim P_1$ and P_n be the predictive distribution of $Y_{n+1}|Y_{1:n}$ for $n \geq 1$, and p_n the corresponding predictive density. The ratio of two consecutive terms of $(p_n)_{n\geq 1}$, say p_n and p_{n-1}, is a copula density; i.e.

$$p_n(y) = c_n(P_{n-1}(y), P_{n-1}(y_n))\, p_{n-1}(y),$$

for some sequence of copula densities (c_n); indeed, Hahn et al. (2017) show that for conditionally *independent identically distributed (i.i.d.)* random variables, it holds that

$$c(P(y), P(z)) = \frac{\int f(y|\theta)\, f(z|\theta)\pi(\mathrm{d}\theta)}{p(y)p(z)},$$

where $p(y) = \int f(y|\theta)\pi(\mathrm{d}\theta)$, for some $\pi(\theta)$. Note that the result can be shown to hold regardless of the probability law of the observables. It follows from Sklar's Theorem that any multivariate distribution function F, and corresponding density f, can be represented by a copula. The ratio of consecutive terms of $(p_n(y))_{n\geq 1}$ is a copula density

$$\frac{p_n(y)}{p_{n-1}(y)} = \frac{p(y, y_n|y_{1:n-1})}{p(y_n|y_{n-1})p(y|y_{1:n-1})} = c(P_{n-1}(y), P_{n-1}(y_n))$$

The Bayesian update of the predictive density at step $n-1$ can thus be done through a copula density function c_n, that depends solely on the sample size. Standard Bayesian models allow for a closed form copula. Lacking a closed form for c_n, as in the case Dirichlet Process mixtures, Hahn et al. (2017) suggest a recursive estimator for the mixture distribution. It is accurate and fast. The copula allows us to bypass some integrals that need to be evaluated numerically in Newton's algorithm, making the iterations even faster. In this case, using a Gaussian copula,

$$Z(\cdot, P, Y) = \Phi\left(\frac{\Phi^{-1}(P(\cdot)) - \rho\Phi^{-1}(P(Y))}{\sqrt{1-\rho^2}}\right),$$

where Φ is a standard Gaussian CDF, Φ^{-1} its inverse, and Y a sample from $p^* = \int_\Theta N(y|\theta, \sigma^2)\mathrm{d}G^*(\theta)$.

The idea of departing from the exact Bayesian update as specified by the model under consideration, in order to gain computational speed, is not a new idea. Smith and Makov (1978) first suggested to move away from the Bayesian paradigm when computations are infeasible. In their paper they proposed a "quasi-Bayes" solution to deal with a sequential classification problem and do inference on the weights in a finite mixture model. They showed how to adapt Bayesian inference through a series of heuristics and make it computationally tractable. Markov chain Monte Carlo methods have made the issue of computations in Bayesian statistics possible. Yet, nowadays, statistical models are often required to incorporate large amounts of data, making once more computability a relevant problem. The theme is particularly significant when dealing with nonparametric models. The flexibility gained using nonparametrics, which is highly desirable, often comes at a much higher computational cost than the parametric alternatives.

The algorithms of Newton et al. (1998) and Hahn et al. (2017) belong to a broader literature that sees the solution of computational problems in recursive algorithms. The reason is twofold. First, each iteration involves one data point and computational time complexity grows linearly with the dimensionality. This has two advantages: statistical quantities needed in the estimation - e.g. the likelihood - are computed inexpensively, having to deal with one data point at the time; also, by construction, it allows to deal with streaming data. Second, recursive algorithms often arise through a simplification of an existing estimator to make it computationally more manageable. A striking example is the well-known Stochastic Gradient Descent method, which substitutes the expensive inversion of $d \times d$ matrices in a Newton-Raphson algorithm with a deterministic sequence of weights (Robbins and Monro 1951, Sakrison 1965).

Whereas the use of stochastic recursions to calculate the maximum likelihood is common, and the theoretical and numerical properties of these algorithms have been thoroughly investigated - see stochastic gradient descent for example - only a few steps have been done in the "quasi-Bayes sphere". Mostly by Martin and Ghosh (2008), Tokdar et al. (2009), Martin and Tokdar (2009, 2011). Martin and Ghosh (2008) provide a solid theoretical justification to the Newton et al. (1998) algorithm by rigorously formalising the connection with the basic paradigm of stochastic approximation (Robbins and Monro 1951). Furthermore, they studied the large-sample behaviour of the algorithm and establish its consistency under the assumption that the observables are *i.i.d.* plus some technical conditions. They have also studied computational strategies to make it numerically stable and reduce the dependence on the order of the observations.

Another theoretical explanation that justifies the use of these algorithms is given by taking a fully Bayesian stance on the problem. If we were to consider (P_n) as a true sequence of predictive distributions we would be able to characterise, under some appropriate conditions, and for certain updates, the stochastic law of the observables. The (Y_n) sampled recursively from P_n such that $Y_{n+1}|Y_{1:n} \sim P_n$ would have the probability law defined by Berti et al. (2004): it is a form of dependence weaker than exchangeability at "small n" but asymptotically equivalent to it. This characterisation could hint that the class of models summarized by the (P_n): we are renouncing exchangeability only up until a sufficiently large n in order to obtain a sequence of nonparametric predictive distributions that is tractable, and that provides a fast and accurate estimate.

The relative lack of work that has been done on this type of recursive algorithms, despite their many advantages, constitutes an evident gap in the literature that we try to fill in this thesis. In particular, there are still many unanswered research questions that we try to address, for example: under which conditions is it possible to state a convergence argument for a procedure of the type (1.1)? Do algorithms "adapted" from Bayesian inference have common features? Which of these shared features allow for the asymptotic convergence? Current algorithms focus solely on density estimation; is it possible to develop algorithms for more complex models? By complex models we mean that the observables are not *i.i.d.* How do they numerically perform in the multivariate settings? Far from being able to find encompassing answers to all these questions, we tackle some of them, and discuss a few critical issues that make answers difficult in other cases.

A contribution in the thesis is to push these procedures beyond density estimation. To our knowledge, the recursion techniques available in the literature are rarely scaled up to the multivariate setting. Bayesian inference for mixture models has been a fertile starting point to develop new recursive algorithms. Mixture models are then the obvious starting point to work on new problems. The idea is simple. We consider existing Bayesian models and compute, if possible, the predictive density for the first step ($n = 1$). The first update must be a "good" one, so it is used as the building block of a new algorithm. Furthermore, Bayesian procedures inherently have a sequential structure that suggests how to update the prior. We use the Bayesian predictive update at $n = 1$ to "turn" the procedure into an iterative one. We will discuss in the following chapters a series of heuristics on the sequence $(\alpha_n)_{n\geq 1}$ and the functional form of the update Z.

We focus on two Bayesian nonparametric models. The first is the Linear Dependent Dirichlet Process Mixture of De Iorio et al. (2004). Under specific choice of the kernel

we are able to propose a new recursive algorithm for regression with a discrete regressor. The algorithm is an extension of Hahn et al. (2017). It arises from a related model but allows to estimate covariate-dependent distributions. Covariates are not modelled and are considered non stochastic design points. Then we consider two Dirichlet Process Mixtures with joint bivariate Dirichlet process prior on the mixing densities (Walker and Muliere 2003). We introduce a new procedure that allows us to estimate the marginal distributions of two dependent random variables. Here the most interesting aspect is that the recursion consists in a two-step algorithm. We will argue that this structure possibly paves the way for more complex procedures.

Another contribution is application of an existing algorithm to a multivariate inverse problem, namely the inversion of a multivariate Laplace transform. Inverse problems are increasingly often handled as a statistical inferential problem - see Stuart (2010). Two recent papers by Walker (2017a, 2017b) show that iterative procedures fit very well in this setting. Here, the data are abundant because they are numerically sampled from a known distribution. Walker (2017b) introduces a novel iterative procedure to solve a linear system of equations. Walker (2017a) shows that Newton's recursion can be used to invert the Laplace transform. The key intuition is that the density associated with the Laplace transform can be seen as a mixture model with gamma kernel. We extend the work to higher dimensions and try to address some numerical issues that arise if the dimension is higher than two, for example approximating a normalising constant that enters into the estimation. It is an important contribution because almost no work has been done to invert the Laplace transform of a multivariate distribution. The few existing methods might fail to recover a proper distribution function. Our methodology theoretical guarantees to estimate a CDF and scales up very well to an increase in dimension.

The connection to stochastic approximation - henceforth SA - gives an elegant theoretical backing to both Newton et al. (1998) and Hahn et al. (2017), but helps little in proving the convergence of these algorithms. The reason is that SA asymptotic theory for infinite-dimensional objects is still an open research area. In this thesis we discuss some ongoing work that deals with the convergence of a family of recursive algorithms and give some results in this area. We propose a novel technique to study the convergence of (P_n). Assuming that the observables are *i.i.d.*, we show that it is possible to connect the sequences of type (1.1) to fixed-point estimation if the update Z satisfies

$$\mathrm{E}_Y[Z(\cdot, P, Y)] = P(\cdot) \quad \forall(\cdot) \iff P = P^*, \tag{1.2}$$

where P^* is assumed to be the true data-generating distribution; i.e. $Y_n \overset{iid}{\sim} P^*$. SA algorithms are used in fixed point estimation, but the existence of a fixed point is not

directly exploited in the asymptotic theory. We show that if Z satisfies (1.2), it is possible to prove the convergence in the weak topology in this general setting. The key will be a novel martingale argument that applies to probability distribution functions over both finite and infinite dimensional spaces. The fixed point is used in the proof to identify the limit. The body of the asymptotic technique we propose in this thesis, and which is subject to ongoing work, is summarized by Theorems 5.2.2 and 5.2.3 in Chapter 5, as well as all the others theoretical contributions.

Our method has an immediate application: Newton et al. (1998) and Hahn et al. (2017) algorithms have an update that satisfies (1.2). Hence their asymptotic behaviour can be studied applying some of the ideas we propose. We study convergence properties in Chapter 5. Theorem 5.3.3 only establishes the weak convergence of the algorithm in Newton et al. (1998) under milder conditions than those in Tokdar et al. (2009). Some common statistical models, such as Gamma rate mixtures, were ruled by the assumptions in Tokdar et al. (2009): our result extends the range of models for which theoretical guarantees are available. Hahn et al. (2017) prove convergence in the L_1 topology but the (α_n) are constrained by harsher conditions than those discussed here. Theorem 5.4.2 only establishes the weak convergence under some given assumptions, but applies to a much larger family of (α_n). It is an extension since it is well known that the sequence (α_n) is crucial in the numerical performance of SA algorithms (Kushner and Yin 2003).

To our knowledge, this result could fill an evident gap in the SA asymptotic literature. Furthermore the application of Theorems 5.2.2 and 5.2.3 to these two algorithms highlight an important strength of our method: the set of assumptions can be verified in a broad set of statistical problems. For example, we check these assumptions for the algorithms we introduce in this thesis and study their theoretical convergence. This is an important advantage over the "standard" SA asymptotic theory which is based on a methodology linked to ordinary differential equations (ODE) stability (Kushner and Clark 1978, Ljung 1978). Even for finite-dimensional problem, applying the ODE method to a statistical estimator is not straightforward; see the proof of Theorem 3.4 in Martin and Ghosh (2008). This claim holds in statistical work. SA is applied in many fields, such as signal processing, queuing theory, where ODEs arise naturally.

1.1 Outline of the Thesis

The thesis is structured in a way that all the materials needed for the theoretical contributions discussed in the previous section is given upfront. Hence, background and

methodological contributions are presented first, then the asymptotic theory and a discussion.

Chapter 2 discusses the background material. It is not meant to be an encompassing review: we sketch only the key ideas that are needed across all the following chapters. In the first section we review the concept of mixture models and discuss Bayesian nonparametric inference for mixture models, introducing the reader to the notion of Dirichlet Process and Dirichlet Process Mixture, DPM. Both Newton et al. (1998) and Hahn et al. (2017) propose their algorithms starting from inference in DPMs, therefore it is important to stress some key features of this well-known model. Section 2.2 deals with the two recursive procedures mentioned in this introduction: the one due to Newton et al. (1998) and the one due to Hahn et al. (2017). Here we present the algorithms, highlight the critical issues, and discuss the connection with two major fields in mathematics that allow us to explain some of their theoretical properties: stochastic approximation and fixed-point theory. We also describe numerical issues and solutions linked to the implementation of these methodologies. Section 2.3 contains asymptotic results available for the algorithms discussed above.

Chapter 3 is dedicated to the two new algorithms we propose. The novelty here is observations that are not assumed to be independent identically distributed. The first algorithm, which we call Recursive Regression (RR) applies a regression with a covariate that can assume finitely many values (Section 3.1). The second one, which we call Two Steps Predictive Recursion (TSPR) to the joint nonparametric estimation of two distributions (Section 3.2). The chapter is divided in two major sections, each presenting one algorithm. Each section has an identical structure: we first explain how the algorithm is built and related to current theory, then we present several illustration to show that they perform well in a variety of numerical problems.

Chapter 4 contains a new methodology to invert the Laplace transform of a multivariate probability distribution. It is an extension of the univariate work of Walker (2017a). We first describe the current methodologies, then develop our procedure, which relies on Newton et al. (1998) algorithm, and deal with numerical issues that arise in the multivariate setting. Several illustrations are then discussed, as well as a comparison with the alternatives.

Chapter 5 collects all the asymptotic results of the thesis. We propose a novel martingale argument that allows to study large-sample properties of the family of recursive algorithms presented in the thesis. First, we introduce our argument to prove the convergence of a one-dimensional sequence of random variables (Section 5.1). Then we extend

the results of Section 5.1 to a sequence of probability distributions. Moving from a one dimensional to an infinite dimensional sequence is delicate, this is the subject of Section 5.2.1. In Section 5.3 we apply our new method to the algorithm introduced by by Newton et al. (1998). Section 5.4 is dedicate the study of the algorithms that have an update defined by a copula: the algorithm by Hahn et al. (2017) (Subsection 5.4.1), the Recursive Regression (Subsections 5.4.2) and the Two Steps Predictive Recursion (Subsection 5.4.3). We prove under a given set of assumptions that both algorithms converge weakly to the data-generating distribution. The results included in this chapter that concerns Newton et al. (1998) algorithm, have already been published in Cappello and Walker (2018); the rest is part of an ongoing research.

Each chapter includes a discussion specific to the topics touched. Chapter 6 provides a very brief discussion of the contributions of the thesis. The focus is in particular on some possible extensions and future works.

Chapter 2

Background

Chapter 2 contains a review of some of the key ideas needed to understand the material covered in the subsequent chapters. It is not a comprehensive review as we focus solely on the topics that are used across the whole thesis. The theory required only in specific contexts, for example, nonparametric regression, will be discussed in the appropriate chapter.

We discuss only nonparametric models. We first start by defining nonparametric mixture models and discussing Bayesian inference for mixtures. We then review the two recursive procedures mentioned in Chapter 1 and introduced in Newton et al. (1998) and Hahn et al. (2017). The emphasis is on their connection to a given Bayesian model and on their properties. These algorithms need to be evaluated numerically: we describe a few numerical strategies that are used to make the algorithms more stable numerically. We link these algorithms to stochastic approximation and fixed-point theory: these two research areas allow for a rigorous descriptions of some of their properties. The last section details the asymptotic theory for these two algorithms.

2.1 Mixture Models and Bayesian Inference

Mixture models are well-studied in statistics. They allow a description of a heterogeneous population and can be thus applied in a variety of settings, including density estimation, clustering and classification. The many books written on the topic certify the great interest, see for example Titterington et al. (1985), Lindsay (1995), Böhning and Seidel (2003), McLachlan and Peel (2004) and Frühwirth-Schnatter (2006).

Let $(\mathcal{Y}, \mathcal{B})$ and $(\Theta, \mathcal{A})$ be two measurable spaces; we call them respectively the sample and parameter space. In this thesis we assume all the spaces to be Euclidean. The observables take values in the sample space. The parameter $\theta \in \Theta$ indexes a parametric

family of densities on $(\mathcal{Y}, \mathcal{B})$ with respect to a dominating σ-finite measure ν on $(\mathcal{Y}, \mathcal{B})$. We denote this family by $\mathcal{K} := \{k(y|\theta) : \theta \in \Theta\}$. $k(y|\theta)$, an element of the family, is called *kernel* of the mixture. Denote by $\mathcal{F}$ the set of all possible probability densities on $(\mathcal{Y}, \mathcal{B})$ with respect to the dominating σ-finite measure ν. Clearly, $\mathcal{K} \subset \mathcal{F}$.

Any convex linear combination of elements of $\mathcal{K}$ is an element of $\mathcal{F}$. Such a convex combination does not generally belong to $\mathcal{K}$. A *nonparametric mixture density* is defined through an infinite linear combination of elements of $\mathcal{K}$. Let $\mathcal{G}$ be the class of probability measures on $(\Theta, \mathcal{A})$. For a fixed $G \in \mathcal{G}$, a *mixture density* is defined as

$$f(y|G) = \int_{\Theta} k(y|\theta) dG(\theta), \tag{2.1}$$

and G is called *mixing measure*. A *mixture model* is a class of mixture density of the form (2.1) indexed by G. We denote it by $\mathcal{M}$, $\mathcal{M} := \{f(y|G) : G \in \mathcal{G}\}$.

The popularity of mixture models is mostly due to their flexibility. The kernel determines whether $f(y|G)$ is continuous or discrete. For example, the Gaussian kernel gives an $a.s.$ continuous $f(y|G)$. A mixture models is also determined by the choice of the mixing parameters. For example a Gaussian kernel with $\theta = \mu$ leads to a *location mixture*, with $\theta = \sigma^2$ to a *scale mixture*, and with $\theta = (\mu, \sigma^2)$ to a *scale-location mixture*. If G admits the Radon-Nykodim derivative with respect to a dominating σ-finite measure μ on $(\Theta, \mathcal{A})$, $f(y|G) = \int_{\Theta} k(y|\theta) g(\theta) \mu(d\theta)$. Analogously, a discrete probability measure G such that $G = \sum_{n=1}^{\infty} w_i \delta_{\theta_i}$, (2.1) reduces to a countable mixture $f(y|G) = \sum_{i=1}^{\infty} w_i k(y|\theta_i)$ *almost surely (a.s.)*.

Doing statistical inference with mixtures, one is usually interested in an estimate of either $f(y|G)$ or G. The former has been extensively studied - see Böhning and Seidel (2003) for a review. It is the primary interest for applications such as density estimation where mixtures are still widely used. The latter gives information useful in classification problems, for example. Estimating G is a challenging problem, primarily because the observables take value in $\mathcal{Y}$ while G is a probability measure on $(\Theta, \mathcal{A})$, which is a latent space. Some of the methods available are nonparametric maximum likelihood (Laird 1978, Lindsay et al. 1983), deconvolution (Fan 1991) and the recursive algorithms proposed by Newton et al. (1998) (see Subsection 2.2.1).

A useful, probabilistically equivalent, representation of a mixture model is through a hierarchical model:

$$\theta_i \overset{iid}{\sim} G, \qquad Y_i|\theta_i \overset{iid}{\sim} k(\cdot|\theta_i), \qquad i \in \{1, ..., n\}.$$

Suppose we are interested in doing Bayesian inference on a mixture model. Then G would be a random probability measure, and a Bayesian statistician would need to define a prior distribution on it, which we denote by Π. The hierarchical representation is thus completed as:

$$G \sim \Pi, \qquad \theta_i|G \overset{iid}{\sim} G, \qquad Y_i|\theta_i \overset{iid}{\sim} k(\cdot|\theta_i), \qquad i \in \{1, ..., n\}. \tag{2.2}$$

Under the assumption that $(\Theta, \mathcal{A})$ is a Polish space, Ferguson (1973) introduced the *Dirichlet Process* (DP), one of the most popular Bayesian nonparametric prior. The Dirichlet Process is a distritution on probability measures. In other words, the realizations of the process are probability distributions. The definition of Ferguson (1973) is given in terms of the finite dimensional distributions of the measure over any finite partition of the space $(\Theta, \mathcal{A})$. A random probability measure G is said to be Dirichlet process distributed with parameters $c > 0$ and $G_0 \in \mathcal{G}$, if the random probability vector $(G(A_1), ..., G(A_k))$ is such that

$$(G(A_1), ..., G(A_k)) \sim Dir_k(c\, G_0(A_1), ..., c\, G_0(A_k)),$$

for any $k \in \mathbb{N}$ and any partition $(A_1, ..., A_k)$ of Θ. We write it as $G \sim \mathrm{DP}(c, G_0)$ and call c the *scale* parameter, and G_0 the *base measure*, and $\mathrm{E}[G] = G_0$. The Dirichlet Process has many properties desirable in a Bayesian prior: the parameters are interpretable and it is analytically tractable being conjugate. It has full weak support, with respect to all distributions absolutely continuous with respect to G_0 (Ferguson 1973).

Two alternative characterisations make further attractive properties explicit. The first one is given by Blackwell and MacQueen (1973). They give a characterization of the process through its predictive distributions. The characterization of Blackwell and MacQueen (1973) is said a *Pólya-urn sequence*, being a continuous generalization of the famous Pólya urn scheme. Given $\theta_i|G \overset{iid}{\sim} G$ with $i \in \{1, ..., n\}$, $G \sim \mathrm{DP}(c, G_0)$, the predictive distribution of $\theta_{n+1}|\theta_{1:n}$ is $P(\theta_{n+1}|\theta_{1:n}) := \mathrm{E}[G(\theta_{n+1})|\theta_{1:n}]$, and is given by

$$\theta_{n+1}|\theta_{1:n} \sim \frac{c}{c+n} G_0 + \frac{1}{c+n} \sum_{i=1}^{n} \delta_{\theta_i}. \tag{2.3}$$

Blackwell and MacQueen (1973) prove that a sequence (θ_n) having predictive distributions defined by (2.3) is exchangeable with a Dirichlet Process $\mathrm{DP}(c, G_0)$ de Finetti directing measure. The predictive distribution $P(\theta_{n+1}|\theta_{1:n})$ is a convex linear combination of the base measure and the empirical distribution function, which offers a natural interpretation of the DP parameters. The base measure G_0 is equal to $P_0(\cdot) = \mathrm{E}[G(\cdot)]$, the predictive distribution when no observations are available. The concentration parameter can be interpreted as a precision parameter: the larger c is, the more a DP prior is concentrated

around its mean, G_0. The predictive distribution in (2.3) is useful for computations and exploited in several Markov Chain Monte Carlo schemes. See Hjort et al. (2010).

Another important feature of the process is that it samples discrete distributions with probability one. The proof is given in Ferguson (1973) and Blackwell and MacQueen (1973). Sethuraman (1994) highlights the discreteness of the Dirichlet process through the so-called *stick-breaking representation*. If $G \sim \mathrm{DP}(c, G_0)$, then, *a.s.*,

$$G = \sum_{i=1}^{\infty} w_i \delta_{\theta_i},$$

where the weights are such that

$$w_i = v_i \prod_{j<i} (1 - v_j), \qquad \forall ji > 1,$$

with $w_1 = v_1$ and $v_i \overset{iid}{\sim} Beta(1, c)$, and the atoms are θ_i *i.i.d.* according to G_0, indipendently of the (v_i).

The stick-breaking construction is useful in applications. Several sampling schemes for the DP have been proposed exploiting this construction; for example the truncation of the series in Muliere and Tardella (1998). Nonparametric prior processes that generalize the Dirichlet Process can be also constructed through a stick-breaking construction (Ishwaran and James 2001). Prior processes beyond the Dirichlet process are given in Pitman and Yor (1997) and in De Blasi et al. (2015).

If we assume that the the prior on the mixing measure G in (2.1) is a DP, the corresponding mixture model is said to be a *Dirichlet Process Mixture Model (DPM)*. DPMs were first studied by Ferguson (1983) and Lo (1984) as a way to define a prior on the space of continuous distributions. The density $f(y|G)$ in (2.1) can be rewritten using the stick-breaking construction. A DPM density can be written *a.s.* as

$$f(y|G) = \sum_{i=1}^{\infty} w_i k(y|\theta_i), \tag{2.4}$$

where (θ_i) and (w_i) are defined above.

Bayesian inference in DPMs is not analytically tractable: the posterior distributions cannot be written in closed form. Several MCMC schemes to compute posterior quantities are available in the literature. The vast majority involves auxiliary latent variables.

Among the most popular schemes for DPMs: the Gibbs sampler of Escobar (1994), Escobar and West (1995); the one of Neal (2000); the slice-sampler (Neal 2003, Walker 2007, Kalli et al. 2011).

On the other hand, the first predictive mixing density ($n = 1$) can be written in closed-form. We assume that G_0, the DP base measure, is absolutely continuous with respect to the dominating σ-finite measure ν on $(\Theta, \mathcal{A})$ with density g_0. Given an observation y_1, the predictive density of $\theta_2|y_1$ can be written as

$$g_1(\theta) = \frac{c}{c+1} g_0(\theta) + \frac{1}{c+1} \frac{k(y_1|\theta) g_0(\theta)}{\int_\Theta k(y_1|\theta) g_0(\theta) d\theta}. \tag{2.5}$$

Trivially, (2.5) follows from the Blackwell and MacQueen (1973) construction. The subscript in g_1 denotes the number observations we have conditioned upon. The expression (2.5) is crucial in the development of the recursive algorithms described next.

2.2 Recursive Algorithms

The Pólya urn sequence is a basic example of the inherent sequential structure of a Bayesian predictive update. In this section we present two recursive algorithms that depart from a proper Bayesian inference in order to gain computational tractability. These procedures deal with cases that are not analytically tractable, trying to exploit this inherent sequential structure. Both algorithms discussed in this section have been proposed for DPMs.

The rationale of what follows is incredibly simple. A predictive distribution at $n = 1$, such as (2.5), for example, is determined by the underlying statistical model; hence it is reasonable to assume that it meaningfully describes the way the observations are incorporated in the prior. The idea is to make use of what is believed to be a good update and modify it to ensure convergence: the update must become less relevant as the sample size increases. This simple principle has lead to at least two procedures that are fast and numerically accurate. Under the assumptions that the data are *i.i.d.* from a given distribution, both algorithms converge to the sampling distribution.

In this section we first present the two algorithms proposed by Newton et al. (1998) and Hahn et al. (2017). The estimates obtained through these algorithms need to be calculated numerically on a grid of points: in Subsection 2.2.3 we describe how to implement the schemes effectively. We then link these recursive schemes to stochastic approximation and fixed-point theory: these two subjects offer a rigorous theoretical backing to the

algorithms and clarify some of their numerical and asymptotic properties.

2.2.1 Recursive Algorithm of Newton et al. (1998)

Newton et al. (1998) introduce a recursive algorithm that gives a fast nonparametric estimate of the mixing density g (or the distribution G) of a density of the type (2.1). The goal is to approximate a predictive distribution G_n, or its corresponding density g_n, of a DPM once a random sample $(y_1, \ldots, y_N)$ is collected. Other relevant references are Newton and Zhang (1999), Quintana and Newton (2000) and Newton (2002). The algorithm is referred by the authors as predictive recursion, we will refer to it as the Newton-Quintana-Zhang (NQZ) algorithm.

Newton-Quintana-Zhang (NQZ) algorithm. (Newton et al. 1998). Fix a prior guess density $g_0(\theta)$ and a user-supplied weight sequence (α_n), with $\alpha_n \in (0, 1)$. Given a random sample $(y_1, \ldots, y_N)$ from a mixture density of the form (2.1), an estimate $g_N(\theta)$ can be obtained by computing recursively

$$g_n(\theta) = (1 - \alpha_i) g_{n-1}(\theta) + \alpha_n \frac{k(y_n|\theta) g_{n-1}(\theta)}{\int_\Theta k(y_n|\theta) g_{n-1}(\theta) \mathrm{d}\theta}, \tag{2.6}$$

for $n = 1, ..., N$, where N is the number of observations available for the estimate.

A few remarks. First of all, the resemblance with the DPM predictive update defined in (2.5) is striking. Indeed, the NQZ algorithm is an exact Bayesian procedure for a DPM $n = 1$ if one assigns to G a $\mathrm{DP}(c, G_0)$ prior, with $c = (1 - \alpha_1)/\alpha_1$ and with G_0 the CDF corresponding to g_0, as base measure. There is one scenario in which the NQZ estimate and the Bayesian one are the same. If a Bayesian statistician had to delete all her statistical analysis after each observation y_n and keep a single quantity, she would reasonably keep the predictive density $g_n(\theta)$. Then once a new observation y_{n+1} is collected, she could update her prior belief assigning to G a $\mathrm{DP}(c_n, G_n)$ prior, with $c_n = (1 - \alpha_{n+1})/\alpha_{n+1}$ and G_n as base measure. In this scenario, the NQZ estimate would be equal to the Bayesian one.

However, the departure from the Bayesian inference as described by a DP mixture model is clear. We have highlighted that the Bayesian nonparametric estimate given by a DPM, and g_n as defined by (2.6), are equal only in the scenario we have just described. Furthermore, g_n cannot be a posterior quantity of a model that characterises exchangeable observations because it is order dependent, thus it is not function of the order statistics (Tokdar et al. 2009). The exchangeability of the observables is lost: Martin and Ghosh

(2008) and Martin (2009) show some numerical illustrations in which the NQZ algorithm leads to a poor approximation of posterior distribution of a DPM. Nonetheless, they show several numerical examples displaying the accuracy of the algorithm to validate its use as an estimator in this type of deconvolution problems.

Lastly, the sequence (α_n) plays a crucial role. Newton (2002) underlines that it is a user supplied sequence that has no particular interpretation in terms of the model. It tunes the speed with which observations are incorporated. It is also extremely important to prove the consistency of the estimate: some conditions on (α_n) will be detailed in the next section.

A lot of work has been done on this procedure. Newton and Zhang (1999) show an application to the problem of missing data. Martin (2009) studies several ways to make the algorithm more numerically stable. Martin and Tokdar (2011) employ it in a semi-parametric model. Walker (2017a) uses it for the inversion of the Laplace transform of a univariate distribution. In Chapter 4, we extend this work to a multidimensional setting (Cappello and Walker 2018).

2.2.2 Recursive Algorithm of Hahn et al. (2017)

The reference Hahn et al. (2017) contains two distinct contributions, both of them much relevant in the following chapters. The first result is a novel characterisation of a Bayesian predictive update, the second is a recursive algorithm to estimate a density $f(y|G)$ of the form (2.1).

Let (p_n) be a sequence of predictive densities such that $Y_n|y_{1:n-1} \sim p_{n-1}$. The first result is a clever representation of the ratio of two consecutive predictive densities, say p_{n-1} and p_n, in terms of a copula density. Such a representation allows to compute p_n directly through a multiplicative update without having to compute the posterior first. Note that no assumption on the law of the observables is necessary: (2.7) is a direct consequence of basic probability rules, the definition of predictive density and copula theory. We do not want to enter into details of copula theory - Nelsen (2007) is the standard reference on the topic. Let us solely remind that, by Sklar's theorem, any multivariate CDF $F = F(y_1, \dots, y_d)$, with marginal $F_1, \dots, F_d$, can be written in terms of a copula distribution:

$$F(y_1, \dots, y_d) = C(F_1(y_1), \dots, F_d(y_d))$$

and if F has density f, we can define a corresponding copula density c:

$$f(y_1, \ldots, y_d) = c(F_1(y_1), \ldots, F_d(y_d)) f_1(y_1) \times \ldots \times f_d(x_d)$$

Consider p_{n-1}, the predictive density of Y_n given $(y_1, \ldots, y_{n-1})$. Once the observation y_n is collected, it is possible to calculate $p_n(y)$ through the following relationship

$$p_n(y) = p_{n-1}(y) \;\; c_n(P_{n-1}(y), P_{n-1}(y_n)), \tag{2.7}$$

where c_n is a copula density function. In (2.7) we are using the dependence between Y_n and Y_{n+1} conditionally on $(y_1, ..., y_{n-1})$ that follows from a given Bayesian model and the corresponding law of the observables. Hahn et al. (2017) discuss a Bayesian model in which the observables are conditionally *i.i.d.*, the extension to any law of the observables follows directly. That is all we need to know.

A few remarks on c_n. Its arguments depend on the sample but c_n does not: it is a function only of the sample size. If we assume (Y_n) to be exchangeable, de Finetti representation theorem ensures the convergence of the corresponding sequence of predictive distributions $\mathbb{P}$-*a.s.*, where $\mathbb{P}$ is the exchangeable law governing (Y_n): we must expect that the update in (2.7) becomes irrelevant for large n. Hence, it must hold that $c_n(\cdot, \cdot) \to 1$ $\mathbb{P}$-*a.s.* as $n \to \infty$. In other words c_n must converge to the independence copula density, which is defined as $c(u, v) = 1 \;\; \forall u, v \in [0, 1]$.

A closed form for c_n does not always exist. As a rule of thumb, it does exist for those models that have the predictive distributions available in closed form. Note that this claim is a conjecture: no theory supports this statement. For example: let $Y_n|\mu \overset{iid}{\sim} N(\cdot|\mu, \sigma^2)$, with $\mu \sim N(\cdot|0, \tau^{-1})$ and $\sigma^2 > 0$ known: the predictive densities of this model imply a sequence of Gaussian copula densities. Recall that a Gaussian copula density is defined as:

$$c_\rho(u, v) = (1 - \rho^2)^{-1/2} \exp\left(\frac{2\rho\Phi^{-1}(u)\Phi^{-1}(v) - \rho^2\Phi^{-1}(u)^2 - \rho^2\Phi^{-1}(v)^2}{2(1 - \rho^2)} \right),$$

with $u, v \in [0, 1]$. The Gaussian copula density has a single parameter, ρ, whose interpretation corresponds to the correlation parameter of a bivariate Gaussian distribution. A sequence of correlation parameters identifies a sequence of Gaussian copula densities. In the example we are considering $\rho_n = (\tau + n)^{-1}$. Note that since $\rho_n \to 0$ as $n \to \infty$, $c_{\rho_n}(\cdot, \cdot) \to 1$ as $n \to \infty$.

In the spirit of Newton et al. (1998), Hahn et al. (2017) consider Bayesian inference in DPMs. The predictive density $p_1(y) = \mathrm{E}[f_1(y|G)]$ is available in closed form. The ratio $p_1(y)/p_0(y)$ must be a copula because of (2.7). The authors show that if $f(y|G)$ is a

Gaussian location mixture with a DP($c\,G_0$) prior on the mixing distribution, the copula density is a mixture of the independence copula density and the Gaussian copula density, which we denote by c_ρ. Therefore

$$p_1(y) = (1-\beta)\,p_0(y) + \beta p_0(y)\,c_\rho(P_0(y), P_0(y_1)),$$

where β is some function of the stick-breaking weights of the DP and $p_0(y) = \int_\Theta k(y|\theta,\sigma^2)\mathrm{d}G_0(\theta)$, with $\sigma^2 > 0$ fixed. If one considers a different mixture model, she will get a different copula. For example, a Gaussian scale-location mixture leads to a Student-t copula density. In a spirit similar to the NQZ algorithm, they propose the following algorithm to estimate the mixture density.

Hahn-Martin-Walker (HMW) Algorithm (Hahn et al. 2017) Fix a prior guess density $p_0(y)$ and a user-supplied weight sequence (α_n), with $\alpha_n \in (0,1)$. Given an *i.i.d.* sample $(y_1,\ldots,y_N)$, an estimate of the sampling distribution p_N is obtained computing recursively

$$p_n(y) = (1-\alpha_i)\,p_{n-1}(y) + \alpha_i\,p_{n-1}(y)\,c_\rho(P_{n-1}(y), P_{n-1}(y_n)), \tag{2.8}$$

for $n = 1, ..., N$; where N is the number of observations available for the estimate, and c_ρ denotes a Gaussian copula with fixed parameter $\rho \in (0,1)$.

