Bayesian Statistical Learning and Applications

Julyan Arbel

Habilitation à Diriger les Recherches

Spécialité : Mathématiques Appliquée et Informatique

Université Grenoble-Alpes

Inria Grenoble Rhône-Alpes

Bayesian Statistical Learning and Applications

Julyan Arbel

Rapporteurs	
Alessandra Guglielmi	Professor, Politecnico di Milano
Éric Moulines	Professeur, École polytechnique & Académie des Sciences
Yee Whye Teh	Professor, University of Oxford & DeepMind

Membres du jury	
Stéphane Girard	Directeur de recherche, Inria Grenoble Rhône-Alpes
Anatoli Juditsky	Professeur, Université Grenoble Alpes
Adeline Samson	Professeur, Université Grenoble Alpes

Remerciements, Acknowledgements

Ce mémoire doit beaucoup à de nombreuses personnes.

Je suis très reconnaissant à Alessandra Guglielmi, Éric Moulines et Yee Whye Teh d'avoir accepté d'en être rapporteurs. C'est un honneur et un plaisir de savoir ses travaux lus et appréciés par des chercheurs que l'on estime. Je vous remercie d'avoir consacré du temps à ce manuscript, et de vous être déplacé jusqu'à Grenoble dans une période chargée en enseignements (et qui plus est, le lendemain d'un semi-marathon pour un de vous). Alessandra, merci d'avoir accepté d'ouvrir le bal avec un exposé. Merci Adeline Samson, Anatoli Juditsky et Stéphane Girard d'avoir accepté de faire partie du jury.

Ghislaine Gayraud et Judith Rousseau, puis Kerrie Mengersen, vous avez su me démontrer par l'exemple pendant ma thèse combien l'exercice de la recherche était passionnant, et aussi et surtout, me soutenir dans les moments difficiles. Je vous en remercie.

Igor Prünster, merci pour ta confiance, et pour l'opportunité de rejoindre le Collegio Carlo Alberto, d'abord à Turin puis à Milan, pour un long postdoc fait de nouvelles collaborations, de découvertes (d'espèces notamment !) et d'une nouvelle culture. La supervision laissant une part belle à l'autonomie m'a permis de développer des liens profonds avec mes collègues du Collegio, Antonio Lijoi, Bernardo Nipoti, Guillaume Kon Kam King, Stefano Favaro, Pierpaolo De Blasi, Matteo Ruggiero, Antonio Canale, Bertrand Lods et Giovanni Pistone.

Je remercie le comité de sélection des *Fellowships* de l'*Alan Turing Institute* de ne pas m'avoir classé. Cela m'a permis a posteriori de ne connaître le *brexit* que de "l'extérieur".

This thesis owes a lot to many people.

I am very grateful to Alessandra Guglielmi, Éric Moulines and Yee Whye Teh for agreeing to be rapporteurs. It is an honour and a pleasure to know that one's work is read and appreciated by researchers we highly think of. Thank you for devoting time to this manuscript, and for travelling to Grenoble in a period full of teachings (and what's more, the day after a half-marathon for one of you). Alessandra, thank you for agreeing to open the ball with a presentation. Thank you, Adeline Samson, Anatoli Juditsky and Stéphane Girard for agreeing to be part of the jury.

Ghislaine Gayraud and Judith Rousseau, as well as Kerrie Mengersen: you showed me by example during my Ph.D. thesis how much exciting research was, and also and above all, thanks for supporting me in difficult times.

Igor Prünster, thank you for your confidence, and for the opportunity to join the Collegio Carlo Alberto, first in Turin and then in Milan, for a long postdoc made of new collaborations, (species) discoveries, and a new culture. The supervision, leaving a lot of room for autonomy, has allowed me to develop deep links with my colleagues at the Collegio, Antonio Lijoi, Bernardo Nipoti, Guillaume Kon Kam King, Stefano Favaro, Pierpaolo De Blasi, Matteo Ruggiero, Antonio Canale, Bertrand Lods and Giovanni Pistone.

I thank the Fellowship Selection Committee of the Alan Turing Institute for not ranking me. This allowed me a posteriori to know the brexit only "from the outside".

Merci Florence Forbes, Stéphane Girard et Jean-Baptiste Durand pour votre accueil à Inria et pour avoir soutenu ma candidature, ainsi que Judith Rousseau, Kerrie Mengersen, Christian Robert, Peter Müller and Igor Prünster. Mon arrivée à Grenoble et mes collaborations et discussions avec vous et d'autres grenoblois, dont Emmanuel Barbier, Michel Dojat, Alexis Arnaud, Pablo Mesejo, Jakob Verbeek, Julien Mairal, Maria Laura Delle Monache, Eugenio Cinquemani, Wilfried Thuiller, Sophie Achard, Simon Barthelmé, Michael Blum, Adeline Samson, m'ont permis d'élargir mes centres d'intérêts en statistiques: valeurs extrêmes, copules, modèles graphiques dont les champs de Markov, apprentissage profond, expectation-propagation, ainsi que des applications en neuroimagerie, au traffic routier, en écologie comme la modélisation jointe de distributions d'espèce, etc. Jakob, Wilfried et Eugenio, merci de m'avoir gentiment, mais de manière récurrente, incité à passer l'HDR.

Thank you Florence Forbes, Stéphane Girard and Jean-Baptiste Durand for your welcome to Inria and for supporting my application, as well as Judith Rousseau, Kerrie Mengersen, Christian Robert, Peter Müller and Igor Prünster. My position at Inria and my collaborations and discussions with you and others in Grenoble, including Emmanuel Barbier, Michel Dojat, Alexis Arnaud, Pablo Mesejo, Jakob Verbeek, Julien Mairal, Maria Laura Delle Monache, Eugenio Cinquemani, Wilfried Thuiller, Sophie Achard, Simon Barthelmé, Michael Blum, Adeline Samson, have allowed me to broaden my interests in statistics: extreme values, copulas, graphic models including Markov fields, deep learning, expectation-propagation (yes Simon, I should program it one day to understand EP) as well as applications in neuroimaging, road traffic, ecology such as joint species distribution models, etc. Jakob, Wilfried and Eugenio, thank you for having me kindly, but repeatedly, encouraged to take the HDR.

J'ai eu la chance de pouvoir échanger et collaborer avec de nombreux étudiants ces deux dernières années à Inria, Marta Crispino, Hongliang Lü, Riccardo Corradin, Łukasz Rajkowski, Mariia Vladimirova, Verónica Muñoz Ramírez, Fabien Boux, Caroline Lawless, Aleksandra Malkova, Michał Lewandowski, Daria Bystrova, Giovanni Poggiato, Sharan Yalburgi. J'apprécie votre dynamisme, et me sais chanceux de vous avoir (eu) ici à Inria. J'ai également beaucoup appris dans nos discussions lors des visites des "moins jeunes", Bernardo Nipoti (présent chez Mistis 'every other week' grâce à Ulysses, entre autres), Guillaume Kon Kam King, Olivier Marchal, Rémi Bardenet, Jean-Bernard Salomond, Botond Szabó, Eric Marchand, Robin Ryder, Hien Nguyen, Nicolas Lartillot, Alisa Kirichenko, et Matteo Sesia.

I have had the opportunity to exchange and collaborate with many students over the past two years in Inria, Marta Crispino, Hongliang Lü, Riccardo Corradin, Łukasz Rajkowski, Mariia Vladimirova, Verónica Muñoz Ramírez, Fabien Boux, Caroline Lawless, Aleksandra Malkova, Michał Lewandowski, Daria Bystrova, Giovanni Poggiato, Sharan Yalburgi. I appreciate your dynamism, and I know I'm lucky to have (had) you here at Inria. I also learned a lot in our discussions with the visits of more seniors ones, Bernardo Nipoti (present at Mistis "every other week" thanks to Ulysses, among others), Guillaume Kon Kam King, Olivier Marchal, Rémi Bardenet, Jean-Bernard Salomond, Botond Szabó, Eric Marchand, Robin Ryder, Hien Nguyen, Nicolas Lartillot, Alisa Kirichenko, and Matteo Sesia.

Merci Stephen Walker et Peter Müller pour votre accueil à UT Austin au printemps 2017, ces trois mois ont été extrêmement productifs pour moi; merci Matti Vihola, Éric Parent, Didier Fraix-Burnet pour vos invitations à présenter des cours de statistique Bayésienne en école d'été, et Richard Nickl, Hanne Kekkonen, Fabrizio Ruggeri, Bernardo Nipoti, Roberto Cassarin, Raffaele Argiento, Pierre Chainais, Alice Cleynen, Françis Sillion, Kerrie Mengersen, Antonio Lijoi, Matteo Ruggiero, Mame Diarra Fall, Bruno Gaujal, Jim Griffin, Silvia Montagna, Fabrizio Leisen, Jean-François Cœurjolly, Eric Marchand, Rebecca Steorts, Anne-Laure Fougères, Adeline Samson, Bas Kleijn, Sara Wade, Aurore Lavigne, Jean-Bernard Salomond, Célestin Kokonendji, Florence Forbes, Benjamin Guedj, Christophe Biernacki, Igor Prünster, Nicolas Chopin, François Caron, Michele Guindani, pour vos invitations à présenter en conférence ou en séminaire. Merci Michele Guindani et Hien Nguyen de m'avoir proposé de rejoindre les équipes de vos journaux. Merci Pierre Jacob, Jérôme Le et Robin Ryder pour la mise en orbite de Statisfaction en 2010 (quoi il y a 9 ans déjà ?!?) qui a été pendant ma thèse, et reste encore aujourd'hui, un excellent moyen d'expression.

Je remercie particulièrement mes coauteurs, déjà tous cités plus haut, et dont les textes sont repris dans ce manuscript parfois à l'insu de leur plein gré! J'ai appris énormément sur vos sujets; et surtout, ces projets communs ont été autant de moments inoubliables passés ensemble.

Thank you Stephen Walker, Peter Müller for your welcome to UT Austin in the spring of 2017, these three months have been extremely productive; thank you Matti Vihola, Éric Parent, Didier Fraix-Burnet for your invitations to present Bayesian statistics courses in summer schools, and Richard Nickl, Hanne Kekkonen, Fabrizio Ruggeri, Bernardo Nipoti, Roberto Cassarin, Raffaele Argiento, Pierre Chainais, Alice Cleynen, Françis Sillion, Kerrie Mengersen, Antonio Lijoi, Matteo Ruggiero, Mame Diarra Fall, Bruno Gaujal, Jim Griffin, Silvia Montagna, Fabrizio Leisen, Jean-François Cœurjolly, Eric Marchand, Rebecca Steorts, Anne-Laure Fougères, Adeline Samson, Bas Kleijn, Sara Wade, Aurore Lavigne, Jean-Bernard Salomond, Célestin Kokonendji, Florence Forbes, Christophe Biernacki, Igor Prünster, Nicolas Chopin, François Caron, Michele Guindani, for your invitations to present at conferences or seminars. Thank you Michele Guindani and Hien Nguyen for offering me to join editorial boards of great journals. Thank you Pierre Jacob, Jérôme Le and Robin Ryder for putting Statisfaction into orbit in 2010 (what, it's already been nine years?!?). This blog has been an excellent means of expression during my Ph.D. thesis, and still is today.