The HMW algorithm shares with NQZ most of the features. The predictive densities (p_n) are not the ones obtained through a DPM and the estimates are order dependent. A strength of the algorithm is that if the objective of interest is the sampling density, the HMW algorithm is much faster than NQZ as we do not need to evaluate numerically two integrals. When the dimension of the observables is higher than one, evaluating the normalising constant in (2.6) is problematic. The HMW algorithm offers an elegant way to avoid numerical integration.

HMW is proposed having in mind a more restrictive class of densities: DP-location mixtures of Gaussian densities with fixed σ^2, whereas NQZ covers a broader spectrum of mixture models. The authors try to address the issue in two ways: first, they show through numerical simulations that the algorithm successfully recovers probability distributions that are not Gaussian location mixtures; second, they hint that similar algorithms can be built for other mixture models, for example DP-scale location Gaussian mixtures.

The user-supplied sequence of weights is again deterministic. In this case (α_n) has a

clearer interpretation. Indeed, from (2.8) it is clear that

$$c_n(P_{n-1}(y), P_{n-1}(y_i)) = (1 - \alpha_n) + \alpha_n \, c_\rho(P_{n-1}(y), P_{n-1}(y_n))].,$$

which is a *copula mixture density*. The weights (α_n) determines the linear combination between the independence and the Gaussian copula density. The sequence (α_n) can be tuned to achieve convergence because we must impose that the copula mixture converges to an independence copula density. We come back to this point in Section 2.3 when discussing the asymptotic theory.

2.2.3 Algorithms Implementation

Both NWZ and HMW must be calculated on a grid of points properly defined. It is not possible to evaluate these algorithms only at an arbitrary point of interest. We explain below the reasons.

NQZ algorithm: Suppose one wishes to evaluate $g_n(\hat{\theta})$ with $\hat{\theta} \in \Theta$, where g_n is the mixing density estimate given by the NQZ algorithm. At each iteration n, for $n = 1, \ldots, N$, the scheme (2.6) requires to compute the normalizing constant $\mathbb{I}_n := \int_\Theta k(y_n|\theta) g_{n-1}(\theta) \mathrm{d}\theta$. However $\mathbb{I}_n$ needs to be approximated numerically, for example through a quadrature rule. It is therefore necessary to be able to evaluate g_n at enough points to ensure that $\mathbb{I}_n$ is well approximated. The grid of points needs to satisfy two requirements: first, it needs to cover the support of the target density; second, it needs to be fine enough to prevent that a large interpolation error arises. It is safe to argue that the choice of the grid does not constitute a major limitation of the procedure: the many numerical illustrations in the literature displaying the accuracy of NQZ's estimates support this claim. See references in Subsection 2.2.1. The accuracy of quadrature rules deteriorates at dimension $d \geq 2$. We tackle this issue in Chapter 5.

HMW algorithm: One of the main advantages of HMW is that no integrals need to be evaluated if the goal is to estimate the mixture density. There are still two numerical issues we need to account for. First, (2.8) defines a density but the arguments of the copula are CDFs. A solution is to consider HMW at the distribution scale. Numerical differentiation can be used if one is then interested in a density estimate. Second, we still require a grid: at step n, P_{n-1} needs to be evaluated at y_n, i.e. $P_{n-1}(y_n)$. Hence, it is necessary to know P_{n-1} at points other than the one of interest. A simple strategy is to use the grid $[y_{(1)}, \ldots, y_{(N)}]$ defined by the order statistics: this choice ensures that the CDF is known at all points that are used in HMW.

Both NQZ and HMW are dependent on the order of observations. Tokdar et al. (2009) study a Rao-Blackwellization of the Newton et al. (1998) algorithm. It consists in a permutation averaged estimate. Given a recursive estimate P_N, it can be defined as

$$\widehat{P_N}(\cdot) = \mathrm{E}[P_N(\cdot)|y_{(1)}, \ldots, y_{(n)}],$$

making $\widehat{P_N}(\cdot)$ function of the order statistics. For large N, computing $\widehat{P_N}(\cdot)$ is computationally unfeasible; in practice, one resorts to a Monte Carlo estimates. First, one randomly permutes the observations, say k times, and gets k sets with the same elements but in a different order. Then she computes k estimates: the algorithm is applied once for each set of observations. The last step consists in averaging the k estimates. Tokdar et al. (2009) show that the asymptotic properties remain unchanged and the algorithm benefits in terms of stability. For small sample, the order dependence could be detected because the estimate is not smooth and "bumpy". The Monte-Carlo estimates of $\widehat{P_N}$ helps overcoming this problem. Monte Carlo averaging is a staple of many iteration procedures (Kushner and Yin 2003).

The decay rate of (α_n) is constrained by the theoretical boundaries that are necessary to prove the convergence of the algorithm. The sequence (α_n) has also a relevant numerical effect which should not be overlooked. Consider for example the sequences defined by $\alpha_n^* = c_1 (1+n)^{-p}$ and $\alpha_n^{\#} = c_2 (1+n)^{-p}$ with $c_1 >> c_2$ and $p > 0$. Now, (α_n^*) and $(\alpha_n^{\#})$ have the same convergence rate but lead to two estimates that can differ substantially. One should keep in mind that a quick decay to zero gives a smooth estimate because the update is discounted by a smaller weight more quickly, but makes it more difficult to incorporate information: the order dependence is more extreme. If we consider two sequences with different rates, this effect is likely to be even more evident. Therefore flexibility in the choice of the sequence (α_n) is important.

There are several ways to aid the algorithms achieving faster convergence. A simple but effective way is *random restart* (Fahlman 1988): the data are divided into two sets, the first being used to "train" the algorithm, which is then restarted using observations of the second set. The restart loosens the dependence on the initial value. See also Ghannadian et al. (1996) and Hu et al. (1997). Another very effective technique is *Polyak averaging* (Polyak and Juditsky 1992): given the recursive sequence (ξ_n), an auxiliary sequence is considered, $\hat{\xi_n} = \sum_{i=1}^{n} \xi_i$. Then $\hat{\xi_n}$ converges both numerically and theoretically faster than ξ_n in several problems (Kushner and Yin 2003). Stochastic gradient descent is commonly used this way.

There are many other effective strategies. Chapter 11 of Kushner and Yin (2003)

provides an excellent source.

2.2.4 Stochastic Approximation

Robbins and Monro (1951) deal with the recursive estimation of the root of an unknown function which cannot be evaluated directly but only estimated via noisy observations at unknown locations. The algorithm was called by the authors stochastic approximation (SA). Here we just sketch the original work in order to introduce the paper by Martin and Ghosh (2008) who study the link between that the NQZ algorithm and SA. The reader will easily see that the same line of reasoning applies to HMW. For a deeper introduction to SA, we refer to Chapter 1 of Kushner and Yin (2003) and Lai (2003).

Suppose one wants to estimate the value $\hat{\xi}$ such that $h(\hat{\xi}) = 0$. If one knew $h(\cdot)$, iterative algorithms would be available, for example Newton-Raphson. Robbins and Monro (1951) suppose that h is unknown and build a sequential algorithm that converges to $\hat{\xi}$. For each input ξ, we suppose to be observing $x = h(\xi) + \epsilon$ in lieu of $h(\xi)$, where ϵ is an error term. The error term arises because h is unknown and it is observed at unknown locations: hence the authors pose the problem as one of a function observed up to an error. The algorithm goes as follows: fix an initial ξ_0 and a sequence (α_n), assume that $x_n = h(\xi_{n-1}) + \epsilon_n$ for all $n \geq 1$. A sequence (ξ_n) can be calculated for all $n \geq 1$ through

$$\xi_n = \xi_{n-1} + \alpha_n x_n. \tag{2.9}$$

Under appropriate assumptions on the behaviour of $h(\cdot)$ - boundedness, continuity and monotonicity - Robbins and Monro shows that under the following conditions on (α_n),

$$\alpha_n > 0, \qquad \sum_{n=1}^{\infty} \alpha_n = \infty, \qquad \sum_{n=1}^{\infty} \alpha_n^2 < \infty, \tag{2.10}$$

it is possible to establish the convergence of ξ_n to $\hat{\xi}$ as $n \to \infty$.

Starting from this initial work the number of applications has grown and SA has become an important research area in applied mathematics, informatics, signal processing and statistics. The assumptions on h used by Robbins and Monro (1951) have been weakened, allowing for a more complex dependence between ξ_n and x_n, and as a consequence in most of current applications h is not fixed but it is defined through a sequence (h_n) that converges to h. The study of the convergence of the SA method has largely focused on finite-dimensional sequences, for example $\xi \in \mathbb{R}^d$ with d finite. The classical SA asymptotic theory relies on the *martingale difference noise* assumption: there exists

a function h_n such that

$$\mathrm{E}[\xi_n|\xi_{1:n-1}, \xi_0] = h_n(\xi_{n-1}),$$

i.e. $x_n = h_n(\xi_{n-1}) + \epsilon_n$ and the (ϵ_n) is a martingale difference sequence. Kushner and Yin (2003) explain that this model is still relevant in the literature and applies to all the problems where $x_n = F(\xi_n, \psi_n)$, with (ψ_n) mutually independent for some given function F. The convergence theory under this assumption can be dealt using the theory of martingales, in particular martingale inequalities; for example; Fabian (1960) and Gladyshev (1965) use semi-martingale, Robbins and Siegmund (1971) introduce the notion of *almost supermartingale* (see next section).

However, the main tool currently used for SA convergence theory is based on the *ordinary differential equation (ODE) method* developed in Kushner and Clark (1978) and Ljung (1978). In this framework the limit point, if it exists, is a *global asymptotically stable point* of the ODE $\dot{x} = h(x)$. To understand the connection between SA and ODE is worth considering the assumption that the error terms sequence (ϵ_n) is a martingale difference. Under this assumption, it follows that for large n the term $\alpha_n\epsilon_n$, which is implicit in (2.9), is negligible and $h_n \approx h$, which implies that (2.9) is equivalent to

$$\frac{\xi_n - \xi_{n-1}}{\alpha_n} \approx h(\xi_{n-1}). \tag{2.11}$$

Trivially (2.11) is asymptotically equivalent to an ODE. This explains the reason why the asymptotic behaviour of SA algorithms is studied using ODE stability theory (Kushner and Yin 2003). The ODE method is applied also to recursive sequences that violate the zero martingale assumption but such a setting is beyond our current scope; see Kushner and Yin (2003) and Lai (2003).

Newton et al. (1998) and Newton (2002) recognised the connection to SA but do not rigorously investigate it. Martin and Ghosh (2008) formalise this interesting theoretical backing. Recall that the NQZ algorithm gives a scheme to compute a sequence of mixing densities (g_n) on $(\Theta, \mathcal{A})$. One can see that the update of NQZ fits into the martingale difference noise framework. Indeed, rewriting everything with the notation used in Subsection 2.2.1, we have that $x_n(\theta) = z(\theta, g_{n-1}, y_n)$ where $z(\cdot, \cdot, \cdot)$ is defined as

$$z(\theta, g_{n-1}, y_n) = \frac{k(y_n|\theta)g_{n-1}(\theta)}{\int_\Theta k(y_n|\theta)g_{n-1}(\theta)d\theta} - g_{n-1}(\theta).$$

The observations $(y_1, ..., y_n)$ are the stochastic components of the scheme, which are required to be sample independently. An analytical expression of the sequence (ϵ_n) is not available, but it is not necessary. Under the assumption that a true mixing measure G^* ex-

ists and $Y_n \overset{iid}{\sim} f(y|G^*)$, with $f(y|G^*)$ defined in (2.1), and the cardinality of Θ is a finite d, i.e. $|\Theta| = d$, Martin and Ghosh (2008) show that G^* is the asymptotically stable ODE solution associated with the scheme NQZ and convergence can be proven using ODE theory.

Unfortunately, it is important to further stress that the use of ODE theory allows us to prove the convergence of NQZ only if θ takes a finite number of values, i.e. $|\Theta| = d$. To our knowledge, SA convergence theory for infinite dimensional functions is still not well studied in general. There are convergence results for sequences taking values in Hilbert spaces both for linearly bounded updates (Révész 1973, Salov 1980, Kushner and Shwartz 1985), nonlinear ones (Yin and Zhu 1990, Seidler et al. 2017), and finite-dimensional approximations (Yin 1992). An interesting paper by Dieuleveut et al. (2016) covers sequences lying in reproducing kernel Hilbert space (*RKHS*): despite the existence of a class of methods to represent a probability distributions as an element of a *RKHS* (Smola et al. 2007), that reference cannot be easily applied to probability distribution functions. The authors cover a least squares optimization design and their results seem to be tied to it.

2.2.5 Fixed Point Iterations

Fixed-point theory has been connected to both NQZ (Martin and Ghosh 2008, Walker 2017a) and Chapters 4 and 5 (published in Cappello and Walker 2018); and HMW (Hahn et al. 2017) algorithms. These papers show that the updates of the algorithm are related to operators that admit a fixed-point. Only Walker (2017a) exploits such a representation to prove asymptotic convergence but the proof is not straightforward and a few points need to be cleared out.

Let us consider a metric space (X, m) and an operator $T : X \to X$. A *fixed point* of T is a point $x^* \in X$ such that $T(x^*) = x^*$. The theory of fixed-points is a well studied area in mathematics - see Granas and Dugundji (2013) for an introduction. Estimation of the fixed-point of an operator is often done through an iterative algorithm. The well known Banach fixed-point theorem states that if T is a *contraction (or nonexpansive) operator* (i.e. $m(T(x), T(y)) \leq m(x, y)$, $\forall x, y \in X$), then T admits one unique fixed-point x^*, and the fixed-point is the limit of a sequence defined by an algorithm that is initialized at an arbitrary $x_0 \in X$ and that repeats iteratively $x_n = T(x_{n-1})$. Several iterations that relax the assumption on T have then appeared. Interesting for our purposes is the research on nonexpansive operators and bounded sequences $(x_n)_{n\geq 1}$ originated from the work of Mann (1953), Krasnosel'skii (1955) and Ishikawa (1974).

All the procedures above are deterministic. Starting from Bharucha-Reid et al. (1976),

the theory has been extended to random fixed-point operators. We limit our discussion to sketch of the basic heuristics of the deterministic argument. The Mann (1953) iterative scheme is often written as

$$x_n = (1 - \alpha_n)x_{n-1} + \alpha_n T x_{n-1}, \tag{2.12}$$

with T a nonexpansive mapping. The convergence of Mann-type iterations can be proven under conditions analogous to (2.10); see for example Ishikawa (1974). However, (2.10) is not necessary in this setting: Chidume (1981) proves a convergence theorem that does not require $\sum_n \alpha_n = \infty$ (Theorem 1 in Chidume, 1981).

Mann (1953) original formulation is not exactly (2.12) but involves a modified iteration process. He introduces an infinite triangular matrix A

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ a_{2,1} & a_{2,2} & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n,1} & a_{n,2} & \cdot & a_{n,n} & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \tag{2.13}$$

with conditions

$$a_{i,j} \geq 0, \qquad \forall i, j; \qquad a_{i,j} = 0, \qquad \forall j > i; \qquad \sum_{j=1}^{i} a_{i,j} = 1, \qquad \forall i. \tag{2.14}$$

Given a starting value x_0, the fixed point of T is the limit of a sequence defined by $x_n = T(v_n)$ with $v_n = \sum_{k=1}^{n} a_{n,k} x_k$. Two remarks: the conditions on the weights (2.14) is more general than (2.10). The use of an infinite matrix A and a modified iteration process is a tractable alternative representation of the algorithm (2.12). It applies to both NQZ and HMW which can be rewritten using a stick-breaking construction.

2.3 Asymptotic Theory for Recursive Algorithms

Stochastic approximation and fixed-point theory give a rigorous theoretical backing to both NQZ and HMW. However they do not give viable tools to directly prove asymptotic convergence when either $\mathcal{Y}$ or Θ, or both of them, are not finite.

We now direct our attention to the asymptotic arguments given for these two algorithms. Most of the papers on NQZ are collected in Ryan Martin's PhD Thesis (2009).

Hahn et al. (2017) study the convergence of HMW.

2.3.1 Convergence Results for the NQZ algorithm

The proof of convergence in the weak topology of the predictive recursion is due to the work of Martin, Ghosh and Tokdar. The key reference on the topic is Tokdar et al. (2009). Martin and Tokdar (2009) and Martin and Tokdar (2011) include several important additions. The convergence of NQZ had been long standing, other important references are Ghosh and Tokdar (2006) and the paper by Martin and Ghosh (2008), which we have already discussed.

The assumption is that the observations are sampled *i.i.d.* from a mixture distribution $F^*(\cdot|G^*)$ corresponding to a density $f^*(\cdot|G^*)$ defined in (2.1). Tokdar et al. (2009) prove that the sequence of CDFs (G_n) calculated through NQZ converges weakly, the sequence of mixture densities $(f_n(y|G_n))$, defined in (2.1), converges with respect to the Kullback-Leibler divergence. The proof exploits a martingale representation of the Kullback-Leibler (KL) divergence between g_n and g^*, the true mixing density. Recall that the KL-divergence between two densities p and g is defined as

$$KL(p, g) = \int \log\left(\frac{p(y)}{g(y)}\right) p(y)\mathrm{d}y,$$

which is an asymmetric function depending on the ordering of the arguments.
Tokdar et al. (2009) consider the following assumptions:

TMG1. $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$
TMG2. The map $G \to \int k(y|\theta)\mathrm{d}G(\theta)$ is injective, making the mixing distribution G identifiable from the mixture.
TMG3. $\forall y \in \mathcal{Y}$, the map $\theta \to k(y|\theta)$ is bounded and continuous.
TMG4. $\forall \epsilon > 0$ and any compact $\mathcal{Y}_0 \subset \mathcal{Y}$, there exists a compact $\Theta_0 \subset \Theta$ such that $\int_{\mathcal{Y}_0} k(y|\theta)\mathrm{d}y < \epsilon$ for all $\theta \in \Theta_0$.
TMG5. There exists a constant $B < \infty$ such that $\forall \theta_1, \theta_2, \theta_3 \in \Theta$

$$\int_{\mathcal{Y}} \left(\frac{k(y|\theta_1)}{k(y|\theta_2)}\right)^2 k(y|\theta_3)\mathrm{d}y < B.$$

A few comments. *TMG1* is the same as (2.10). *TMG2* is satisfied by some important mixture classes (Lindsay 1995) - e.g Gaussian kernel with mean θ and known variance $\sigma^2 > 0$. Yet it rules out important cases such as Gaussian scale-location mixtures. However, it is a standard requirement to prove the convergence of mixtures; see for example Barron et al. (1999) for DPM posterior consistency. *TMG3* and *TMG4* are

also commonly required to prove mixtures consistency. *TMG3* demands the kernel to be well-behaved, for example neither degenerated nor unbounded. *TMG4* sets a condition on the support of the true distribution: the probability is not concentrated in the tails of the distribution. *TMG5* is more critical: it requires Θ to be a compact set. Whereas this is not a concern for several applications, this assumption rules out several popular kernels - e.g. Gaussian kernel with mean θ and known variance $\sigma^2 > 0$. *TMG5* is often employed in the literature for mixture models because it allows us to prove some desirable properties. For example Bruni and Koch (1985) show that scale-location mixtures of Gaussian distributions are identifiable under the constraint that Θ is compact. Such a constraint is not negligible though.

We elaborate on the KL divergence sequence defined in this reference. Let $K_n := KL(g^*, g_n)$. It can be shown that

$$K_n - K_0 = -\sum_{i=1}^{n} \int_{\Theta} \log \left[1 + \alpha_i \left(\frac{k(Y_i|\Theta)}{p_{n-1}(Y_i)} - 1 \right) \right] g^*(\theta) \mathrm{d}\theta. \tag{2.15}$$

The key is to use the following Taylor expansion of the logarithm: $\log(1+x) = x - x^2 R(x)$, where $R(x)$ denotes the remainder of the Taylor expansion and satisfies the inequality $0 \leq R(x) \leq \max\{1, (1+x)^{-2}/2\}$. Plugging the expansion into (2.15) allows to rewrite it as the sum of three series. Under assumptions *TMG1* and *TMG5*, Tokdar et al. (2009) show that these series are either martingales or bounded, hence convergent. Tokdar et al. (2009) Theorem 2 proves that $\overline{K}_n := KL(f^*, f_n)$ converges to 0, F^*- *a.s.*, as $n \to \infty$. A tightness argument that uses *TMG2*-*TMG4* is used to prove that G_n converges weakly to G^*, F^*- *a.s.*, (Tokdar et al. 2009, Theorem 3).

Walker (2017a) gives an alternative proof of convergence, given later completed in Chapter 5; see also Chapter 5 for an analogous multivariate argument (Cappello and Walker 2018). In these two references, the authors consider a Gamma scale mixture. The key idea follows from noticing that if one picks $g_0(\theta) = Ga(\theta|a, b)$, where $Ga(\cdot|a, b)$ denotes a Gamma density with shape parameter a and rate parameter b, g_n can be rewritten as a mixture of 2^n independent Gamma densities (Lemma 1 and 2 in Walker 2017a). Under a mild assumption, $\mathrm{E}|Y| < \infty$, one can show tightness without needing *TMG3*-*TMG5*.

Martin and Tokdar (2009) give a more general proof. They show convergence under misspecification; that is, when the true density f^* is outside the mixture model. The most interesting part is that $(\overline{K}_n)$ is shown to be an *almost supermartingale*, the notion introduced in Robbins and Siegmund (1971) briefly mentioned in the previous section.

Definition 1. *(Robbins and Siegmund 1971). Let $(M_n)_{n\geq 1}$ be a sequence of non-negative*

random variables adapted to a filtration $(\mathcal{F}_n)_{n\geq 1}$. *If there exists a sequence of non-negative random variables* $(\beta_n, \xi_n, \chi_n)_{n\geq 1}$ *such that* $\forall n \geq 1$, $M_n, \beta_n, \xi_n, \chi_n$ *are* $\mathcal{F}_n$ *measurable and*

$$E[M_n|\mathcal{F}_{n-1}] \leq (1+\beta_{n-1})M_{n-1} + \xi_{n-1} - \chi_{n-1}, \tag{2.16}$$

then (M_n) *is an almost supermartingale.*

Robbins and Siegmund (1971) study the convergence of an almost supermartingale.

Theorem 2.3.1. *(Robbins and Siegmund 1971, Theorem 1). Let* (M_n) *be an almost supermartingale and suppose* $\sum_{n=1}^{\infty}\beta_n < \infty$ *and* $\sum_{n=1}^{\infty}\xi_n < \infty$. *Then* (M_n) *converges a.s. and* $\sum_{n=1}^{\infty}\chi_n < \infty$.

It is possible to apply Theorem 2.3.1 to $(\overline{K}_n)$. The idea is that for, $x \approx 0$, the Taylor expansion of $\log(1+x)$ is an inequality. Namely, $\log(1+x) \geq x - 2x^2$. *TMG1-TMG4* are still in place, as well as a stronger version of *TMG5*. We refer to Martin and Tokdar (2009) for details.

2.3.2 Convergence Results for the HMW algorithm

Hahn et al. (2017) prove convergence of the algorithm (2.8) with respect to the Kullback-Leibler divergence. Recall that HMW targets directly the mixture density and is developed under the assumption that the kernel is Gaussian. The Gaussian copula follows from the kernel chosen, even if it does not directly appear in (2.8).

The observations are assumed to be *i.i.d.* from a density $p*$, such that $P^*(y) = \int_{\infty}^{y} p^*(t)\mathrm{d}t$ belongs to $\hat{\mathcal{P}}$, the class of Gaussian location mixture distributions. The argument employed in the proof is similar to the one in Martin and Tokdar (2009). Define $K_n := KL(p^*, p_n)$, which is the sequence of KL-divergences between the estimated predictive densities (p_n) and p^*

$$K_n(p^*, p_n) - K_n(p^*, p_{n-1}) = -\int \log[1 + \alpha_n \left[c_\rho(P_{n-1}(y), P_{n-1}(Y_n)) - 1\right] p^*(y)dy. \tag{2.17}$$

Then (2.17) can be decomposed in two series using the Taylor expansion $\log(1+x) \geq x - 2x^2$, given that $x \approx 0$, and shown to be an almost supermartingale. The two series can be bounded in order to apply (2.3.1). The bounds use some properties of the Gaussian distribution: although a Gaussian distribution does not appear directly in (2.8), a clever representation of the Gaussian copula as a mixture model makes it explicit,

$$c_\rho(u,v) = \int \psi_\theta(u)\psi_\theta(v)N(\theta|0,\rho)\mathrm{d}\theta,$$

with $N(\theta|0,\rho)$ denoting a Normal density function, and

$$\psi_\theta(u) = \frac{N(\Phi^{-1}(u)|\theta, 1-\rho)}{N(\Phi^{-1}(u)|0,1)}.$$

In order to study the convergence, *TMG2-TMG4* are not necessary. *TMG1* is more restrictive: α_n is assumed to be equal to $a(n+1)^{-1}$, with $0 < a < (2\rho+2)/(7\rho+1)$. Such a restriction on the (α_n) is somewhat restrictive: the sequence (α_n) has a numerically relevant effect when the sample is small; see discussion in Subsection 2.2.3. Furthermore, in the article the authors hint that other type of copulas can be used to build alternative recursive algorithms, for example the Student-t copula. However, the extension to other copula density functions does not seem straightforward.

Chapter 3

Recursive Procedures for Conditional Distributions

This chapter contains the first original contributions of the thesis. In Chapter 2 we have presented two algorithms, NQZ and HMW, and highlighted their strengths. Both NQZ and HMW have been proposed for a model where the observables are *i.i.d.* or exchangeable. An extension to these type of algorithms is to incorporate covariates into the sampling model.

Regression is one of the key problems in statistics. It is reasonable to believe that the strengths of the recursive algorithms discussed in the previous chapters extend to this more complex setting. The stochastic approximation framework naturally allows to estimate a regression function; see Lai (2003), Kushner and Yin (2003). Following Robbins and Monro (1951), many recursive algorithms targeting the regression function have been proposed; see Kiefer et al. (1952) for an algorithm to estimate the maximum of a regression function in a least squares setting, Révész proposed a method to nonparametrically estimate a regression function at any given point; see Révész (1973) and Révész (1977). Other methods then followed (Mokkadem et al. 2007, Dieuleveut et al. 2016).

In many applications, the regression function may be a poor summary of the relationship between a response y and the predictor x and the conditional distributions can be much more informative (Efromovich 2007); however, to our knowledge, no algorithms within a SA framework target conditional probability distributions. The gap in the literature is particularly surprising given that the strengths of recursive algorithms seem particularly suitable in this setting. This gap has also constituted the motivation for the work in this chapter.

In particular, we consider a *fixed regression design* where the regressor x takes values in $\mathcal{X}$, the covariate space, with $|\mathcal{X}| = m$ finite, and a continuous response $Y \in \mathcal{Y}$, the sample space. Let $\mathcal{P}$ denote the family of distribution functions on $(\mathcal{Y}, \mathcal{B})$. Suppose for all $n \geq 1$ $Y_n|x_n \overset{ind}{\sim} P(\cdot|x_n)$ with $P(\cdot|x_n) \in \mathcal{P}$ and $x_n \in \mathcal{X}$. Our goal is to estimate $P(\cdot|x)$ only for $x \in \mathcal{X}$. Note that here and throughout the rest of the text, with an abuse of notation, we refer to $P(\cdot|x)$ as a *conditional distribution function*. In fact, being the covariate x fixed by design, $P(\cdot|x)$ is not a conditional distribution function, but a probability indexed by x: by $Y|x \sim P(\cdot|x)$ we mean that Y has distribution $P(\cdot|x) \in \mathcal{P}$ when observed for the covariate value x. In this chapter we introduce two new recursive algorithms that allow a fast estimation of a set of conditional distribution functions. Again more properly, we should say that the target is a family of CDF indexed by $x \in \mathcal{X}$.

What we have in mind is a recursive algorithm that extends the algorithm (1.1) introduced in Chapter 1 to a regression design. Recursive algorithms allow to deal with streaming data: we assume to be in such a setting and denote a sample by $(y_n, x_n)_{n=1:N}$. Note that such a notation is slightly unsual in an experiment where $|\mathcal{X}| = m$, as it could bring one to think that the pedix indexes an element of $\mathcal{X}$. Although, the notation seems reasonable for streaming data: in this chapter, the pedix indexes the order in which observations are collected. Suppose I am interested in an estimate of $P(\cdot|x)$, where $x \in \mathcal{X}$ denotes a design point of interest. Given a sample $(y_n, x_n)_{n=1:N}$, an initial guess $P_0(\cdot|x) \in \mathcal{P}$, an estimate $P_N(\cdot|x)$ could be obtained by repeating recursively

$$P_n(\cdot|x) = (1 - \alpha_n)P_{n-1}(\cdot|x) + \alpha_n \bar{Z}(\cdot, x, P_{n-1}(\cdot|x), P_{n-1}(\cdot|x_n), y_n, x_n) \tag{3.1}$$

for $n = 1, \ldots, N$, where $\alpha_n \in (0, 1)$ is a sequence of user-supplied weight, and $\bar{Z}$ is a CDF. Clearly, (3.1) extends (1.1): the two are equivalent if $\mathcal{X}$ has a single element. The update $\bar{Z}$ generalizes Z in (1.1): it includes both the observed covariate, x_n, and the one of interest, x. The two algorithms we propose in this chapter fit into this general framework. The reason will become clear in Chapter 5: the asymptotic theory that we develop for the sequence (P_n) defined in Chapter 1 will easily extend to this setting.

Another desiderata is to develop algorithms related to a Bayesian predictive update, in the spirit of NQZ and HMW. There are good reasons to continue working with Bayesian nonparametric models. In our opinion, recursive procedures with a Bayesian interpretation have the potential to alleviate a lot of the computational burden associated with an exact inference. These issues are particular severe when dealing with a nonparametric regression model. Given that both NQZ and HMW have been proposed in the context of mixture models, we focus our attention to Bayesian inference in mixture of regression

models. We view data as a random sample from a density of the form

$$f(y|G_x, x) = \int_\Theta k(y|x, \theta) dG_x(\theta), \tag{3.2}$$

where $k(y|x, \theta)$ is a covariate dependent kernel, $G_x(\theta)$ is a covariate dependent mixing distribution, $\theta \in \Theta$ the mixing parameter; see Chapter 2 for a review of mixture models. We should emphasize that both the general algorithm (3.1) and the density (3.2) are suitable to model a regression design more complex than the one under investigation in this chapter, where we assume $|\mathcal{X}| < \infty$, deterministic regressor; and, we will further simplify the setting in the coming sections. However, it seems reasonable to fit the work in this chapter into a general framework in order to highlight the potential for future work.

The first algorithm we propose (Section 3.1) deals with a covariate that can assume finitely many values; for example a regressor describing the number of years a PhD student takes to write his thesis. The second algorithm (Section 3.2) targets a categorical covariate, although it is restricted to two categories, male and female for example. The two main sections of the chapter present two different proposals for the CDF $\bar{Z}$ in the algorithm (3.1). In each section the structure is specular: we first review the theory underlying the two models we consider; we then construct the algorithm; lastly we present some numerical illustrations to show that the algorithms perform well in a variety of problems. The theoretical convergence is studied in Chapter 5.

3.1 Recursive Nonparametric Predictive Regression

The algorithm we introduce in this section is the direct extension of HMW to a regression design with a discrete regressor $x \in \mathcal{X}$ with $|\mathcal{X}| = m$. In Subsection 2.2.2 we point out that the copula characterisation of sequence of predictive densities given by Hahn et al. (2017) holds to a greater generality than that presented in that reference. Indeed, it can be extended to a regression design.

Let $Y_1|x_1 \sim p_0(y|x_1)$ and $p_{n-1}(y|x)$ be the predictive density of $Y_n|x_n = x, y_{1:n-1}, x_{1:n-1}$, with $x \in \mathcal{X}$ the design point of interest, and $P_{n-1}(\cdot|x)$ the corresponding predictive distribution. Given (y_n, x_n), the predictive density of $Y_{n+1}|x_{n+1} = x, y_{1:n}, x_{1:n}$, i.e. $p_n(y|x)$, can be calculated as

$$\frac{p_n(y|x)}{p_{n-1}(y|x)} = \frac{p(y, y_n|x, x_n, y_{1:n-1}, x_{1:n-1})}{p_{n-1}(y|x)p_{n-1}(y_n|x_n)} = c_n\Big(P_{n-1}(y|x), P_{n-1}(y_n|x_n), x, x_{1:n}\Big), \tag{3.3}$$

where c_n denotes a copula density function. The first equality in (3.3) follows from

basic probability rules, the second equality follows from Sklar's theorem and copula theory (Nelsen 2007, Subsection 2.2.2 for a review). To elaborate, let us remind for any multivariate distribution F, with marginals $(F_1, \dots, F_d)$, that admits probability densities f and $(f_1, \dots, f_d)$ respectively, a copula density function is defined as

$$\frac{f(u_1, \dots, u_d)}{f_1(u_1) \dots f_d(u_2)} = c(F_1(u_1), \dots, F_d(u_d)).$$

Hence, (3.3) trivially follows. Expression (3.3) is a direct, but insightful, extension of Hahn et al. (2017) to a regression design. The key difference is that the copula density carries information across the group I am sampling from. Recall that the covariate $x \in \mathcal{X}$ is deterministic and it simply indexes a finite vector of distributions. Other properties will be discussed in the next section.

3.1.1 A parametric model

Our prime interest is Bayesian inference for mixture of regression models. However, to examine the properties of the copula density characterization, it seems useful to start by considering the parametric case. Recall that Hahn et al. (2017) show several parametric models for which it is possible to write the sequence of copula densities (c_n) in closed form. Analogously, we show that for Bayesian inference for a linear regression model, we can write the copula density in closed form.

Let $Y_n|\beta, x_n \overset{ind}{\sim} N(\beta x_n, \sigma^2)$, with a Gaussian prior distribution $\beta \sim N(\beta|0, \tau^{-1})$ for a fixed $\tau^{-1} = \sigma^2$, and $x_n \in \mathcal{X}$ a non-stochastic design point. Suppose we collect streaming observations. For any $n \geq 1$, the predictive density $p_n(y|x)$, where $x \in \mathcal{X}$ is the design point of interest, is available in closed-form:

$$p_n(y|x) = N\left(y \middle| \frac{x(\sum_{i=1}^n x_i y_i/\sigma^2)}{\tau + \sum_{i=1}^n x_i^2/\sigma^2}, \frac{\sum_{i=1}^n x_i^2 + \tau\sigma^2 + x^2}{\tau + \sum_{i=1}^n x_i^2 \sigma^2}\right),$$

where $N(\cdot|\mu, \sigma^2)$ denotes the density $N(\mu, \sigma^2)$ computed at $(\cdot)$. Fairly simple computations show that

$$p_n(y|x)/p_{n-1}(y|x) = c_{\rho_n(x, x_{1:n})}(P_{n-1}(y|x), P_{n-1}(y_n|x_n)),$$

where $c_\rho(u, v)$ denotes a Gaussian copula density

$$c_\rho(u, v) = (1 - \rho^2)^{-1/2} \exp\left(\frac{2\rho\Phi^{-1}(u)\Phi^{-1}(v) - \rho^2\Phi^{-1}(u)^2 - \rho^2\Phi^{-1}(v)^2}{2(1 - \rho^2)}\right).$$

Here, the parameter $\rho_n(x, x_{1:n})$ is defined by

$$\rho_n^2(x, x_{1:n}) = \frac{x_n^2 x^2}{(x_n^2 + \sum_{i=1}^{n-1} x_i^2 + \tau) + (x^2 + \sum_{i=1}^{n-1} x_i^2 + \tau)}. \tag{3.4}$$

Note that the copula is independent from the responses $y_{1:n}$ but it now includes the covariates $x_{1:n}$. The other key features highlighted by Hahn et al. (2017) hold though. First, it includes what we called for the density estimator in Chapter 2, a "sample size effect", represented by $\sum_{i=1}^{n-1} x_i^2$ at the denominator. The reason of this sample size interpretation is that $\sum_{i=1}^{n-1} x_i^2$ is strictly increasing with n. Second, note that we are modelling a partially exchangeable sequence: the predictive distributions are convergent by de Finetti's representation theorem, which implies that the copula density must converge to the independence copula density, i.e $c(\cdot, \cdot) = 1$. This is true for any model that has convergent predictive distributions. In the example above, $\rho(x, x_n)$ converges to zero as n increases, leading to the independence copula.

Another important difference with the setting discussed by Hahn et al. (2017) is that the sequence of copulas corresponding to $(p_n(\cdot|x))$ depends on $x \in \mathcal{X}$. In the illustration discussed in this section, the Gaussian copula density is parametrized by $\rho_n(x, x_{1:n})$.

3.1.2 Bayesian Inference for Mixture Regression

Here we review some Bayesian nonparametric regression models based on mixtures. Given that we consider a continuous response Y, all the kernels are assumed to be continuous; while here the regressor x is discrete. We first consider the mixture model:

$$Y_n|\beta_n, x_n \overset{ind}{\sim} k(y|\beta_n x), \qquad \beta_n|G \overset{iid}{\sim} G \qquad G|c, G_0 \sim \mathrm{DP}(cG_0), \tag{3.5}$$

with the mixture density represented through the stick-breaking construction of the DP,

$$f(y|G, x) = \sum_{j=1}^{\infty} w_j k(y|\beta_j^* x), \qquad a.s. \tag{3.6}$$

where the regression coefficients β_j^* and the weights w_j are sampled using the scheme of Sethuraman (1994), described in Section 2.1. The density (3.6) is evaluated in the design point x of interest. Escobar and West (1992) first introduced this model also in the context of a fixed design matrix: it is an infinite mixture of linear models, usually called a DP mixture of linear models. It is not the first Bayesian nonparametric regression model, see Cifarelli and Regazzini (1978); but it is the first model employing DP mixtures.

The Dependent Dirichlet Process introduced by MacEachern (1999, 2000) has been

also introduced in the context of fixed design regression. The Dependent DP extends the DP by introducing covariates in both the weights and the atoms. The author assumes that the mixing distribution in a mixture model of the form (3.5), is indexed by the covariate x. Let us denote it by G_x. The Dependent DP defines a prior on $(G_x)_{x\in\mathcal{X}}$ when one wishes to enforce dependence across the covariates. Formally, a vector $(G_x, x \in \mathcal{X})$, where $|\mathcal{X}| = m < \infty$, is Dependent Dirichlet Process distributed with parameters $c(x) > 0$ and $G_{0,x}$ if, *a.s.*,

$$G_x(\cdot) = \sum_{j=1}^{\infty} w_j(x)\delta_{\theta_j(x)}(\cdot),$$

where the weights are

$$w_i(x) = v_i(x)\prod_{j<i}(1 - v_j(x)), \qquad \forall j > 1,$$

with $w_1(x) = v_1(x)$ and $v_i(x) \overset{iid}{\sim} Beta(1, c(x))\ \forall j > 0$, and the $(\theta_j(x))$ are independent across j, for all j, and $\theta_j(x)$ is a stochastic process on $\mathcal{X}$ with marginals $G_{0,x}$ for $x \in \mathcal{X}$. The covariate dependence across the weights is given by $c(x)$; the covariate dependence across the atoms naturally arises because $\theta_j(x)$ is a stochastic process on $\mathcal{X}$. The Dependent DP rarely appears in the literature with both the weights and the atoms dependent on the covariate. In the rich literature based on, and expanding the Dependent DP construction, we just mention a few papers: De Iorio et al. (2004), Dunson and Park (2008) and Antoniano-Villalobos et al. (2014).

The work of De Iorio et al. (2004) is of particular interest for us. In this reference only the atoms are covariate dependent, i.e. $w_j(x) \equiv w_j$ for all $x \in \mathcal{X}$. In particular we consider a model the authors discuss, the so-called single-p DDP mixture, with a Gaussian kernel. Then, for a fixed x, the conditional density is

$$f(y|G_x, x) = \sum_{j=1}^{\infty} w_j N\left(y\middle|\mu_j(x), \sigma_j^2\right) \qquad a.s.$$

De Iorio et al. (2004) show that the single-p DDP mixture is equivalent to the DPM of linear models up to a transformation $m(\cdot)$ of the x

$$f(y|G, x) = \sum_{j=1}^{\infty} w_j N\left(y\middle|\beta_j m(x), \sigma_j^2\right) \qquad a.s.$$

which is equivalent to (3.6) for $m(x) = x$.

3.1.3 A Recursive Predictive Regression Model

Having extended, as in Hahn et al. (2017), the copula based characterization to a linear regression, we continue the analogy and consider a nonparametric mixture regression. The sequence of Gaussian copula densities defined by the parameter (3.4) allows us to calculate all the predictive distributions without having to update the posterior. We assume that a random sample is available from a mixture density of the form (3.2) with $G_x \equiv G$ for all $x \in \mathcal{X}$: the choice is motivated by the equivalence between (3.5) and the single-p DDP (De Iorio et al. 2004) for a transformation $m(x) = x$.

3.1.3.1 Algorithm Construction

Set $k(y|\beta x) = N(y|\beta x, \sigma^2)$, a Gaussian kernel with mean parameter βx and fixed variance σ^2. Note that the construction holds also for unknown variance. So,

$$f(y|G, x) = \int N(y|\beta x, \sigma^2)\, \mathrm{d}G(\beta),$$

with prior on G given by a DP, $G \sim \mathrm{DP}(c\, G_0)$. The model is fully specified as in (3.5). The predictive density $p_n(y|x)$ for this model is not analytically tractable for $n \geq 2$. Yet, p_0 and p_1 are available and this is all we need to get a recursive algorithm; see also Newton et al. (1998) and Hahn et al. (2017). Given (y_1, x_1),

$$c(P_0(y|x), P_0(y_1|x_1), x, x_1) = \frac{p_1(y|x)}{p_0(y|x)} = \frac{\mathrm{E}_G[f(y|G, x) f(y_1|G, x_1)]}{p_0(y|x) p_0(y_1|x_1)}, \tag{3.7}$$

where $p_0(y|x) = \int N(y|\beta x, \sigma^2)\, \mathrm{d}G_0(\beta)$ is the prior guess. The quantities in the integral are bounded, we can use Fubini-Tonelli twice and show that (3.7) is equal to

$$\alpha \, \frac{\int_B N(y|\beta x, \sigma^2) N(y_1|\beta x_1, \sigma^2)\, \mathrm{d}G_0(\beta)}{p_0(y|x) p_0(y_1|x_1)} + (1 - \alpha),$$

where $\alpha = \sum \mathbb{E}[w_j^2]$. The copula density that relates $p_1(y|x)$ and $p_0(y|x)$ is a mixture of a Gaussian copula density and the independence copula. The correlation in the Gaussian copula can be shown to be equal to (3.4) with $n = 1$. Hence, $p_1(y|x)$ can be computed as

$$p_1(y|x) = (1 - \alpha) p_0(y|x) + \alpha \; c_{\rho(x,x_1)}\Big(P_0(y|x), P_0(y_1|x_1) \Big) \; p_0(y|x). \tag{3.8}$$

A trivial extension is to consider the Gaussian kernel $k(y|\beta, x) = N(y|\beta x, \sigma^2)$, with $(\beta, \sigma^2)|G \overset{iid}{\sim} G$ and $G \sim \mathrm{DP}$. The model is a Bayesian linear regression with σ^2 unknown and random such that σ^2 is inverse-Gamma distributed. Simple calculations lead to the same update but with a Student-t copula density in place of the Gaussian one.

We can now provide a recursive update for general n in a spirit similar to Newton et al. (1998) and Hahn et al. (2017). The Bayesian predictive update defined by the first term (3.8) must be "good". Hence, it is repeated iteratively. Let us discuss the inputs of the algorithm. First, we need to choose a deterministic sequence of weights (α_n) taking values in $(0, 1)$. The desiderata for (α_n) are the usual ones: first, the sequence (α_n) needs to go eventually to zero in order to drive the convergence to the independence copula; second, we have to ensure that the decay to zero is not abrupt in order to move away from the initial guess. In other words, (α_n) incorporates the "sample size effect" highlighted in the parametric case: in the long run, the observations are incorporated with decreasing weight. Conditions on the (α_n) are formalized in Chapter 5.

Second, the function $\rho(\cdot, \cdot)$ is crucial. In short, it defines how to update $p_{n-1}(y|x)$ once we have observed (y_n, x_n). The parametric ρ, defined in (3.4), assumes linearity and furthermore it carries the "sample size effect" which is now going to be set by the sequence (α_n), see Subsection 3.1.1. Hence, the correlation function (3.4) we have introduced for linear regression needs to be modified. We do not need it to change with n now, as that role is played by α_n. Regression models often employ the notion of distance between covariates; see for example Aitchison and Aitken (1976) who use a L_2 distance. Using the notion of distance implies the reasonable assumption that covariate values, closer to the one we want to predict are more informative. The desiderata is that the correlation goes to zero as a pair of covariates move away from each other: This translates into a copula that moves towards the independence copula. We propose the following function:

$$\rho(x, x') = \frac{\rho_0 \, \tau}{\tau + |x - x'|^p}. \tag{3.9}$$

Hence we need a setting of (τ, ρ_0, p) and we will discuss it in Subsection 3.1.3.2.

We propose the following algorithm. We define it to target a CDF rather than a density as we will prove the asymptotic convergence of the sequence of distributions. We will refer to the algorithm targeting the CDF as the one on a *distribution scale*, analogously *density scale* for a density estimator. The connection to the copula is clarified after the definition of the algorithm.

Recursive Regression (RR). For all $x \in \mathcal{X}$, fix an initial guess $P_0(y|x) \in \mathcal{P}$, a decreasing deterministic sequence of weights $(\alpha_n) \in (0, 1)$, the correlation function $\rho : \mathcal{X} \times \mathcal{X} \to [0, 1]$ defined in (3.9), and a CDF $H_{\rho(\cdot,\cdot)}$ defined as

$$H_{\rho(x,x')}(P(y|x), P(y'|x')) = \Phi\left(\frac{\Phi^{-1}(P(y|x)) - \rho(x, x')\Phi^{-1}(P(y'|x'))}{\sqrt{1 - \rho(x, x')^2}}\right),$$

where (y', x') denotes an observation. Given observations $(y_n, x_n)_{n=1:N}$, an estimate of $P_N(y|x)$ is given by repeating recursively

$$P_n(y|x) = (1 - \alpha_n)P_{n-1}(y|x) + \alpha_n H_{\rho(x,x_n)}(P_{n-1}(y|x), P_{n-1}(y_n|x_n)), \tag{3.10}$$

for $n = 1, \ldots, N$, and for all $x \in \mathcal{X}$.

A few remarks. Note that $H_{\rho(x,x')}$ corresponds the Bayesian predictive update given in (3.8) on a distribution scale with the parameter ρ defined by (3.9). Indeed,

$$H_{\rho(x,x')}(P(y|x), P(y'|x')) = \int_{-\infty}^{y} c_{\rho(x,x')}\Big(P(t|x), P(y'|x')\Big)\ p(t|x)\mathrm{d}t,$$

where $c_{\rho(x,x')}$ is a Gaussian copula with paramater $\rho(x, x')$. For example, suppose $x' = x$, then ρ defined in (3.9), is equal to ρ_0. Also, it is possible for very small N that the covariate value x is not observed: the borrowing of information is tuned by the parameters (τ, ρ_0, p).

Different choices of the kernel $k(y|\beta, x)$ in (3.6) will correspond to different definitions of $H_{\rho(\cdot,\cdot)}$ and $\rho(\cdot, \cdot)$. A trivial extension is to consider the Gaussian kernel $k(y|\beta_j x) = N(y|\beta_j x, \sigma_j^2)$ with $(\beta_j, \sigma_j^2)|G \overset{iid}{\sim} G$ and $G \sim \mathrm{DP}$. Let (y', x') be one observation, RR can be still applied with $H_{\rho(x,x')}$ defined as

$$H_{\rho(x,x')}(P(y|x), P(y'|x')) = \mathcal{T}_\nu\left(\frac{\mathcal{T}_\nu^{-1}(P(y|x)) - \rho(x, x')\mathcal{T}_\nu^{-1}(P(y'|x'))}{\sqrt{(1 - \rho(x, x')^2)\frac{\nu + \mathcal{T}_\nu^{-1}(P(y'|x'))^2}{\nu+1}}}\right),$$

where $\mathcal{T}_\nu$ denotes a Student-t CDF with ν degrees of freedom. Note that RR is consistent with the general framework we have conceptualized in (3.1), setting

$$\bar{Z}(y, x, P(\cdot|x), P(\cdot|x'), y', x') = H_{\rho(x,x')}(P(y|x), P(y'|x')).$$

For this particular algorithm we favour the notation on the right hand side, since it makes explicit the way the inputs enter into the function.

As other iterative procedures we have discussed, RR is not order-independent. The lack of symmetry, which is often undesirable in an estimator, is sacrificed in order to gain computational speed and ease of implementation. The trade-off is acceptable as we do not loose much accuracy for large n; see Subsection 3.1.4 and Chapter 5. The algorithm moreover is highly relevant in the case when the observations are taken sequentially, which in regression design is a practical problem.

3.1.3.2 Algorithm Implementation

Similarly to HMW and NQZ, RR needs to be calculated on a grid of points: expression (3.10) cannot be implemented to give a function at each iteration but need to be calculated at chosen points. Let $\mathcal{X} := \{x'^1, \ldots, x'^m\}$: one needs to decide in which points to calculate the algorithm RR; see Subsection 2.2.3. Our proposal shares the advantage of HMW: no integral needs to be approximated. In Subsection 2.2.3 we suggest a simple strategy to define a grid using the order statistics. In this case, we would consider the $n \times m$ matrix defined by $[y_{(1)}, \ldots, y_{(N)}] \times [x'^1, \ldots, x'^m]$. We refer to Subsection 2.2.3 for details. The grid can be made finer if necessary.

The order dependence can be tackled as we have discussed for NQZ and HMW: Monte-Carlo averaging, see Subsection 2.2.3. In that subsction, we refer to the idea of "smoothness" of the estimate: based on our numerical simulations, smoothness is only particularly relevant for small sample sizes; say less than 100 observations. Our experience is that ten to twenty Monte Carlo averaging iterations suffice.

For a large sample size, e.g. more than 100 observations per category, restarting the algorithm also offers a viable improvement to the algorithm. Empirical simulations suggest that such a procedure may benefit the numerical convergence. In particular, we use the fact that the best possible update of $P_n(y|x)$, is achieved when I am updating the CDF from which I have observed one sample. We thus suggest to employ the order-dependency as an advantage: fix $x \in \mathcal{X}$, when interested in predicting $P(y|x)$, we first update $P_0(y|x)$ using $\{(y_{j_1}, x_{j_1}), \ldots, (y_{j_m}, x_{j_m})\}$, being $j_n \in J := \{n \in \{1, \ldots, N\} : x_n \neq x\}$; then restart (3.10) and use the remaining observations. Whereas Monte Carlo averaging is effective in both scenarios, restart is beneficial solely when a large sample size is available.

Finally, we have proposed to use either a Gaussian or a Student-t update. Let us remind that the two distributions are related to two different models: the Gaussian update to a Gaussian location mixture, the Student-t update to a Gaussian scale-location mixture. If we use the algorithm RR as an estimator, the reasons to prefer one or the other, follow directly from well known properties of the two distributions. On the one hand, the Student-t distribution offers the flexibility of an additional parameter, ν, regulating the degrees of freedom. On the other hand, the Gaussian update is much faster to calculate, at least using a standard software such as R and Matlab. We find that generally the method is robust to both choices.

The parameters (ρ_0, τ, p) tune the borrowing of information. ρ_0 defines the maximum correlation, it is achieved when the predicted covariate value and the observed one

match. It should be kept fairly high, say above 0.95. τ and p regulates the borrowing of information across covariate values. These are problem specific. Again, they should ensure enough borrowing of strength, especially given the fact that the update is discounted by the weights (α_n). Due to the decay of (α_n), correlations below 0.4 tend to lead to an ineffective update after a few iterations.

3.1.3.3 Numerical Examples

We investigate numerical properties of RR. The first two examples are designed to study the large-sample convergence, the third is a small sample illustration. Some inputs are shared across the three examples: the parameters of the function ρ are $\rho_0 = 0.99, \tau = 5$, and $p = 1$, the sequence of weights is such that $\alpha_n = (1+n)^{-0.7}$. In Chapter 5 we prove that this choice of weights meets the necessary condition to establish asymptotic convergence. Although our procedure is proposed for fixed-design regression, in this simulation study the covariate values are randomly generated. Specifically, we sample the value of the regressor x from a distribution we will specify in each example and then treat it as a deterministic quantity.

Example 1. We consider simulated data from the following regression model with the covariate in $\mathcal{X} := \{1, \dots, 4\}$, with

$$Y_n|x_n \overset{ind}{\sim} wN(\beta_1 x_n, \sigma_1^2) + (1-w)N(\beta_2 x_n, \sigma_2^2),$$

and $\beta_1 = -0.5$, $\beta_2 = 1$, $\sigma_1^2 = 1.4$, $\sigma_2^2 = 1$ and $w = 0.5$. Covariates are sampled through a uniform discrete on $\mathcal{X}$ and $N = 2500$. We take $P_0(y|x) = N(y|0, 2)$ for all x. The RR estimates is given by the Student-t update with degrees of freedom $\nu = 30$. Running time is about two and half minutes. Figure 3.1 plots the RR estimates (black), the true distributions (blue) and $P_0(y|x)$(green), and displays CDFs conditional on $x = 1$ and $x = 4$. The distributions are clearly identified.

Example 2. In this example we simulate data from a different regression model

$$Y_n|x_n \overset{ind}{\sim} N(\beta x_n, \sigma^2 x_n),$$

with covariate in $\mathcal{X} := \{1, \dots, 5\}$ and sampled from a uniform discrete on $\mathcal{X}$, $\beta = 0.5$, $\sigma^2 = 0.5$ and $N = 2500$. Again, the RR estimates is given by the Student-t update with degrees of freedom $\nu = 30$. Running time is about four minutes with a restart. Figure 3.2 plots the RR estimates (black), the true distributions (blue) and $P_0(y|x)$ (green) and plots the two distributions conditional on the two extreme covariates $x = 1$ and $x = 5$. Again, we are able to recover the true distributions perfectly.

Example 3. We simulate new data from the model described in Example 1. Now, we want to show a numerical evidence of the borrowing of information and highlight a few numerical properties of RR. We restrict $\mathcal{X} = \{1, 2\}$ and we sample 25 observations, 20 from $x = 1$ and 5 from $x = 2$. Figure 3.3 plots the RR estimates (dashed black line), the HMW estimates (dashed green line), the true CDFs (blue line) and the empirical distribution (red line). Recall that HMW is identical to RR when we do not use any data from the other category - i.e. no borrowing of information. To make the estimates smoother and more order independent, we plot the average estimate given by 20 Monte Carlo permutations of the data, both for RR and HMW.

It is clear that the HMW estimates follow more closely the empirical distribution function in the presence of a larger datasets; which is entirely reasonable. RR gives an estimate of the under-represented group which matches more closely the true distribution. Group 1 also exhibits generally a better performance of the predictive regression. It is a comforting result because it is an empirical evidence of the borrowing of information.

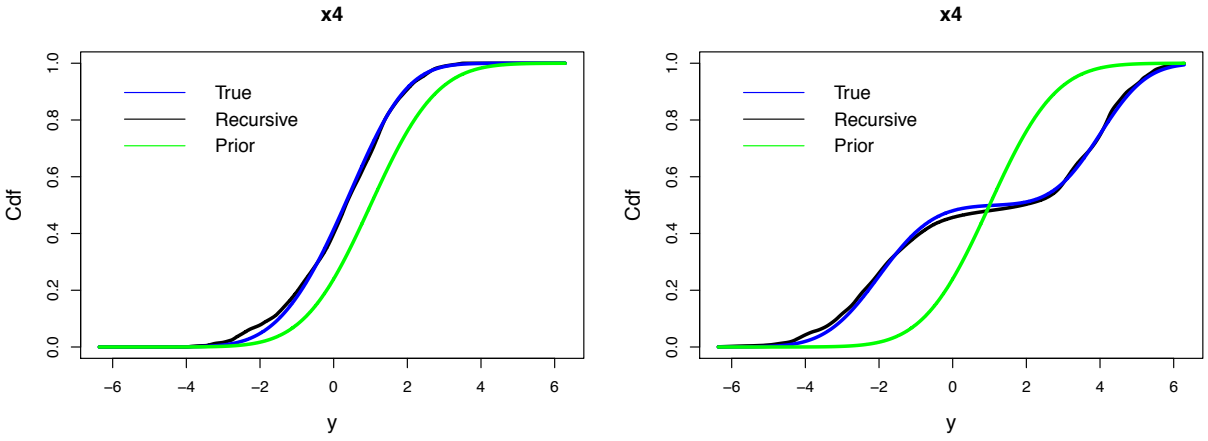


Figure 3.1: RR estimates of the predictive distributions $P(y|x)$ for $x = 1$ and $x = 4$ in Example 1: P_0 (green), recursive estimate (black) and true distribution (blue).

3.1.4 Illustrations

We numerically compare the recursive regression to the Bayesian nonparametric discrete regression of De Iorio et al. (2004). We argued that the single-p DDP mixture studied by De Iorio et al. is closely related to the mixture model our algorithm is based upon. Next to the Bayesian inference given by the single-p DDP mixture, we compare our recursive regression to the discrete kernel regression model proposed by Aitchison and Aitken (1976).

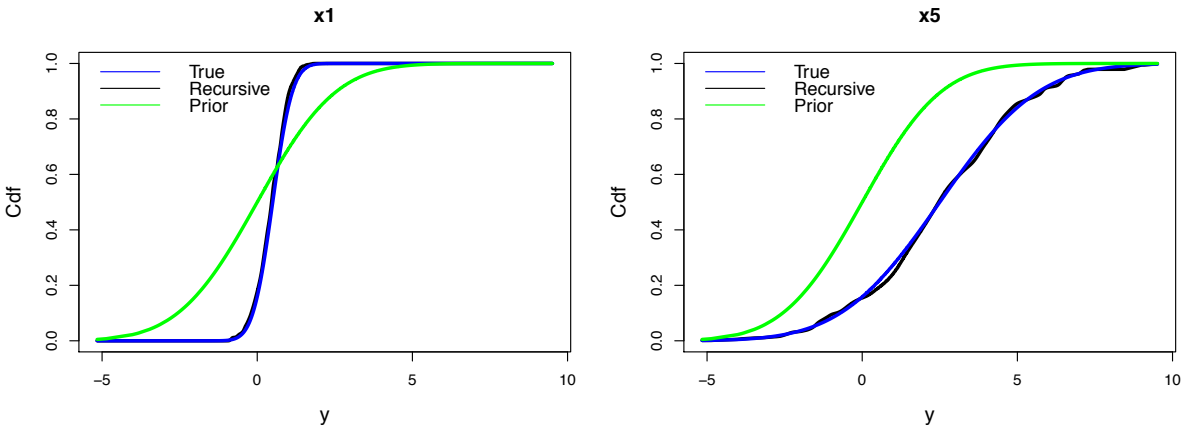


Figure 3.2: RR estimates of the predictive distributions $P(y|x)$ for $x = 1$ and $x = 4$ in Example 2: P_0 (green), recursive estimate (black) and true distribution (blue).

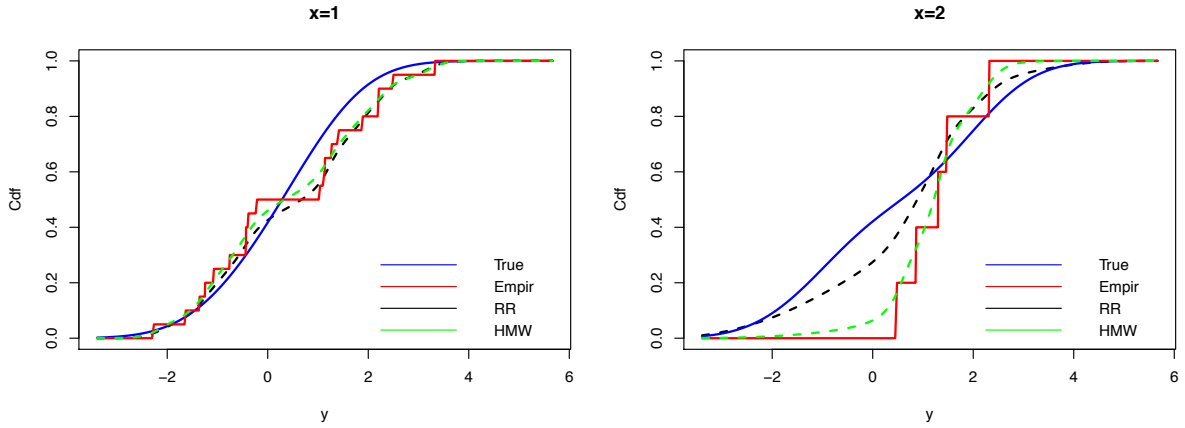


Figure 3.3: Estimates of the predictive distributions $P(y|x)$ for $x = 1$ and $x = 2$ in Example 3: recursive estimate given by RR (black), recursive estimate given by HMW (green), empirical CDF (red) and true distribution (blue).

All the methods are implemented in `R`. The `DPpackage` (Jara et al. 2011) has a built-in function, `LDDPdensity`, which allows to do inference in the single-p DDP mixture model (LDDPM). The function `LDDPdensity` inputs are the parameters of the LDDPM; in particular it is assumed that the observations $Y_n|x_n$ are sampled from a mixture density

$$f(y|G_x, x) = \int N(y|\beta x, \sigma^2) \mathrm{d}G(\beta, \sigma^2),$$

where G is random and such that $G|cG_0 \sim \mathrm{DP}(cG_0)$ with

$$G_0(\beta, \sigma^2) = N(\beta|\mu_b, s_b)\Gamma(\sigma^2|\tau_1/2, \tau_2/2),$$

$$c|a_0, b_0 \sim Ga(c|a_0, b_0); \qquad \mu_b|m_0, S_0 \sim N(\mu_b|m_0, S_0),$$

$$sb|\nu, \Psi \sim IW(sb|\nu, \Psi); \qquad \tau_2|\tau_{s1}, \tau_{s2} \sim Ga(\tau_2|\tau_{s1}/2, \tau_{s2}/2).$$

We will specify the hyperparameters chosen for our simulation study in the two subsections. To implement the kernel regression we use the `np` package developed by Hayfield et al. (2008), specifically the function `npcdist` to estimate conditional distribution functions. Bandwidth and kernel types are picked through the built in data-driven procedure. In our case, the discrete nature of the data leads to the kernel regression of Aitchison and Aitken (1976).

3.1.4.1 Repeated regressions simulation

We build a simulation model where the same experiment is run multiple times. The data are sampled from a mixture of Student-t and Gaussian densities with covariate dependent weights, i.e.

$$f(y|x) = \left(1 - \frac{s_1 + x - x_{(1)}}{s_2 + x - x_{(1)}}\right) \mathcal{T}_\nu(y|\beta_a x, \sigma_a^2) + \frac{s_1 + x - x_{(1)}}{s_2 + x - x_{(1)}} N(y|\beta_b x, \sigma_b^2).$$

We consider the following quantities: the covariate in $\mathcal{X} := \{0.8, 1, 1.3, 1.5\}$, $\beta_a = -0.75$, $\beta_b = 1$, $\sigma_a^2 = 0.5$, $\sigma_b^2 = 1$, $\nu = 3$, $s_1 = 1/3$ and $s_2 = 1$. We sample the covariates from a uniform discrete on $\mathcal{X}$. Each experiment produces 250 observations and 50 experiments are run. For each run, we estimate the vector of conditional CDFs $[P_n(y|x = 0.8), P_n(y|x = 1), P_n(y|x = 1.3), P_n(y|x = 1.5)]$. The estimates are given by RR, LDDPM and kernel regression. To compare the accuracy, we compute for each estimate the Kolmogorov–Smirnoff and the Cramer von Mises statistics. We are not going to do any tests: we use these two statistics simply as a measure of distance between two distributions. In particular, we compute the distance between the estimates $[P_n(y|x), x \in \mathcal{X}]$ and the true distributions $[P^*(y|x_1), x \in \mathcal{X}]$. In this respect, Kolmogorov–Smirnoff is simply a L_∞ distance between curves, and Cramer von Mises a L_2 distance.

We fix the following estimation parameters. For RR: $P_0(y|x) = N(y|0, 3)$ for all $x \in \mathcal{X}$ and $\alpha_n = (n + 1)^{4/5}$. The update is given by a Student t distribution with $\nu = 8$, $\rho_0 = 0.98$, $\tau = 1$, and $p = 1$. Each estimate $P_n(y|X)$ is given by the average of 10 Monte Carlo iteration (see Subsection 3.1.3.2). For LDDPM: the nine parameters and hyperparameters of the prior distributions of the Bayesian nonparametric regression model have been chosen to make the prior less informative as possible $(a_0 = 8, b_0 = 1, m_0 = 0, S_0 = 4, \tau_1 = 6.01, \tau_{s1} = 6.01, \tau_{s2} = 2.01, \nu = 1, \psi^{-1} = 0.25)$. The Markov Chain Monte Carlo has a burn in of 5000 iterations, and then 20000 scans are considered and subsampled one every four values; the final chain size is therefore 5000. For kernel regression: we use the in built procedure in the R package `np`.

Figures 3.4 and 3.5 plot the distributions of the two distances considered. The accuracy of the three methods is quite comparable. Figure 3.5 shows a slightly better overall performance of RR with respect to the Cramer von Mises metric. Recall that the Cramer von Mises distance represents a more aggregate estimation of a distribution being related to the L_2 distance: this result hints that our algorithm gives a better average estimates of the different quantiles of a distribution. The poorer, even if still solid performance, with respect to the Kolmogorov–Smirnoff statistics, is justified by the fact that a RR estimate is generally less smooth, hence more penalized by the L_∞ distance. A comparable performance is achieved at a computational speed much faster than the LDDPM: 18 seconds for a single RR estimates, whereas a single run of the Bayesian regression model takes about 9.8 minutes. The kernel estimate takes about 2.5 seconds.

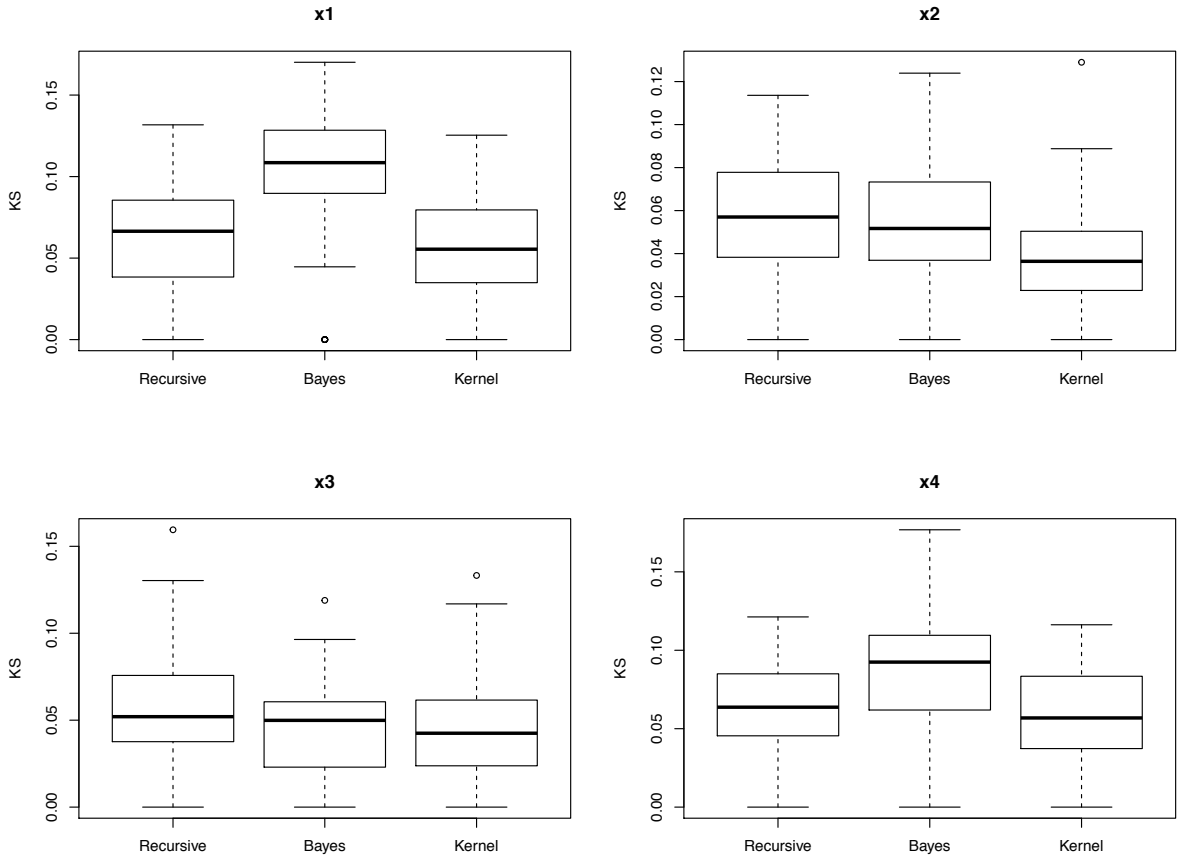


Figure 3.4: Distributions of Kolomorov Smirnov statistics for the four groups and the three estimator: RR (Recursive), the single-p DDP of De Iorio et al. (2004) (Bayes) and the kernel regression of Aitchison and Aitken (1976) (Kernel)

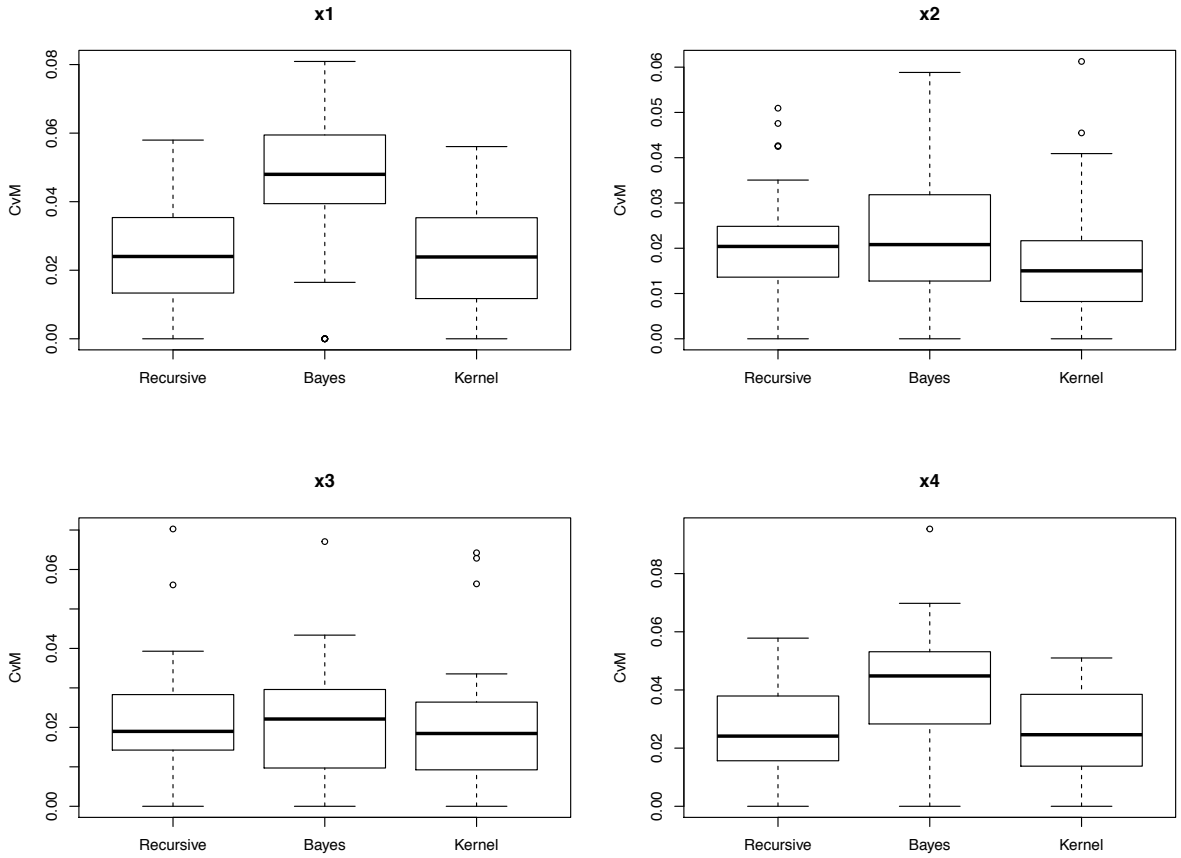


Figure 3.5: Distributions of Cramer von Mises statistics for the four groups and the three estimator: RR (Recursive), the single-p DDP of De Iorio et al. (2004) (Bayes) and the kernel regression of Aitchison and Aitken (1976) (Kernel)

3.1.4.2 Serum Immunoglobulin concentration study

The Serum Immunoglobulin concentration study reports concentrations of serum immunoglobulin (IgG) from children of ages ranging from 6 months to 6 years (Isaacs et al. 1983). Monitoring IgG concentration is important because it measures the amount of antibodies in the blood. It is reasonable to expect that as a child is growing, the amount of antibodies will increase. Our interest here is to estimate the distributions of IgG for different age classes.