I would particularly like to thank my coauthors, all of whom are already mentioned above, whose texts are included in this manuscript, sometimes without them even knowing. I have learned a lot about your subjects, but above all these joint projects have been unforgettable moments spent together.

Enfin, merci à ma famille; merci Christelle, mon *matching prior*[1], pour ton soutien constant; merci Jeanne et Léonie, nos deux observations pseudo-aléatoires chéries, nos starlettes pipelettes adorées qui savez si bien nous soustraire à nos équations et autres écrans d'ordinateur et nous ramener à la vraie vie.

Finally, thank you to my family; thank you Christelle, my matching prior[2], for your constant support; thank you Jeanne and Léonie, our two cherished pseudo random observations, our beloved chattering starlets who know so well how to take us away from our equations and computer screens and bring us back to real life.

[1] En paraphrasant légèrement Xian dans son Choix Bayésien.
[2] By slightly paraphrasing Xian in his Bayesian Choice.

Contents

Introduction

Contents

Bayesian inference refers to Thomas Bayes (170?-1761) who first used conditional probability to describe the probability of an event given prior knowledge of conditions related to the event, known as Bayes' theorem and published as [47]. A more general version of Bayes' theorem was introduced by Pierre-Simon Laplace (1749-1827). Bayesian inference used to be called "inverse probability" due to the fact that inference goes backward from effects to causes, or from observations to parameters in statistics terms. In the beginning of the twentieth century, alternative approaches referred to as frequentist statistics were developped and essentially superseded Bayesian methods. Bayesian inference research and applications attracted again a lot of attention in the eighties, mostly due to the advent of Markov chain Monte Carlo algorithms.

Bayesian inference has found applications in a broad range of fields. This chapter introduces the scientific context of my research activities within Bayesian Nonparametrics and Bayesian Machine Learning (Section 0.1) and then summarizes my contributions (Section 0.2) arranged in three parts: Chapter 1. Bayesian Nonparametric Mixture Modeling, Chapter 2. Approximate Bayesian Inference and Chapter 3. Distributional Properties of Statistical and Machine Learning Models.

0.1 Scientific context

0.1.1 Bayesian Nonparametrics (BNP)

The Bayesian nonparametric choice comes from two main ideas. The Bayesian choice [284], on the one hand, which assumes that each physical system is subject to uncertainty, and models the data and the parameters by probability distributions. Nonparametric models [324], which are of particular interest when the parametric hypothesis proves to be too restrictive. At the intersection of these two concepts, Bayesian nonparametric methods provide a flexible methodology within a solid probabilistic framework [see monographs 157, 131]. Historically, Bayesian nonparametrics started in the early 1970s in California with the work of members of the Berkeley Department of Statistics. More specifically, the Dirichlet process, defined by [117], has given rise to a number of methodological and applied contributions. Most of the current BNP tools are based on extensions of the Dirichlet process.

The nonparametric denomination, commonly adopted, is rather to be understood as massively parametric: indeed, we consider large dimensional, or infinite dimensional, parameter spaces such as functional spaces. A Bayesian requires formulating a joint probability distribution on observations and parameters. This joint distribution consists of (i) the sample distribution of the model, which describes the degree of confidence given to the data for a particular value of the parameters, and (ii) the prior distribution, which can be interpreted in several ways, notably as information available on the parameters before experiment, or as regularization. The prior distribution takes the opposite of the so-called classical approach, where the parameters are assumed to be fixed. There is an optimal way to update the prior from the observations, which is Bayes' theorem. This learning rule defines the posterior distribution, which describes all the information available on the parameters, after experience. The main strengths of the non-parametric Bayes approach are:

- to define a natural framework for quantifying uncertainty, via the posterior distribution,
- to naturally avoid overfitting by regularizing the parameters thanks to the prior distribution specification,
- to adapt to the data complexity thanks to models whose number of parameters increases with the size of the data.

The flexibility of Bayesian nonparametric methods comes at a price, such as constructing prior distributions and sampling from posterior distributions over infinite dimensional parameter spaces.

Designing priors

Building a prior distribution on an infinite dimensional space may appear like contradictory: on the one hand, quantifying the information in the form of a distribution requires to know many fine aspects of the parameter to be modeled, while on the other hand the nonparametric choice stems from the desire to relax parametric assumptions. As a balance, an acceptable prior should take the form of a *default*

prior which 'is proposed by some automatic mechanism that is not in favor of any particular parameter values, and has low information content compared with the data' [131]. Key features of the prior, such as mean value, concentration around the mean value, should be available as hyperparameters to be set by the user, while the bulk of the prior is derived by some automatic mechanism. Such a prior does not need be noninformative. However, it should have large support in the sense that it spreads over the whole parameter space without overconcentrating in some parts of the space.

There is a wide range of tools for building prior distributions on infinite dimensional spaces. For functional spaces, stochastic processes such as Gaussian processes or processes with positive increments, or developments in function bases are typically used. In the context of probability measurement spaces, prior distributions are generally based on discrete processes that can be roughly divided into two main categories. First species sampling models (SSM), a class that is generally considered too general to be directly usable because key features mentioned above are not always tractable. This is why tractable special cases attracted a lot of attention, including stick-breaking processes, Gibbs-type processes, the Pitman–Yor process, or the Dirichlet process already mentioned. The second large family is obtained by normalizing so-called completely random measures (CRM), also known as Lévy processes. The advantage of this family over the first is that the moments of every order of the processes are generally known which can be very useful from a posterior sampling perspective. All these generalizations of the Dirichlet process offer a greater flexibility. In particular in the context of mixture models, they can induce a prior on the number of components that is less informative than the Dirichlet process.

Sampling from posteriors

Practical Bayesian inference requires to sample from the posterior distribution. Before the advent of Markov chain Monte Carlo methods [287], conjugacy used to be essential since it enabled to update by closed form expressions a prior into a posterior. Conjugacy still plays a central role in BNP for simplifying the posterior.

By essence, a BNP posterior distribution is infinite dimensional and cannot be sampled from directly. Reducing the parameter space from infinite to finite dimension is necessary, being it by analytically integrating out some parts of the parameters, or by approximating it, for example by truncating infinite summations. The aim is to break up the parameter into (a finite number of) finite-dimensional parts whose posterior distributions are tractable. The properties of the prior are important as they may suggest appropriate integration or approximation of the posterior. For instance, a Gaussian process prior induces a joint probability at a finite number of positions in the form of a simple multivariate normal.

Marginal methods Efficient posterior computation techniques often incorporate analytic integration of infinite dimensional parts of the parameter and posterior sampling. For instance, density estimation under a Dirichlet process mixture can be recast in a hierarchical model by introducing a latent (allocation) variable for

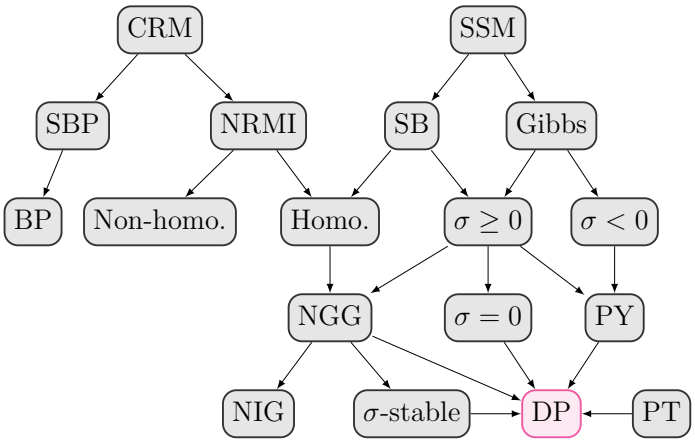


FIGURE 1: Some random probability distributions often used as priors. An arrow indicates that the target is a special case, a limit case or a normalization of its origin. CRM: Completely random measures. SSM: Species sampling models. SBP: Stable Beta process. NRMI: Normalized random measure with independent increments. SB: stick-breaking process. Gibbs: Gibbs-type process. BP: Beta process. (Non-) Homo: (Non-) Homogeneous NRMI. σ: discount parameter of Gibbs-type process. NGG: Normalized generalized Gamma process. PY: Pitman–Yor process. NIG: Normalized inverse Gaussian process. σ-stable: normalized σ-stable process. DP: Dirichlet process. PT: Pólya-tree.

each observation and integrate out analytically the Dirichlet process, given the latent variables, turning the problem into a finite dimensional one. Such techniques that rely on analytic integration of (part of) the process are termed *marginal methods*. The Blackwell–MacQueen Pólya urn scheme is a Markov chain Monte Carlo sampling scheme that relies on predictive distributions. For a description in the context of mixture modeling, see [220, 112] for the Dirichlet process, [164] for general stick-breaking priors, e.g., the two parameter Poisson–Dirichlet process prior, and [A1] for Gibbs-type process priors.

Conditional methods Techniques that directly sample trajectories of the process are termed *conditional methods*. Some are exact: they rely on augmenting the parameter space with some random variables, conditional on which only a finite number of parameters need be sampled and posterior sampling becomes feasible. These include the slice sampler [333] and the retrospective sampler [259]. Some are approximate: they consist in approximating the process by removing some infinite-dimensional part of it, preferably some part of least importance. When a stick-breaking representation is avalaible, we can mention the blocked Gibbs sampler [164], ϵ approximations of processes ([236] for the Dirichlet process, [A6] for the Pitman–Yor process). Without a stick-breaking representation, for example with normalized random measure with independent increments (NRMI), one typically resorts to the Ferguson and Klass algorithm or inverse Lévy method [119]. This is

usually considered as a slow sampling technique due to some heavy numerical evaluation, but it has the advantage to sample the elements of the process in (strictly) decreasing order (i.e. in order of least importance) while stick-breaking strategies sample them in *stochastic* decreasing order only. See [34, 43, 94, 111, 149, 148, 251, 249, 252, 248, 247, 335, 29, 257].

R packages Many R packages can solve BNP computational problem: DPpackage [173, 174] is a bundle of functions for many Bayesian nonparametric models based essentially on the Dirichlet process and Pólya tree, BNPmix package [70] addresses density estimation with the Dirichlet process and Pitman–Yor process mixing measures via marginal algorithms, while BNPdensity package [43], [S3] focusses on density estimation under normalized random measures with independent increments (see Figure 1).

Approximate Bayesian computation (ABC) This short review of computational aspects is far from exhaustively listing all techniques available for posterior sampling, leaving aside particle filtering or sequential Monte Carlo for instance [72]. However, let us mention approximate Bayesian computation (ABC), the most prominent strategy in situations where the likelihood function is known but not tractable, or when the likelihood function has entirely unknown form. In that case, one cannot exactly sample from the posterior in an inexpensive manner, or at all. In such situations, a sample from an approximation of the posterior may suffice in order to conduct the user's desired inference. Such a sample can be drawn via ABC. The ABC paradigm originates from the works of [291, 316, 274]; see [315] for details. Stemming from the initial listed works, there are now numerous variants of ABC methods. Some good reviews of the current ABC literature can be found in the expositions of [224, 332, 216, 184, 285]. The volume [302] provides a comprehensive treatment regarding ABC methodologies. The core philosophy of ABC is to define a quasi-posterior by comparing data with plausibly simulated replicates. The comparison is traditionally based on some summary statistics, the choice of which being regarded as a key challenge of the approach. In recent years, data discrepancy measures bypassing the construction of summary statistics have been proposed by viewing data sets as empirical measures. Examples include the Wasserstein distance [52], the Kullback–Leibler divergence [176], the energy distance [S4].