The dataset includes 298 observations, a continuous response (IgG concentration) and a covariate (age). In the original dataset age is continuous: we discretise it in order to get an interval covariate by grouping children in six peers. We round each observation to the closest year and a half. For example, $x = 3.2$ is assigned to the category of kids that are three years and a half old, i.e. $x = 3.5$. Similarly if a kid ages is $x = 2.9$, it is assigned to the category of kids that are two years and a half old, i.e. $x = 2.5$. The

data is partitioned in the following six groups: $n_{x=0.5} = 62$, $n_{x=1.5} = 68$, $n_{x=2.5} = 39$, $n_{x=3.5} = 40$, $n_{x=4.5} = 49$ and $n_{x=5.5} = 40$.

We fix the following parameters: $P_0(y|x) = N(y|0, 4)$ for all x, the update is given by a Gaussian copula with $\rho_0 = 0.98, \tau = 2, p = 1$, $\alpha_n = (n+1)^{3/4}$. The hyperparameters in the Bayesian prior are set to be less informative as possible, with the exception of m_0, which is chosen bigger than zero to help the convergence of the MCMC ($a_0 = 8, b_0 = 1, m_0 = 2, S_0 = 5, \tau_1 = 6.01, \tau_{s1} = 6.01, \tau_{s2} = 2.01, \nu = 1, \psi^{-1} = 0.2$). The Markov Chain Monte Carlo has a burn in of 5000 iterations, and then 20000 scans are considered and subsampled one every four values; the final chain size is therefore 5000. The bandwidth and the kernel are determined using the built in procedure and are the only data dependent quantities in the problem. The kernel is always picked to be the discrete kernel proposed by Aitchison and Aitken (1976).

Figure 3.6 plots the estimates for the six categories given by the three methodologies, along with the empirical distribution functions computed for each group. For RR we plot an average of 20 Monte Carlo iterations. Both kernel regression and RR track pretty closely the empirical distributions. The RR estimate for kids older than five years is quite distinct and does not pick the jump of the empirical distribution. It seems a positive feature, being the distribution of all the other groups unimodal. It is a bit surprising the behaviour of the Bayesian model, which underestimates the distributions for low valued covariates and overestimates it for high valued ones. Both the kernel regression and the recursive procedure have an opposite effect.

Estimation time is 2.5 seconds for the Monte Carlo RR, since the inversion of the Gaussian distribution function much faster than the Student-t used in the previous example, 13.3 minutes for the LDDPM and 2 seconds for the kernel regression.

3.2 A Two-Step Recursive Predictive Algorithm

We introduce a second recursive algorithm that allows to estimate jointly the marginal CDFs of two dependent random variables. In practice, we consider a setting in which two different but somewhat related groups are observed, and suppose that the observations collected in one group are useful to infer statistical quantities of interest of the other one.

In particular, consider two sequences of observables: $(Y_n^{(a)})$ and $(Y_n^{(b)})$ both taking values in $(\mathcal{Y}, \mathcal{B})$. We assume that a finite sample for each group is available from densities

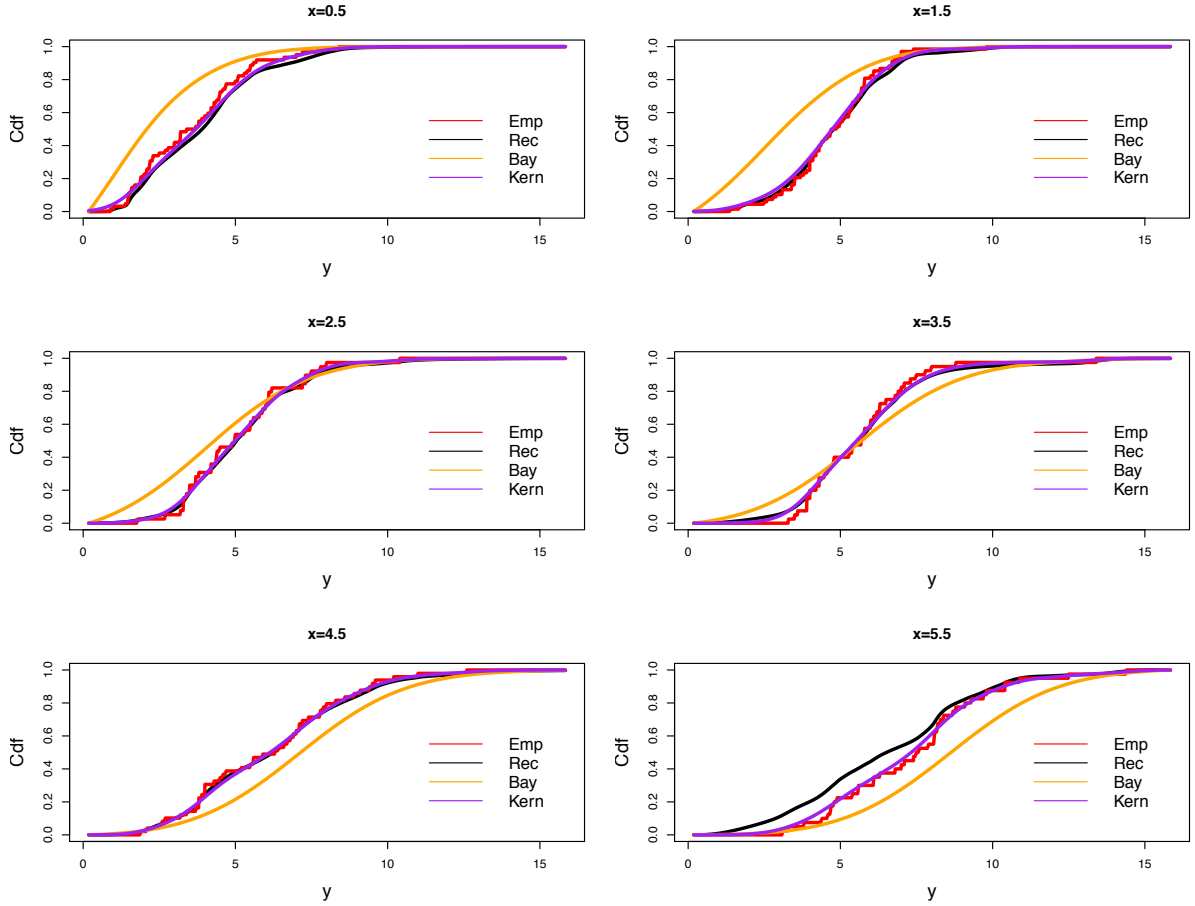


Figure 3.6: IgG distribution $P(y|x)$ for the six age groups $(x = 0.5, x = 1.5, x = 2.5, x = 3.5, x = 4.5, x = 5.5)$ estimated through the four procedures: recursive estimate RR (Rec, black), single-p DDP of De Iorio et al. (2004) (Bay, orange), kernel regression of Aitchison and Aitken (1976) (Kern, purple) and empirical CDF (Emp, red).

of the form

$$f^{(a)}(y|G^{(a)}) = \int_{\Theta} N(y|\theta, \sigma^2) dG^{(a)}(\theta), \qquad (3.11)$$
$$f^{(b)}(y|G^{(b)}) = \int_{\Theta} N(y|\theta, \sigma^2) dG^{(b)}(\theta),$$

where $y \in \mathcal{Y}$ and the joint prior on $G^{(a)}$ and $G^{(b)}$ is such that

$$(G^{(a)}, G^{(b)}) \sim BiDP(c, r, G_0),$$

where $BiDP(c, r, G_0)$ denotes the bivariate Dirichlet Process, introduced in Walker and Muliere (2003). Note that the mixtures above fit into the general framework (3.2): we have fixed $k(y|x, \theta) \equiv k(y|\theta)$, and x indexing G_x in (3.2) determines whether we are considering group (a) or (b).

We assume the sample is collected as $(y_n, x_n)_{n=1}^N$, with x_n denoting which one of the two groups an observation y_n has been sampled from. Clearly there are many ways to estimate the densities above, or the corresponding CDFs. If we were able to associate a numerical value to each group, RR (3.10) would already be a good candidate. Although note that if x is simply a numerical label, RR could not be used because we are using a notion of distance in the algorithm, see (3.9). A possible alternative could also be the method due to De Iorio et al. (2004). The new algorithm can be applied to a setting where the covariate is categorical. For example, we collect a continuous response from a group of male and one of female.

The Bayesian predictive update of the bivariate Dirichlet Process given (y_1, x_1) is the building block of the new algorithm. In order to continue we proceed as follows: we first review the bivariate Dirichlet Process in the next subsection. The rest of the section completely mirrors Section 3.1: we build the algorithm, show how to implement it and its numerical convergence (3.2.2). The large sample behaviour is studied in Chapter 5.

3.2.1 The Bivariate Dirichlet Process

Let $G^{(a)}$ and $G^{(b)}$ be two random probability measures on $(\Theta, \mathcal{A})$. Walker and Muliere (2003) fix two desiderata for a prior on $(G^{(a)}, G^{(b)})$: first, $G^{(a)}$ and $G^{(b)}$ are marginally $\mathrm{DP}(c, G_0)$ distributed; second, the dependence is such that for all sets $\mathrm{A} \in \mathcal{A}$

$$Corr(G^{(a)}(A), G^{(b)}(A)) = \rho > 0, \tag{3.12}$$

with ρ a parameter that describes the dependence between the random distributions, and $Corr$ the linear correlation. The bivariate Dirichlet Process satisfies this two criteria. Its construction makes an elegant use of the Dirichlet-multinomial point process introduced by Lo (1986).

Definition 2. *Let $G \sim DP(c\,G_0)$, with $c > 0$ the concentration parameter, G_0 the base measure, and let r be a non negative integer. A point process N on the real line is said to be Dirichlet-multinomial with parameters (c, r, G_0) if for any k and any partition $(B_1, \ldots, B_k)$ of the real line*

$$(N(B_1), \ldots, N(B_k)) \sim \mathcal{M}(r; G(B_1), \ldots, G(B_k)),$$

where $\mathcal{M}(n; p_1, \ldots, p_k)$ denotes the Multinomial distribution with parameters $n; p_1, \ldots, p_k$
.

It is easy to see that the joint distribution $\pi(G, N)$ is $\mathrm{DP}(c, G_0)\,\mathcal{M}(N; r, G)$. The

bivariate Dirichlet Process ($BiDP$) construction is given extending the above

$$\pi(G^{(a)}, N, G^{(b)}) = \mathrm{DP}(c, G_0)\, \mathcal{M}(N; r, G)\, \mathrm{DP}(c + r, G_r), \tag{3.13}$$

where

$$G_r = \frac{cG_0 + N}{c + r}.$$

It is easy to verify that (3.13) implies both (3.12) and equal marginal DP distributions. Simple calculations allow us to write r as a function of ρ: $r = c\rho/(1-\rho)$.

A key result is that both $G^{(a)}$ and $G^{(b)}$ are conditionally conjugate. Suppose that we sample n observations from $G^{(a)}$ and m from $G^{(b)}$. The posterior of $G^{(b)}$ is

$$G_m^{(b)} | N, y_1^{(b)}, \dots y_m^{(b)} \sim \mathrm{DP}\left(c + m + r, \frac{cG_0 + m\widehat{G_m^{(b)}} + N}{c + n + r} \right), \tag{3.14}$$

where $\widehat{G_m^{(b)}}$ is the empirical distribution function of $(y_1^{(b)}, \dots y_m^{(b)})$. Note that (3.13) implies that the posterior of N carries information on $(y_1^{(a)}, \dots y_n^{(a)})$. Indeed, it holds that

$$\mathrm{E}[N(A) | G_n^{(a)}] = rG_n^{(a)}(A),$$

for all $A \in \mathcal{A}$. The equation above makes explicit how the samples from one group enters into the posterior distribution of the other one. A symmetric argument can be given for $G_n^{(a)}$.

One final key remark. The posterior (3.14) suggests that, as $n \to \infty$, the effect of the dependence chosen a priori becomes less relevant. It is reasonable to think so. The more the right category is observed, the less we care about the other. A nice formalisation of this remark is given in the paper through a joint Polya urn scheme. See Walker and Muliere (2003) for more details.

3.2.2 A Two-step Recursive Algorithm

3.2.2.1 Algorithm Construction

We consider the mixture model defined in (3.11). Note that the prior predictive distribution is the same for the two groups because $G^{(a)}$ and $G^{(b)}$ have equal marginal distributions. We assume to be observing streaming data. The pedix indexes the order at which I am observing the data. Suppose that one observes (y_1, x_1) with $x_1 = 1$, meaning that y_1

belongs to group (a). One can easily update the prior predictive distribution, which we denote by $P_0^{(a)}$, using HMW

$$P_1^{(a)}(y) = (1-\alpha_1)P_0^{(a)}(y) + \alpha_1 Z(y, P_0^{(a)}, y_1),$$

where

$$Z(y, P^{(a)}, Y) = \Phi\left(\frac{\Phi^{-1}(P(y)) - \rho\Phi^{-1}(P(Y))}{\sqrt{1-\rho^2}}\right),$$

with $\rho \in (0,1)$. The update Z is exactly the one defined in Subsection 2.2.2. It is defined here on the distribution function scale.

To compute $P_1^{(b)}$, we want to exploit the dependence between the two groups given by the Bivariate Dirichlet Process. The $BiDP$ gives us a way to compute the posterior distribution $G_1^{(b)}$, and consequently the corresponding predictive distribution, which we can then use to calculate $P_1^{(b)}$. Given G_0, the base measure of the $BiDP$, we want to compute $\mathrm{E}[G^{(b)}|y_1, x_1]$. Recall that the posterior (3.14) is given conditionally on the latent variable N, which needs to be integrated out. The construction (3.13) allows us to compute the expected value $E[N|G_1^{(a)}]$ and $\mathrm{E}[G_1^{(a)}]$ and we will make use of it. Now

$$\begin{aligned} p_1^{(b)} &= \int N(y|\theta,\sigma^2)\mathrm{dE}[G^{(b)}|y_1,x_1] \\ &= \int N(y|\theta,\sigma^2)\frac{c}{c+r}\mathrm{d}G_0(\theta) + \int N(y|\theta,\sigma^2)\frac{r}{c+r}\mathrm{dE}[G_1^{(a)}(\theta)] \\ &= \int N(y|\theta,\sigma^2)\frac{c}{c+r}\mathrm{d}G_0(\theta) + \frac{r}{c+r}p_1^{(a)} \\ &= \frac{c}{c+r}p_0^{(b)} + \frac{r}{c+r}\,p_1^{(a)}. \end{aligned} \tag{3.15}$$

The last step can be rewritten on the distribution scale

$$P_1^{(b)} = \frac{c}{c+r}P_0^{(b)} + \frac{r}{c+r}P_1^{(a)}.$$

The update $P_1^{(b)}$ exploits the prior expected dependence given by the $BiDP$. The recursive algorithm we propose has the HMW algorithm as the first step: when we observe the right group there is no reason to change an estimator that works really well. The second step will be given by something that mimics $P_1^{(b)}$. Note that the two steps are sequential: we need to have calculated $P_1^{(a)}$ to obtain $P_1^{(b)}$.

In order to propose a recursive algorithm, the user-supplied sequence of weights plays once more a fundamental role. The goal is to be consistent with the idea that, as $n \to \infty$ and $m \to \infty$, the observations from the other group matter less and less; this is

consistent with the interpretation of the posterior distribution (3.14). We introduce a second sequence of weights (β_n), with $\beta_n \in (0,1)$. A way to mirror (3.14) is to impose that β_n goes to zero at a faster rate than α_n. Doing so, the second step becomes irrelevant much earlier along the sequence and the algorithm maintains the interpretation of the posterior (3.14). This discussion will be formalised in Chapter 5 in which we study the large sample properties. We define the new algorithm below.

Two Steps Predictive Recursion (TSPR). Choose two initial guesses $P_0^{(a)}$, $P_0^{(b)} \in \mathcal{P}$, the correlation parameter $\rho \in (0,1)$, two decreasing sequences of weights (α_n) and (β_n) taking values in $(0,1)$ such that $\beta_n/\alpha_n \to 0$ as $n \to \infty$. Given observations $(y_n, x_n)_{n=1}^N$, the estimates $P_N^{(a)}(y)$ and $P_N^{(b)}(y)$ are given by

If $x_n = 1$, first compute

$$P_n^{(a)}(y) = (1-\alpha_n)\, P_{n-1}^{(a)}(y) + \alpha_n Z(y, P_{n-1}^{(a)}, y_n), \tag{3.16}$$

then

$$P_n^{(b)}(y) = (1-\beta_n)\, P_{n-1}^{(b)}(y) + \beta_n P_n^{(a)}(y), \tag{3.17}$$

else, if $x_n = 0$, use (3.16) *for* $P_n^{(b)}(y)$ *and* (3.17) *for* $P_n^{(a)}(y)$.

Repeated recursively for $n = 1, ..., N$.

TSPR shares most the properties of the related procedures. It is order-dependent, fast to compute, and easy to implement. We will see that it is also numerically accurate and theoretically convergent. The use of two sequential steps is its striking feature. It is obviously nothing new when we look at the broader literature of recursive procedures - see for example Ishikawa (1974) - but it is a first example among the algorithms we are employing in the thesis. How to implement the algorithms (3.16)-(3.17) has already been largely covered in the thesis: see Subsections 2.2.3 and 3.1.3.2.

3.2.2.2 Numerical Examples

We conduct simulation studies to investigate the numerical properties of TSPR. We want to show that the algorithm converges numerically. Although TSPR handles a categorical covariate that can assume two values, in this simulation study we sample data from regression models where the covariate is numerical, which we randomly generate from a uniform discrete. The numerical labels of the two groups have been chosen to have mean parameters comparable to the simulation studies in a 2015 unpublished report by Hahn, Martin and Walker (arXiv:1508.07448v3), where

the HMW algorithm has been first proposed. The parameters of the HMW algorithm are set as follows: $\alpha_n = (n+1)^{-4/5}$, $\rho = 0.9$. The sequence (β_n) is set as $\beta_n = (1+n)^{(-1.01)}$.

Example 4. We simulate data from the following model with a covariate in $\mathcal{X} := \{0.6, 1.2\}$

$$Y_n|x_n \overset{ind}{\sim} wN(y|\beta_1 x_n, \sigma_1^2) + (1-w)N(y|\beta_2 x_n, \sigma_2^2),$$

with $w = 1/2$, $\beta_1 = -1.5$, $\beta_2 = 2$, $\sigma_1^2 = 1.3^2$, $\sigma_2^2 = 1$. We label the observations sampled with $x = 0.6$ as group (a), group (b) if $x = 1.2$. We sample 1500 observations, $\#x_{(a)} = 736$, $\#x_{(b)} = 764$. $P_0^{(a)}(y) = P_0^{(b)}(y) = N(y|0, 3)$.
Figure 3.7 plots the TSPR estimate (black line), the true distribution (blue line) and P_0(green line). The true CDFs are perfectly recovered.

Example 5. We consider a new regression model and simulate data from it,

$$Y_n|x_n \overset{ind}{\sim} N(y|x_n, x_n),$$

where the covariate takes value in $\mathcal{X} := \{0.6, 1.2\}$. We label the observations sampled with $x = 0.6$ as group (a), group (b) if $x = 1.2$. We sample 1000 observations, $\#x_{(a)} = 486$, $\#x_{(b)} = 514$. $P_0^{(a)}(y) = P_0^{(b)}(y) = N(y|0, 3)$ for both categories.
Figure 3.8 plots the TSPR estimate (black line), the true distribution (blue line) and P_0 (green line). The true CDFs are perfectly recovered.

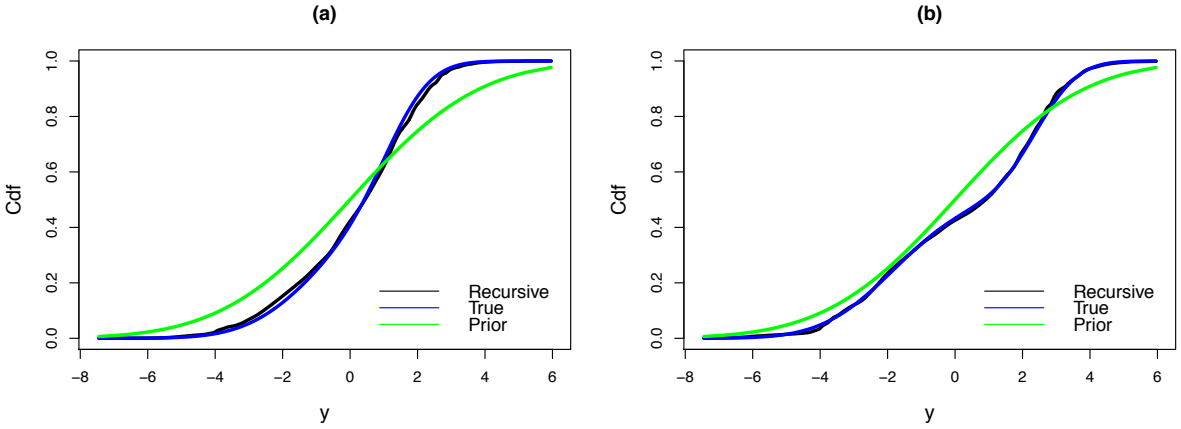


Figure 3.7: TSPR estimates of the predictive distributions $P^{(a)}$ (Panel (a)) and $P^{(b)}$ (Panel (b)) in Example 4: the prior guess P_0 (green), the TSPR estimate (black) and true distribution (blue).

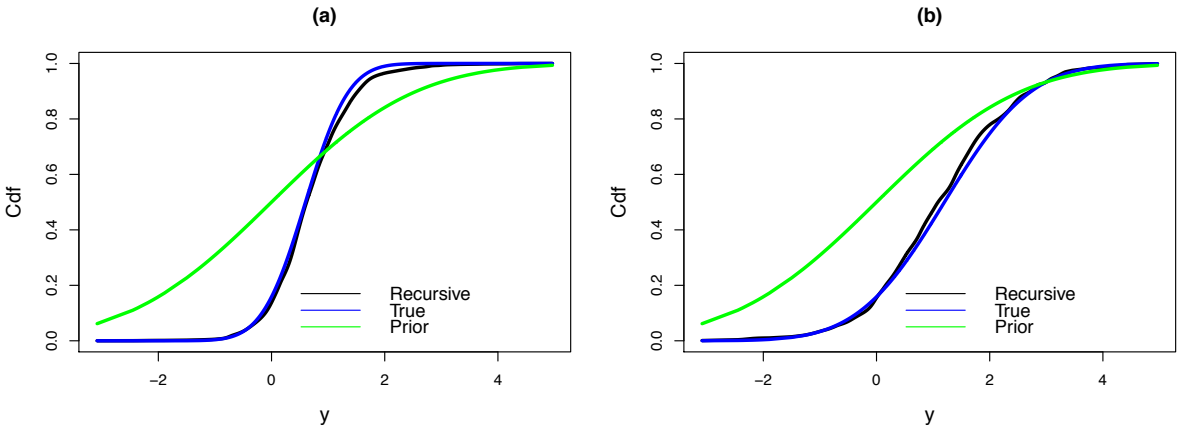


Figure 3.8: TSPR estimates of the predictive distributions $P^{(a)}$ (Panel (a)) and $P^{(b)}$ (Panel (b)) in Example 5: the prior guess P_0 (green), the TSPR estimate (black) and true distribution (blue).

3.3 Discussion

Recursive procedures allow a fast and accurate estimation of statistical quantities. In this thesis we are focusing of a particular family of algorithms : those ones that mimics the Bayesian predictive update suggested by some given models, for example NQZ and HMW. In this chapter we have moved beyond the estimation of a single CDF and proposed two recursive procedures to estimate a finite set of covariate-dependent distributions, assuming data are sampled from different groups and one want to exploit the borrowing of information across groups.

Both estimators are inspired by a Bayesian procedure and carry some of the desirable features of Bayesian methods; such as the borrowing of information from one distribution to another. Furthermore, they exploit the inherent sequential nature of a Bayesian model. We have displayed through several illustrations the good numerical properties of the algorithms. Theoretical guarantees are also available and are going to be discussed in Chapter 5.

Numerical performance is highly competitive with competing models. Both the recursive regression and the two-step predictive recursion are very fast, they require a small amount of coding and minimal CPU requirement. As they are calculated numerically on a grid, which translates into several cell by cell primitive operations (e.g. addition, multiplication), they are both largely parallelisable. Indeed one can simply assign portions of the grid to different CPU cores. We did not exploit this feature in the present chapter, but it could be a source of investigation in the future. The possibility of parallelisation is especially relevant for the recursive regression: as the

number of covariates and/or covariates' values, the computational burden goes up linearly.

To our opinion, the recursive regression algorithm is the one that has more potential for developments, both in term of applications, theory and possibilities to extend it. The reason is simply because it allows the covariate x to assume an arbitrary large number of values, whereas TSPR is restricted to two groups. An obvious extension of RR is ANOVA. Another interesting application for the recursive regression is quantitative risk assessment - for example health response to different dose levels. Canale et al. (2017) is a very recent working paper trying to tackle this problem, showing that it is still much relevant. To deal with quantitative risk assessment the model is a bit more complicated, but RR seems to fit well in this context.

The most interesting direction of investigation for future work is to enlarge/change the assumptions on the space of covariates. Ideally dealing with x continuous is of great interest, although the problem is difficult. RR loses a lot of accuracy when x is continuous. Intuitively, the function $\rho(\cdot, \cdot)$ should play a fundamental role. A problem with a naive use of RR in a continuous covariate setting is that we are not estimating a regression function. The most interesting aspect of the TSPR algorithm is showing that it is possible to be accurate using two recursive algorithms that interact. If it was too ambitious to have one recursion carrying information on the conditional distributions, regression function etc., a multi-steps algorithm would be worth exploring. DPMs of linear regressions carry a lot of information not used in RR, so the same model may be a good starting point. Other BNP regressions could also be the starting point to develop new recursive algorithms.

Chapter 4

Laplace Inversion for Multivariate Probability Distributions

This is the most applied chapter of the thesis. We deal with an important problem: the inversion of the Laplace transform of a multivariate distribution. We show that a suitable adapted version of the NQZ algorithm (Newton et al. 1998) is very efficient and competitive with alternative methodologies. Using a statistical method for an inverse problem has also some notable advantages which we will emphasise across the chapter; these advantages are well understood in the literature (Stuart 2010).

We start by explaining the underlying motivation of the chapter. The rest of the chapter is organized as follows: Section 4.2 gives an overview of current methodologies available in the literature. Several alternatives are listed in this chapter due to its strongly numerical nature. Section 4.3 explains how to apply the predictive recursion in this problem and how to solve some numerical problems that arise at dimension higher than one. Section 4.4 includes four illustrations in which we compare our procedure to Gaver-Stehfest algorithm. It presents inversions of two and three dimensional distributions. Finally, we conclude in Section 4.5 with a brief discussion. The content of this chapter is published in Cappello and Walker (2018).

4.1 Motivating Application

Inverting the Laplace transform of a probability distribution function is an important problem in many scientific areas, motivated by cases where the Laplace transform has a simple form yet the underlying probability distribution does not. Much of the literature and theory deal with the one dimensional setting. However, there are many applications, such as aggregate claims in actuarial science (Jin and Ren 2010, Goffard

et al. 2015), queuing problems in computer systems (Choudhury et al. 1994, Gelenbe 2006) and finance, such as options pricing (Fusai 2004, Cai and Kou 2012), which require multivariate inversions. It is also a relevant problem in statistics when we assume a model described solely through a parametric family of Laplace transforms; for example the positive stable distribution (Feller 1971).

Many of the inversion methods for the one dimensional case can be extended to the multivariate setting. However, a concern with many current one dimensional solutions is that they do not guarantee a numerical estimate which is a probability distribution function. Solutions can be numerically unstable, for example, decreasing in places or failing to reach 1 at the upper limit of the support of the distribution. This problem can be compounded as the dimension increases. We argue, therefore, that a solution which provides a numerical estimate which is a distribution function is essential.

How a univariate procedure is extended to multiple dimensions is also crucial. A naive reiteration of the univariate algorithm does not usually work in practice, at least for dimensions higher than two. The quality of the approximation given by the polynomial expansions, on which most of the alternatives rely upon, deteriorates, while computational costs increase exponentially. Other numerical approximations employed in the literature, such as quadrature integration and differentiation, also suffer from the curse of dimensionality.

In the one dimensional case, Walker (2017a) shows that this specific inverse problem can be treated as a statistical inference problem. The algorithm proposed in this work mimics NQZ and was demonstrated to be a highly competitive one, which guaranteed a numerical solution as a distribution function. In fact, the recursive procedure is a statistical estimator of a distribution. The idea of the chapter is to extend the methodology to cover the multivariate setting. The goal is to extend the benefits seen for univariate to distributions to any dimension. Doing so, we tackle a few numerical issues that arise.

4.2 Literature Review

The Laplace transform of the distribution function $G(\underline{\theta})$ with density function $g(\underline{\theta})$, and $\underline{\theta} \in (0, \infty)^d$ for integer $d \geq 1$, is given by

$$F(\underline{y}) = \int_{[0,\infty)^d} e^{-\underline{\theta}\ \underline{y}}\, g(\underline{\theta})\ \mathrm{d}\underline{\theta}, \tag{4.1}$$

where $\underline{y} \in (0,\infty)^d$.

The vast majority of inversion methods are based on polynomial approximations. A broad class of the methodologies share a common style of solution and Abate and Whitt (2006) present a unified framework; in the bivariate case a numerical solution takes the form

$$g_{n,m}(\theta_1,\theta_2) = \frac{1}{\theta_1\,\theta_2}\sum_{l=0}^{m}\sum_{k=0}^{n}\zeta_l\,\omega_k F\left(\frac{\alpha_l}{\theta_1},\frac{\beta_k}{\theta_2}\right), \quad \theta_1,\theta_2 > 0. \tag{4.2}$$

Here (n,m) represent the level of accuracy.

The idea is that $g_{n,m}(\theta_1,\theta_2)$ can be seen as the result of a two-step process, where a one-dimensional procedure is applied twice. The methodologies differ in the way that the parameters $(\zeta,\omega,\alpha,\beta)$ are defined. Among the well established methods, we have the *Gaver- Stehfest* algorithm, see Gaver Jr (1966) and Stehfest (1970); the *Fourier* series method with Euler summation, see Dubner and Abate (1968); and the *Talbot* algorithm, see Talbot (1979). The latter technique is based on the Bromwich contour integral.

Abate and Whitt (2006) suggest that a combination of two algorithms, along with two different truncation orders, might be the most accurate strategy. The two-dimensional Gaver-Stehfest algorithm, which we will compare with our own method, has $n = 2m$, with $\alpha_k = \beta_k = k\log 2$, $\omega_0 = 0$ and, for all $k \geq 1$,

$$\omega_k = \zeta_k = \log 2(-1)^{k+n/2} \times \sum_{j=\lceil (k+1)/2\rceil}^{\min\{k,n/2\}} \frac{j^{n/2}(2j)!}{(n/2-j)!j!(j-1)!(k-j)!(2l-k)!}\,.$$

Note that $\alpha_k = \beta_k$ and $\omega_k = \zeta_k$ for all k since we are applying the same algorithm to both the inner and outer loops. If we were to apply an alternative procedure, let us say only to the outer loop, for example, we would need to redefine β_k and ζ_k.

For the Euler method, $\alpha_k = \beta_k = [M\ln 10]/3 + \pi i k$, $\omega_k = \zeta_k = (-1)^k\epsilon_k$, with

$$\epsilon_0 = \tfrac{1}{2}, \quad \epsilon_k = 1 \quad \text{for} \quad 1 \leq k \leq M, \quad \epsilon_{2M} = \tfrac{1}{2}^M,$$

and

$$\epsilon_{2M-k} = \epsilon_{2M-k+1} + 2^{-M}\binom{M}{k}, \quad \text{for} \quad 0 < k < M.$$

With β_k being a complex number, we take the real part of the Laplace transform when estimating the density.

The Talbot algorithm would have $\beta_0 = 2M/5$, $\zeta_0 = e^{\beta_0}/5$

$$\alpha_k = \beta_k = \frac{2k\pi}{5} \cot\left(\frac{k\pi}{M} + i\right), \quad \text{for} \quad 1 < k < M,$$

and

$$\omega_k = \zeta_k = \left\{1 + i\left[\frac{k\pi}{M}\right]\left[1 + \left\{\cot\left(\frac{k\pi}{M}\right)\right\}^2\right] - i\cot\left(\frac{k\pi}{M}\right)\right\}\frac{2}{5}e^{\beta_k}, \quad 1 < k < M.$$

Here both coefficients are complex.