0.1.2 Bayesian Machine Learning

Machine learning is a fairly recent research field which requires skills from statistics and computer science. Neural networks (NNs), and their deep counterparts [139], have largely been used in many research areas such as image analysis [198], signal processing [142], or reinforcement learning [301], just to name a few. The impressive performance provided by such machine learning approaches has greatly motivated research that aims at a better understanding the driving mechanisms behind their effectiveness. In particular, the study of the NNs distributional properties through Bayesian analysis has recently gained much attention.

Distributional properties of Bayesian neural networks Studying the distributional behaviour of feedforward networks has been a fruitful avenue for understanding these models, as pioneered by the works of Radford Neal [241, 242] and David MacKay [221]. The first results in the field addressed the limiting setting when the number of units per layer tends to infinity, also called the wide regime. [242] proved that a single hidden layer neural network with normally distributed weights tends in distribution in the wide limit either to a Gaussian process [281] or to an α-stable process, depending on how the prior variance on the weights is rescaled. In recent works, [226] and [204] extend the result of Neal to more-than-one-layer neural networks: when the number of hidden units grows to infinity, deep neural networks (DNNs) also tend in distribution to the Gaussian process, under the assumption of Gaussian weights for properly rescaled prior variances. For the rectified linear unit (ReLU) activation function, the Gaussian process covariance function is obtained analytically [82]. For other nonlinear activation functions, [204] use a numerical approximation algorithm. This Gaussian process approximation is used for instance by [152] for improving neural networks training strategies. [255] extend the results by proving the Gaussian process limit for convolutional neural networks.

Various distributional properties are also studied in NNs regularization methods. The *dropout* technique [306] was reinterpreted as a form of approximate Bayesian variational inference [188, 124]. While [124] built a connection between dropout and the Gaussian process, [188] proposed a way to interpret Gaussian dropout. They suggested *variational dropout* where each weight of a model has its individual dropout rate. *Sparse variational dropout* [235] extends variational dropout to all possible values of dropout rates, and leads to a sparse solution. The approximate posterior is chosen to factorize either over rows or individual entries of the weight matrices. The prior usually factorizes in the same way, and the choice of the prior and its interaction with the approximating posterior family are studied by [162]. Performing dropout can be used as a Bayesian approximation but, as noted by [106], it has no regularization effect on infinitely-wide hidden layers. Recent work by [56] provides the expression of the first two moments of the output units of a one layer NN. Obtaining the moments is a first step towards characterizing the full distribution. However, the methodology of [56] is limited to the first two moments and to single-layer NNs, while we address the problem in more generality for deep NNs.

0.2 Research activity

My long term research objective is to contribute methodological and theoretical advances in Bayesian statistical learning. This section describes my process towards this goal.

0.2.1 Research conducted before and during my Ph.D.

I have started doing research during my final study internship from École polytechnique in the summer of 2006 at the University of Liverpool. I worked with Hugh Morton on *knot theory*, a field from topology concerned among other things with classifying knots up to invariance transformations. My contribution was to derive

some new relations between the crossing index of a knot (the minimal number of crossings in a planar representation of a knot) and its arc index (the minimal number of half planes needed to represented the knot in an open-book structure).

The next year, my first encounter with Bayesian statistic in an academic context was at the University of Valencia, Spain. Thanks to Nicolas Chopin, I had the chance to discover this research field by collaborating with José Miguel Bernardo, one of the founding fathers[3] of the International Society for Bayesian Analysis (ISBA). During this project, I have proposed a Bayesian model for studying the positive predicted value (PPV) of a medical test. Given a diagnostic test on a certain disease, the PPV represents the probability of having the disease given a positive test result.

Moving back to Paris, I decided to work on the problem of level set estimation with Judith Rousseau for my M2 thesis. At the beginning of my Ph.D., supervised by Judith Rousseau and Ghislaine Gayraud, I then moved to the infinite normal means model [A14], where I derived asymptotic frequentist properties of generic Bayesian procedures based on sieve priors. The main result is that these priors give rise to minimax adaptive posterior concentration rates. The result can be applied to different models, as illustrated in the article on the density, regression, nonlinear autoregression and Gaussian white noise models. We also prove in the paper that a well-behaved posterior distribution for a global loss can have a pathological behaviour for a pointwise loss (in the white noise model).

Later, I started to work on Bayesian nonparametric models based on discrete random probability models after attending a summer school course on Bayesian nonparametrics given by Peter Müller in 2010 at University of California, Santa Cruz. I am grateful to Sonia Petrone for hosting me for two months in 2011 and for fruitful discussions on this new line of research. At that time, I wanted to apply my work to some concrete data. I met with Kerrie Mengersen, from Queensland University of Technology, who was visiting Crest in Paris, and she introduced me to Australian ecologists colleagues of her. They had soil microbial data collected in Antarctica and wanted to understand how an environmental factor, a contaminant called Total Petroleum Hydrocarbon, would affect species diversity in the soil. I developped a covariate-dependent process as a prior distribution for modelling species data partially replicated at different locations. Such data can be represented as a frequency matrix giving the number of times each species is observed in each site. The aim is to study the impact of additional factors (covariates), for instance environmental factors, on the data structure, and in particular on the diversity. This work was published in [A13] with a focus on the application, and in [A10] with a complete description of the methodology and of theoretical results, such as moments and full support of the process.

As part of my work as a statistician at Insee (national statistics bureau of France) in parallel of my Ph.D., I developped with Vianney Costemalle a Bayesian model for describing migration flows from dissimilar datasets [A11], in French.

[3] Whose "ISBA Fellow status has been permanently revoked" in the meantime...

0.2.2 Contributions since my Ph.D.

After my Ph.D., I moved to Turin for a postdoc funded with the ERC of Igor Prünster. I feel extremely lucky to have worked there: at the time, Igor Prünster statistics group at the Collegio Carlo Alberto was composed of a number of experts of BNP including Antonio Lijoi, Matteo Ruggiero, Pierpaolo De Blasi, Stefano Favaro, Bernardo Nipoti and Antonio Canale. I also benefited a lot from discussions with Bertrand Lods and Giovanni Pistone, while Guillaume Kon Kam King joined the group in 2015. My research there flourished in a number of directions, including Bayesian nonparametric mixture modeling and approximate Bayesian inference, the first two chapters of this manuscript. I stayed three years in Italy, including a semester at the University of Bocconi in Milan when Igor Prünster moved there.

I then moved to Inria Grenoble Rhône-Alpes in 2016 for a research position (Chargé de recherche). I joined MISTIS team (soon to be called Statify) with Florence Forbes, Stéphane Girard and Jean-Baptiste Durand as permanent members. I wanted to take the opportunity of this new research environment to delve into new research areas. I got interested in Machine Learning with a Deep Learning reading group organized at Inria. Discussions with the team members opened some new research interests such as Markov random fields with Florence Forbes, extreme value theory and copulas with Stéphane Girard. I am also grateful to Inria for financial support of a three-month visit to Stephen Walker at the University of Texas at Austin in 2017 where I started to work on the sub-Gaussian property of random variables [A7]. Joint work with Steven started during this visit is still ongoing.

The rest of this manuscript is devoted to my main contributions to Bayesian statistical learning since my Ph.D. It is divided into three main chapters presenting :

- **Chapter 1. Bayesian Nonparametric Mixture Modeling**
- **Chapter 2. Approximate Bayesian Inference**
- **Chapter 3. Distributional Properties of Statistical and Machine Learning Models**

The borders between the categories are rather blurred and could be easily crossed depending on where the focus is put, should it be on applications, on models, on posterior sampling strategies, or whatsoever. Looking back, the three chapters roughly correspond to arXiv Statistics categories as follows:

- **Chapter 1. stat.AP - Applications & stat.ME - Methodology**
- **Chapter 2. stat.CO - Computation**
- **Chapter 3. stat.ML - Machine Learning & stat.TH - Statistics Theory**

But for the sake of readability, I will keep my original and more explicit chapter titles...

Chapter 1. Bayesian Nonparametric Mixture Modeling

This chapter is devoted to mixture models which are central to Bayesian nonparametrics. The original motive for mixture modeling is the desire to expand the spectrum of available distributions. For example, Quetelet [277] used mixtures in the nineteenth century to describe asymmetric or multimodal distributions. My contributions described in this chapter are concerned with mixtures of some kernel $\mathcal{K}(\cdot \mid \theta)$ by a nonparametric mixing measure P following some (nonparametric) prior $\mathcal{Q}$ as described in the *Designing priors* paragraph of Section 0.1.1. Such mixtures can be formulated as follows

$$\begin{aligned} P &\sim \mathcal{Q}, \\ f(\cdot) &= \int \mathcal{K}(\cdot \mid \theta) P(\mathrm{d}\theta). \end{aligned}$$

If P is a discrete random probability measure as those described in Figure 1, then the random function f may take the form of a countable infinite mixture. Depending on the choice of kernel, f can model a broad spectrum of functional objects: density, hazard rate, survival function or regression function.

My work on mixtures has focussed on *methodological aspects*, including on survival analysis [A12], [C5], robust analysis [S6]; on *computational questions*, such as recasting Dirichlet process mixtures into the sequential quasi Monte Carlo framework of [129] in [P8], [D4], and contributing to the R package BNPdensity [S3] with Guillaume Kon Kam King and other colleagues from the Collegio Carlo Alberto, designed for posterior sampling from mixtures of NRMI (see Figure 1); and finally on *applications*, including ecotoxicological risk assessment [C3], industrial applications [C2], and image segmentation [S2]. In this chapter, I focus essentially on survival analysis [A12], image segmentation [S2], and ecotoxicological applications [C3].

Section 1.1 Survival analysis: modeling hazard functions with mixtures [A12]. Bayesian nonparametric inferential procedures based on Markov chain Monte Carlo marginal methods typically yield point estimates in the form of posterior expectations. Though very useful and easy to implement in a variety of statistical problems, these methods may suffer from some limitations if used to estimate non-linear functionals of the posterior distribution. The main goal of [A12] is to develop a novel methodology that extends a well-established marginal procedure designed for hazard mixture models, in order to draw approximate inference on survival functions that is not limited to the posterior mean but includes, as remarkable examples, credible intervals and median survival time. Our approach relies on a characterization of the posterior moments that, in turn, is used to approximate the posterior distribution by means of a technique based on Jacobi polynomials.

Section 1.2 Image segmentation: mixtures with hidden Markov random fields [S2]. One of the central issues in statistics and machine learning is how to select an adequate model that can automatically adapt its complexity to the observed data. We focus on the issue of determining the structure of clustered data, both in terms of finding the appropriate number of clusters and of modelling the right dependence

structure between the observations. Bayesian nonparametric (BNP) models, which do not impose an upper limit on the number of clusters, are appropriate to avoid the required guess on the number of clusters but have been mainly developed for independent data. In contrast, Markov random fields (MRF) have been extensively used to model dependencies in a tractable manner but usually reduce to finite cluster numbers when clustering tasks are addressed. Our main contribution is to propose a general scheme to design tractable BNP-MRF priors that combine both features: no commitment to an arbitrary number of clusters and a dependence modelling. A key ingredient in this construction is the availability of a stick-breaking representation which has the three-fold advantage to allowing us to extend standard discrete MRFs to infinite state space, to design a tractable estimation algorithm using variational approximation and to derive theoretical properties on the predictive distribution and the number of clusters of the proposed model. This approach is illustrated on a challenging natural image segmentation task for which it shows good performance with respect to the literature.

Section 1.3 Ecotoxicological application to species sensitivity distribution modeling [C3]. We revisit a classical method for ecological risk assessment, the Species Sensitivity Distribution (SSD) approach, in a Bayesian nonparametric framework. SSD is a mandatory diagnostic required by environmental regulatory bodies from the European Union, the United States, Australia, China etc. Yet, it is subject to much scientific criticism, notably concerning a historically debated parametric assumption for modelling species variability. Tackling the problem using nonparametric mixture models, it is possible to shed this parametric assumption and build a statistically sounder basis for SSD.

Chapter 2. Approximate Bayesian Inference

This chapter is concerned in large parts with computational aspects of Bayesian inference, see the *Sampling from posteriors* paragraph of Section 0.1.1 for an overview of the existing methods. More specifically, I present *conditional approaches* in the form of truncation-based approximations for the Pitman–Yor process [A6] and for completely random measures [A8] (see Figure 1). Then I move to a marginal approach based on approximations of the predictive distribution of Gibbs-type processes [A1]. Finally, I describe an approximate Bayesian computation (ABC) algorithm using the energy distance as data discrepancy.

Section 2.1 Truncation-based approximations: Pitman–Yor [A6]. We consider approximations to the popular Pitman–Yor process obtained by truncating the stick-breaking representation. The truncation is determined by a random stopping rule that achieves an almost sure control on the approximation error in total variation distance. We derive the asymptotic distribution of the random truncation point as the approximation error ϵ goes to zero in terms of a polynomially tilted positive stable random variable. The practical usefulness and effectiveness of this theoretical result is demonstrated by devising a sampling algorithm to approximate functionals of the ϵ-version of the Pitman–Yor process.

Section 2.2 Truncation-based approximations: normalized random measures [A8]. Completely random measures (CRM) represent the key building block of a wide variety of popular stochastic models and play a pivotal role in modern Bayesian Nonparametrics. A popular representation of CRMs as a random series with decreasing jumps is due to
[119] T. S. Ferguson and M. J. Klass. A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643, 1972.
This can immediately be turned into an algorithm for sampling realizations of CRMs or more elaborate models involving transformed CRMs. However, concrete implementation requires to truncate the random series at some threshold resulting in an approximation error. The goal of this paper is to quantify the quality of the approximation by a moment-matching criterion, which consists in evaluating a measure of discrepancy between actual moments and moments based on the simulation output. Seen as a function of the truncation level, the methodology can be used to determine the truncation level needed to reach a certain level of precision. The resulting moment-matching Ferguson and Klass algorithm is then implemented and illustrated on several popular Bayesian nonparametric models.

Section 2.3 Approximating the predictive weights of Gibbs-type priors [A1]. This section presents the results of [A1] which is a follow-up to [A9], extending results from the Pitman–Yor process to Gibbs-type processes. Gibbs-type random probability measures, or Gibbs-type priors, are arguably the most "natural" generalization of the celebrated Dirichlet process prior. Among them the Pitman–Yor process prior certainly stands out in terms of mathematical tractability and interpretability of its predictive probabilities, which made it the natural candidate in a plethora of applications. Given a random sample of size n from an arbitrary Gibbs-type prior, we show that the corresponding predictive probabilities admit a large n approximation, with an error term vanishing as $o(1/n)$, which maintains the same desirable features as the predictive probabilities of the Pitman–Yor process prior. Our result is illustrated through an extensive simulation study, which includes an application in the context of Bayesian nonparametric mixture modeling.

Section 2.4 Approximate Bayesian computation based on the energy distance [S4]. Approximate Bayesian computation (ABC) has become an essential part of the Bayesian toolbox for addressing problems in which the likelihood is prohibitively expensive or entirely unknown, making it intractable. ABC defines a quasi-posterior by comparing observed data with simulated data, traditionally based on some summary statistics, the elicitation of which is regarded as a key difficulty. In recent years, a number of data discrepancy measures bypassing the construction of summary statistics have been proposed, including the Kullback–Leibler divergence, the Wasserstein distance and maximum mean discrepancies. Here we propose a novel importance-sampling (IS) ABC algorithm relying on the so-called *two-sample energy statistic*. We establish a new asymptotic result for the case where both the observed sample size and the simulated data sample size increase to infinity, which highlights to what extent the data discrepancy measure impacts the asymptotic pseudo-posterior. The result holds in the broad setting of IS-ABC methodologies, thus generalizing previous

results that have been established only for rejection ABC algorithms. Furthermore, we propose a consistent V-statistic estimator of the energy statistic, under which we show that the large sample result holds. Our proposed energy statistic based ABC algorithm is demonstrated on a variety of models, including a Gaussian mixture, a moving-average model of order two, a bivariate beta and a multivariate g-and-k distribution. We find that our proposed method compares well with alternative discrepancy measures.

Chapter 3. Distributional Properties of Statistical and Machine Learning Models

This chapter is concerned with general distributional properties of statistical and machine learning models. These properties include the sub-Gaussian and sub-Weibull properties of random variables [A7], [A2], [S5]. This sub-Weibull property defined in [S5] is then used in [A4], [P5], [P6] for characterizing the prior distribution of neural network units with Gaussian weight priors and under ReLU-like nonlinearity function; see Section 0.1.2 for an overview of Bayesian neural networks. We conclude the chapter by a presentation of novel theoretical properties of a large class of asymmetric copulas [A3].

Section 3.1 Sub-Gaussian and sub-Weibull properties [A7], [A2], [S5]. The line of research on the sub-Gaussian property of random variables started in 2017 while I was refereeing a machine learning conference paper (COLT). It was stating some conjectures about the sub-Gaussian property and optimal proxy variance of the beta distribution (see Section 3.1). I liked this conjecture because the problem was very simply stated. We solved this conjecture with my friend Olivier Marchal [A7]. This short note opened many more questions: once, Hien Nguyen asked me whether the methodology of [A7] could be used for other bounded support random variables such as Bernoulli, binomial, Kumaraswamy or triangular distributions. Of course such bounded random variables are *de facto* sub-Gaussian, but how to characterize the optimal sub-Gaussian proxy variance remains in general an open question. Another question is how to characterize *strict* sub-Gaussianity, defined by a proxy variance equal to the (standard) variance? With Olivier Marchal and Hien Nguyen, we address these questions in [A2]. A particular focus is given to the relationship between strict sub-Gaussianity and symmetry of the distribution. In particular, we demonstrate that symmetry is neither sufficient nor necessary for strict sub-Gaussianity. In contrast, simple necessary conditions on the one hand, and simple sufficient conditions on the other hand, for strict sub-Gaussianity are provided.

Section 3.2 Understanding priors in Bayesian neural networks [A4], [P5], [P6]. I started to be interested in neural networks in 2018 also as a consequence of attending a reading group on Deep Learning at Inria and of refereeing for machine learning conferences. The ICLR paper [226] dealing with distributional properties of neural networks in the infinite width regime (ie when the number of neurons or units per layer tends to infinity) and discussions with Thibaud Rahier have been quite inspirational for me and were in a sense the starting point of the internship and then Ph.D. work of Mariia Vladimirova with Jakob Verbeek and Pablo Mesejo focussing on understanding priors in Bayesian neural networks [A4], [P5], [P6]. We investigate deep

Bayesian neural networks with Gaussian weight priors and a class of ReLU-like nonlinearities. Bayesian neural networks with Gaussian priors are well known to induce an $\mathcal{L}^2$, "weight decay", regularization. Our results characterize a more intricate regularization effect at the level of the unit activations. Our main result establishes that the induced prior distribution on the units before and after activation becomes increasingly heavy-tailed with the depth of the layer. We show that first layer units are Gaussian, second layer units are sub-exponential, and units in deeper layers are characterized by sub-Weibull distributions [S5]. Our results provide new theoretical insight on deep Bayesian neural networks, which we corroborate with simulation experiments.

Section 3.3 Dependence properties of asymmetric copulas [A3]. Discussions with my office mate Stéphane Girard at the intersection of our scientific interests (extreme value theory on the one hand of the office and Bayesian analysis on the other hand) led us to propose a postdoc offer on extreme value theory and copula modeling from a Bayesian perspective. We had the chance to hire Marta Crispino and we started to work in 2018 on copulas. In [A3], we study a broad class of asymmetric copulas introduced by [207] as a combination of multiple—usually symmetric—copulas. The main thrust of the paper is to provide new theoretical properties including exact tail dependence expressions and stability properties. A subclass of Liebscher copulas obtained by combining comonotonic copulas is studied in more details. We establish further dependence properties for copulas of this class and show that they are characterized by an arbitrary number of singular components. Furthermore, we introduce a novel iterative representation for general Liebscher copulas which *de facto* insures uniform margins, thus relaxing a constraint of Liebscher's original construction. Besides, we show that this iterative construction proves useful for inference by developing an Approximate Bayesian computation (ABC) sampling scheme. This inferential procedure is demonstrated on simulated data and is compared to a likelihood-based approach in a setting where the latter is available.