Despite having been widely studied and used in applications, little is known about the convergence properties of these algorithms. Convergence is given in the univariate case and the theory can be extended when the univariate algorithms are applied iteratively. Results are in Kuznetsov (2013) for the *Gaver- Stehfest* and Weideman (2006) for the *Talbot*, while Dubner and Abate (1968) discuss convergence of the underlying Fourier series.

There are a number of other key classes of inversion procedures. Choudhury et al. (1994) extends the work of Abate and Whitt (1992) to multivariate settings; the inversion here is based on a Fast Fourier Transform (FFT). Another popular class of method is based on a Laguerre series representation: Abate et al. (1998) present a multivariate version; a more recent method employing Laguerre Polynomials is given by Mustapha and Dimitrakopoulos (2010). Finally, moment–based methods are studied and a recent paper following this approach can be found in Mnatsakanov (2011), who extends the univariate procedure in Mnatsakanov and Ruymgaart (2003). The methodology is somewhat related to our own: the target function is seen as the mixing distribution of a Binomial distribution. However, it is restricted to distributions on $[0, 1]$.

Some relevant contributions have appeared into the actuarial science literature, see in particular, Jin and Ren (2010) and Jin et al. (2016). The former is based on a FFT algorithm. The authors show how to use the tilting method introduced by Grübel and Hermesmeier (1999) in a multivariate context. Tilting is an efficient way to reduce the error associated with FFTs. The latter is a moment–based method.

Goffard et al. (2015) target a bivariate probability distribution. It employs a clever decomposition of the target density - the polynomials are orthogonal with respect to a well-studied class of probability measures (natural exponential family with quadratic variance function, Morris 1982). The polynomials are then easily computable. It is

interesting that their method is related to both Abate et al. (1998), Mnatsakanov (2011) and Jin et al. (2016). We refer to the paper for a more detailed discussion.

A drawback of both Abate et al. (1998) and Goffard et al. (2015) is that they require the calculation $(n+1) \times (m+1)$ partial derivatives of order ranging from 0 to $n+m$. Abate et al. (1998) show how to get approximate partial derivatives using Cauchy contour integration. Goffard et al. (2015) argue that through specific parametrization of the densities f_{ν_1} and f_{ν_2} it is possible to simplify the expression. While this seems to be true for insurance applications, the prime interest of the authors, the claim does not hold for several examples of interest in statistics, such as the positive stable distribution.

4.3 A Multidimensional Inversion Method

In a spirit similar to Walker (2017a), we discuss how to treat the Laplace inversion as a inferential problem in the multivariate setting . The key intuition is to recognize that the Laplace transform density; i.e.

$$f(\underline{y}) = \frac{\partial^2 F(\underline{y})}{\partial y_1 \dots \partial y_d},$$

given in (4.3), is the model we have widely discussed in this thesis: it is a mixture model with independent gamma kernels

$$f(\underline{y}) = \int_{[0,\infty)^d} \left\{ \left(\prod_{i=1}^{d} \theta_i \right) e^{-\sum_{i=1}^{d} \theta_i \, y_i} \right\} g(\underline{\theta}) \, \mathrm{d}\underline{\theta}. \tag{4.3}$$

Under the assumption we are able to sample from $f(\underline{y})$, NQZ is a good candidate to estimate the mixing density function $g(\underline{\theta})$. The procedure fits well in this context. It is generally really fast and accurate. Since we are dealing with a problem where we have a theoretically unlimited supply of observations the accuracy should be guaranteed. Walker (2017a) shows the positive performance in the univariate case. Here we describe and illustrate the extension to the multivariate case.

We redefine NQZ below and apply it to our multivariate problem. It is a bit repetitive but it allows to make a quick reference to some components of the algorithm that need to be discussed. There are indeed a couple of numerical issues related to the multivariate implementation. Given deterministic sequence of weights (α_n) and $i.i.d.$ samples $(\underline{y}_n)_{n=1:N}$

from $f(\underline{y})$, an initial estimate $g_0(\underline{y})$ is updated iteratively through

$$g_n(\underline{\theta}) = g_{n-1}(\underline{\theta}) \left[1 - \alpha_n + \alpha_n \frac{\prod_{i=1}^{d} \theta_i e^{-\underline{\theta}\,\underline{y}'_n}}{\int_{[0,\infty)^d} \prod_{i=1}^{d} \theta_i e^{-\underline{\theta}\,\underline{y}'_n}\, g_{n-1}(\underline{\theta})\ \mathrm{d}\underline{\theta}} \right], \tag{4.4}$$

for $n = 1, \ldots, N$; for some sufficiently large n. A useful starting point is $g_0(\underline{\theta})$ as a product of d independent gamma density functions. The evaluation of (4.4) relies on a grid of points $(\underline{\theta}_j)_{j=1}^{M}$ and the recursion (4.4) is repeated n times for each point of the M^d array. The distribution version, which will form the basis of the convergence theory, is given by

$$G_n(\underline{\theta}) = (1 - \alpha_n)\, G_{n-1}(\underline{\theta}) + \alpha_n \frac{\int_{\mathbf{0}}^{\underline{\theta}} \prod_{i=1}^{d} \theta_i e^{-\underline{\theta}\,\underline{y}'_n}\ \mathrm{d}G_{n-1}(\underline{\theta})}{\int_{[0,\infty)^d} \prod_{i=1}^{d} \theta_i e^{-\underline{\theta}\,\underline{y}'_n}\ \mathrm{d}G_{n-1}(\underline{\theta})}. \tag{4.5}$$

Numerical errors arise both when evaluating the integral in the denominator of (4.5) and the numerator of (4.5). The denominator is particularly critical for the procedure at a dimension higher than one. In this respect, the choice of the grid is important; a large M is needed in order to calculate

$$\mathbb{I}_n = \int_{[0,\infty)^d} \prod_{i=1}^{d} \theta_i\, e^{-\underline{\theta}\,\underline{y}'_n}\, g_{n-1}(\underline{\theta})\ \mathrm{d}\underline{\theta}, \tag{4.6}$$

regardless the numerical approximation employed. Numerical evaluation of (4.6) constitutes a major challenge of the algorithm in high dimensional settings. For a 2–dimensional problem it is practical to use the trapezoidal rule. To simplify the notation, denote by (s, t) the two coordinates and by $[s_k, t_j]_{1 \le k,j \le M}$ the grid, (4.6) can be approximated by

$$\widehat{\mathbb{I}}_n = \tfrac{1}{4} \sum_{k=1}^{M} \sum_{j=1}^{M} (s_k - s_{k-1})(t_j - t_{j-1}) s_k t_j e^{-x_{n,1} s_k - x_{n,2} t_j} [g_{n-1}(s_k, t_j) + g_{n-1}(s_{k-1}, t_{j-1})]. \tag{4.7}$$

The underlying assumption in (4.7) is that s_M and t_M are a good substitutes for ∞ as upper extremes in the integral. Alternative numerical approximation rules, such as Simpson's 3/8 rule, have comparable performance.

It is well known that the efficiency of the approximation (4.7) decreases as the dimensionality increases. Alternative quadrature methodologies also suffer the same "curse of dimensionality". In additional to this, the computational complexity increases to the power of the dimension. However, $\mathbb{I}_n$ can be alternatively evaluated via Monte Carlo integration. A key property of Monte Carlo methods is that the convergence rate is independent of the dimensionality of the integral (Hammersley and Handscomb 1964). We can exploit the mixture model construction; $\underline{\theta} e^{-\underline{\theta}\,\underline{y}'_n}$ is proportional to the kernel of

a multivariate gamma density, and we can sample independently $(\theta_1, \ldots, \theta_d)$ from d independent gamma distributions with a common shape parameter equal to two and scale parameters $(y_{n,1}, \ldots, y_{n,d})$, and approximate

$$\widehat{\mathbb{I}}_n = L^{-1} \sum_{l=1}^{L} \frac{g_{n-1}(\theta_{1,l}, \ldots, \theta_{d,l})}{(y_{n,1} \ldots y_{n,d})^2}, \tag{4.8}$$

for some suitably large L.

This method can be unstable when at least one of the $y_{n,j}$ is too small or, to a lesser extent, when $x_{n,j}$ are all too high. More specifically, the issue here is that we are sampling observations which do not lie in probability where g_n concentrates its mass. A solution that appears to solve this problem is importance sampling. Importance sampling makes possible to decide where to concentrate more mass. An efficient importance sample distribution could be

$$q(\underline{\theta}) = (1 - w) \times g_0(\underline{\theta}) + w \times \prod_{i=1}^{d} \mathrm{Ga}(\theta_i | 2, y_{n,i})$$

for some $0 < w < 1$; e.g. $w = \frac{1}{2}$. To elaborate on our proposal $q(\underline{\theta})$: the first component allows to sample from areas of the support where g_n concentrates mass; the second one ensures that enough importance weights stay away from zero. The proposal $q(\underline{\theta})$ aids the full exploration of the support. The underlying assumption is that g_0 is a good guess. The assumption about g_0 is reasonable, because in practice it could be based on a number of primary iterations: even if the algorithm has not converged yet, the estimate should already be closer to the correct density. If we were to choose $q(\underline{\theta}) = g_0$, the risk would be to have only a few $\underline{\theta}_l$ (those in the tails of g_0), absorbing all the importance weights.

An alternative way to better explore the support quasi Monte Carlo sampling; see Niederreiter (1992) for an introduction. It can be combined with importance sampling. Quasi Monte Carlo integration is a deterministic counterpart to Monte Carlo integration, in the sense that the points in which the integrand is evaluated are not random but deterministic. The deterministic points are generated through an algorithm which defines a low discrepancy sequence; for example, Halton and Sobol sequences. The idea is that low discrepancy sequences should allow the exploration of the support more thoroughly. It translates into a faster rate of convergence; Quasi Monte Carlo has a rate $\mathcal{O}(N^{-1} \log N^k)$, for some constant k whereas Monte Carlo has a rate $\mathcal{O}(N^{-1/2})$, (Caflisch 1998). If the integrand is smooth, Quasi Monte Carlo integration is more accurate than Monte Carlo, especially when dealing with multidimensional integrals, because the

sample points are more close to being uniformly distributed than those sampled from random generators, (Morokoff and Caflisch 1995).

A further issue using any type of Monte Carlo integration is that g_n is known up to a finite number of points. A solution to this problem is interpolation. For example, in the 2 dimensional case, from the grid $[s_k, t_j]_{1\leq k,j\leq M}$ we can evaluate $g_n(s,t)$ using linear interpolation. We see that also simply taking the closest point to the sampled value provides good accuracy. Numerical simulations run on gamma distributed random variables suggest that up to 7 dimensions (4.8) can be evaluated without incurring any large approximation errors (max 10% relative error).The experiment mimics Example 4 for increasing dimensions ($\#samples = 100000$). A more rigorous theoretical study could be source of future work.

4.4 Numerical Examples

All the numerical strategies discuss so far in the thesis apply to this chapter. In particular we refer to Subsections 2.2.3 and 3.1.3.2. A grid is required. Monte-Carlo permutation and restart are beneficial. A few further comments are necessary in this setting.

A direct consequence of the use of a grid is that the vast majority of operations of the algorithm (4.5) are standard cell by cell primitive operations, which can be run in parallel. A single iteration can be split into an arbitrary number of cores, having assigned to each core a portion of the grid. Only the evaluation of the integral $\widehat{\mathbb{I}}_n$ needs to collect all the information on a single core, where it is required to sum all the grid values. Since all the calculations are primitive, a useful computation tool is offered by a graphic card. See for example Lee et al. (2010).

To compare our methodology with another, the Gaver–Stehfest, Talbot and Fourier–series methods are the most widely applied inversion algorithms. Abate and Whitt (2006) suggest that a combination of two procedures is the most efficient approach when dealing with a dimension higher than 1. Our experience is that this is not the case for the density functions we consider. Given that the Gaver–Stehfest method involves only real quantities and to us is the most accurate and numerically stable, we select this one for a comparative study, specifically the two–step Gaver–Stehfest algorithm.

All the examples which follow have been implemented in Matlab on a standard

Macbook Pro. We report the running time of the algorithms. Note that it should not be taken as a reference for the speed of convergence. Our aim is to highlight a positive feature of our algorithm: it scales up well with the dimension. This property is important because the methodologies based on polynomial approximations are known to scale poorly in multivariate settings.

The comparison of our procedure with any other methodologies is not of easy interpretation. Our algorithm is stochastic whereas all the alternatives are deterministic. The goal of the comparison is simply to have a reference and investigate in which kind of problems our method is competitive. To avoid making the notation too heavy in the figures, we use s, t and r in lieu of θ_1, θ_2 and θ_3, and x, y and z in lieu of y_1, y_2 and y_3.

Example 1. *Bivariate gamma.* We construct a bivariate gamma density as a mixture of two conditionally independent gamma densities with common shape parameter a and rate parameter γ, which is assigned a standard exponential distribution. Therefore,

$$g(s,t) = \int_0^\infty \frac{\gamma^a}{\Gamma(a)} s^{a-1} e^{-\gamma s} \frac{\gamma^a}{\Gamma(a)} t^{a-1} e^{-\gamma t} e^{-\gamma} \ \mathrm{d}\gamma,$$

and so

$$F(x,y) = \int \frac{\gamma^a}{(x+\gamma)^a} \frac{\gamma^a}{(y+\gamma)^a} e^{-\gamma} \mathrm{d}\gamma,$$

with

$$f(x,y) = \int \frac{a^2 \gamma^{2a} e^{-\gamma}}{(x+\gamma)^{1+a}(y+\gamma)^{1+a}} \ \mathrm{d}\gamma.$$

To sample from this, we take γ a sample from a standard exponential and x and y are sampled from independent Pareto; i.e. $x = \gamma \ (u_1^{-1/a} - 1)$ and $y = \gamma \ (u_2^{-1/a} - 1)$, with u_1 and u_2 independent uniform samples from $(0,1)$. We choose $a = 7$, $g_0 = \mathrm{Ga}(s|2, \frac{1}{4})\mathrm{Ga}(t|2, \frac{1}{4})$, $s_M = t_M = 100$, $M = 500$ equal spacing, $r = 0.7$, $N = 10,000$.

Table 4.1 and Fig. 4.1 summarize the results. For the recursive procedure, we compute the Monte Carlo average of 10 estimates. The computing time is about 2 minutes per run, much faster than the Gaver–Stehfest which takes about 25 minutes. Relative errors are reported within brackets below the estimates.

The precision of the recursive method stands out clearly both numerically and visually. It manages to recover quite successfully also the tails of the true distribution $G(s,t)$. The two dimensional Gaver–Stehfest method is in this case both inaccurate and slow. For both methods, the relative errors are higher close to the origin and in the tails. This may be due to numerical error when interpolating to compute the CDF. The

maximum relative error of the recursive estimate is about 5%, whereas it reaches 10% for the Gaver–Stehfest method.

Example 2. *Bivariate positive stable distribution.* Here we consider two identically independent positive stable distributed variables. The Laplace transform is given in Feller (1971) and is equal to

$$F(x, y) = e^{-bx^a - by^a}$$

with $0 < a < 1$, $b = \gamma^a / \cos(a\pi/2)$ and $\gamma > 0$. The Laplace density $f(x, y)$ is the product of two Weibull univariate densities which we can sample by taking u_1 and u_2 independent samples from a uniform $(0, 1)$, and then take $x = [(-\log u_2)b]^{(1/a)}$ and $y = [(-\log u_2)b]^{(1/a)}$. In the example we consider the same parameters of Ridout (2009); i.e. $a = 0.9$ and $\gamma = 1$. We choose a non-informative starting point, $g_0(s, t) = 1/(s_M t_M)$, $s_M = t_M = 50$, $M = 1,000$ equal spacings, $N = 10,000$, $r = 0.7$.

Our method takes again about two minutes, comparable to the Stehfest which is around one minute (there is not need to compute an integral to account for the dependence). We report here a single run. Fig. 4.2 displays the estimated densities for the two methods. Since we do not know the true density, we use the estimates given by the two methods to compute the Laplace transform, which we do know. A comparison is given in Table 4.2.

The estimated mode of $g_n(s, t)$ is $(6.15, 6.10)$, which is similar to the one dimensional estimate of Ridout (5.95). The Stehfest method yields $(6.05, 6.05)$. The recursive method gives a really smooth estimate, without the need to average over several iterations. On the other hand, it emerges clearly from Fig. 4.2b that the Stehfest estimate is not a density, as it assumes negative values. Our technique gives instead a very smooth density, see Fig. 4.2a.

Example 3. *Downton Bivariate Exponential.* We consider another bivariate distribution: Downton's bivariate exponential distribution (1970). The Laplace transform is defined as:

$$F(x, y) = \frac{\mu_1 \mu_2}{(\mu_1 + x)(\mu_2 + y) - \rho xy}.$$

We sample from the density associated to $F(x, y)$ with rejection sampling: we sample from two independent identically exponential distributed random variables ($\lambda = 0.5$). We choose $\mu_1 = \mu_2 = 0.5$ and $\rho = 0.2$, $g_0(s, t) = \mathrm{Ga}(s|1, 0.2)\,\mathrm{Ga}(t|1, 0.2)$, $s_M = t_M = 25$, $M = 1,000$ equal spacings, $N = 8,000$, $r = 0.7$.

Table 4.1: Cumulative distribution function of a bivariate gamma for selected values.

		s=5	s=10	s=20	s=50	s=80
t=5						
	True	.2058	.2692	.2786	.2791	.2791
	Stehfest	.2034	.2842	.2925	.2932	.2938
		(-.0117)	(.0557)	(.0499)	(.0505)	(.0527)
	Recursive	.1990	.2681	.2848	.2863	.2863
		(-.0330)	(-.0041)	(.0223)	(.0258)	(.0258)
t=10						
	True	.2692	.4418	.5055	.5131	.5131
	Stehfest	.2842	. 4676	.5355	.5411	.5422
		(.0557)	(.0584)	(.0593)	(.0546)	(.0567)
	Recursive	.2707	.4432	.5139	.5256	.5257
		(.0056)	(.0032)	(.0166)	(.0244)	(.0246)
t=20						
	True	.2786	.5055	.6599	.7089	.7105
	Stehfest	.2925	.5355	.6988	.7427	.7450
		(.0499)	(.0593)	(.0589)	(.0477)	(.0486)
	Recursive	.2895	.5198	.6549	.6974	.6985
		(.0391)	(.0283)	(-.0076)	(-.0162)	(-.0169)
t=50						
	True	.2791	.5131	.7089	.8453	.8652
	Stehfest	.2932	.5411	.7427	.8979	.9354
		(.0505)	(.0546)	(.0477)	(.0622)	(.0811)
	Recursive	.2910	. 5317	.6971	.8158	.8444
		(.0426)	(.0363)	(-.0166)	(-.0349)	(-.0240)
t=80						
	True	.2791	.5131	.7105	.8652	.9000
	Stehfest	.2938	.5422	.7450	.9354	.9886
		(.0527)	(.0567)	(.0486)	(.0811)	(.0984)
	Recursive	.2910	.5318	.6980	.8419	.9288
		(.0426)	(.0364)	(-.0176)	(-.0269)	(.0320)

Note: The values of the cumulative distribution functions for the three methods have been calculated by numerical integration of the corresponding density function. The grid is $[0, 100] \times [0, 100]$*, with* $M = 500$ *equal spacings. We have followed this procedure also for the true distribution. The true distribution is a bivariate gamma, with shape* $a = 1$ *and dependent rate parameters* $\theta \sim exp(1)$*. The start* $g_0(s,t) = Ga(s|2, \frac{1}{4})Ga(t|2, \frac{1}{4})$*. Relative errors w.r.t the true between brackets. See example 1 for more details.*

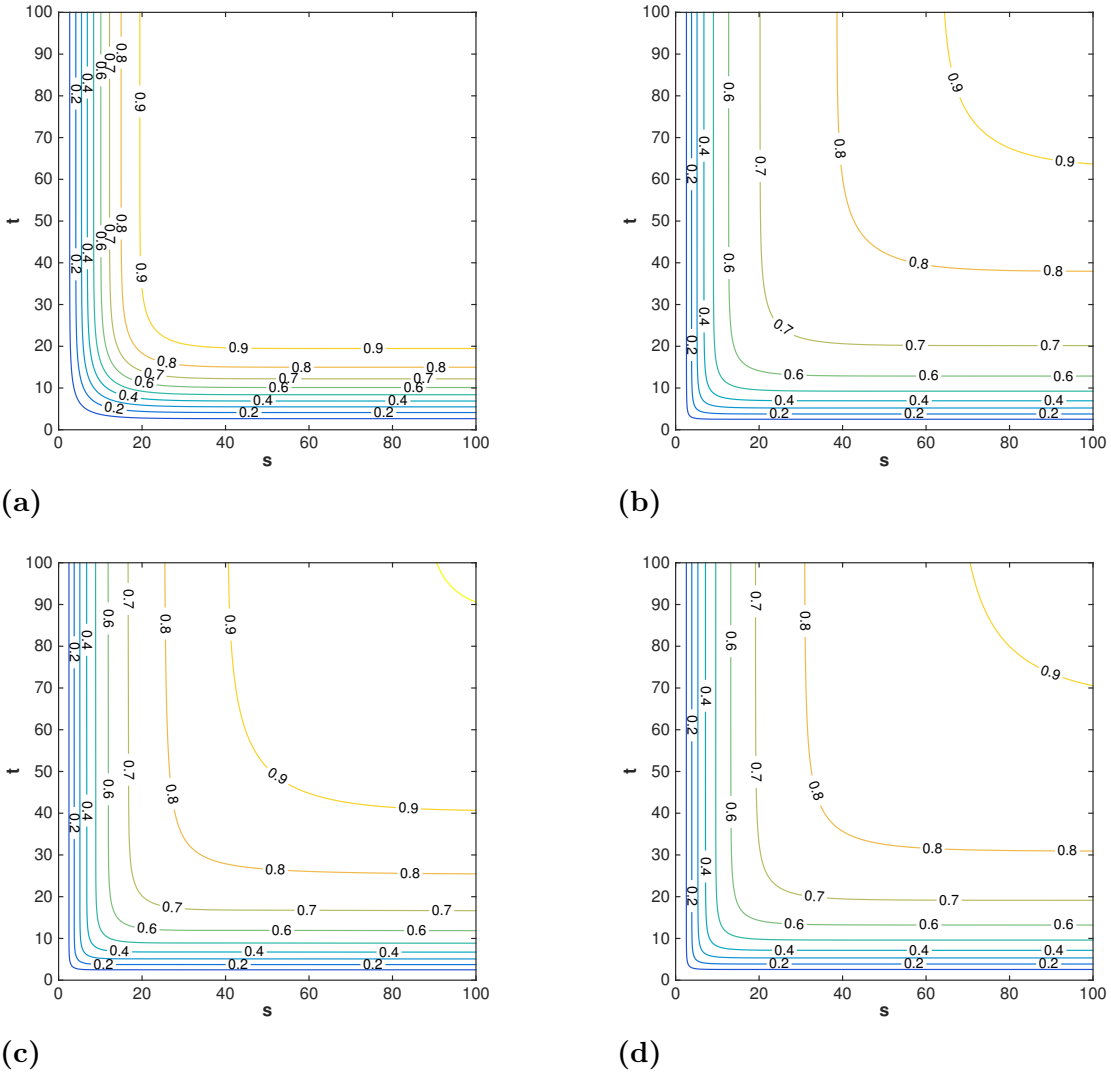


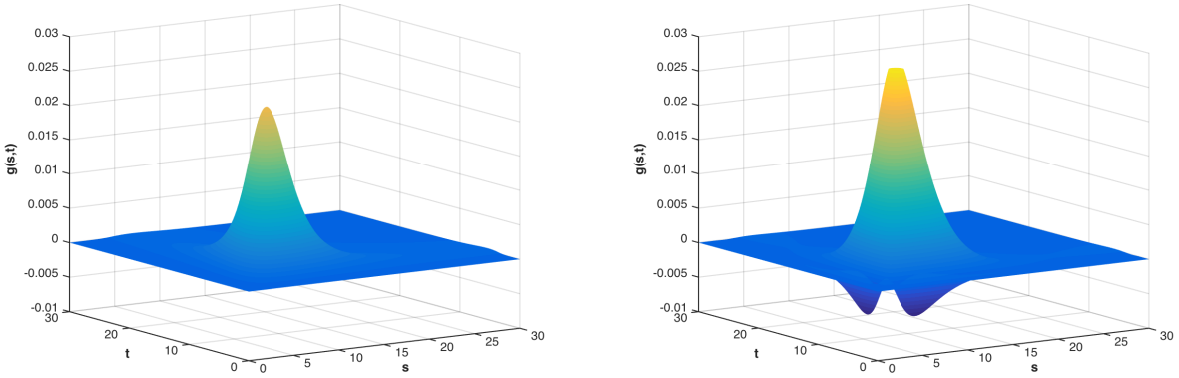
Figure 4.1: Contour plots of the *prior* $G_0(s,t)$ (Panel 4.1a), *estimated* $G_n(s,t)$ (Panel 4.1b), *Gaver–Stehfest* (Panel 4.1c) and the *true* $G(s,t)$ (Panel 4.1d): Example 1

Note: The values of the cumulative distribution functions for the three methods have been calculated by numerical integration of the corresponding density function. The grid is $[0,100]\times[0,100]$*, with* $M=500$ *equal spacings. We have followed this procedure also for the true distribution. The true distribution is a bivariate gamma, with shape* $a=1$ *and dependent rate parameters* $\gamma\sim exp(1)$*. The start* $g_0(s,t)=Ga(s|2,\frac{1}{4})Ga(t|2,\frac{1}{4})$*. See Example 1 for more details.*

Table 4.2: Laplace transform of a bivariate positive stable for selected values (x, y)

		x=0	x=0. 2	x=0.4	x=0.8
y=0					
	True	1	.2267	.0607	.0054
	Stehfest	1	.2281	.0621	.0055
		(0)	(.0062)	(.0231)	(.0185)
	Recursive	1	.2161	.0620	.0072
		(0)	(-.0468)	(.0214)	(.3333)
y=0.2					
	True		.0496	.0135	.0012
	Stehfest		.0520	.0142	.0013
			(.0484)	(.0519)	(.0833)
	Recursive		.0470	.0137	.0016
			(-.0524)	(.0148)	(.3333)
y=0.4					
	True			.0037	0
	Stehfest			.0039	.0003
				(.0540)	(//)
	Recursive			.0040	0
				(.0811)	(0)
y=0.8					
	True				0
	Stehfest				”
	Recursive				”

Note: The values of the Laplace transform for the three methods have been calculated by numerical integration of the corresponding density function. The grid is $[0, 50] \times [0, 50]$*, with* $M = 1,000$ *equal spacings. The true Laplace transform is a bivariate positive stable, with parameters* $a = 0.9$ *and* $\gamma = 1$*. The start* $g_0(s, t) = 1/50^2$*. Relative errors w.r.t the true between brackets. See Example 2 for more details.*



(a) **(b)**

Figure 4.2: Surface plots of the densities *recursive* $g_n(s,t)$ (Panel 4.2a) and *Gaver–Stehfest* $g_N(s,t)$(Panel 4.2b): Example 2
Note: A plot of the true density is not provided as no closed form is available. For the recursive method: sample size $n = 10,000$*, grid* $[0,50]^2$*, with* $M = 1,000$ *number of spacings. The bivariate positive stable is made up of two independent positive stables with parameters* $a = 0.9$ *and* $\gamma = 1$ *(Ridout (2009)). Relative errors w.r.t the true between brackets. See Example 2 for more details.*

We report an average of 10 runs for the recursive algorithm, our algorithm takes about one minutes and half, the Gaver–Stehfest 50 seconds. Table 4.3 collects the results. No figure is given as it is not informative. The Gaver–Stehfest method displays a better performance than the recursive algorithm. The relative error of our procedure remains anyway low (maximum 5%). The poorer performance can be explained by the shape of the exponential: the sample $(x_n, y_n)_{n=1}^N$ is close to zero and might lead to the larger numerical error evaluating $\widehat{\mathbb{I}}_n$. The issue was already discussed in Section 3.

Example 4. *Trivariate gamma distribution.* Here we consider the joint distribution of three Gamma distributed random variables: $g(s,t,r|a,b) = \text{Ga}(s|a,b)\text{Ga}(t|a,b)\text{Ga}(r|a,b)$. Sampling from the Laplace transform density corresponding to $g(s,t,r|a,b)$is straightforward: as it is equivalent to sample a triplet of independent and identically distributed Pareto variables. To our knowledge, this is the only attempt to invert a 3–dimensional distribution available in the literature.

We choose $a = 5$ and $b = 2$, $s_M = t_M = r_M = 30$, $M = 300$ equal spacings, $N = 5,000$, $g_0(s,t,r|a,b) = \text{Ga}(s|2,0.5)\text{Ga}(t|2,0.5)\text{Ga}(r|2,0.5)$. We calculate $\widehat{\mathbb{I}}_n$ with Quasi Monte Carlo integration and use importance sampling as suggested in Section 2 with

$$q(s,t,r) = \tfrac{1}{2}g_0(s,t,r) + \tfrac{1}{2}str\exp(-xs - yt - zr).$$

We use a Halton low–discrepancy sequence to generate $50,000$ points. The parameters

Table 4.3: Cumulative distribution function of a Downton Exponential for selected values.

		s=0.5	s=1	s=4	s=8	s=10
t=0.5						
	True	.0537	.0945	.1987	.2208	.2226
	Stehfest	.0532	.0936	.1971	.2193	.2210
		(-.0093)	(-.0095)	(-.0081)	(-.0068)	(-.0072)
	Recursive	.0521	.0959	.1953	.2130	.2148
		(-.0298)	(.0148)	(-.0171)	(-.0353)	(-.0350)
t=1						
	True	.0945	.1665	.3520	.3921	.3954
	Stehfest	.0936	.1650	.3494	.3898	.3931
		(-.0095)	(-.0090)	(-.0074)	(-.0059)	(-.0058)
	Recursive	.0990	.1647	.3567	.3880	.3914
		(.0476)	(-.0108)	(.0134)	(-.0105)	(-.0101)
t=4						
	True	.1987	.3520	.7597	.8546	.8632
	Stehfest	.1971	.3494	.7575	.8541	.8628
		(-.0081)	(-.0074)	(-.0029)	(-.0006)	(-.0005)
	Recursive	.1975	.3542	.7643	.8489	.8556
		(-.0060)	(.0063)	(.0061)	(-.0067)	(-.0088)
t=8						
	True	.2208	.3921	.8546	.9663	.9768
	Stehfest	.2193	.3898	.8541	.9690	.9798
		(-.0068)	(-.0059)	(-.0006)	(.0028)	(.0031)
	Recursive	.2162	.3887	.8459	.9491	.9670
		(-.0208)	(-.0087)	(-.0102)	(-.0178)	(-.0100)
t=10						
	True	.2226	.3954	.8632	.9768	.9875
	Stehfest	.2210	.3931	.8628	.9798	.9909
		(-.0072)	(-.0058)	(-.0005)	(.0031)	(.0034)
	Recursive	.2184	.3928	.8664	.9815	.9915
		(-.0189)	(-.0066)	(.0037)	(.0048)	(.0041)

Note: The values of the cumulative distribution functions for the three methods have been calculated by numerical integration of the corresponding density function. The grid is $[0, 25] \times [0, 25]$*, with* $M = 500$ *equal spacings. We have followed this procedure also for the true distribution. The true distribution is a Downton bivariate exponential with* $(\mu_1 = \mu_2 = 0.5, \rho = 0.2)$*. The start* $g_0(s, t) = Ga(s|1, 0.2)\, Ga(t|1, 0.2)$*. Relative errors w.r.t the true between brackets. See Example 3 for more details.*

are standard; i.e. we skip the first 1,000 values, keep every 101st point, and then apply reverse–radix scrambling to further decrease the correlation. Each point is assigned with probability $\frac{1}{2}$ either to $g_0(s,t,r)$ or to $s\,t\,r\,\exp(-xs-yt-zr)$.

The running time is about 45min for our procedure, while the Gaver–Stehfest runs for about 8.5h. Table 4.4 reports the value of the cumulative distribution function $G_N(s,t,r)$ corresponding to $g_N(s,t,r)$.

Table 4.4: Cumulative distribution function of a trivariate gamma for selected values.

		s=1	s=3	s=5	s=7		s=1	s=3	s=5	s=7	
t=1											
	True	.0001	.0020	.0027	.0028	r=1	.0028	.0378	.0514	.0528	r=5
	Stehfest	.0003	.0026	.0037	.0041		.0037	.0389	.0552	.0607	
		(2)	(.3000)	(.3704)	(.4643)		(.3214)	(.0291)	(.0739)	(.1496)	
	Recursive	.0001	.0015	.0020	.0022		.0035	.0417	.0511	.0531	
		(0)	(-.2500)	(-.2593)	(-.2143)		(.2500)	(.1032)	(-.0058)	(.0057)	
t=3											
	True	.0271		.0368	.0378		.0378		.6925	.7121	
	Stehfest	.0275		.0389	.0424		.0424		.5760	.6276	
		(.0148)		(.0571)	(.1217)		(.1217)		(-.1682)	(-.1187)	
	Recursive	.0247		.0356	.0375		.0380		.6568	.6903	
		(-.0886)		(-.0326)	(-.0079)		(.0053)		(-.0516)	(-.0306)	
t=5											
	True	.0368	.4958		.0514		.0514	.6125		.9405	
	Stehfest	.0389	.4066		.0601		.0601	.6276		.8892	
		(.0571)	(-.1799)		(.1693)		(.1693)	(.0247)		(-.0545)	
	Recursive	.0281	.5167		.0536		.0425	.6844		.8867	
		(-.2364)	(.0422)		(.0428)		(-.1732)	(.1174)		(-.0572)	
t=7											
	True	.0378	.5099	.6925	.7121		.0528	.7121	.9672	.9949	
	Stehfest	.0424	.4430	.6276	.6838		.0655	.6838	.9688	1.005	
		(.1217)	(-.1312)	(-.0937)	(-.0397)		(.2405)	(-.0397)	(.0017)	(.0102)	
	Recursive	.0283	.5285	.6715	.7090	r=3	.0428	.7024	.8994	.9505	r=7
		(-.2513)	(.0365)	(-.0303)	(-.0044)		(-.1894)	(-.0136	(-.0701)	(-.0446)	

Note: The values of the cumulative distribution functions for the three methods have been calculated by numerical integration of the corresponding density function. The grid is $[0,100] \times [0,100] \times [0,100]$*, with* $M = 300$ *equal spacings. We have followed this procedure also for the true distribution.* $g(s,t,r) = Ga(s|5,2)\, Ga(t|5,2)\, Ga(r|5,2)$*.* $g_0(s,t) = Ga(s|1,0.2)\, Ga(t|1,0.2)$*. Relative errors w.r.t the true between brackets. See Example 4 for details.*

A first observation is that our recursive procedure scales up much better to the increase in dimension. We could not use a finer grid, lacking the computer power to evaluate the Gaver–Stehfest estimate. Furthermore, the accuracy of the recursive method remains satisfactory, though it fails slightly to portray the tails. However, here the alternative does not provide a density function, taking negative values. The true mode is $(2, 2, 2)$, the recursive estimate has this estimated at $(1.8, 2, 1.9)$, which again outperforms the Gaver–Stehfest value of $(1.7, 1.7, 1.7)$.

Example 5. *Bivariate lognormal distribution.* We consider two identically distributed, independent lognormal distributed random variables. It is notorious that the Laplace transform of the lognormal distribution is not available in closed form. A closed form approximation, $\widetilde{F}(x)$, is given in Rojas-Nandayapa (2008) and then analysed by Asmussen et al. (2016). Asmussen et al. (2016) prove that $\widetilde{F}(x)$ is asymptotically equivalent to the true Laplace transform $F(x)$. The univariate approximation is

$$\widetilde{F}(x) = \frac{1}{\sqrt{1 + \mathcal{W}(x\sigma^2)}} \exp\left(-\frac{1}{2\sigma^2}\mathcal{W}(x\sigma^2)^2 - -\frac{1}{\sigma^2}\mathcal{W}(x\sigma^2)\right),$$

where $\mathcal{W}(\cdot)$ is the Lambert function defined as the solution of $x = \mathcal{W}(x)\exp^{\mathcal{W}(x)}$.

Asmussen et al. (2016) give an expansion of the error term using $\widetilde{F}(x)$ that allows to construct a Monte-Carlo estimator of $F(x)$. Such an estimator allows a study on how the error term depends on σ^2. The authors point out that for small values of σ^2, the approximation error is negligible. We fix $\sigma^2 = 1$ and treat $\widetilde{F}(x)$ as the true Laplace transform of a lognormal distribution with variance equal one. We fix mean equal to zero. Since we want to deal with a bivariate inversion we have that $F(x, y) \approx \widetilde{F}(x)\widetilde{F}(y)$.

To implement our methodology, we sample from the density corresponding to $F(x, y)$, which we denote by $\widetilde{f}(x)\widetilde{f}(y)$, with rejection sampling. We use the following quantities: $g_0 = \text{Ga}(s|\frac{3}{2}, \frac{1}{5})\text{Ga}(t|\frac{3}{2}, \frac{1}{5})$, $s_M = t_M = 30$, $M = 1000$ equal spacing, $w_n = (1 + n)^{-0.7}$, $N = 1000$. Note that the sample N is much smaller than previous examples: it was chosen in order to speed computational time as the rejection sampling is quite slow in this example as evaluating $\widetilde{f}(x)\widetilde{f}(y)$ is very slow. For the same reason we do not have a comparison: all the alternatives require to evaluate $\widetilde{F}(x)\widetilde{F}(y)$ at each point of the grid, which is not computationally feasible.

Table 4.5 reports the comparison between our estimate and the true distribution for selected values. The recursive estimate is obtained through 20 Monte Carlo averages to make it smoother. The accuracy is worst than the other examples considered, but the

Table 4.5: Cumulative distribution function of a Lognormal for selected values.

		s=0.5	s=1.5	s=2.5	s=4	s=9
t=0.5						
	True	.0629	.1650	.2055	.2303	.2473
	Recursive	.0655	.1803	.2194	.2400	.2512
		(.0422)	(.0933)	(.0673)	(.0424)	(-.0155)
t=1.5						
	True	.1650	.4330	.5394	.6043	.6490
	Recursive	.1750	.4680	.5739	.6360	.6724
		(.0622)	(.0809)	(.0641)	(.0524)	(.0361)
t=2.5						
	True	.2055	.5394	.6719	.7528	.8085
	Recursive	.2196	.5818	.7131	.7917	.8392
		(.0685)	(.0786)	(.0612)	(.0516)	(.0379)
t=4						
	True	.2303	.6043	.7528	.8434	.9059
	Recursive	.2430	.6449	.7893	.8763	.9297
		(.0552)	(.0667)	(.0484)	(.0390)	(.0263)
t=9						
	True	.2473	.6490	.8085	.9059	.9729
	Recursive	.2546	.6780	.8303	.9224	.9798
		(.0295)	(.0446)	(.0269)	(.0183)	(.0070)

Note: The values of the cumulative distribution functions for the three methods have been calculated by numerical integration of the corresponding density function. The grid is $[0, 30] \times [0, 30]$*, with* $M = 1000$ *equal spacings. We have followed this procedure also for the true distribution. The true distribution is a bivariate lognormal with* $(\mu_1 = \mu_2 = 0.5, \sigma_1^2 = \sigma_1^2 = 1)$*. The start* $g_0(s,t) = Ga(s|1.5, 0.2)\, Ga(t|1.5, 0.2)$*. Relative errors w.r.t the true between brackets. See Example 3 for more details.*

relative error is anyway maximum 10%. The higher relative error might be due to the much smaller sample size considered.

4.5 Discussion

We have proposed to use NQZ to invert the Laplace transform of a multivariate distribution function with positive support. The method, which extends Walker (2017a), fills an evident gap in the literature: there is no method directly targeting a multivariate probability distribution of dimension higher than two. In addition to this, none of the alternatives give a theoretical guarantee to obtain a numerical estimate which is a probability distribution.

The chapter offers a nice illustration of an applied problem where a recursive algorithm works really well. Most of the advantages of recursive procedures are displayed: the procedure is easy to implement and scales up much better to an increase in dimension than competing methods. The accuracy seems to be also less affected by the extra dimension. In this particular application, a recursive algorithms makes possible to deal with certain examples that are not feasible using the alternatives. The number of grid points to evaluate increases to the power of the dimension. To make it relevant in high dimensions, i.e. more than four, more efficient computational tools have to be considered. Our algorithm is highly parallelisable, thus many well studied approaches can be used. In this respect, GPU computing would seem appropriate and to be explored.

Chapter 5

Asymptotic Theory for Recursive Procedures

In this chapter we study the large-sample properties of the family of recursive algorithms discussed in this thesis and present some preliminary results that allow to analyze their asymptotic behaviour. We have shown in the previous chapters that the recursive algorithms we have discussed can be seen as stochastic approximation (SA) procedures. However, SA asymptotic theory does not offer a viable method to prove convergence of infinite-dimensional functions. The key idea in the chapter is a novel asymptotic argument that allows us to study the convergence of a sequence of continuous distribution functions defined recursively. This argument relies on conditions very familiar to the stochastic approximation literature and is based on a novel martingale argument. We first deal with the general framework described in the introduction and then apply our results to the algorithms presented in the previous chapters.

In order to introduce our asymptotic argument, we first discuss the convergence of a one-dimensional sequence (X_n). This one-dimensional example is useful to introduce and explain the martingale argument which, along with a fixed point condition, is crucial for our approach. As the same example can be analysed with SA asymptotic theory, we draw an analogy. This will be the subject of Section 5.1. In Section 5.2 we establish the weak convergence of the sequence (P_n) defined in the Introduction under appropriate conditions (Theorems 5.2.2 and 5.2.3).

In Section 5.3 we show that the results of Section 5.2 can be applied to the convergence of the NQZ algorithm for a general family of kernel functions. We first study the theoretical convergence for the application of NQZ to the Laplace inversion presented in Chapter 4. A contribution in this section is that we are able to prove its asymptotic behaviour under less restrictive requirements that those

currently employed in the literature. In Section 5.4 we focus on HMW and the two algorithms we have introduced in Chapter 3, the Recursive Regression and the Two Steps Predictive Recursion. We showed that they perform well in several numerical illustrations but have not proven yet their theoretical convergence. Section 5.4.2 deals with the Recursive Regression, Section 5.4.3 with the Two Steps Predictive Recursion. Section 5.3 contains the convergence of the NQZ algorithm. The proof holds under somewhat restrictive conditions. We conclude with a discussion (Section 5.5).

5.1 One-dimensional Convergence

Consider a sequence of random variables (X_n) such that $X_n \in \mathcal{X} = [0,1]$ for all $n \in \mathbb{N}$. We define (X_n) through a recursive algorithm: given observations $Y_n \overset{iid}{\sim} P^*$, a probability distribution on a Euclidean space $\mathcal{Y}$, update an initial point $X_0 \in [0,1]$ iteratively by

$$X_n = (1-\alpha_n)X_{n-1} + \alpha_n \eta(X_{n-1}, Y_n), \tag{5.1}$$

where $\alpha_n \in (0,1)$ for all $n \in \mathbb{N}$ and $\eta : [0,1] \times \mathcal{Y} \to [0,1]$, which we assume to be *increasing*. The convergence of (X_n) has to be considered as an illustrative example. The definition (5.1) mimics the type of algorithms we have discussed in the thesis, with appropriate restrictions given that we are dealing with one-dimensional random variables.

First, note that the convergence of (X_n) can be dealt with using the SA convergence theory. Recall from Subsection 2.2.4 that the main tool used in SA is based on the ordinary differential equation (ODE) method developed in Kushner and Clark (1978) and Ljung (1978). In this framework the limit point, if it exists, is a global asymptotically stable point of the ODE $\dot{x} = h(x)$. In our example, (5.1) is an algorithm targeting the root of $h(x) = \mathrm{E}_Y[\eta(x,Y)] - x$. Also, it is a "projected algorithm" - i.e. (X_n) is constrained in a compact set - and conditions of Theorem 5.2.3 in Kushner and Yin (2003) need to hold in order to prove convergence.

We assume the following conditions hold:

A1. $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$.
A2. The function $\mathrm{E}_Y[\eta(X,Y)]$ has a unique fixed point X^*, i.e.

$$\mathrm{E}_Y[\eta(X,Y)] = X \iff X = X^*. \tag{5.2}$$

Both conditions are consistent with SA asymptotic theory: *A1* is standard in the literature on SA algorithms; *A2* replaces the more common condition that the limit

point X^* is a global asymptotically stable point for $\dot{x} = h(x)$, with X^* being a unique fixed point. Uniqueness of a fixed point and global asymptotically stability are equivalent notions in a large set of problems, both in mathematics, see Hartman and Olech (1962), Amann (1976), Lin (1998), and in statistics, see Delyon et al. (1999).

The use of martingale theory to prove convergence of a SA sequence is not new and it arises naturally from the assumption $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$. The proof usually involves simple martingale inequalities; for example, Fabian (1960) and Gladyshev (1965) bound a finite-dimensional sequence by a semi-martingale. In our method we do not resort to any martingale inequalities but simply use the fact that (X_n) takes values in a compact set which is partitioned by the fixed point X^*.

Define the set $A := \{x \in \mathcal{X} : x > \mathrm{E}_Y[\eta(x, Y)]\}$ and denote by A' its complementary and δA the boundary set. Note that a direct consequence of *A2* is that $\delta A = \{X^*\}$. In order to prove the result below, we assume that the sequence (α_n) is defined as $\alpha_n = (c+n)^{-d}$ with $d \in [1/2, 1]$, which trivially satisfies *A1*. This choice is not particularly restrictive as it is reasonable to think that any sequence $(\widehat{\alpha}_n)$ satisfying *A1* is such that for large n $\widehat{\alpha}_n \leq (c_1 + n)^{-d_1}$ and $\widehat{\alpha}_n \geq (c_2 + n)^{-d_2}$ for a given $d_1, d_2 \in [1/2, 1]$ and $c_1, c_2 > 0$, which implies we can bound below and above a sequence (X_n) defined through $(\widehat{\alpha}_n)$. Without loss of generality we fix $c = 0$ to simplify the notation.

Theorem 5.1.1. *Under* A1 *and* A2*, we have that $X_n \to X^*$, P^*–a.s., as $n \to \infty$.*

Proof. Suppose (X_n) moves *infinitely often* (*i.o.*) between A and A'. Indeed, if there exists a $\hat{n}$ such that for all large n it is that $X_n \in A$ (or A'), then $(X_n)_{n>\hat{n}}$ will be a submartingale (supermartingale), and it is easy to see that it will be convergent, P^*–a.s. Let us define the martingale S_N by

$$S_N = \sum_{n=1}^{N} \Big[X_n - X_{n-1} - \alpha_n \left(\mathrm{E}_Y[\eta(X_{n-1}, Y_n) | \mathcal{G}_{n-1}] - X_{n-1} \right) \Big],$$

where $\mathcal{G}_{n-1} = \sigma(Y_{1:n-1})$. By *A1*,

$$\sum_{n=1}^{\infty} (X_n - X_{n-1})^2 < \infty$$

which implies $\mathrm{Var}(S_N) < \infty$; hence S_N converges P^*–a.s. to a finite r.v. S by the martingale convergence theorem.

Now, to prove that (X_n) is Cauchy, we have to establish that for any given $\epsilon > 0$, there exists a $\bar{n}$ such that for all $M, N > \bar{n}$ we have $|X_M - X_N| < \epsilon$. We assume that (X_n) does not converge (otherwise the theorem follows). Note that the lim inf and lim sup must be

bounded and respectively strictly smaller and larger than the point $X^* = \mathrm{E}_Y[Y|\eta(X^*, Y]$, given that (X_n) moves *i.o.* from A to A' and *A2*. Let $\overline{X} = \limsup X_n$ and $\underline{X} = \liminf X_n$. If $\epsilon \geq \overline{X} - \underline{X}$, it is trivial to show that the condition holds.

Suppose w.l.o.g $X_M \geq X^*$ and $X_N \geq X^*$. Fix $\epsilon > 0$ such that $X^* + \epsilon \leq \overline{X}$ and consider the set $A_\epsilon := \{x \in \mathcal{X} : x - X^* \geq \epsilon\}$. If both X_M and X_N do not belong to A_ϵ, we have that $|X_M - X_N| < \epsilon$. We now consider the nontrivial scenario in which at least one of the two does belong to A_ϵ.

For each sample path, we can construct two subsequences (X_{n_k}) and (X_{m_k}) converging to $X^* + \epsilon$ satisfying the following conditions,

1. (X_{n_k}) and (X_{m_k}) take values in A_ϵ.

2. It holds that $n_k \leq m_k$ for all $k \geq 1$. Also for all $n_k \leq n \leq m_k$, $X_n \in A_\epsilon$.

The existence of the subsequences is given by the fact that all points in $(\underline{X}, \overline{X})$ are accumulaton points.

By construction, $X_{n_k} - X_{m_k} \to 0$ P^*–*a.s.* as $k \to \infty$. Also $S_{n_k} - S_{m_k} \to 0$, which in turn implies

$$\sum_{n=n_k+1}^{m_k} \alpha_n \Big\{ \mathrm{E}[\eta(X_{n-1}, Y_n)|\mathcal{G}_{n-1}] - X_{n-1} \Big\} \to 0 \quad \text{as } k \to \infty. \tag{5.3}$$

Note that for all $x \in A_\epsilon$ there exists $d > 0$ such that it holds that either $x - \mathrm{E}_Y[\eta(x, Y)] \geq d$ or $x - \mathrm{E}_Y[\eta(x, Y)] \leq d$. Let us suppose w.l.o.g. that the former holds.

Given that $X_{n-1} - \mathrm{E}_Y[\eta(X_{n-1}, Y)|\mathcal{G}_{n-1}] \geq d$ for all X_n such that $n_k + 1 \leq n \leq m_k$, (5.3) implies that

$$\sum_{n=n_k+1}^{m_k} \alpha_n \to 0 \quad \text{as } k \to \infty, \tag{5.4}$$

must hold. Otherwise it must be the that (X_n) converges. From (5.4), it follows that $(m_k - n_k)/(m_k)^d \to 0$ as $k \to \infty$, which in turn implies that $(m_k - n_k)/(m_k) \to 0$ and $n_k/m_k \to 1$. Then, it follows that

$$(m_k - n_k) \sum_{n=n_k+1}^{m_k} \alpha_n^2 \leq \frac{(m_k - n_k)^2}{(n_k + 1)^{2d}} \to 0 \tag{5.5}$$

as $k \to \infty$. Now, for any $N < M$,

$$|X_M - X_N| \leq |X_M - X_{n_{k'}}| + |X_{n_{k'}} - X_{m_{k^*}}| + |X_{m_{k^*}} - X_N|,$$

where m_{k^*} is the smallest number in (m_k) greater than N, and $n_{k'}$ is the largest number in the subsequence (n_k) smaller than M. The center term on the right side goes to zero

by construction. Let us consider $|X_M - X_{n_{k'}}|$,

$$(X_M - X_{n_{k'}})^2 \leq (M - n_{k'}) \sum_{n=n_{k'}+1}^{M} \alpha_n^2 \leq (m_{k'} - n_{k'}) \sum_{n=n_{k'}+1}^{m_{k'}} \alpha_n^2, \tag{5.6}$$

with $m_{k'}$ being by construction such that $n_{k'} \leq M \leq m_{k'}$. The first inequality is Cauchy Schwartz, the second is given by construction. The rightmost term in (5.6) goes to zero by (5.5) as $k \to \infty$, which in turns implies that $|X_M - X_{n_{k'}}|$ goes to 0. The same argument applies to $|X_{m_{k^*}} - X_N|$. Hence $|X_{m_{k^*}} - X_N|$, $|X_M - X_{n_{k'}}|$ both go to zero and (X_n) is a Cauchy sequence with probability one since the proof applies to every sample path.
Note that the other scenarios follow from this proof: if X_M, X_N are both smaller that $X^* - \epsilon$, the same argument applies; if $X_M \geq X^*$ and $X_N < X^*$, it is possible to split the problem in two.
Since (X_n) lies in a compact set and it is Cauchy, it is convergent. To show that the limit is X^*, note that X_n can be rewritten as

$$X_n = X_0 + \sum_{i=1}^{n} \alpha_i \{\mathrm{E}_Y[\eta(X_{i-1}, Y_i)] - X_{i-1}\} + \sum_{i=1}^{n} \alpha_i \{\eta(X_{i-1}, Y_i) - \mathrm{E}_Y[\eta(X_{i-1}, Y_i)]\}. \tag{5.7}$$

Now $U_n := \eta(X_{n-1}, Y_n) - \mathrm{E}_Y[\eta(X_{n-1}, Y)]$ is a bounded martingale difference sequence. Since $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$, the last term is convergent P^*–*a.s.* by the martingale convergence theorem. Since X_n is convergent, it must be that also the central term is convergent. By *A2* and $\sum_{n=1}^{\infty} \alpha_n = \infty$, X^* is the only accumulation point because the central term must accumulate around zero. □

The novelty here is the use of the martingale S_N to prove that (X_n) is Cauchy. Note that under the assumption that X_n moves *i.o.* between A and A' along with the fact that $|X_n - X_{n-1}| \to 0$ as $n \to \infty$ P^*– *a.s.*, imply that a limit point is on δA; i.e. X^* is an accumulation point, without the need to use the argument (5.7).

5.2 Weak Convergence of (P_n)

We study the convergence of a sequence (P_n) of probability distribution function analogous to the one described in Chapter 1. Let $\mathcal{Y}$ be an Euclidean space and $\mathcal{P}$ be the family of *absolutely continuous* probability distributions on $(\mathcal{Y}, \mathcal{A})$ with respect to the Lebesgue measure, where $\mathcal{A}$ denotes the σ-algebra of Borel sets of $\mathcal{Y}$. Suppose we want to estimate $P^* \in \mathcal{P}$ given N observations $(Y_n)_{n=1:N} \in \mathcal{Y}^N$ such that $Y_n \overset{iid}{\sim} P^*$ for $1 \leq n \leq N$. A recursive estimate P_N can be obtained as follows: choose a prior estimate $P_0 \in \mathcal{P}$, a deterministic sequence of weights (α_n) with $\alpha_n \in (0, 1)$ for all n, repeat iteratively for

$n = 1, \ldots, N$

$$P_n(y) = (1 - \alpha_n)P_{n-1}(y) + \alpha_n Z(y, P_{n-1}, Y_n), \tag{5.8}$$

where $Z(y, P_{n-1}, Y_n)$ is a CDF on $\mathcal{Y}$ evaluated at y. Note that for a given Y the function $Z(\cdot, \cdot, Y) : \mathcal{P} \to \mathcal{P}$: it defines how to map a given $P \in \mathcal{P}$ given the observation Y to another element of $\mathcal{P}$. A key property we require Z to have is that it must admit a unique fixed point P^*, i.e.

$$\int_{\mathcal{Y}} Z(y, P, y')p^*(y)\mathrm{d}y' = P(y) \quad \forall y \in \mathcal{Y} \iff P = P^*.$$

We formalize this assumption in the following. Note that we often refer to Z as the *update*. Without loss of generality we fix $\mathcal{Y} = \mathbb{R}^+$.

Weak convergence is proven in three steps. First, we prove that (P_n) is a *relatively compact sequence*; this means that each subsequence (P_{n_k}) has a further subsequence that converges to a probability distribution. We then show that the sequence $(P_n(y))$ converges for all $y \in \mathcal{Y}$ with probability one. Relative compactness and pointwise convergence imply convergence in the weak topology to a probability distribution, which we denote by P_∞ (Billingsley 1999). Finally, we prove that $P_\infty = P^*$ *a.s.*

Let Π_n be the distribution of P_n. To prove relative compactness we show the sequence (P_n) is *tight*, in the sense that the sequence of distributions (Π_n) is tight, and invoke Prohorov's Theorem. By construction, for all n, P_n is a continuous CDF on $[0, \infty)$ because P_0 and Z are in $\mathcal{P}$, hence it belongs to $C = C[0, \infty)$, the space of *real valued* continuous functions defined on $[0, \infty)$. (Π_n) is then a sequence of probability distributions on $C[0, \infty)$. The study of weak convergence of a random sequence taking values on a function space is a well established topic, see Billingsley (1999). A criteria for tightness of a sequence of random functions on $C[0, \infty)$ is given by Theorem 4 in Whitt (1970). Let $C_R = C[0, R]$ be the space of continuous function on $[0, R]$ and let $w_f^R : (0, R] \to [0, \infty)$ be the *modulus of continuity* of an arbitrary function $f \in C_R$ defined by

$$w_f^R(\delta) = \sup_{0 \le s,t \le R : |s-t| < \delta} |f(s) - f(t)|.$$

Theorem 5.2.1. *(Whitt 1970) Let* (Π_n) *be a sequence of probability measures on* $C[0, \infty)$. *The sequence* (Π_n) *is tight if and only if these two conditions hold:*

(i) For each $t \ge 0$ *and each positive* η, *there exists a compact set* K *in* $\mathbb{R}$ *such that*

$$\Pi_n(f \in C : f(t) \in K_t) > 1 - \eta \qquad \forall n \in \mathbb{N}$$

(ii) For each $R \geq 1$ and positive ϵ and η, there exists a δ, with $0 < \delta < 1$, and integer n_0 such that

$$\Pi_n(f \in C : w_f^R(\delta) \geq \epsilon) \leq \eta \qquad \forall n > n_0$$

A few remarks. First, in the reference Whitt (1970) Theorem 5.2.1 applies to a space $C[0, \infty)$ of functions valued in any complete separable space metric (E, m), where m is a metric on E. In our setting we have assumed that $E = \mathbb{R}$ because $\mathcal{Y} \subseteq \mathbb{R}$. Second, the elements $f \in C$ we are interested in are probability distributions functions: so condition (i) is trivially verified by $K_t = [0, 1]$ for all $t \geq 0$. To prove tightness of (Π_n) we need to check (ii).

The critical part in the proof is to show that the sequence $(P_n(y))$ converges. We do that by showing that the sequence is Cauchy. The proof of Theorem 5.1.1 does not apply to (P_n) because the updates $Z(\cdot, P, Y)$ and $\eta(X, Y)$ are defined on different domains: it is possible to define a fixed point condition for $Z(\cdot, P, Y)$ but it is not to define a set with properties equivalent to the sets - for example A and A_ϵ - used in the proof of Theorem 5.1.1. The reason appears clear when one thinks that the entire probability distribution P_{n-1} is needed to calculate $P_n(y)$. No comparison with the SA theory is given in this section because, to our knowledge, SA asymptotic theory does not cover the type of sequence we are interested in.

We assume that *A1* and the following conditions hold:

A2.* The function $\mathrm{E}_Y[Z(\cdot, P, Y)]$ has a unique fixed point, which is P^*, the data-generating distribution, i.e.

$$\mathrm{E}_Y[Z(y, P, Y)] = P(y) \quad \forall y \in \mathcal{Y} \iff P \equiv P^*. \tag{5.9}$$

A3. Given $h \geq 0$, $R > 0$ and s_1, s_2 and s_3 in $\mathcal{Y}$ such that $0 \leq s_1 \leq s_2 \leq s_3 \leq R$, $|s_1 - s_2| \leq h$ and $|s_2 - s_3| \leq h$. There exists a constant C such that

$$|P_0(s_2) - P_0(s_1)||P_0(s_3) - P_0(s_2)| \leq Ch^2. \tag{5.10}$$

A4. Given $h \geq 0$, $R > 0$ and s_1, and s_2 in $\mathcal{Y}$ such that $0 \leq s_1 \leq s_2 \leq R$, $|s_1 - s_2| \leq h$. If there exists a constant C such that

$$\mathrm{E}[|P_n(s_2) - P_n(s_1)|^2] \leq Ch^2. \tag{5.11}$$

then, it also holds that

$$\mathrm{E}[|Z_n(s_2, P_n, Y) - Z(s_1, P_n, Y)|^2] \leq Ch^2. \tag{5.12}$$

*A2** requires the uniqueness of a fixed point for all $y \in \mathcal{Y}$: it is the obvious infinite-dimensional counterpart of *A2*. In the following section we give examples of statistical estimators that satisfy *A2**. *A3* and *A4* are two assumptions that will be used in the tightness proof. *A3* is a condition on the initial guess P_0: it is easy to see that it holds for most of the $P_0 \in \mathcal{P}$; it does trivially for all the continuous distributions that admit a bounded density. *A4* is more delicate: we require it for to prove tightness in such a general setting. Note that the expected value is calculated with respect to the joint distribution of Y and P_n. It seems reasonable and it holds for some commonly employed recursive updates in the literature, for example the empirical distribution, which on the other hand is not continuous. However, it is difficult to prove it holds for the algorithms we have discussed in this thesis: in the following section we discuss some alternative conditions that allow us to prove tightness for specific algorithms. A more encompassing condition for the general framework (5.8) is the subject of ongoing work.

Theorem 5.2.2. *Under* A1, A3 *and* A4, *the sequence* (Π_n) *is tight.*

Proof. Condition (i) of Theorem 5.2.1 holds trivially: we are left to show that (ii) holds. We prove the following condition in place of (ii): fix $R > 0$ and $h \geq 0$, s_1, s_2 and s_3 such that $0 \leq s_1 \leq s_2 \leq s_3 \leq R$, and $|s_1 - s_2| \leq h$ and $|s_2 - s_3| \leq h$, there exists a constant C such that

$$\mathrm{E}[|P_n(s_2) - P_n(s_1)||P_n(s_3) - P_n(s_2)|] \leq Ch^2, \tag{5.13}$$

holds for all $n \in \mathbb{N}$. The equivalence between (ii) and (5.13) follows from the proof of Theorem 13.5 in Billingsley (1999) and Theorem 2.1 in Rao and Sethuraman (1975). These references apply to the space D of cadlag function, hence also to C.

Let us denote by

$$P(s_1, s_2) = P(s_2) - P(s_1),$$

and

$$\Delta(P(s_1, s_2), Y) = Z(s_2, P, Y) - Z(s_1, P, Y).$$

The proof proceeds by induction. The bound (5.13) holds for $n = 0$ by *A3*. We assume that it holds for $n - 1$ and denote the bound by Ch^2. We choose a C large enough such that *A4* holds for $n - 1$. Note that the the quadratic bound implies that both

$$\mathrm{E}[|P_{n-1}(s_2) - P_{n-1}(s_1)|^2] < Ch^2 \quad \text{and} \quad \mathrm{E}[|P_{n-1}(s_3) - P_{n-1}(s_2)|^2] < Ch^2.$$

To complete the induction proof we have to bound

$$\mathrm{E}[|P_n(s_2) - P_n(s_1)||P_n(s_3) - P_n(s_2)|]$$

Now,

$$\begin{aligned}
&\mathrm{E}[|P_n(s_2) - P_n(s_1)||P_n(s_3) - P_n(s_2)|] \\
&\leq (1-\alpha_n)^2\mathrm{E}[P_{n-1}(s_1,s_2)P_{n-1}(s_2,s_3)] + \alpha_n(1-\alpha_n)\mathrm{E}[P_{n-1}(s_1,s_2)\Delta(P_{n-1}(s_2,s_3),Y_n)] + \\
&\quad + \alpha_n^2\mathrm{E}[\Delta(P_{n-1}(s_2,s_3),Y_n)\Delta(P_{n-1}(s_1,s_2),Y_n)] + \alpha_n(1-\alpha_n)\mathrm{E}[P_{n-1}(s_2,s_3)\Delta(P_{n-1}(s_1,s_2),Y_n)]
\end{aligned}$$

The first term is bounded by induction. Consider the second term, by Cauchy Schwartz

$$\mathrm{E}[P_{n-1}(s_1,s_2)\Delta(P_{n-1}(s_2,s_3),Y_n)] \leq \sqrt{|P_{n-1}(s_1,s_2)|^2}\sqrt{|\Delta(P_{n-1}(s_2,s_3),Y_n)|^2}$$

which is bounded by *A4* and induction. The same applies to the third term. The fourth term is bounded by *A4*. □

Let us now fix $y \in \mathcal{Y}$ and consider the sequence $(P_n(y))$. As explained at the beginning of the section, the proof of Theorem 5.2.3 is different from that of Theorem 5.1.1. A similarity in the proof is due to the fact that we employ a martingale argument to establish the statement below. The definition of the martingale and its convergence will appear different from the martingale defined in Theorem 5.1.1. Some of the mathematical details that follow the convergence are shared across the two proofs. The analogy arises solely because the sequence of weights (α_n) of the two examples can be identical. We assume once more that the sequence (α_n) is defined as $\alpha_n = (n)^{-d}$ with $d \in [1/2, 1]$. See the previous section for a discussion.

Theorem 5.2.3. *Under* A1, A2* – A4, *the sequence* $(P_n(y))$ *is Cauchy for all* $y \in \mathcal{Y}$.

Proof. By Theorem 5.2.2 and Prohorov's Theorem, a tight sequence of distribution function (Π_n) on a Euclidean space has a subsequence (Π_{n_k}) that converges weakly as $k \to \infty$. This also implies that there exists a subsequence (P_{n_k}) that converges weakly to a probability distribution as $k \to \infty$. Let us define

$$M_N = \sum_{n=1}^{N}\left[\xi_n - \xi_{n-1} - \alpha_n\left(\int h(y)\mathrm{d}Z(y,P_{n-1}) - \xi_{n-1}\right)\right],$$

where $\xi_n = \int h(y)\mathrm{d}P_n(y)$, $h \in C_b(\mathcal{Y})$, i.e. it belongs to the space of continuous and bounded function defined on $\mathcal{Y}$, and

$$\int h(y)\mathrm{d}Z(y,P_{n-1}) = \mathrm{E}\left[\int h(y)\mathrm{d}Z(y,P_{n-1},Y)|\mathcal{F}_{n-1}\right],$$

where $\mathcal{F}_{n-1} = \sigma(P_1, \ldots, P_{n-1})$. Note that M_N is a martingale, and it is such that $\text{Var}(M_N) < \infty$ and M_N converges to a finite random variable M P^*–$a.s.$, by the martingale convergence theorem.

Note that the weakly convergent subsequence (P_{n_k}) defines a corresponding subsequence (ξ_{n_k}) convergent P^*-$a.s.$ (Portmanteau Theorem, Theorem 2.1 in Billingsley, 1999). Given that both (ξ_{n_k}) and (M_{n_k}) converge as $k \to \infty$, it follows that

$$\sum_{n=n_k+1}^{n_{k+1}} \alpha_n \left(\int h(y) \mathrm{d}Z(y, P_{n-1}) - \xi_{n-1} \right) \to 0 \tag{5.14}$$

P^*-$a.s.$ as $k \to \infty$. Note that we do not assume neither the existance of a fixed point, nor the compactness of the space in which (P_n) takes value, the subsequence (P_{n_k}) is guaranteed by Prohorov's Theorem.

Also, note that the limit (5.14) holds for all $h \in C_b(\mathcal{Y})$ and a fixed subsequence (P_{n_k}). Assuming for now that (P_n) is not convergent, $\gamma_n^h = \int h(y) \mathrm{d}Z(y, P_{n-1}) - \xi_{n-1}$ also does not converge for all $h \in C_b(\mathcal{Y})$. Also it cannot be that (γ_n) accumulates around 0 for all $h \in C_b(\mathcal{Y})$, since (P_n) is a sequence of probability distributions. It must be that

$$\sum_{n=n_k+1}^{n_{k+1}} \alpha_n \to 0, \tag{5.15}$$

because the convergent subsequence indexed by (n_k) and h are unrelated, and (5.14) must hold for all $h \in C_b(\mathcal{Y})$. Now note that (5.15) implies that $(n_{k+1})/n_k \to 1$ since both $(n_{k+1} - n_k)/(n_{k+1})^d \to 0$ as $k \to \infty$ and $(n_{k+1} - n_k)/(n_{k+1}) \to 0$ follow from (5.15). Then it holds that

$$(n_{k+1} - n_k) \sum_{n=n_k+1}^{n_{k+1}} \alpha_n^2 \leq \frac{(n_{k+1} - n_k)^2}{(n_k + 1)^{2d}} \to 0, \tag{5.16}$$

with (5.15) and (5.16) being equivalent to, respectively, (5.4) and (5.5).

To show that $(P_n(y))$ is Cauchy, the argument is now identical to that in the proof of Theorem 5.1.1. For any $N < M$,

$$|P_M(y) - P_N(y)| \leq |P_M(y) - P_{n_{k*}}(y)| + |P_{n_{k*}}(y) - P_{n_{k'}}(y)| + |P_{n_{k'}}(y) - P_N(y)|,$$

where $n_{k'}$ is the smallest number belonging to the subsequence (n_k) greater than N, and n_{k*} is the largest number in the subsequence (n_k) smaller than M. The center term on the right side goes to zero by construction. Let us consider $|P_M(y) - P_{n_{k*}}(y)|$,

$$(P_M(y) - P_{n_{k*}}(y))^2 \leq (M - n_{k*}) \sum_{n=n_{k*}+1}^{M} \alpha_n^2 \leq (n_{k*+1} - n_{k*}) \sum_{n=n_{k*}+1}^{n_{k*+1}} \alpha_n^2. \tag{5.17}$$

Now, (5.16) implies that the right term in (5.17) is bounded. It also implies that $|P_M(y) - P_{n_{k*}}(y)|$ goes to 0 as $M, n_{k*} \to \infty$.
The same applies to $|P_{n_{k'}}(y) - P_N(y)|$. The statement follows. □

Theorems 5.2.2 and 5.2.3 imply that under *A1*, *A2*-A4* the sequence (P_n) converges weakly P^* – *a.s.* to some probability distribution in $\mathcal{P}$, which we denote by P_∞. We are left to establish that $P_\infty = P^*$ *a.s.*, which we do in Corollary (5.2.1). However, we want to stress the two theorems represent already the bulk of the contribution of the section: we introduce a new methodology to prove the weak convergence of a stochastic approximation sequence taking values in a infinite dimensional probability distribution space. The proof of Theorem 5.2.3 exploits the properties of CDFs and some well known probability theory results valid for $\mathcal{P}$.

We are able to identify the limit of (P_n) thanks to the fixed point condition. An analogous proof appears also in Martin and Ghosh (2008), Walker (2017a) and Cappello and Walker (2018).