Chapter 1

Bayesian Nonparametric Mixture Modeling

Contents

This chapter is based on the following papers and preprints

Section 1.1

[A12] J. Arbel, A. Lijoi, and B. Nipoti. Full Bayesian inference with hazard mixture models. *Computational Statistics & Data Analysis*, 93:359–372, 2016

[C5] J. Arbel, A. Lijoi, and B. Nipoti. *Bayesian Statistics from Methods to Models and Applications*, chapter Bayesian Survival Model based on Moment Characterization, pages 3–14. Springer Proceedings in Mathematics & Statistics, Volume 126. Springer International Publishing, Editors: Sylvia Frühwirth-Schnatter et al., 2015

Section 1.2

[S2] H. Lü, J. Arbel, and F. Forbes. Bayesian Nonparametric Priors for Hidden Markov Random Fields. *Under major revision, Statistics and Computing*, 2019

Section 1.3

[C3] G. Kon Kam King, J. Arbel, and I. Prünster. *Bayesian Statistics in Action*, chapter A Bayesian nonparametric approach to ecological risk assessment, pages 151–159. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017

[S3] J. Arbel, G. Kon Kam King, A. Lijoi, L. E. Nieto-Barajas, and I. Prünster. BNPdensity: Bayesian nonparametric mixture modeling in R. *Submitted*, 2019

[A12] J. Arbel, A. Lijoi, and B. Nipoti. Full Bayesian inference with hazard mixture models. *Computational Statistics & Data Analysis*, 93:359–372, 2016

[C5] J. Arbel, A. Lijoi, and B. Nipoti. *Bayesian Statistics from Methods to Models and Applications*, chapter Bayesian Survival Model based on Moment Characterization, pages 3–14. Springer Proceedings in Mathematics & Statistics, Volume 126. Springer International Publishing, Editors: Sylvia Frühwirth-Schnatter et al., 2015

1.1 Survival analysis: modeling hazard functions with mixtures

1.1.1 Introduction

Most commonly used inferential procedures in Bayesian nonparametric practice rely on the implementation of sampling algorithms that can be gathered under the general umbrella of Blackwell–MacQueen Pólya urn schemes. These are characterized by the marginalization with respect to an infinite-dimensional random element that defines the de Finetti measure of an exchangeable sequence of observations or latent variables. Henceforth these will be referred to as *marginal methods.* Besides being useful for the identification of the basic building blocks of ready to use Markov chain Monte Carlo (MCMC) sampling strategies, marginal methods have proved to be effective for an approximate evaluation of Bayesian point estimators in the form of posterior means. They are typically used with models for which the predictive distribution is available in closed form. Popular examples are offered by mixtures of the Dirichlet process for density estimation [112] and mixtures of gamma processes for hazard rate estimation [165]. While becoming well-established tools, these computational techniques are easily accessible also to practitioners through a straightforward software implementation [173]. Though it is important to stress their relevance both in theory and in practice, it is also worth pointing out that Blackwell–MacQueen Pólya urn schemes suffer from some drawbacks which we wish to address here. Indeed, one easily notes that the posterior estimates provided by marginal methods are not suitably endowed with measures of uncertainty such as posterior credible intervals. Furthermore, using the posterior mean as an estimator is equivalent to choosing a square loss function whereas in many situations of interest other choices such as absolute error or 0–1 loss functions and, as corresponding estimators, median or mode of the posterior distribution of the survival function, at any fixed time point t, would be preferable. Finally, they do not naturally allow inference on functionals of the distribution of survival times, such as the median survival time, to be drawn. A nice discussion of these issues is provided by [126] where the focus is on mixtures of the Dirichlet process: the authors suggest complementing the use of marginal methods with a sampling strategy that aims at generating approximate trajectories of the Dirichlet process from its truncated stick-breaking representation.

The aim is to propose a new procedure that combines closed-form analytical results arising from the application of marginal methods with an approximation of the posterior distribution which makes use of posterior moments. The whole machinery

is developed for the estimation of survival functions that are modeled in terms of hazard rate functions. To this end, let F denote the cumulative distribution function (CDF) associated to a probability distribution on $\mathbb{R}^+$. The corresponding survival and cumulative hazard functions are denoted as

$$S(t) = 1 - F(t) \quad \text{and} \quad H(t) = -\int_{[0,t]} \frac{\mathrm{d}F(s)}{F(s-)},$$

for any $t > 0$, respectively, where $F(s-) := \lim_{\varepsilon \downarrow 0} F(s-\varepsilon)$ for any positive s. If F is absolutely continuous, one has $H(t) = -\log(S(t))$ and the hazard rate function associated to F is, thus, defined as $h(t) = F'(t)/[1 - F(t-)]$. It should be recalled that survival analysis has been one of the most relevant areas of application of Bayesian nonparametric methodology soon after the groundbreaking contribution of [117]. A number of papers in the '70s and the '80s have been devoted to the proposal of new classes of priors that accommodate for a rigorous analytical treatment of Bayesian inferential problems with censored survival data. Among these it is worth mentioning the neutral to the right processes proposed in [100] and used to define a prior for the CDF F: since they share a conjugacy property they represent a tractable tool for drawing posterior inferences. Another noteworthy class of priors has been proposed in [156], where a beta process is used as a nonparametric prior for the cumulative hazard function H has been proposed. Also in this case, one can considerably benefit from a useful conjugacy property.

As already mentioned, the plan consists in proposing a method for full Bayesian analysis of survival data by specifying a prior on the hazard rate h. The most popular example is the gamma process mixture that has been originally proposed in [107] and generalized in later work by [218] and [168] to include any mixing random measure and any mixed kernel. Recently [212] have extended such framework to the context of partially exchangeable observations. The uses of random hazard mixtures in practical applications have been boosted by the recent developments of powerful computational techniques that allow for an approximate evaluation of posterior inferences on quantities of statistical interest. Most of these arise from a marginalization with respect to a completely random measure that identifies the de Finetti measure of the exchangeable sequence of observations. See, e.g., [165]. Though they are quite simple to implement, the direct use of their output can only yield point estimation of the hazard rates, or of the survival functions, at fixed time points through posterior means. The main goal of the present paper is to show that a clever use of a moment-based approximation method does provide a relevant upgrade on the type of inference one can draw via marginal sampling schemes. The takeaway message is that the information gathered by marginal methods is not confined to the posterior mean but is actually much richer and, if properly exploited, can lead to a more complete posterior inference. To understand this, one can refer to a sequence of exchangeable survival times $(X_i)_{i\geq 1}$ such that $\mathbb{P}[X_1 > t_1, \ldots, X_n > t_n \mid \tilde{P}] = \prod_{i=1}^n \tilde{S}(t_i)$ where $\tilde{P}$ is a random probability measure on $\mathbb{R}^+$ and $\tilde{S}(t) = \tilde{P}((t,\infty))$ is the corresponding random survival function. Given a suitable sequence of latent variables $(Y_i)_{i\geq 1}$, a closed-form expression for

$$\mathbb{E}[\tilde{S}^r(t) \mid \boldsymbol{X}, \boldsymbol{Y}], \qquad \text{for any } r \geq 1, \quad \text{and } t > 0, \tag{1.1}$$

with $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, will be provided. Our strategy consists

in approximating the posterior distribution of $\tilde{S}(t)$, at each instant t, and relies on the fact that, along with the posterior mean, marginal models allow to straightforwardly estimate posterior moments of any order of $\tilde{S}(t)$. Indeed, an MCMC sampler yields a sample from the posterior distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$: this can be used to integrate out the latent variables appearing in (1.1) and obtain a numerical approximate evaluation of the posterior moments $\mathbb{E}[\tilde{S}^r(t) \mid \boldsymbol{X}]$. These are finally used to deduce, with almost negligible effort, an approximation of the posterior distribution of $\tilde{S}(t)$ and, in turn, to estimate some meaningful functionals of $\tilde{S}(t)$.

It is to be mentioned that one could alternatively resort to a different approach that boils down to the simulation of the trajectories of the completely random measure that defines the underlying random probability measure from its posterior distribution. In density estimation problems, this is effectively illustrated in [250], [248] and [43]. As for hazard rates mixtures estimation problems, one can refer to [168], [252] and [247]. In particular, [168] provides a posterior characterization that is the key for devising a [119] representation of the posterior distribution of the completely random measure which enters the definition of the prior for the hazards. Some numerical aspects related to the implementation of the algorithm can be quite tricky since one needs to invert the Lévy intensity to simulate posterior jumps and a set of suitable latent variables need to be introduced in order to sample from the full conditionals of the hyperparameters. These aspects are well described and addressed in [247].

The section is organized as follows. In Section 1.1.2 hazard mixture models are briefly reviewed together with some of their most important properties. Furthermore, explicit expressions characterizing the posterior moments of any order of a random survival function are provided both for general framework and for the extended gamma process case. Section 1.1.3 is dedicated to the problem of approximating the distribution of a random variable on $[0, 1]$, provided that the first N moments are known. In particular, a convenient methodology based on Jacobi polynomials is described in Section 1.1.3 and, then, implemented in Section 1.1.3 in order to approximate random survival functions. Its performance is tested through a thorough numerical investigation. The focus of Section 1.1.4 is the use of the introduced methodology for carrying out Bayesian inference on survival functions. Specifically, the algorithm is presented in Section 1.1.4 whereas simulated data and a real two-sample dataset on leukemia remission times are analysed in Sections 1.1.4 and 1.1.4 respectively.

1.1.2 Hazard mixture models

A well-known nonparametric prior for the hazard rate function within multiplicative intensity models used in survival analysis arises as a mixture of *completely random measures* (CRMs). To this end, recall that a CRM $\tilde{\mu}$ on a space $\mathbb{Y}$ is a boundedly finite random measure that, when evaluated at any collection of pairwise disjoint sets $A_1, \ldots, A_d$, gives rise to mutually independent random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_d)$, for any $d \geq 1$. Importantly, CRMs are almost surely discrete measures [192]. A detailed treatment on CRMs can also be found in [90]. With reference to Theorem 1 in [189], it is assumed that $\tilde{\mu}$ has no fixed atoms, which in turn implies the existence of

a measure ν on $\mathbb{R}^+ \times \mathbb{Y}$ such that $\int_{\mathbb{R}^+\times\mathbb{Y}} \min\{s,1\}\nu(\mathrm{d}s,\mathrm{d}y) < \infty$ and

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{Y}} f(y)\tilde{\mu}(\mathrm{d}y)}\right] = \exp\left(-\int_{\mathbb{R}^+\times\mathbb{Y}} \left[1-\exp\left(-s\,f(y)\right)\right]\nu(\mathrm{d}s,\mathrm{d}y)\right), \tag{1.2}$$

for any measurable function $f : \mathbb{Y} \to \mathbb{R}$ such that $\int_{\mathbb{Y}} |f|\ \mathrm{d}\tilde{\mu} < \infty$, with probability 1. The measure ν is termed the *Lévy intensity* of $\tilde{\mu}$. For our purposes, it will be useful to rewrite ν as

$$\nu(\mathrm{d}s,\mathrm{d}y) = \rho_y(s)\,\mathrm{d}s\,c\,P_0(\mathrm{d}y),$$

where P_0 is a probability measure on $\mathbb{Y}$, c a positive parameter, and $\rho_y(s)$ is some transition kernel on $\mathbb{Y}\times\mathbb{R}^+$. If $\rho_y = \rho$, for any y in $\mathbb{Y}$, the CRM $\tilde{\mu}$ is said *homogeneous*. Henceforth, it is further assumed that P_0 is non-atomic. A well-known example corresponds to $\rho_y(s) = \rho(s) = \mathrm{e}^{-s}/s$, for any y in $\mathbb{Y}$, which identifies a so-called *gamma CRM*. With such a choice of the Lévy intensity, it can be seen, from (1.2), that for any A such that $P_0(A) > 0$, the random variable $\tilde{\mu}(A)$ is gamma distributed, with shape parameter 1 and rate parameter $cP_0(A)$. If $k(\,\cdot\,;\cdot\,)$ is a transition kernel on $\mathbb{R}^+\times\mathbb{Y}$, a prior for h is the distribution of the random hazard rate (RHR)

$$\tilde{h}(t) = \int_{\mathbb{Y}} k(t;y)\tilde{\mu}(\mathrm{d}y), \tag{1.3}$$

where $\tilde{\mu}$ is a CRM on $\mathbb{Y}$. It is worth noting that, if $\lim_{t\to\infty}\int_0^t \tilde{h}(s)\mathrm{d}s = \infty$ with probability 1, then one can adopt the following model

$$\begin{aligned} X_i \mid \tilde{P} &\overset{\text{i.i.d.}}{\sim} \tilde{P} \\ \tilde{P}((\,\cdot\,,\infty)) &\overset{\mathrm{d}}{=} \exp\left(-\int_0^{\cdot}\tilde{h}(s)\,\mathrm{d}s\right) \end{aligned} \tag{1.4}$$

for a sequence of (possibly censored) survival data $(X_i)_{i\geq 1}$. This means that $\tilde{h}$ in (1.3) defines a random survival function $t \mapsto \tilde{S}(t) = \exp(-\int_0^t \tilde{h}(s)\mathrm{d}s)$. In this setting, [107] characterize the posterior distribution of the so-called *extended gamma process*: this is obtained when $\tilde{\mu}$ is a gamma CRM and $k(t;y) = \mathbb{I}_{(0,t]}(y)\,\beta(y)$ for some positive right-continuous function $\beta : \mathbb{R}^+ \to \mathbb{R}^+$. The same kind of result is proved in [218] for *weighted gamma processes* corresponding to RHRs obtained when $\tilde{\mu}$ is still a gamma CRM and $k(\,\cdot\,;\,\cdot\,)$ is an arbitrary kernel. Finally, a posterior characterization has been derived in [168] for any CRM $\tilde{\mu}$ and kernel $k(\,\cdot\,;\,\cdot\,)$.

We shall quickly display such a characterization since it represents the basic result our construction relies on. For the ease of exposition we confine ourselves to the case where all the observations are exact, the extension to the case that includes right-censored data being straightforward and detailed in [168]. For an n-sample $\boldsymbol{X} = (X_1,\ldots,X_n)$ of exact data, the likelihood function equals

$$\mathcal{L}(\tilde{\mu};\boldsymbol{X}) = e^{-\int_{\mathbb{Y}} K_{\boldsymbol{X}}(y)\tilde{\mu}(\mathrm{d}y)}\prod_{i=1}^{n}\int_{\mathbb{Y}} k(X_i;y)\tilde{\mu}(\mathrm{d}y), \tag{1.5}$$

where $K_t(y) = \int_0^t k(s;y)\mathrm{d}s$ and $K_{\boldsymbol{X}}(y) = \sum_{i=1}^n K_{X_i}(y)$. A useful augmentation suggests introducing latent random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ such that the joint distribution of $(\tilde{\mu}, \boldsymbol{X}, \boldsymbol{Y})$ coincides with

$$e^{-\int_{\mathbb{Y}} K_{\boldsymbol{X}}(y)\tilde{\mu}(\mathrm{d}y)} \prod_{i=1}^n k(X_i; Y_i)\tilde{\mu}(\mathrm{d}Y_i)\, Q(\mathrm{d}\tilde{\mu}), \tag{1.6}$$

where Q is the probability distribution of the completely random measure $\tilde{\mu}$, characterized by the Laplace transform functional in (1.2) [see for instance 90]. The almost sure discreteness of $\tilde{\mu}$ implies there might be ties among the Y_i's with positive probability. Therefore, the distinct values among $\boldsymbol{Y}$ are denoted as $(Y_1^*, \ldots, Y_k^*)$, where $k \leq n$, and, for any $j = 1, \ldots, k$, $C_j := \left\{ l \,:\, Y_l = Y_j^* \right\}$ with $n_j = \#C_j$ as the cardinality of C_j. Thus, the joint distribution in (1.6) may be rewritten as

$$e^{-\int_{\mathbb{Y}} K_{\boldsymbol{X}}(y)\tilde{\mu}(\mathrm{d}y)} \prod_{j=1}^k \tilde{\mu}(\mathrm{d}Y_j^*)^{n_j} \prod_{i \in C_j} k(X_i; Y_j^*)\, Q(\mathrm{d}\tilde{\mu}). \tag{1.7}$$

We introduce, also, the density function

$$f(s \,|\, \kappa, \xi, y) \propto s^\kappa \, \mathrm{e}^{-\xi s}\, \rho_y(s)\, \mathbb{I}_{\mathbb{R}^+}(s) \tag{1.8}$$

for any $\kappa \in \mathbb{N}$ and $\xi > 0$. The representation displayed in (1.7), combined with results concerning disintegrations of Poisson random measures, leads to prove the following

Proposition 1.1.1. (168) *Let $\tilde{h}$ be a RHR as defined in (1.3). The posterior distribution of $\tilde{h}$, given $\boldsymbol{X}$ and $\boldsymbol{Y}$, coincides with the distribution of the random hazard*

$$\tilde{h}^* + \sum_{j=1}^k J_j k(\,\cdot\,; Y_j^*), \tag{1.9}$$

where $\tilde{h}^*(\,\cdot\,) = \int_{\mathbb{Y}} k(\,\cdot\,; y)\, \tilde{\mu}^*(\mathrm{d}y)$ and $\tilde{\mu}^*$ is a CRM without fixed points of discontinuity whose Lévy intensity is

$$\nu^*(\mathrm{d}s, \mathrm{d}y) = e^{-sK_{\boldsymbol{X}}(y)} \rho_y(s)\mathrm{d}s\, cP_0(\mathrm{d}y).$$

The jumps $J_1, \ldots, J_k$ are mutually independent and independent of $\tilde{\mu}^*$. Moreover, for every $j = 1, \ldots, k$, the distribution of the jump J_j has density function $f(\,\cdot \,|\, n_j, K_{\boldsymbol{X}}(Y_j^*), Y_j^*)$ with f defined in (1.8).

See [215] for an alternative proof of this result. The posterior distribution of $\tilde{h}$ displays a structure that is common to models based on CRMs, since it consists of the combination of two components: one without fixed discontinuities and the other with jumps at fixed points. In this case, the points at which jumps occur coincide with the distinct values of the latent variables $Y_1^*, \ldots, Y_k^*$. Furthermore, the distribution of the jumps J_j depends on the respective locations Y_j^*.

Beside allowing us to gain insight on the posterior distribution of $\tilde{h}$, Proposition 1.1.1 is also very convenient for simulation purposes. See, e.g., [165]. Indeed, (1.9) allows obtaining an explicit expression for the posterior expected value of $\tilde{S}(t)$ (or, equivalently, of $\tilde{h}(t)$), for any $t > 0$, conditionally on the latent variables $\boldsymbol{Y}$. One can, thus,

integrate out the vector of latent variables $\boldsymbol{Y}$, by means of a Gibbs type algorithm, in order to approximately evaluate the posterior mean of $\tilde{S}(t)$ (or $\tilde{h}(t)$). As pointed out in next section, a combination of Proposition 1.1.1 and of the same Gibbs sampler we have briefly introduced actually allows moments of $\tilde{S}(t)$, of any order, to be estimated. We will make use of the first N of these estimated moments to approximate, for each $t > 0$, the posterior distribution of $\tilde{S}(t)$ and therefore to have the tools for drawing meaningful Bayesian inference. The choice of a suitable value for N will be discussed in Section 1.1.3.

As pointed out in the Introduction, one can, in line of principle, combine Proposition 1.1.1 with the Ferguson and Klass representation to undertake an alternative approach that aims at simulating the trajectories from the posterior distribution of the survival function. This can be achieved by means of a Gibbs type algorithm that involves sampling $\tilde{\mu}^*$ and Y_j^*, for $j = 1, \ldots, k$, from the corresponding full conditional distributions. Starting from the simulated trajectories one could then approximately evaluate all the posterior quantities of interest. The latter is an important feature of the method based on the Ferguson and Klass representation, that is shared only in part by our proposal. Indeed, extending the moment-based procedure to estimate functionals of $\tilde{S}(t)$, although achievable in many cases of interest, is not always straightforward. For instance, in order to carry out inference based on the posterior distribution of the random hazard rate $\tilde{h}(t)$, one should start with the estimation of the posterior moments of $\tilde{h}(t)$ and adapt accordingly the methodology which throughout the paper is developed for $\tilde{S}(t)$. An illustration, with an application to survival analysis, is provided in [247] and it appears that the approach, though achievable, may be difficult to implement. The main non-trivial issues one has to deal with are the inversion of the Lévy measure, needed to sample the jumps, and the sampling from the full conditionals of the hyperparameters. The latter has been addressed by [247] through a clever augmentation scheme that relies on a suitable collection of latent variables. The approach based on the simulation of trajectories is an example of non-marginal, or *conditional*, method since it does not rely on the marginalization with respect to the mixing CRM $\tilde{\mu}$.

In the next sections, attention will be mainly devoted to marginal methods with the aim of showing that they allow for a full Bayesian inference, beyond the usual evaluation of posterior means. The required additional effort to accomplish this task is minimal and boils down to computing a finite number of posterior moments of $\tilde{S}(t)$, at a given t. An approximate evaluation of these moments can be determined by resorting to (1.9) which yields closed-form expressions for the posterior moments of the random variable $\tilde{S}(t)$, conditionally on both the data $\boldsymbol{X}$ and the latent variables $\boldsymbol{Y}$.

Proposition 1.1.2. *For every $t > 0$ and $r > 0$,*

$$\mathbb{E}[\tilde{S}^r(t) \mid \boldsymbol{X}, \boldsymbol{Y}] = \exp\left\{-c \int_{\mathbb{R}^+ \times \mathbb{Y}} \left(1 - \mathrm{e}^{-rK_t(y)s}\right) \mathrm{e}^{-K_{\boldsymbol{X}}(y)s} \rho(s)\mathrm{d}s P_0(\mathrm{d}y)\right\} \times \prod_{j=1}^{k} \frac{1}{B_j} \int_{\mathbb{R}^+} \exp\left\{-s\left(rK_t(Y_j^*) + K_{\boldsymbol{X}}(Y_j^*)\right)\right\} s^{n_j} \rho(s)\mathrm{d}s,$$

where $B_j = \int_{\mathbb{R}^+} s^{n_j} \exp\left\{-sK_{\boldsymbol{X}}(Y_j^*)\right\} \rho(s)\mathrm{d}s$, for $j = 1, \ldots, k$.

Although the techniques that will be described hold true for any specification of $\tilde{\mu}$ and kernel $k(\,\cdot\,;\cdot\,)$, the proposed illustration will focus on on the extended gamma process case [107]. More specifically, we consider a kernel $k(t;y) = \mathbb{I}_{(0,t]}(y)\beta$, with $\beta > 0$. This choice of kernel is known to be suitable for modeling monotone increasing hazard rates and to give rise to a class of random hazard functions with nice asymptotic properties [96]. Moreover, without loss of generality, it is assumed that $X_1 > X_2 > \ldots > X_n$. For notational convenience, one further sets $X_0 \equiv \infty$, $X_{n+1} \equiv 0$, $\xi_l \equiv \sum_{i=1}^{l} X_i$, for any $l \geq 1$, and $\xi_0 \equiv 0$. The next Corollary displays an expression for the conditional moments corresponding to this prior specification.

Corollary 1.1.1. *For every $t > 0$ and $r > 0$,*

$$\mathbb{E}[\tilde{S}^r(t) \mid \boldsymbol{X}, \boldsymbol{Y}] = \prod_{i=0}^{n} \exp\left\{-c \int_{X_{i+1}\wedge t}^{X_i \wedge t} \log\left(1 + r\,\frac{t-y}{\xi_i - iy + 1/\beta}\right) P_0(\mathrm{d}y)\right\} \times \prod_{j=1}^{k} \left(1 + r\,\frac{(t - Y_j^*)\mathbb{I}_{[Y_j^*,\infty)}(t)}{\sum_{i=1}^{n}(X_i - Y_j^*)\mathbb{I}_{[Y_j^*,\infty)}(X_i) + 1/\beta}\right)^{-n_j}. \quad (1.10)$$

By integrating out the vector of latent variables $\boldsymbol{Y}$ in (1.10) one obtains an estimate of the posterior moments of $\tilde{S}(t)$. To this end one can resort to a Gibbs algorithm whose steps will be described in Section 1.1.4.