Corollary 5.2.1. *Under* A1, A2* – A4*, the sequence* (P_n) *converges weakly to* P^**,* P^**–a.s.*

Proof. By Theorem 5.2.2 we have a subsequence of (P_{n_k}) convergent to a probability distribution, say P_∞. Since $(P_n(y))$ is a Cauchy sequence lying in a compact set (Theorem 5.2.3): we have that $(P_n(y))$ is convergent for all $y \in \mathcal{Y}$. Also, the limit point must be $P_\infty(y)$. Note that $P_n(y)$ can be rewritten as

$$P_n(y) = P_0(y) + \sum_{i=1}^{n} \alpha_i \Big(Z(y, P_{i-1}, Y_i) - E_Y[Z(y, P_{i-1}, Y_i)] \Big) + \\ + \sum_{i=1}^{n} \alpha_i \Big(E_Y[Z(y, P_{i-1}, Y_i)] - P_{i-1} \Big). \tag{5.18}$$

Now, $Z(y, P_{n-1}, y_i) - E_Y[Z(y, P_{n-1}, Y_n)]$ is martingale difference and converges P^*–*a.s.* to a finite random variable (martingale convergence theorem). Since $\sum_n \alpha_n^2 < \infty$, we have that the term $\sum_{i=1}^n \alpha_i \left(Z(y, P_{i-1}, y_i) - E_Y[Z(y, P_{i-1}, Y_i)] \right)$ also converges P^*–*a.s.* to a finite random variable. Since $P_n(y)$ converges, it must be that the last term in (5.18) is convergent. In particular $(E_Y[Z(y, P_{n-1}, Y_i)] - P_{n-1})$ must accumulate around zero because $\sum_n \alpha_n = \infty$. The limit $P_\infty(y)$ must then be a fixed point of $E_Y[Z(y, P_{n-1}, Y_i)]$. Then *A2** guarantees that the only fixed point for all $y \in \mathcal{Y}$ is P^*. □

5.3 NQZ Algorithm

The NQZ algorithm was introduced in Subsection 2.2.1. It is an algorithm proposed to estimate the mixing distribution G of a mixture model. The assumption is that the observables are *i.i.d.* from a density p^* of the form

$$p(y) = \int_\Theta k(y|\theta) dG(\theta),$$

where $k(y|\theta)$ is the kernel of the mixture, i.e. a known density with respect to the dominating measure μ on $(\mathcal{Y}, \mathcal{A})$, Θ is a latent parameter space and G is a probability distribution on $(\Theta, \mathcal{B})$, and assume $G \in \mathcal{G}$ denote the class of probability distributions on $(\Theta, \mathcal{B})$ *absolutely continuous* with respect to the Lebesgue measure. The restriction of $\mathcal{G}$ to continuous distribution functions is needed to apply the asymptotic method introduced in Section 5.2.

Given N observations, the algorithm gives an estimate of G. Given an initial guess $G_0 \in \mathcal{G}$, N observations $(Y_n)_{n=1:N}$, an estimate G_N is obtained repeating iteratively the following

$$G_n(\theta) = (1 - \alpha_n) G_{n-1}(\theta) + \alpha_n \frac{\int_{-\infty}^{\theta} k(y_n|\theta') \mathrm{d}G_{n-1}(\theta')}{p_{n-1}(y_n)}, \tag{5.19}$$

where $p_{n-1}(y_n) := \int_\Theta k(y_n|\theta) \mathrm{d}G_{n-1}(\theta)$. In this case we have that

$$Z(\theta, G, Y) = \frac{\int_{-\infty}^{\theta} k(Y|\theta') \mathrm{d}G(\theta')}{p(Y)},$$

with $p(Y) = \int_\Theta k(Y|\theta) \mathrm{d}G(\theta)$. However, the striking difference between (5.8) and (5.19) is that $(G_n) \in \mathcal{G}$ is a sequence of probability distributions on a latent space, whereas the observations that feed the algorithm (5.8) take values on $\mathcal{Y}$.

We have reviewed the asymptotic theory for NQZ in Subsection 2.3.1. The key reference on the topic is Tokdar et al. (2009). Recall that Tokdar et al. (2009) prove that (G_n) converges weakly P^*- *a.s.* under a set of assumptions which can easily be violated by common mixture models, for example Gaussian location mixtures. See Subsection 2.3.1 for a discussion on the conditions in Tokdar et al. (2009). Note that Tokdar et al. (2009) also assume that G^* is *absolutely continuous* with respect to a $\sigma-$ finite measure ν on $(\Theta, \mathcal{A})$.

It is still possible to analyse the large sample properties of the predictive recursion

using the asymptotic arguments presented in this chapter. The two key components to verify are that the condition *A2** holds and that the sequence is tight.

Note that NQZ has been connected before to fixed point estimation: the link was mentioned in Martin and Ghosh (2008), Martin and Tokdar (2009) and later formalised by Walker (2017a),

$$\mathrm{E}_Y\left[Z(\theta, G, Y)\right] = G(\theta). \tag{5.20}$$

Now (5.20) is a Fredholm equation of first kind. For certain $k(y|\theta)$ and G, Fredholm equations of the first kind admit a unique solution. We do not want to enter into details here, see Corduneanu (1991), because we prefer to focus on an argument that has a more direct statistical interpretation.

If the assumptions of Theorem 5.2.2 hold, then tightness follows by the statement of theorem. It is also possible to establish tightness for specific kernels, such as a Gamma kernels, with an alternative proof. We illustrate this alternative argument for a Gamma kernel in Subsection 5.3.1.

5.3.1 Inversion of the Laplace Transform.

We consider the application of NQZ to the inversion of the Laplace transform, this was the subject of Chapter 4. We present the argument in the multidimensional setting which was the subject of Chapter 4. Recall that we considered $\mathcal{Y} = \mathbb{R}^{+d}$ and $\Theta = \mathbb{R}^{+d}$. The key intuition was to recognize that the Laplace transform density; i.e.

$$p(\underline{y}) = \frac{\partial^2 P(\underline{y})}{\partial y_1 \dots \partial y_d},$$

given in (5.21) is a mixture model with independent gamma kernels

$$p(\underline{y}) = \int_{[0,\infty)^d} \left\{ \left(\prod_{i=1}^{d} \theta_i \right) e^{-\sum_{i=1}^{d} \theta_i \, y_i} \right\} g(\underline{\theta}) \, \mathrm{d}\underline{\theta}. \tag{5.21}$$

Inverting a Laplace transform $P(\underline{y})$ is equivalent to estimate the mixing distribution of (5.21) and NQZ can be used for this purpose. The uniqueness of the Laplace transform implies the uniqueness of the fixed point (5.20), which is a condition analogous to *A2** for this specific problem.

We show a proof of tightness alternative to Theorem 5.2.2. This alternative proof is interesting to display some properties of NQZ and to give a theoretical guarantee of recovering a distribution function. We argued that this type of theoretical guarantee was the

feature that distinguish our proposal from all the alternative methodologies. This proof is given by Walker (2017a) and Cappello and Walker (2018) in the multidimensional setting.

To make the notation lighter we deal with the case $d = 2$. We first present a decomposition of $g_n(\theta_1, \theta_2)$ when $g_0(\theta_1, \theta_2) = \text{Ga}(\theta_1|a, b)\text{Ga}(\theta_2|c, d)$. $(y_{i1}, y_{i2})_{i=1:n}$ are i.i.d. samples from $f(y_1, y_2)$.

Lemma 5.3.1. *Given the power set S_n of $\{1, \ldots, n\}$, which includes the empty set,*

$$g_n(\theta_1, \theta_2) = \sum_{A \in S_n} q_{n,A}\, Ga\left(\theta_1 \left| a + |A|, b + \sum_{i \in A} y_{i1}\right.\right)\, Ga\left(\theta_2 \left| c + |A|, d + \sum_{i \in A} y_{i2}\right.\right),$$

where $q_{n+1,A} = (1 - \alpha_{n+1})q_{n,A}$ and

$$q_{n+1,A\cup(n+1)} = \alpha_{n+1} \frac{q_{n,A}\phi(y_{n+1,1}, A)\phi(y_{n+1,2}, A)}{\sum_{A \in S_n} q_{n,A}\phi(y_{n+1,1}, A)\phi(y_{n+1,2}, A)},$$

with

$$\phi(y_1, A) = (a + |A|)\left(b + \sum_{i \in A} y_{i,1}\right)^{a+|A|} \times \left(b + \sum_{i \in A} x_{i,1} + y_1\right)^{-a-|A|-1},$$

and, equivalently, $\phi(y_2, A)$ using (y_2, c, d) in place of (y_1, a, b).

Proof. The proof relies on the notion that a product of gamma kernels is a gamma kernel with updated parameters. Let us consider first the case $n = 1$. For all θ_1 and θ_2, g_1 is given by a mixture of two gamma distributions

$$g_1(\theta_1, \theta_2) = (1 - \alpha_1)\text{Ga}(\theta_1|a, b)\,\text{Ga}(\theta_2|c, d) + \alpha_1\text{Ga}(\theta_1|a + 1, b + y_{1,1})\,\text{Ga}(\theta_2|c + 1, d + y_{1,2}).$$

The case $n = 2$ is equivalent to the one above: now the mixture would be composed by four components. Carrying on this reasoning, we obtain that g_n is a mixture of 2^n independent gamma kernels with parameters all available in closed form. The extension to the power set is obtained from this representation. □

Lemma 5.3.1 will be used below to prove that the sequence (G_n) is tight. Let w_G^R denote the *modulus of continuity* of a function f in $C[0, R]^2$, the space of continuous functions defined on $[0, R]^2$

$$w_f^R(\delta_1, \delta_2) = \sup_{\substack{|\theta_{1,2} - \theta_{1,1}| < \delta_1 \\ |\theta_{2,2} - \theta_{2,1}| < \delta_2}} \{|f(\theta_{1,2}, \theta_{2,2}) - f(\theta_{1,1}, \theta_{2,1})|\}$$

and $0 < \theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2} < R$. To prove tightness of a sequence of probability measures (Π_n) governing (G_n) we use theorem 7.3 in Billingsley (1999). Since $G_n(0) = 0$ for all n,

tightness occurs if the following condition is verified:

$$H = \lim_{\substack{\delta_1 \to 0 \\ \delta_2 \to 0}} \limsup_n \Pi_n(w^R_{G_n}(\delta_1, \delta_2) > \epsilon) = 0 \quad \forall \epsilon > 0. \tag{5.22}$$

To clarify the heavy notation of $\theta_{1,2}$: the first subscript denotes which dimension we are considering, the second one is an index. The equivalent between condition (5.22) and (ii) in Theorem 5.2.1 is trivial.

Now, note that

$$|G(\theta_{1,2}, \theta_{2,2}) - G(\theta_{1,1}, \theta_{2,1})| = \left| \int_{\theta_{1,1}}^{\theta_{1,2}} \int_{\theta_{2,1}}^{\theta_{2,2}} g_n(\theta_1, \theta_2) \, \mathrm{d}\theta_2 \mathrm{d}\theta_1 \right|,$$

and for $\Delta = (\theta_{1,2} - \theta_{1,1}) \ (\theta_{2,2} - \theta_{2,1})$, we can write the following,

$$H \leq \lim_{\Delta \to 0} \lim_n s\Delta \sup \Pi_n \left(\sup_{\theta_1, \theta_2} g_n(\theta_1, \theta_2) > \epsilon \right) \leq s\epsilon^{-1} \lim_{\Delta \to 0} \lim_n \Delta \sup \mathrm{E} \sup_{\theta_1, \theta_2} g_n(\theta_1, \theta_2),$$

where the first inequality is true for some $s > 0$ and the second by the Markov inequality. We are thus left only with proving that the expected value is finite.

Considering the product of independent gamma distributions in Lemma 1, the modal value is

$$\frac{a(a-1)^{a-1} \exp^{1-a}}{\Gamma(a)} \frac{b}{a} \frac{c(c-1)^{c-1} \exp^{1-c}}{\Gamma(a)} \frac{d}{c},$$

which is bounded by $bd/(ac)$. Hence, by Lemma 1

$$\begin{aligned} \mathrm{E} \sup_{\theta_1, \theta_2} g_n(\theta_1, \theta_2) \quad &\leq \textstyle\sum_{A \in S_n} q_{n,A} \mathrm{E} \left\{ \frac{b + \sum_{i \in A} y_{i1}}{a - 1 + |A|} \frac{d + \sum_{i \in A} y_{i2}}{c - 1 + |A|} \right\} \\ &\leq \max \left\{ \tfrac{b}{a-1} \tfrac{d}{c-1}, \mu \right\}, \end{aligned}$$

where $\mu = E(y_{i1}, y_{i2})$, with both terms finite if $\mathrm{E}[\underline{Y}] < \infty$. This tightness argument is now available as an alternative to Theorem 5.2.2.

Theorem 5.3.2. *Under* A1 *and* $E[\underline{Y}] < \infty$*, the sequence* (G_n)*, defined applying NQZ to the mixture model* (5.21)*, converges weakly to* G^**,* P^* *– a.s.*

Proof. The sequence (G_n) is tight: the proof is given above. We can then apply Theorem 5.2.3 to prove that $(G_n(\theta))$ is Cauchy for all $\theta \in \Theta$. Corollary 5.2.1 gives the result since the uniqueness of the Laplace transform guarantees the uniqueness of the fixed point. □

5.3.2 General NQZ Algorithm

We first discuss the fixed point condition. Uniqueness of the fixed point (5.20) can be also proven with a fully statistical approach. The key is to make an ad hoc assumption on the relationship between the latent and the sample spaces: the probability distributions on $(\Theta, \mathcal{A})$ must be identifiable. We report here for completeness a sketch of the proof given in Walker (2017a). Let $M(G, \theta)$ be defined as

$$M(G, \theta) := \mathrm{E}_Y \left[Z(\theta, G, Y) \right].$$

Now, note that the integrand above is positive and bounded by one; so Fubini-Tonelli applies, and $M(G, \theta)$ can be rewritten as

$$M(G, \theta) = \int_{\mathcal{Y}} \int_{-\infty}^{\theta} k(y|\theta') \mathrm{d}G(\theta') \frac{p^*(y)}{p(y)} \mathrm{d}y, \tag{5.23}$$

and consequently $M(G^*, \theta)$ is

$$M(G^*, \theta) = \int_{\mathcal{Y}} \int_{-\infty}^{\theta} k(y|\theta') \mathrm{d}G^*(\theta') \mathrm{d}y = G^*(\theta).$$

If we assume that the family of densities indexed by θ,

$$p(y|G^*, \theta) = (G^*(\theta))^{-1} \int_{-\infty}^{\theta} k(y|\theta') dG^*(\theta'),$$

is complete, G^* is the only fixed point, i.e.

$$\mathrm{E}_Y \left[Z(\theta, G, Y) \right] = G(\theta) \quad \forall \theta \iff G \equiv G^*,$$

which makes condition *A2** verified. We denote the assumption of completeness by *A5*. Completeness of the family is somewhat related to an identifiability assumption of the mixture model. *A5* reconciles (5.19) and the framework in (5.8). Note that whereas the condition *A5* has a clear statistical interpretation, it can be unnecessary. The example in the previous subsection is an example. Treating (5.20) as a Fredholm equation of first kind formalizes these scenarios.

Another alternative proof of tightness is given in Theorem 5.2.2. In order to apply Theorem 5.2.2, the condition *A4* needs to hold. One can imagine that it is possible, for most bounded kernels, to show that *A4* holds. It is subject of current work to formalize it in a condition. We now state the theorem below assuming that *A4* holds.

Theorem 5.3.3. *Under* A1, A3 – A5, *the sequence* $(G_n)_{n\geq 1}$ *defined in* (5.19) *satisfies* $G_n \to G^*$ *weakly* P^* *- a.s. as* $n \to \infty$.

Proof. Follows by Corollary 5.2.1. □

We have thus been able to prove the convergence of (G_n) under less restrictive conditions than Tokdar et al. (2009) and Martin and Tokdar (2009). Some important family of mixtures, such as Gamma shape mixture, which was ruled out by these works, have now theoretical convergence guarantees.

5.4 Algorithms based on the copula

We study the asymptotic behaviour of HMW (defined in Subsection 2.2.2), RR (defined in Section 3.1) and TSPR (defined in Section 3.2) and prove that they converge weakly under some given conditions. The assumptions of Theorem 5.2.2 are difficult to verify for these family of algorithms: since they all share a similar update, we first assume each sequence we study to be tight, then prove it at the end of the section (Subsection 5.4.4). The proof of tightness is given assuming that the observations take value in a compact set, say $[-M, M]$ with arbitrary large M. We deals with two possible assumptions: M either known or unknown. Hence, convergence holds only for distribution with a compact support. Unbounded support is left for future work. Whereas this is not cause of concerns for applications, a more general formulation is an interesting open question.

5.4.1 HMW Algorithm

The HMW algorithm was described in Subsection 2.2.2 and the definition was given on a density scale. We define it in here in the distribution function scale to align it with the setting of this chapter. Let $\hat{\mathcal{P}}$ denote the class of Gaussian location mixture probability distributions. Observations are assumed to be *i.i.d.* $Y_n \overset{iid}{\sim} P^*$ with $P^* \in \hat{\mathcal{P}}$. As observations are collected, an initial guess $P_0(y) \in \hat{\mathcal{P}}$, is updated recursively through (5.8), where Z is defined as

$$Z(y, P, Y) = \Phi\left(\frac{\Phi^{-1}(P(y)) - \rho\Phi^{-1}(P(Y))}{\sqrt{1-\rho^2}}\right), \tag{5.24}$$

where Φ denotes the standard Gaussian distribution function, Φ^{-1} its inverse, and $\rho \in (0, 1)$.

Hahn et al. (2017) investigate the asymptotic convergence properties of the sequence (P_n) defined by (5.8) – (5.24) and prove that it converges with respect to the Kullback–Leibler divergence. The authors use the notion of *almost supermartingale* to bound the Kullback–Leibler distance. Despite the fact they study a stronger mode of convergence, the conditions are somewhat restrictive. In particular, Hahn et al. require the weights (α_n) to be defined for all n as

$$\alpha_n = a(n+1)^{-1},$$

with $0 < a < (2\rho+2)/(7\rho+1)$. The choice of the sequence (α_n) is crucial in determining the numerical performance of any SA algorithms. Whereas this assumptions satisfies *A1*, it is desirable to keep (α_n) as flexible as possible. See Subsection 2.3.2 for more details on the large sample behaviour of a HMW sequence.

It is easy to see that HMW falls into the general framework (5.8). Interestingly *A2** holds for the update (5.24): the proof is available in a 2015 unpublished report by Hahn, Martin and Walker (arXiv:1508.07448v3). We provide it here for completeness since it is an unpublished result.

Lemma 5.4.1. *When restricted to distribution functions P supported on $\mathbb{R}$, $E_Y[Z(y,P,Y)]$, where Z is defined by* (5.24)*, admits a unique fixed point P^**

$$E_Y[Z(y,P,Y)] = P(y) \quad \forall y \iff P \equiv P^*. \tag{5.25}$$

Proof. $\mathrm{E}_Y[Z(y,P^*,Y)] = P^*(y)$ for all y can be seen from a simple change of variable, putting $z = \Phi^{-1}(P^{*-1}(Y'))$. Uniqueness is given using the fact that $H_\rho(y,P,Y')$ is a distribution function in Y, conditioned on $Y' = y'$. Both marginals are equal to P, hence the only way that $\mathrm{E}_Y[Z(y,P,Y)] = P(y)$ can be obtained is if we integrate the conditional distribution with respect to P^*. □

Theorem 5.4.2. *Under* A1 *and assuming that (P_n) is tight, the sequence $(P_n)_{n\geq 1}$ defined in* (5.8) – (5.24) *satisfies $P_n \to P^*$ weakly P^* - a.s. as $n \to \infty$.*

Proof. Theorems 5.2.3 applies to (P_n). *A2** follows from Lemma 5.4.2. The statement follows from Corollary 5.2.1. □

5.4.2 Recursive Regression Algorithm

We prove that the recursive regression algorithm converges weakly to the true distribution under some given conditions. Recall that the recursive regression is a generalization of HMW: the two are equivalent if $\mathcal{X}$, the space of covariates, has only one element. Therefore it is natural to expect that some of the results given in the previous subsection apply here. The setting is quite different and the argument is more delicate: we will

emphasise the differences. The update of the recursive regression algorithm is different from Z but we are able to draw a connection. Below the definition of the recursive regression algorithm.

Recursive Regression (RR). For all $x \in \mathcal{X}$ fix an initial guess for $P_0(y|x) \in \hat{\mathcal{P}}$, a decreasing deterministic sequence of weights $(\alpha_n) \in (0,1)$, the correlation function $\rho : \mathcal{X} \times \mathcal{X} \to [0,1]$ defined by

$$\rho(x, x') = \frac{\rho_0 \, \tau}{\tau + |x - x'|^p},$$

Given observations $(y_n, x_n)_{n=1}^N$, an estimate of $P_N(y|x)$ is given by repeating recursively

$$P_n(y|x) = (1 - \alpha_n) P_{n-1}(y|x) + \alpha_n H_{\rho(x,x_n)}(P_{n-1}(y|x), P_{n-1}(y_n|x_n)), \tag{5.26}$$

for $i = 1, \ldots, n$, and for all $x \in \mathbb{X}$, where $H_{\rho(\cdot,\cdot)}$ is a CDF defined as

$$H_{\rho(x,x')}(P(y|x), P(y'|x')) = \Phi\left(\frac{\Phi^{-1}(P(y|x)) - \rho(x,x')\Phi^{-1}(P(y'|x'))}{\sqrt{1 - \rho(x,x')^2}}\right),$$

where (y', x') denotes one observation.

Suppose (Y_n) to be conditionally *i.i.d.* such that $Y_n|x_n \overset{ind}{\sim} P^*(y|x_n)$, $\forall n \geq 1$ with $x_n \in \mathcal{X} = \{x'^1, ..., x'^m\}$, m finite and $P^*(\cdot|x) \in \hat{\mathcal{P}}$ for all x. Let $\psi(P(\cdot|x), P(\cdot|x'))$ be defined as

$$\psi(P(y|x), P(\cdot|x')) = \mathrm{E}_{Y'|x'} \; [H_{\rho(x,x')}(P(y|x), P(Y'|x'))]. \tag{5.27}$$

ψ is used to define the fixed point in this setting. ψ extends the fixed point of HMW to a regression design. The fact that $\psi(\cdot,\cdot)$ can fit two probability measures constitutes an important difference with the fixed point condition we have defined for HMW. Such a difference will become apparent when discussing the fixed point. However, when $x = x'$ it is that

$$\psi(P(y|x), P(\cdot|x)) = \mathrm{E}_{Y|x}[Z(y, P(\cdot|x), Y)], \tag{5.28}$$

where Z is the one defined in (5.24). We can use some of the results in Section 5.2 because of (5.28).

We assume that the following conditions hold:

A5. $\sum_{i:x_i=x}^{\infty} \alpha_i = \infty$ and $\sum_{i:x_i=x}^{\infty} \alpha_i^2 < \infty$ for all $x \in \mathcal{X}$.

A6. If $P_n(y|x) \to P_\infty(y|x)$ as $n \to \infty$ for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$, it holds that for no

$\gamma \in \mathbb{R}$, $y \in \mathcal{Y}$, and $x \in \mathcal{X}$:

$$\frac{\sum_{i:x_i=x}^{n} \alpha_i}{\sum_{i:x_i \neq x}^{n} \alpha_i} \nrightarrow \gamma \; \frac{\sum_{j:x_j \neq x}[\psi(P_\infty(y|x), P_\infty(\cdot|x_j)) - P_\infty(y|x)]}{[\psi(P_\infty(y|x), P_\infty(\cdot|x)) - P_\infty(y|x)]}.$$

Now, *A5* implies *A1* and specifies how often the deterministic design points are observed. *A6* is a technical condition we use to identify the limit. We are making an assumption regarding the relationship between the sequence of weights, the order of observing different covariate values, the mapping ψ, and the asymptotic limit of $(P_n(y|x))$ (if it exists). Note that the left and right sides in *A6* are two quantities which are totally unrelated. The left side depends solely on how the observed covariates are ordered. The right side is a function of the updates and the asymptotic limit of the sequence $(P_n(y|x))$. Obviously, if the existence of the asymptotic limit is not verified, the condition cannot hold. We are not making any assumption on the limiting value.

A few words on the fixed point in this setting. It will be similar to *A2** but suitably changed to incorporate covariates. The RR update $H_{\rho(x,x')}$ satisfies the following condition: fix $x \in \mathcal{X}$

$$\mathrm{E}_{Y|x'}[H_{\rho(x,x')}(P(y|x), P(Y|x'))] = P(y|x) \quad \forall y \in \mathcal{Y}, x' \in \mathcal{X} \iff P \equiv P^*. \tag{5.29}$$

It is easy to see the proof of (5.29): P^* is a fixed point of $\mathrm{E}_{Y|x}[H_{\rho(x,x')}(P^*(y|x), P^*(Y|x'))]$ by an argument almost identical to that in Lemma 5.4.1; the uniqueness holds by Lemma 5.4.1 when $x = x'$, see (5.28).

We are now ready to state the main result of the section.

Theorem 5.4.3. *Under* A5, A6*, and assuming that* $(P_n(\cdot|x))$ *is tight for all* $x \in \mathcal{X}$*, the sequence* $(P_n(\cdot|x))$ *defined by* (5.26) *satisfies* $P_n(\cdot|x) \to P^*(\cdot|x)$ *weakly* P^**-a.s. as* $n \to \infty$ *for all* $x \in \mathcal{X}$*.*

Proof. For a fixed $x \in \mathcal{X}$, $(P_n(\cdot|x))$ is tight by assumption. For any fixed $y \in \mathcal{Y}$ $(P_n(y|x))$ is Cauchy by Theorem 5.2.3 and thus let $P_\infty(y|x)$ denote the limit.

Now we want to show that $P_\infty(y|x) = P^*(y|x)$ *a.s.* The rationale is similar to that in Corollary 5.2.1, although the argument is more delicate. Let us consider the following function,

$$M(P_{n-1}(y|x), y_n, x_n) = H_{\rho(x,x_n)}(P_{n-1}(y|x), P_{n-1}(y_n|x_n)) - \psi(P_{n-1}(y|x), P_{n-1}(\cdot|x_n)).$$

By (5.26), $P_n(y|x)$ can be rewritten as

$$P_n(y|x) = P_0(y|x) + \sum_{i=1}^{n} \alpha_i M(P_{i-1}(y|x), y_i, x_i) \tag{5.30}$$
$$+ \sum_{i=1}^{n} \alpha_i [\psi(P_{i-1}(y|x), P_{i-1}(\cdot|x_i)) - P_{i-1}(y|x)].$$

The central term on the right side is a martingale; more precisely, a bounded martingale difference. The condition $\sum_{i=1}^{\infty} a_i^2 < \infty$ ensures uniform integrability and so by the martingale convergence theorem, this sequence converges to a finite mean random variable. Since also $(P_n(y|x))$ converges by Theorem 5.2.3, it must be that the last term on the right side does not diverge. We can rewrite it as

$$\sum_{i:x_i=x} \alpha_i [\psi(P_{i-1}(y|x), P_{i-1}(\cdot|x_i)) - P_{i-1}(y|x)] + \sum_{j:x_j \neq x} \alpha_j [\psi(P_{j-1}(y|x), P_{j-1}(\cdot|x_j)) - P_{j-1}(y|x)]$$

Now, $A7$ implies that

$$\sum_{i:x_i=x} \alpha_i [\psi(P_{i-1}(y|x), P_{i-1}(\cdot|x)) - P_{i-1}(y|x)]$$

does not diverge and since $\sum_{i:x_i=x}^{\infty} \alpha_i = \infty$ and the fact that $P_n(y|x) \to P_\infty(y|x)$, we must have that for all (y, x)

$$\psi(P_\infty(y|x), P_\infty(\cdot|x)) - P_\infty(y|x) = 0$$

which in turn implies that $P_\infty(y|x) = P^*(y|x)$, since it is the only points that satisfy this for all y and x, see (5.29). $(P_n(y|x))$ converges pointwise for all $y \in \mathcal{Y}$ to $P^*(y|x)$. Tightness guarantees that the limit is a proper probability distribution. The statement follows. □

5.4.3 Two Steps Predictive Recursion

We have introduced all the asymptotic arguments needed to study the convergence of the recursive algorithms discussed and introduced in this thesis. However, the asymptotic theory of the Two Steps Predictive Recursion requires special care. Suppose that the observables are sample *i.i.d.* from one of the two groups: for each n one observes either $Y_n|x_n = 1 \overset{ind}{\sim} P^{(a)*}$ or $Y_n|x_n = 0 \overset{ind}{\sim} P^{(b)*}$, with $P^{(a)*}$, $P^{(b)*} \in \hat{\mathcal{P}}$. The latent variable x_n is deterministic. Given observations $(y_n, x_n)_{n=1:N}$, TSPR gives an estimate $(P_n^{(a)}, P_n^{(b)})$. We restate the definition of the algorithm below.

Two Steps Predictive Recursion (TSPR). Choose two initial guesses $P_0^{(a)}$,$P_0^{(b)} \in \hat{\mathcal{P}}$, $\rho \in$

$(0,1)$, two decreasing sequences of weights $(\alpha_n)_{n\geq 1}$ and $(\beta_n)_{n\geq 1}$ taking values in $(0,1)$ such that $\beta_n/\alpha_n \to 0$ as $n \to \infty$. Given observations $(y_n, x_n)_{n=1}^N$, the estimates $P_n^{(a)}(y)$ and $P_n^{(b)}(y)$ are given by

If $x_n = 1$, then first compute

$$P_n^{(a)}(y) = (1-\alpha_n)\, P_{n-1}^{(a)}(y) + \alpha_n Z(y, P_{n-1}^{(a)}, y_n), \tag{5.31}$$

where Z is defined in (5.24)*, then*

$$P_n^{(b)}(y) = (1-\beta_n)\, P_{n-1}^{(b)}(y) + \beta_n P_n^{(a)}(y), \tag{5.32}$$

else if $x_n = 0$, use (5.31) *for $P_n^{(b)}(y)$ and* (5.32) *for $P_n^{(a)}(y)$.*

repeated recursively for $n = 1, ..., N$.

The peculiarity of the algorithm is that it enforces somewhat a tree structure: for each new observation we use one of two possible updates. To elaborate on this, suppose w.l.o.g. that given an estimate $P_n^{(a)}$ and an observation (y_{n+1}, x_{n+1}), we want to compute $P_{n+1}^{(a)}$. If $x_{n+1} = 1$, $P_{n+1}^{(a)}$ is computed through (5.31), else through (5.32). It follows that after n observations there are 2^n possible different estimates. A finite sequence $x_{1:n}$ uniquely identifies a path on this tree. Note also that the first update is identical to HWM. The second one was suggested the predictive distribution of a Bivariate DPM. See Section 3.2.

We consider the convergence of the sequence $(P_n^{(a)})$. It will be clear that the same result holds for $(P_n^{(b)})$. Suppose the following two conditions hold:

A7. $\sum_{n\geq 1:x_n=1}^{\infty} \alpha_n = \sum_{n\geq 1:x_n=0}^{\infty} \alpha_n = \infty$, $\sum_{n\geq 1:x_n=1}^{\infty} \alpha_n^2 < \infty$ and $\sum_{n\geq 1:x_n=0}^{\infty} \alpha_n^2 < \infty$.
A8. $\beta_n > 0$, $\sum_{n=1}^{\infty} \beta_n < \infty$.

It is clear that in order to prove any mode of convergence, some paths of the tree need to be made inadmissible through some appropriate conditions. For example, assume we observe only observations sample from $P^{(b)*}$ and consequently always update $P_n^{(a)}$ via (5.32), clearly $P_n^{(a)}$ will not converge to $P^{(a)*}$. To rule out such a scenario, we need to assume that both groups are observed *i.o.* and at a *similar rate*. This notion is formalized by *A7*. Trivially *A7* implies *A1*.

A further remark. The condition $\beta_n/\alpha_n \to 0$ as $n \to \infty$ was included in the definition of TSPR to keep the second step (5.32) interpretable. Recall that (5.32) follows from

the dependence between the two groups assumed a priori: the more we are observing the correct group, the less we want to rely on observations sampled from the other and consequently the less I am going to update my estimate using (5.32). *A7* and *A8* imply $\beta_n/\alpha_n \to 0$ as $n \to \infty$ and are necessary to prove convergence to the true distribution. Note that the convergence of the series of weights implies a strong dependence on the initial point, in our setting *A7* ensures that such the dependence on $P_0^{(a)}$ and $P_0^{(b)}$ is lost as $n \to \infty$. See Section 3.2.2 for a more thorough explanation.

We are now ready to state the main theorem of the section.

Theorem 5.4.4. *Under* A7-A8*, and assuming that* $(P_n^{(a)})$ *is tight, the sequence* $(P_n^{(a)})$ *defined in* (3.16)-(3.17) *satisfies* $P_n^{(a)} \to P^{(a)*}$ *weakly* $P^{*\infty}$*-a.s. as* $n \to \infty$.

Proof. $(P_n^{(a)})$ is tight by assumption. Let us fix $y \in \mathcal{Y}$, it is easy to show that $(P_n^{(a)}(y))$ is Cauchy. For any $N < M$, Note that $P_n^{(a)}(y)$ can be rewritten as

$$|P_M^{(a)}(y) - P_N^{(a)}(y)| = \sum_{i \geq N : x_i = 1}^{M} \alpha_i \left[Z\left(y, P_{i-1}^{(a)}, y_i\right) - P_{i-1}^{(a)}(y) \right] + \tag{5.33}$$
$$+ \sum_{i \geq N : x_i = 0}^{M} \beta_i \left[P_i^{(b)}(y) - P_{i-1}^{(a)}(y) \right].$$

The first term in (5.33) can be bounded by an argument identical to the one used in Theorem 5.2.3. The existence of a convergent subsequence invoked in Theorem 5.2.3 is guaranteed by the fact that any subsequence has a further convergence subsequence because $(P_n^{(a)}(y))$ is relatively compact because it is tight. The second series is trivially bounded since $\sum_n \beta_n < \infty$.

We are now left to show that the limit is $P^{(a)*}$. For every $y \in \mathcal{Y}$ we can rewrite $P_n^{(a)}(y)$ as

$$P_n^{(a)}(y) = P_0^{(a)}(y) + \sum_{i \geq 1 : x_i = 1}^{n} \alpha_i \left\{ Z\left(y, P_{i-1}^{(a)}, y_i\right) - \mathrm{E}_Y \left[Z\left(y, P_{i-1}^{(a)}, Y\right) \right] \right\} + \tag{5.34}$$
$$+ \sum_{i \geq 1 : x_i = 1}^{n} \alpha_i \left\{ \mathrm{E}_Y \left[Z\left(y, P_{i-1}^{(a)}, Y\right) \right] - P_{i-1}^{(a)}(y) \right\} + \sum_{i \geq 1 : x_i = 0}^{n} \beta_i \left[P_i^{(b)}(y) - P_{i-1}^{(a)}(y) \right].$$

The first sum is a zero-difference martingale sequence. The last sum is convergent as $\sum_{n=1}^{\infty} \beta_n < \infty$ (*A8*). Since $\sum_{n \geq 1 : x_n = 1}^{\infty} \alpha_n = \infty$ (*A7*), the third term in (5.34) needs to accumulate around zero. By Lemma 5.4.1 the limit must be $P^{(a)*}(y)$. because it is the only function for which the equality holds.

□

5.4.4 Tightness

Throughout this section we have assumed that the sequences of CDFs under study are tight. We give in this section a proof of this assumption. In particular, let us consider a sequence (P_n) defined by HMW, and let (Π_n) the corresponding sequence of distributions. To prove that the sequence of (Π_n) is tight, we have to impose a restriction on P^*, the true sampling distribution. We assume that $P^* \in \mathcal{P}_M$, the class of continuous probability distributions on $[0, M]$, such that a random variable $Y \sim P^*$ takes values in $(0, M)$, for all $P^* \in \mathcal{P}_M$.

In order to prove that the sequence (Π_n) is *tight*, we show that, asymptotically, Π_n assigns probability one to the set $\mathcal{P}_M$. Indeed recall that in HMW, the update is given by

$$Z(y, P, Y) = \Phi\left(\frac{\Phi^{-1}(P(y)) - \rho\Phi^{-1}(P(Y))}{\sqrt{1-\rho^2}}\right),$$

which is defined on the real line. We need to check is that the limit P_∞, if it exists, is a distribution function on $[0, M]$. Since (P_n) is a sequence of distribution functions by definition, we need to check that

$$P_\infty(0) = 0 \quad \text{and} \quad P_\infty(M) = 1,$$

Note that in order to prove the condition above, we do not need to assume that (P_n) converges, but simply to where $(P_n(M))$ and $(P_n(0))$ converge to. What we are showing is that the sequence (P_n) will asymptotically lie into $\mathcal{P}_M$ with probability one. This implies tightness because the set of probability distribution on a compact set is compact. This can also be seen by Billingsley (1999), Chapter 1, Section 5.

We first prove that the sequence (Π_n) is tight assuming that M is known. Note that if M is known, one chooses $P_0 \in \mathcal{P}_M$, which ensures $P_0(M) = 1$ and $P_0(0) = 0$. Given that the random variable Y is such that $0 < Y_n < M$ for all n, the sequence (P_n) lies in $\mathcal{P}_M$ for all n. (Π_n) is then trivially tight, because the space of probability distributions on a compact is also compact; see also Ruggiero and Walker (2009).

Let us now discuss the case where M is unknown. A consequence is that one cannot assume a priori that $P_0 \in \mathcal{P}_M$. Let $B > M$, consider $(P_n(B))$, a sequence defined through HMW with $0 < Y_n < M$ for all n. We want to show that $(P_n(B))$ converges to 1. Note that

$$\Phi\left(\frac{\Phi^{-1}(P(B)) - \rho\Phi^{-1}(P(Y))}{\sqrt{1-\rho^2}}\right) > \Phi\left(\frac{\Phi^{-1}(P(B)) - \rho\Phi^{-1}(P(M))}{\sqrt{1-\rho^2}}\right),$$

Hence, if we define a new sequence, which we denote by $(\widehat{P}_n(B))$, where all the observations are considered at a value M, this new sequence will always be smaller than $(P_n(B))$. Note that this argument applies because we are able to show that the sequence would lie in the space of probability distributions on $[0, B]$ with probability one. Therefore the argument does not need to apply for any B.

Such a sequence can be constructed through the two sequences defined below:

$$\widehat{P}_n(M) = (1-\alpha_n)\widehat{P}_{n-1}(M) + \alpha_n \Phi\left(\frac{\Phi^{-1}(P_{n-1}(M)) - \rho\Phi^{-1}(P_{n-1}(M))}{\sqrt{1-\rho^2}}\right),$$

$$\widehat{P}_n(B) = (1-\alpha_n)\widehat{P}_{n-1}(B) + \alpha_n \Phi\left(\frac{\Phi^{-1}(P_{n-1}(B)) - \rho\Phi^{-1}(P_{n-1}(M))}{\sqrt{1-\rho^2}}\right),$$

where $(\widehat{P}_n(M))$ is necessary in order to update $\widehat{P}_n(B)$. Since for all n

$$P_n(B) > \widehat{P}_n(B),$$

it is enough to prove that $\widehat{P}_n(B) \to 1$ as $n \to \infty$. $(\widehat{P}_n(B))$ and $(\widehat{P}_n(M))$ are two deterministic sequences that are easier to deal with than the stochastic sequence $(P_n(B))$.

We are able to show through computations that the convergence $\widehat{P}_n(B)$ to 1 holds, see Figure 5.1. The plots in Figure 5.1 are robust to different choices of $\rho \in (0,1)$, (α_n) and $\widehat{P}_0(M), \widehat{P}_0(B) \in (0,1)$, with 0 and 1 excluded, otherwise they are the accumulation points.

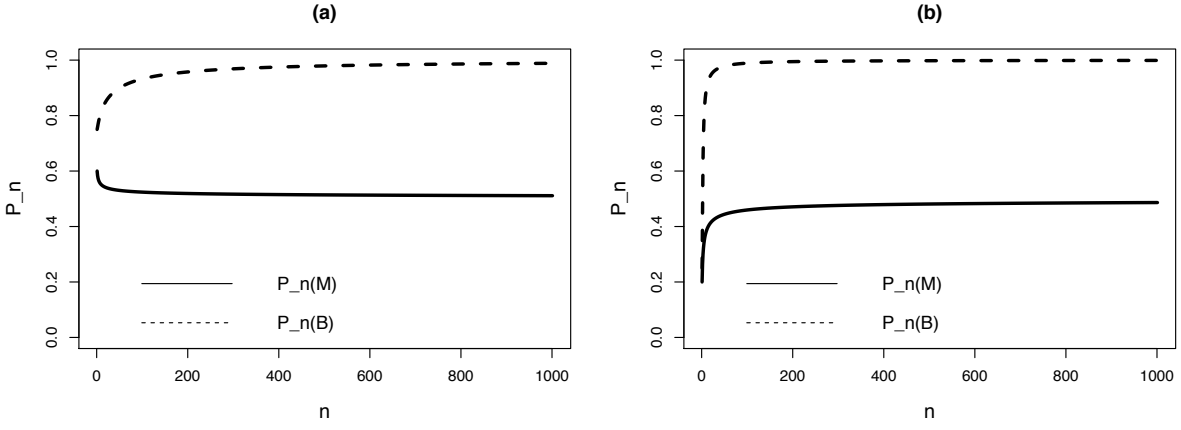


Figure 5.1: Plots of $\widehat{P}_n(B)$ and $\widehat{P}_n(M)$

To support the plots in Figure 5.1, we sketch the heuristics of a proof. The argument is given in two steps. First, one shows that $(\widehat{P}_n(M))$ converges to $1/2$; then one shows that $(\widehat{P}_n(B))$ to 1. Let us now consider $(\widehat{P}_n(M))$. As Figure 5.1 exhibits, there are

two scenarios that depend on whether $\widehat{P}_0(M) \in [0, 1/2)$ or $\widehat{P}_0(M) \in (1/2, 1]$. The case $\widehat{P}_0(M) = 1/2$ is trivial. The sequence has three accumulation points: 0, 1/2 and 1.

If $\widehat{P}_0(M) \in (1/2, 1]$, it will be monotonically decreasing, i.e.

$$\widehat{P}_{n+1}(M) < \widehat{P}_n(M),$$

and

$$\Phi(\lambda \Phi^{-1}(\widehat{P}_n(M))) \geq \frac{1}{2},$$

with $\lambda = \sqrt{(1-\rho)/(1+\rho)} < 1$, which in turns implies that $\widehat{P}_{n+1}(M) \geq 1/2$ for all n. A monotonically decreasing sequence bounded below is convergent, hence $\widehat{P}_n(M) \to 1/2$ as $n \to \infty$. We now need to establish that $(\widehat{P}_n(B))$ to 1. Given that $\widehat{P}_n(M) \to 1/2$, for large n,

$$\widehat{P}_n(B) = (1-\alpha_n)\widehat{P}_{n-1}(B) + \alpha_n \Phi\left(\widehat{\lambda}\Phi^{-1}(P_{n-1}(B)) - \epsilon_{n-1}\right),$$

where $\widehat{\lambda} = \sqrt{1-\rho^2}^{-1} > 1$ and $\epsilon_{n-1} = \rho/\sqrt{1-\rho^2}\Phi(\widehat{P}_{n-1}(M))$. The convergence of $\widehat{P}_n(M)$ implies that $\epsilon_n \to 0$ as $n \to \infty$, which in turn implies that asymptotically

$$\widehat{P}_n(B) = (1-\alpha_n)\widehat{P}_{n-1}(B) + \alpha_n \Phi\left(\widehat{\lambda}\Phi^{-1}(P_{n-1}(B))\right).$$

The above hints that $\widehat{P}_n(B)$ is going to be monotonically increasing for large n since $\widehat{\lambda} > 1$. The convergence is then given by the fact that $(\widehat{P}_n(B))$ has an upper bound, which is 1. The case $\widehat{P}_0(M) \in [0, 1/2)$ can be dealt with a similar argument.

The extension of this proof to RR is straightforward. The extension to TSPR follows because any linear combination between $(P_n^{(a)})$ and $(P_n^{(b)})$, will lie in $\mathcal{P}_M$. The extension of this argument to the case $M = \infty$ is not trivial and is subject of current research. All the simulations in Chapter 3 and Hahn et al. (2017) suggest that convergence is achieved when also with unbounded support. A theoretical guarantee is currently not available.

5.5 Discussion

Chapter 5 deals with the asymptotic theory of the recursive algorithm discussed across the thesis. The algorithms we discussed fit the SA framework and pointed out a gap in the literature: none of the current existing works cover our infinite dimensional setting, which we believe to be a very common problem in the statistical world.

The main results are in Section 5.2. There we conceptualized a general algorithm to

estimate a CDF and studied its asymptotic behaviour. The algorithm is very simple: it consists in a linear combination between the previous step P_{n-1} and a properly defined update. We introduced a novel asymptotic argument: along to a very common conditions on the deterministic sequence (α_n), the key property that must hold is that the expected value of the update must admit an unique fixed point. A martingale argument is used in the proof.

We illustrated our results to several algorithms discussed in this thesis. Our asymptotic results apply to NQZ and allow us to to prove convergence under milder conditions than those currently employed for some given models. We have also shown that the asymptotic theory we have proposed can be extended to a setting where the observables are not *i.i.d.*: the two new algorithms we have introduced, RR and TSPR, are shown to be consistent under some given assumptions.

There are some obvious directions for future work. The first direction is how to deal with tightness. The condition *A4* is not fully satisfactory because it is not straightforward to show that it holds for the copula algorithms. There are some bounds for tightness less restrictive than (5.13). Theorem 13.5 in Billingsley (1999) contains a general version of (5.13),

$$\mathrm{E}[|P_n(s_2) - P_n(s_1)|^{2\beta}|P_n(s_3) - P_n(s_2)|^{2\beta}] \leq (F(s_3) - F(s_1))^{2\alpha}, \tag{5.35}$$

where $\alpha > 1/2$, $\beta \geq 0$ and F is a nondecreasing continuous function on $[0, R]$, and s_1, s_2, s_3 and P_n defined in the proof of Theorem 5.2.2. F does not need to bounded but it is not trivially to define properly. It is the subject of current work.

A novel tightness argument would be an important contribution to give robust theoretical guarantees to RR and TSPR, which now are proven convergent if the sampling distribution has a compact support. We left tightness with an unbounded support as an open problem. Note that it is enough a new proof of tightness to extend our result to P^* being defined on the real line. We have shown in Chapter 3 through several illustrations that both algorithms perform well in a variety of illustration where P^* is defined on the real line. Therefore this does not seem a concern for applications, but a convergence result under milder conditions is highly desirable.

We discussed the connection to stochastic approximation and try to apply our method to some of the traditional SA schemes. We discuss in Chapter 2 that the literature focuses on ODE methods. The limit is seen as a solution of a differential equation. Our approach lies almost to the opposite side of the spectrum: the limit is the fixed point solution of an expectation - see *A2* and *A2**. Whereas it is clear that fixed points are

connected to *stable* solution of corresponding ODE, a direction for future work would be making these heuristics rigorous. The added value would be the possibility to extends SA theory to function spaces that have yet not been studied.

Another line of research is other asymptotic results which are often desirable: stronger mode of convergence, convergence rates and asymptotic normality. We focus briefly on the first one. HMW was proven to be consistent in the KL distance using the idea of almost supermartingale (see Section 2.3.2). If one was interested into a stronger mode of convergence than the weak studied in this chapter, this idea would be the place to start. RR extends HMW to a regression design, although we did not manage to prove the same type of convergence for this algorithm. The proof in Hahn et al. (2017) would actually come along if one was able to prove that

$$\int \left(\int \psi_\theta(P(y|x)) p^*(y|x) \mathrm{d}y - 1 \right) \left(\int \psi_\theta(P(y'|x')) p^*(y'|x') \mathrm{d}y' - 1 \right) N(\theta|0, \rho) \mathrm{d}\theta > 0,$$

where ψ_θ is defined in (5.27). Now consider $X = \int \psi_\theta(P(y|x)) p^*(y|x) \mathrm{d}y$ and $Y = \int \psi_\theta(P(y'|x')) p^*(y'|x') \mathrm{d}y'$. If $X \equiv Y$, the above is true because it is $\mathrm{Var}(X)$. If not, it is the $\mathrm{Cov}(X, Y)$, and it is not straightforward to see under which conditions it holds.

Chapter 6

Discussion and Future Work

This thesis is about recursive algorithms, with a particular emphasis on a family of recursions linked to Bayesian statistics. We have studied their asymptotic properties and numerical performance. We have also enlarged the range of applications. We lack a dataset driven objective but the motivation underlying this thesis is very practical. Recursive algorithms have the potential to address several issues discussed in statistics nowadays, like the development of methodologies that are computationally tractable and sequential. Two algorithms available in the literature, the ones due to Newton et al. (1998) and Hahn et al. (2017), provide an excellent example of the validity of this claim. In this thesis we confirm the effectiveness of this family of algorithms with new empirical - Chapters 3 and 4 - and theoretical arguments - Chapter 5.

It is a very basic rationale that allows us to develop a recursive procedure mimicking a predictive distribution: the first predictive update, considered to be a good one, is repeated iteratively with some suitable adjustments. In this thesis we give some clear guidelines on which adjustments are important. Chapter 5 shows that a recursive estimator needs to satisfy two conditions to be consistent: admitting a unique fixed-point and having a deterministic sequence of weights that carries a sample size effect. In Chapter 3 we propose two new algorithms that successfully apply this simple rationale to more complex statistical problems: estimation of covariate-dependent distributions and regression with regressors taking finitely many values.

In our opinion, the results of this thesis suggest that it is worth exploring the possibility to build new algorithms. We have provided both heuristics (Chapters 3 and 4) and requirements for an algorithm to converge (Chapter 5), that should offer some clear guidelines on how to propose new recursive estimators. Armed with the reassuring results available in the literature and in the thesis, an important research direction is to move even further away from the *i.i.d.* setting. Chapter 3 deals with coovariate-dependent dis-

tributions. Chapter 4 with multivariate random variables whose univariate components are dependent. In the discussion of Chapter 4 we have suggested a few ways in which the recursive regression can be extended. There, the key idea was to develop some multi-steps procedures to "carry more information" at each iteration. Time-series is for example an interesting area to tackle. The order dependence of the algorithm should be irrelevant in this setting. If it was really so, it would cancel out a drawback that these procedures have.

Chapter 5 summarizes the theoretical ideas in the thesis. The content belongs to an ongoing work on the large sample behaviour of a general type of recursive algorithms: those whose update admits a unique fixed point. We are able to state, under some appropriate assumptions, the weak convergence of the algorithm (Theorems 5.2.2 and 5.2.3, Corollary 5.2.1). Our results allow us to study the weak convergence of the algorithm introduced by Newton et al. (1998) under milder conditions that those currently used: some popular mixture models that did not have theoretical convergence guarantees, such as Gamma shape mixtures, are now shown to be consistent.

The convergence theory presented in Chapter 5 fills an evident gap into the literature on recursive procedures. The algorithms we have presented fit the general framework of stochastic approximation (SA). However, despite its popularity, SA for infinite dimensional functions is still an open research area. We proposed in this thesis a novel approach that apply to a sequence of distribution functions. We argued that an interesting area of research is to understand the applicability of our results into a general SA framework and draw a parallel between the conditions we require and the more traditional SA asymptotic theory based on ODE methods. In our opinion, the uniqueness of the fixed point which we require, has a clear statistical interpretation and it is more suitable than the ODE theory in several applications in statistics.

A critical issue in Chapter 5 was how to prove tightness of the sequence of probability measures governing the sequence defined recursively. We have discussed a couple of alternative proofs but a lot of work still needs to be done. It is very difficult to give a general argument that encompasses a broad spectrum of procedures while relying on requirements that are fairly easy to check. We believe that our assumptions are reasonable, although they still remain difficult to check for specific algorithms. Therefore it is important to keep working in this area.

There are some other obvious research directions if one was interested in continuing our work on asymptotic theory for recursive procedures. The most relevant one would be to obtain a Central Limit type of theorem. A second research direction is to get the convergence rates. These two results would be relevant in applications to provide

uncertainty quantification of our estimates. For example in inverse problems, the rates could allow to determine how many observations to sample. Lastly, we have not investigated stronger modes of convergence: all the algorithms have been discussed on a CDF scale, if one was interested into estimating densities, stronger asymptotic results would be desirable. The connection with SA could give some intuitions on how to obtain these results since Central Limit Theorems, convergence rates and L_1 convergence are often available for SA algorithms.

Lastly, these recursive algorithms have a good potential in applications. We focus on one: the use of statistical models to tackle numerical problems, in particular inverse problems, which is an emerging area of study. See probabilistic numerics for example. The use of a recursive procedure seems very appropriate in this context: they are fast, easy to implement and parallelisable. Furthermore there is a theoretically unlimited supply of observations since we can sample from known distribution. We have shown a clear example, the inversion of a Laplace transform, where the advantages of using a recursive algorithm are evident. The literature is not limited to our result. See for example Walker (2017b) for an iterative algorithm that allows to solve a system of linear equations. Inverse problems is a very important area in applied mathematics and there are plenty of open problems that are worth pursuing. Among the many, integral equations could be a good place to start - see for example the Fredholm equation.

Bibliography

Abate, J., Choudhury, G. L. and Whitt, W. (1998), 'Numerical inversion of multidimensional Laplace transforms by the Laguerre method', *Performance Evaluation* **31**, 229–243.

Abate, J. and Whitt, W. (1992), 'The Fourier-series method for inverting transforms of probability distributions', *Queueing Systems* **10**, 5–87.

Abate, J. and Whitt, W. (2006), 'A unified framework for numerically inverting Laplace transforms', *INFORMS Journal of Computing* **18**, 408–421.

Aitchison, J. and Aitken, C. G. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**, 413–420.

Amann, H. (1976), 'Fixed point equations and nonlinear eigenvalue problems in ordered banach spaces', *SIAM Review* **18**, 620–709.

Antoniano-Villalobos, I., Wade, S. and Walker, S. G. (2014), 'A Bayesian nonparametric regression model with normalized weights: a study of hippocampal atrophy in Alzheimer's disease', *Journal of the American Statistical Association* **109**, 477–490.

Asmussen, S., Jensen, J. L. and Rojas-Nandayapa, L. (2016), 'On the Laplace transform of the lognormal distribution', *Methodology and Computing in Applied Probability* **18**, 441–458.

Barron, A., Schervish, M. J., Wasserman, L. et al. (1999), 'The consistency of posterior distributions in nonparametric problems', *Annals of Statistics* **27**, 536–561.

Berti, P., Pratelli, L. and Rigo, P. (2004), 'Limit theorems for a class of identically distributed random variables', *Annals of Probability* **32**, 2029–2052.

Bharucha-Reid, A. et al. (1976), 'Fixed point theorems in probabilistic analysis', *Bulletin of the American Mathematical Society* **82**, 641–657.

Billingsley, P. (1999), *Convergence of Probability Measures*, Wiley, Chicago.

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

Böhning, D. and Seidel, W. (2003), 'Editorial: recent developments in mixture models', *Computational Statistics and Data Analysis* **41**, 349–357.

Bruni, C. and Koch, G. (1985), 'Identifiability of continuous mixtures of unknown Gaussian distributions', *Annals of Probability* **13**, 1341–1357.

Caflisch, R. E. (1998), 'Monte Carlo and Quasi-Monte Carlo methods', *Acta Numerica* **7**, 1–49.

Cai, N. and Kou, S. (2012), 'Pricing Asian options under a hyper-exponential jump diffusion model', *Operations Research* **60**, 64–77.

Canale, A., Durante, D. and Dunson, D. (2017), 'Convex mixture regression for quantitative risk assessment', *arXiv preprint arXiv:1701.02950* .

Cappello, L. and Walker, S. G. (2018), 'A Bayesian motivated Laplace inversion for multivariate probability distributions', *Methodology and Computing in Applied Probability* **to appear**, 1–21.

Chidume, C. (1981), 'On the approximation of fixed points of nonexpansive mappings', *Houston Journal of Mathematics* **7**, 3–5.

Choudhury, G. L., Lucantoni, D. M. and Whitt, W. (1994), 'Multidimensional transform inversion with applications to the transient M/G/1 queue', *Annals of Applied Probability* **4**, 719–740.

Cifarelli, D. and Regazzini, E. (1978), Problemi statistici non parametrici in condizioni di scambiabilita parziale e impiego di medie associative, Technical report, Quaderni Istituto di Matematica Finanziaria, Universitá di Torino.

Corduneanu, C. (1991), *Integral Equations and Applications*, Vol. 148, Cambridge University Press Cambridge.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015), 'Are Gibbs-type priors the most natural generalization of the Dirichlet process?', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004), 'An ANOVA model for dependent random measures', *Journal of the American Statistical Association* **99**, 205–215.

Delyon, B., Lavielle, M. and Moulines, E. (1999), 'Convergence of a stochastic approximation version of the EM algorithm', *Annals of Statistics* **27**, 94–128.

Dieuleveut, A., Bach, F. et al. (2016), 'Nonparametric stochastic approximation with large step-sizes', *Annals of Statistics* **44**, 1363–1399.

Downton, F. (1970), 'Bivariate exponential distributions in reliability theory', *Journal of the Royal Statistical Society. Series B* **32**, 408–417.

Dubner, H. and Abate, J. (1968), 'Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform', *Journal of the ACM* **15**, 115–123.

Dunson, D. B. and Park, J.-H. (2008), 'Kernel stick-breaking processes', *Biometrika* **95**, 307–323.

Efromovich, S. (2007), 'Conditional density estimation in a regression setting', *Annals of Statistics* **35**, 2504–2535.

Escobar, M. D. (1994), 'Estimating normal means with a Dirichlet process prior', *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M. D. and West, M. (1992), Computing bayesian nonparametric hierarchical models, Technical report, Institute of Statistics and Decision Sciences, Duke University.

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

Fabian, V. (1960), 'Stochastic approximation methods', *Czechoslovak Mathematical Journal* **10**, 123–159.

Fahlman, S. E. (1988), An empirical study of learning speed in back-propagation networks, Technical report, Carnagie Mellon University.

Fan, J. (1991), 'On the optimal rates of convergence for nonparametric deconvolution problems', *Annals of Statistics* **19**, 1257–1272.

Feller, W. G. (1971), *An Introduction to Probability Theory and its Applications, Vol. 2*, Wiley, New York.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *Annals of Statistics* **1**, 209–230.

Ferguson, T. S. (1983), 'Bayesian density estimation by mixtures of normal distributions', *Recent Advances in Statistics* **24**, 287–302.

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

Fusai, G. (2004), 'Pricing Asian options via Fourier and Laplace transforms', *Journal of Computational Finance* **7**, 87–106.

Gaver Jr, D. P. (1966), 'Observing stochastic processes, and approximate transform inversion', *Operations Research* **14**, 444–459.

Gelenbe, E. (2006), *Computer System Performance Modeling in Perspective: a Tribute to the Work of Professor Kenneth C. Sevcik*, Vol. 1, World Scientific.

Ghannadian, F., Alford, C. and Shonkwiler, R. (1996), 'Application of random restart to genetic algorithms', *Information Sciences* **95**, 81–102.

Ghosh, J. K. and Tokdar, S. T. (2006), Convergence and consistency of Newton's algorithm for estimating mixing distribution, *in* 'Frontiers in Statistics', Imperial College Press 429–443 (J. Fan and H. L Koul, eds.).

Gladyshev, E. (1965), 'On stochastic approximation', *Theory of Probability and Its Applications* **10**, 275–278.

Goffard, P.-O., Loisel, S. and Pommeret, D. (2015), 'Polynomial approximations for bivariate aggregate claims amount probability distributions', *Methodology and Computing in Applied Probability* **19**, 151–174.

Granas, A. and Dugundji, J. (2013), *Fixed Point Theory*, Springer Verlag.

Grübel, R. and Hermesmeier, R. (1999), 'Computation of compound distributions i: Aliasing errors and exponential tilting', *Astin Bulletin* **29**, 197–214.

Hahn, P. R., Martin, R. and Walker, S. G. (2017), 'On recursive Bayesian predictive distributions', *Journal of the American Statistical Association* **to appear**.

Hammersley, J. M. and Handscomb, D. C. (1964), General principles of the Monte Carlo method, *in* 'Monte Carlo Methods', Springer.

Hartman, P. and Olech, C. (1962), 'On global asymptotic stability of solutions of differential equations', *Transactions of the American Mathematical Society* **104**, 154–178.

Hayfield, T., Racine, J. S. et al. (2008), 'Nonparametric econometrics: the np package', *Journal of Statistical Software* **27**, 1–32.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010), *Bayesian Nonparametrics*, Vol. 28, Cambridge University Press.

Hu, X., Shonkwiler, R. and Spruill, M. (1997), Iterative improvement plus random restart stochastic search, Technical report, Georgia Institute of Technology.

Isaacs, D., Altman, D., Tidmarsh, C., Valman, H. and Webster, A. (1983), 'Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for igg, iga, igm.', *Journal of Clinical Pathology* **36**, 1193–1196.

Ishikawa, S. (1974), 'Fixed points by a new iteration method', *Proceedings of the American Mathematical Society* **44**, 147–150.

Ishwaran, H. and James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**, 161–173.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P. and Rosner, G. L. (2011), 'DPpackage: Bayesian semi-and nonparametric modeling in R', *Journal of Statistical Software* **40**, 1–30.

Jin, T., Provost, S. B. and Ren, J. (2016), 'Moment-based density approximations for aggregate losses', *Scandinavian Actuarial Journal* **2016**, 216–245.

Jin, T. and Ren, J. (2010), 'Recursions and fast Fourier transforms for certain bivariate compound distributions', *Journal of Operational Risk* **5**, 19.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011), 'Slice sampling mixture models', *Statistics and Computing* **21**, 93–105.

Kiefer, J., Wolfowitz, J. et al. (1952), 'Stochastic estimation of the maximum of a regression function', *Annals of Mathematical Statistics* **23**, 462–466.

Krasnosel'skii, M. A. (1955), 'Two remarks on the method of successive approximations', *Uspekhi Matematicheskikh Nauk* **10**, 123–127.

Kushner, H. J. and Clark, D. S. (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer Verlag.

Kushner, H. J. and Shwartz, A. (1985), 'Stochastic approximation in Hilbert space: Identification and optimization of linear continuous parameter systems', *SIAM Journal on Control and Optimization* **23**, 774–793.

Kushner, H. and Yin, G. G. (2003), *Stochastic Approximation and Recursive Algorithms and Applications*, Vol. 35, Springer Verlag.

Kuznetsov, A. (2013), 'On the convergence of the Gaver-Stehfest algorithm', *SIAM Journal on Numerical Analysis* **51**, 2984–2998.

Lai, T. L. (2003), 'Stochastic approximation', *Annals of Statistics* **31**, 391–406.

Laird, N. (1978), 'Nonparametric maximum likelihood estimation of a mixing distribution', *Journal of the American Statistical Association* **73**, 805–811.

Lee, A., Yau, C., Giles, M. B., Doucet, A. and Holmes, C. C. (2010), 'On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods', *Journal of Computational and Graphical Statistics* **19**, 769–789.

Lin, M. (1998), 'The uniform zero-two law for positive operators in banach lattices', *Studia Mathematica* **2**, 149–153.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, Vol. 5, Institute of Mathematical Statistics.

Lindsay, B. G. et al. (1983), 'The geometry of mixture likelihoods: a general theory', *Annals of Statistics* **11**, 86–94.

Ljung, L. (1978), 'Strong convergence of a stochastic approximation algorithm', *Annals of Statistics* **22**, 680–696.

Lo, A. Y. (1984), 'On a class of Bayesian nonparametric estimates: I. density estimates', *Annals of Statistics* **12**, 351–357.

Lo, A. Y. (1986), 'Bayesian statistical inference for sampling a finite population', *Annals of Statistics* **14**, 1226–1233.

MacEachern, S. N. (1999), Dependent nonparametric processes, *in* 'ASA proceedings of the section on Bayesian statistical science', American Statistical Association 50–55.

MacEachern, S. N. (2000), Dependent Dirichlet processes, Technical report, Department of Statistics, The Ohio State University.

Mann, W. R. (1953), 'Mean value methods in iteration', *Proceedings of the American Mathematical Society* **4**, 506–510.

Martin, R. (2009), Fast nonparametric estimation of mixing distributions with application to high-fimensional inference, PhD thesis, Purdue University, Department of Statistics.

Martin, R. and Ghosh, J. K. (2008), 'Stochastic approximation and Newton's estimate of a mixing distribution', *Statistical Science* **23**, 365–382.

Martin, R. and Tokdar, S. T. (2009), 'Asymptotic properties of predictive recursion: robustness and rate of convergence', *Electronic Journal of Statistics* **3**, 1455–1472.

Martin, R. and Tokdar, S. T. (2011), 'Semiparametric inference in mixture models with predictive recursion marginal likelihood', *Biometrika* **98**, 567–582.

McLachlan, G. and Peel, D. (2004), *Finite mixture models*, John Wiley and Sons.

Mnatsakanov, R. M. (2011), 'Moment-recovered approximations of multivariate distributions: the Laplace transform inversion', *Statistics and Probability Letters* **81**, 1–7.

Mnatsakanov, R. and Ruymgaart, F. (2003), 'Some properties of moment-empirical cdf's with application to some inverse estimation problems', *Mathematical Methods of Statistics* **12**, 478–495.

Mokkadem, A., Pelletier, M. et al. (2007), 'A companion for the Kiefer–Wolfowitz–Blum stochastic approximation algorithm', *Annals of Statistics* **35**, 1749–1772.

Morokoff, W. J. and Caflisch, R. E. (1995), 'Quasi-Monte Carlo integration', *Journal of Computational Physics* **122**, 218–230.

Morris, C. N. (1982), 'Natural exponential families with quadratic variance functions', *Annals of Statistics* **10**, 65–80.

Muliere, P. and Tardella, L. (1998), 'Approximating distributions of random functionals of Ferguson-Dirichlet priors', *Canadian Journal of Statistics* **26**, 283–297.

Mustapha, H. and Dimitrakopoulos, R. (2010), 'Generalized Laguerre expansions of multivariate probability densities with moments', *Computers and Mathematics with Applications* **60**, 2178–2189.

Neal, R. M. (2000), 'Markov chain sampling methods for Dirichlet process mixture models', *Journal of Computational and Graphical Statistics* **9**, 249–265.

Neal, R. M. (2003), 'Slice sampling', *Annals of Statistics* **31**, 705–767.

Nelsen, R. B. (2007), *An Introduction to Copulas*, 2 edn, Springer Verlag.

Newton, M. A. (2002), 'On a nonparametric recursive estimator of the mixing distribution', *Sankhyā, Series A* **64**, 306–322.

Newton, M. A., Quintana, F. A. and Zhang, Y. (1998), Nonparametric Bayes methods using predictive updating, *in* 'Practical Nonparametric and Semiparametric Bayesian Statistics', Vol. 133 (D. Dey, P. Muller and D. Sinha, eds.), Springer 45–61.

Newton, M. A. and Zhang, Y. (1999), 'A recursive algorithm for nonparametric analysis with missing data', *Biometrika* **86**, 15–26.

Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia.

Pitman, J. and Yor, M. (1997), 'The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator', *Annals of Probability* **25**, 855–900.

Polyak, B. T. and Juditsky, A. B. (1992), 'Acceleration of stochastic approximation by averaging', *SIAM Journal on Control and Optimization* **30**, 838–855.

Quintana, F. A. and Newton, M. A. (2000), 'Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences', *Journal of Computational and Graphical Statistics* **9**, 711–737.

Rao, J. and Sethuraman, J. (1975), 'Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors', *Annals of Statistics* **3**, 299–313.

Révész, P. (1973), 'Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes I', *Studia Scientiarum Mathematicarum Hungarica* **8**, 391–398.

Révész, P. (1977), 'How to apply the method of stochastic approximation in the non-parametric estimation of a regression function 1', *Statistics: A Journal of Theoretical and Applied Statistics* **8**, 119–126.

Ridout, M. (2009), 'Generating random numbers from a distribution specified by its Laplace transforms', *Statistics and Computing* **19**, 439–450.

Robbins, H. and Monro, S. (1951), 'A stochastic approximation method', *Annals of Mathematical Statistics* **22**, 400–407.

Robbins, H. and Siegmund, D. (1971), A convergence theorem for non negative almost supermartingales and some applications., *in* 'Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)', Academic Press, New York 233–257.

Rojas-Nandayapa, L. (2008), Risk probabilities: asymptotics and simulation, PhD thesis, Aarhus Universitet, Faculty of Science, Department of Mathematical Sciences.

Ruggiero, M. and Walker, S. G. (2009), 'Bayesian nonparametric construction of the fleming-viot process with fertility selection', *Statistica Sinica* **19**, 707–720.

Sakrison, D. J. (1965), 'Efficient recursive estimation; application to estimating the parameters of a covariance function', *International Journal of Engineering Science* **3**, 461–483.

Salov, G. I. (1980), 'On a stochastic approximation theorem in a Hilbert space and its applications', *Theory of Probability and Its Applications* **24**, 413–419.

Seidler, J., Žák, F. et al. (2017), 'A note on continuous-time stochastic approximation in infinite dimensions', *Electronic Communications in Probability* **22**, 1–13.

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

Smith, A. and Makov, U. (1978), 'A quasi-Bayes sequential procedure for mixtures', *Journal of the Royal Statistical Society. Series B* **40**, 106–112.

Smola, A., Gretton, A., Song, L. and Schölkopf, B. (2007), A Hilbert space embedding for distributions, *in* 'International Conference on Algorithmic Learning Theory', Springer 13–31.

Stehfest, H. (1970), 'Algorithm 368: Numerical inversion of Laplace transforms', *Communications of the ACM* **13**, 47–49.

Stuart, A. M. (2010), 'Inverse problems: a Bayesian perspective', *Acta Numerica* **19**, 451–559.

Talbot, A. (1979), 'The accurate numerical inversion of Laplace transforms', *IMA Journal of Applied Mathematics* **23**, 97–120.

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

Tokdar, S., Martin, R. and Ghosh, J. (2009), 'Consistency of a recursive estimate of mixing distributions', *Annals of Statistics* **37**, 2502–2522.

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

Walker, S. G. (2017a), 'A Laplace transform inversion method for probability distribution functions', *Statistics and Computing* **27**, 439–448.

Walker, S. G. (2017b), 'An iterative algorithm for solving sparse linear equations', *Communications in Statistics - Simulation and Computation* **46**, 5113–5122.

Walker, S. G. and Muliere, P. (2003), 'A bivariate Dirichlet process', *Statistics and Probability Letters* **64**, 1–7.

Weideman, J. A. C. (2006), 'Optimizing Talbot's contours for the inversion of the Laplace transform', *SIAM Journal on Numerical Analysis* **44**, 2342–2362.

Whitt, W. (1970), 'Weak convergence of probability measures on the function space C[0,infinity)', *Annals of Mathematical Statistics* **41**, 939–944.

Yin, G. (1992), 'On h–valued stochastic approximation: finite dimenstional projections', *Stochastic Analysis and Applications* **10**, 363–377.

Yin, G. and Zhu, Y. (1990), 'On h-valued Robbins-Monro processes', *Journal of Multivariate Analysis* **34**, 116–140.