1.1.3 Approximate inference via moments

Moment-based density approximation and sampling

Recovering a probability distribution from the explicit knowledge of its moments is a classical problem in probability and statistics that has received great attention in the literature. See, e.g., [275], references and motivating applications therein. Our specific interest in the problem is motivated by the goal of determining an approximation of the density function of a distribution supported on $[0,1]$ that equals the posterior distribution of a random survival function evaluated at some instant t. This is a convenient case since, as the support is a bounded interval, all the moments exist and uniquely characterize the distribution, see [280]. Moment-based methods for density functions' approximation can be essentially divided into two classes, namely methods that exploit orthogonal polynomial series [275] and maximum entropy methods [87, 231]. Both these procedures rely on systems of equations that relate the moments of the distribution with the coefficients involved in the approximation. For our purposes the use of orthogonal polynomial series turns out to be more convenient since it ensures faster computations as it involves uniquely linear equations. This property is particularly important in our setting since the same approximation procedure needs to be implemented a large number of times in order to approximate the posterior distribution of a random survival function. Moreover, as discussed in [110], maximum entropy techniques can lead to numerical instability.

Specifically, we work with Jacobi polynomials, a broad class which includes, among others, Legendre and Chebyshev polynomials. They are well suited for the expansion of densities with compact support contrary to other polynomials like Laguerre

and Hermite which can be preferred for densities with infinite of semi-infinite support [see 275]. While the classical Jacobi polynomials are defined on $[-1,1]$, a suitable transformation of such polynomials is considered so that their support coincides with $[0,1]$ and therefore matches the support of the density we aim at approximating. That is, we consider a sequence of polynomials $(G_i)_{i\geq 0}$ such that, for every $i \in \mathbb{N}$, G_i is a polynomial of order i defined by $G_i(s) = \sum_{r=0}^{i} G_{i,r} s^r$, with $s \in [0,1]$. The coefficients $G_{i,r}$ can be defined by a recurrence relation [see for example 311]. Such polynomials are orthogonal with respect to the L^2-product

$$\langle F, G \rangle = \int_0^1 F(s)G(s)w_{a,b}(s)\mathrm{d}s,$$

where

$$w_{a,b}(s) = s^{a-1}(1-s)^{b-1}$$

is named *weight function* of the basis. Moreover, without loss of generality, the G_i's can be assumed to be normalized and, therefore, $\langle G_i, G_j \rangle = \delta_{ij}$ for every $i, j \in \mathbb{N}$, where δ_{ij} is the Kronecker symbol. Any univariate density f supported on $[0,1]$ can be uniquely decomposed on such a basis and therefore there is a unique sequence of real numbers $(\lambda_i)_{i\geq 0}$ such that

$$f(s) = w_{a,b}(s) \sum_{i=0}^{\infty} \lambda_i G_i(s). \tag{1.11}$$

Let us now consider a random variable S whose density f has support on $[0,1]$. Its raw moments will be denoted by $\mu_r = \mathbb{E}[S^r]$, with $r \in \mathbb{N}$. From the evaluation of $\int_0^1 f(s)\, G_i(s)\, \mathrm{d}s$ it follows that each λ_i coincides with a linear combination of the first i moments, specifically $\lambda_i = \sum_{r=0}^{i} G_{i,r}\mu_r$. Then, the polynomial approximation method consists in truncating the sum in (1.11) at a given level $i = N$. This procedure leads to a methodology that makes use only of the first N moments and provides the approximation

$$f_N(s) = w_{a,b}(s) \sum_{i=0}^{N} \left(\sum_{r=0}^{i} G_{i,r}\mu_r \right) G_i(s). \tag{1.12}$$

It is important to stress that the polynomial expansion approximation (1.12) is not necessarily a density as it might fail to be positive or to integrate to 1. In order to overcome this problem, the density π_N proportional to the positive part of f_N, i.e. $\pi_N(s) \propto \max(f_N(s), 0)$, will be considered. An importance sampling algorithm [see, e.g., 287] will be used to sample from π_N. This is a method for drawing independent weighted samples (ϖ_ℓ, S_ℓ) from a distribution proportional to a given non-negative function, that exempts us from computing the normalizing constant. More precisely, the method requires to pick a proposal distribution p whose support contains the support of π_N. A natural choice for p is the Beta distribution proportional to the weight function $w_{a,b}$. The weights are then defined by $\varpi_\ell \propto \max(f_N(S_\ell), 0)/p(S_\ell)$ such that they add up to 1.

An important issue related to any approximating method refers to the quantification of the approximating error. As for the described polynomial approach, the error can be assessed for large N by applying the asymptotic results in [5]. Specifically, the convergence $f_N(s) \to f(s)$ for $N \to \infty$, for all $s \in (0,1)$, implies $\pi_N(s) \to f(s)$

for $N \to \infty$. Thus, if S_N denotes a random variable with distribution π_N, then the following convergence in distribution to the target random variable S holds:

$$S_N \xrightarrow{\mathcal{D}} S \text{ as } N \to \infty.$$

However, here the interest is in evaluating whether few moments allow for a good approximation of the posterior distribution of $\tilde{S}(t)$. This question will be addressed by means of an extensive numerical study in the next section. See [111] and [110] for a similar treatment referring to functionals of neutral-to-the-right priors and Dirichlet processes respectively.

Numerical study

In this section the quality of the approximation procedure described above is assessed by means of a simulation study. The rationale of the analysis consists in considering random survival functions for which moments of any order can be explicitly evaluated at any instant t, and then compare the true distribution with the approximation obtained by exploiting the knowledge of the first N moments. This in turn will provide an insight on the impact of N on the approximation error. To this end three examples of random survival functions will be considered, namely $\tilde{S}_j$ with $j = 1, 2, 3$. For the illustrative purposes of this Section, it suffices to specify the distribution of the random variable that coincides with $\tilde{S}_j$ evaluated in t, for every $t > 0$. Specifically, we consider a Beta, a mixture of Beta, and a normal distribution truncated to $[0, 1]$, that is

$$\begin{aligned}
\tilde{S}_1(t) &\sim \text{Beta}\left(\frac{S_0(t)}{a_1}, \frac{1 - S_0(t)}{a_1}\right), \\
\tilde{S}_2(t) &\sim \frac{1}{2}\,\text{Beta}\left(\frac{S_0(t)}{a_2}, \frac{1 - S_0(t)}{a_2}\right) + \frac{1}{2}\,\text{Beta}\left(\frac{S_0(t)}{a_3}, \frac{1 - S_0(t)}{a_3}\right), \\
\tilde{S}_3(t) &\sim \text{t}\mathcal{N}_{[0,1]}\left(S_0(t), \frac{S_0(t)(1 - S_0(t))}{a_4}\right),
\end{aligned}$$

where $S_0(t) = \mathrm{e}^{-t}$ and we have set $a_1 = 20$, $(a_2, a_3) = (10, 30)$ and $a_4 = 2$. Observe that, for every $t > 0$, $\mathbb{E}[\tilde{S}_1(t)] = \mathbb{E}[\tilde{S}_2(t)] = S_0(t)$ but the same does not hold true for $\tilde{S}_3(t)$.
For each $j = 1, 2, 3$, the first 10 moments of $\tilde{S}_j(t)$ were computed on a grid $\{t_1, \ldots, t_{50}\}$ of 50 equidistant values of t in the range $[0, 2.5]$. The choice of working with 10 moments will be motivated at the end of the section. The importance sampler described in Section 1.1.3 was then used to obtain samples of size 10 000 from the distribution of $\tilde{S}_j(t_i)$, for each $j = 1, 2, 3$ and $i = 1, \ldots, 50$. In Figure 1.1, for each $\tilde{S}_j$, we plot the true mean as well as the 95% highest density intervals for the true distribution and for the approximated distribution obtained by exploiting 10 moments. Notice that the focus is not on approximating the mean since moments of any order are the starting point of our procedure. Interestingly, the approximated intervals show a very good fit to the true ones in all the three examples. As for the Beta case, the fit is exact since the Beta-shaped weight function allows the true density to be recovered with the first two moments. As for the mixture of Beta, exact and approximated intervals can hardly be distinguished. Finally, the fit is pretty good also for the intervals in the

truncated normal example. Similarly, in Figure 1.2 the true and the approximated densities of each $\tilde{S}_j(t)$ are compared for fixed t in $\{0.1, 0.5, 2.5\}$. Again, all the three examples show a very good pointwise fit.

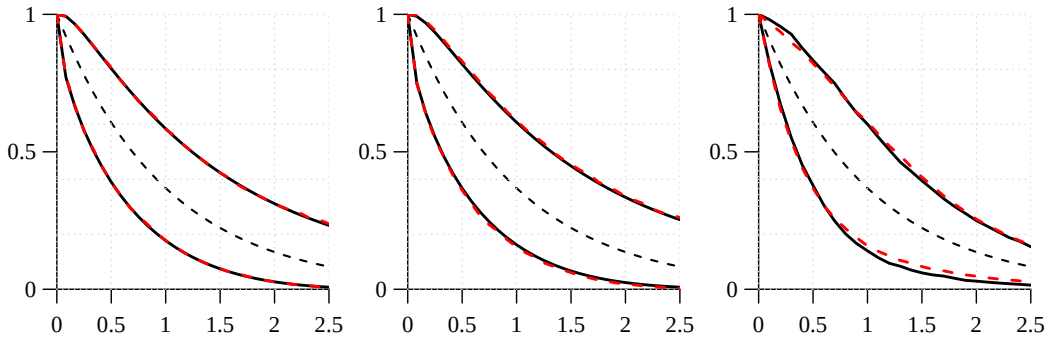


FIGURE 1.1: Mean of $\tilde{S}_j(t)$ (dashed black) and 95% highest density intervals for the true distribution (solid black) and the approximated distribution (dashed red) for the Beta ($j = 1$), mixture of Beta ($j = 2$) and truncated normal ($j = 3$) examples (left, middle and right, respectively).

This section is concluded by assessing how the choice of N affects the approximation error. To this end, for each instant t on the grid, the true and approximated distributions of $\tilde{S}_j(t)$ are compared by computing the integrated squared error (L^2 error) between the two. Thus the average of these values is considered as a measure of the overall error of approximation. The results are illustrated in Figure 1.3. As expected, the approximation is exact in the Beta example. In the two other cases, it can be observed that the higher is the number of exploited moments, the lower is the average approximation error. Nonetheless, it is apparent that the incremental gain of using more moments is more substantial when N is small whereas it is less impactful as N increases: for example in the mixture of Beta case, the L^2 error is 2.11, 0.97, 0.38 and 0.33 with N equal to 2, 4, 10 and 20 respectively. Moreover, when using a large number of moments, e.g. $N > 20$, some numerical instability can occur. These observations suggest that working with $N = 10$ moments in (1.12) strikes a good balance between accuracy of approximation and numerical stability.

1.1.4 Bayesian inference

In this section the characterization of the posterior moments of $\tilde{S}(t)$ provided in Proposition 1.1.2 is combined with the approximation procedure described in Section 1.1.3. The model specification (1.4) is completed by assuming an extended gamma prior for $\tilde{h}(t)$, with exponential base measure $P_0(\mathrm{d}y) = \lambda \exp(-\lambda y)\mathrm{d}y$, and considering the hyperparameters c and β random. Finally we choose for both c and β independent gamma prior distributions with shape parameter 1 and rate parameter $1/3$ (so to ensure large prior variance) and set $\lambda = 1$. Given a sample of survival times $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, the first N moments of the posterior distribution of $\tilde{S}(t)$ are estimated for t on a grid of q equally-spaced points $\{t_1, \ldots, t_q\}$ in an interval $[0, M]$.

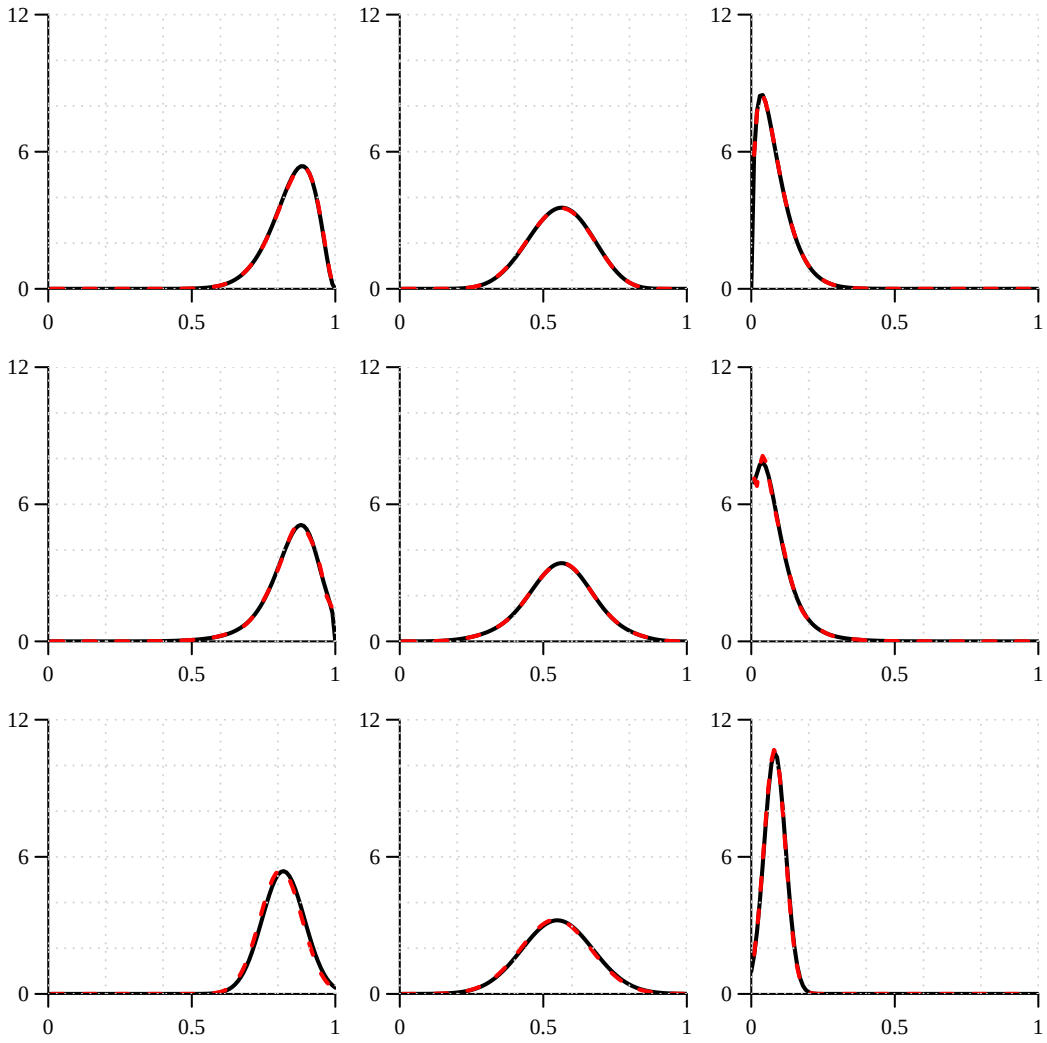


FIGURE 1.2: True density (solid black) and approximated one (dashed red) at time values $t = 0.1$ (left column), $t = 0.5$ (middle column) and $t = 2.5$ (right column), for the Beta ($j = 1$, top row), mixture of Beta ($j = 2$, middle row) and truncated normal ($j = 3$, bottom row) examples.

Such estimates are then exploited to approximate the posterior distribution of $\tilde{S}(t_i)$ for $i = 1, \ldots, q$. This allows us to devise an algorithm for carrying out full Bayesian inference on survival data. In the illustrations the focus will be on the estimation of the median survival time and, at any given t in the grid, of the posterior mean, posterior median, posterior mode and credibility intervals for $\tilde{S}(t)$. The same approach can be, in principle, used to estimate other functionals of interest.

Algorithm

The two main steps needed in order to draw samples from the posterior distribution of $\tilde{S}(t)$, for any $t \in \{t_1, \ldots, t_q\}$, are summarized in Algorithm 1.1.1. First a Gibbs sampler is performed to marginalize the latent variables $\boldsymbol{Y}$ and the hyperparameters (c, β) and therefore, for every $i = 1, \ldots, q$, an estimate for the posterior moments $\mathbb{E}[\tilde{S}^r(t_i)|\boldsymbol{X}]$, with $r = 1, \ldots, N$, is obtained. The algorithm was run for

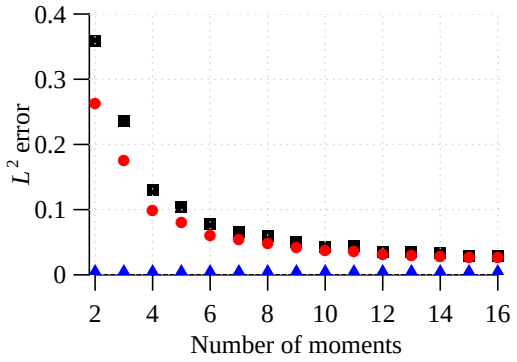


FIGURE 1.3: Average across t of the L^2 error between the true and the approximated densities of $\tilde{S}_j(t)$, in the Beta example (blue triangles), the mixture of Beta (red dots) and the truncated normal example (black squares). The approximation is exact in the Beta example.

$l_{\max} = 100\,000$ iterations, with a burn-in period of $l_{\min} = 10\,000$. Visual investigation of the traceplots of the parameters, in the illustrations of Sections 1.1.4 and 1.1.4, did not reveal any convergence issue. The second part consists in sampling from the posterior distribution of $\tilde{S}(t_i)$, for every $i = 1, \ldots, q$, by means of the importance sampler described in Section 1.1.3. Specifically $\ell_{\max} = 10\,000$ values were sampled for each t_i on the grid.

Algorithm 1.1.1 Posterior sampling

Part 1. Gibbs sampler

1: set $l = 0$ and admissible values for latent variables and hyperparameters, i.e. $\{Y_1 = Y_1^{(0)}, \ldots, Y_n = Y_n^{(0)}\}$, $c = c^{(0)}$ and $\beta = \beta^{(0)}$

2: while $l < l_{\max}$, set $l = l + 1$, and

- update $Y_j = Y_j^{(l)}$ for every $j = 1, \ldots, n$
- update $c = c^{(l)}$ and $\beta = \beta^{(l)}$
- if $l > l_{\min}$, compute

$$\mu_{r,t}^{(l)} = \mathbb{E}[\tilde{S}^r(t) \mid \boldsymbol{X}, \boldsymbol{Y}^{(l)}, c^{(l)}, \beta^{(l)}] \tag{1.13}$$

for each $r = 1, \ldots, N$ and for each t in the grid

3: for each $r = 1, \ldots, N$ and each t define $\hat{\mu}_{r,t} = \frac{1}{l_{\max} - l_{\min}} \sum_{l=l_{\min}+1}^{l_{\max}} \mu_{r,t}^{(l)}$

Part 2. Importance sampler

1: for each t, use (1.12) and define the approximate posterior density of $\tilde{S}(t)$ by $f_{N,t}(\,\cdot\,) = w_{a,b}(\,\cdot\,) \sum_{i=0}^{N} \left(\sum_{r=0}^{i} G_{i,r} \hat{\mu}_{r,t} \right) G_i(\,\cdot\,)$, where $\hat{\mu}_{0,t} \equiv 1$

2: draw a weighted posterior sample $(\varpi_{\ell,t}, S_{\ell,t})_{\ell=1,\ldots,\ell_{\max}}$ of $\tilde{S}(t)$, of size $\ell_{\max}$, from $\pi_{N,t}(\,\cdot\,) \propto \max\left(f_{N,t}(\,\cdot\,), 0\right)$ by means of the important sampler described in Section 1.1.3

The drawn samples allow us to approximately evaluate the posterior distribution of $\tilde{S}(t_i)$, for every $i = 1, \ldots, q$. This, in turn, can be exploited to carry out meaningful Bayesian inference (Algorithm 1.1.2). As a remarkable example, we consider the median survival time, denoted by m. The identity for the cumulative distribution

function of m

$$\mathbb{P}\left(m \leq t|\boldsymbol{X}\right) = \mathbb{P}\big(\tilde{S}(t) \leq 1/2|\boldsymbol{X}\big)$$

allows us to evaluate the CDF of m at each time point t_i as $c_i = \mathbb{P}\big(\tilde{S}(t_i) \leq 1/2|\boldsymbol{X}\big)$. Then, the median survival time m can be estimated by means of the following approximation:

$$\hat{m} = \mathbb{E}_{\boldsymbol{X}}[m] = \int_0^\infty \mathbb{P}[m > t|\boldsymbol{X}]\,\mathrm{d}t \approx \frac{M}{q-1}\sum_{i=1}^{q}(1-c_i) \tag{1.14}$$

where the subscript $\boldsymbol{X}$ in $\mathbb{E}_{\boldsymbol{X}}[m]$ indicates that the integral is with respect to the distribution of $\tilde{S}(\cdot)$ conditional to $\boldsymbol{X}$. Equivalently,

$$\hat{m} \approx \sum_{i=1}^{q} t_i(c_{i+1} - c_i), \tag{1.15}$$

with the proviso that $c_{q+1} \equiv 1$. Moreover, the sequence $(c_i)_{i=1}^q$ can be used to devise credible intervals for the median survival time, cf. Part 1 of Algorithm 1.1.2. Note that both in (1.14) and in (1.15) the integrals on the left-hand-side are approximated by means of simple Riemann sums and the quality of such an approximation clearly depends on the choice of q and on M. Nonetheless, our investigations suggest that if q is sufficiently large the estimates we obtain are pretty stable and that the choice of M is not crucial since, for t_i sufficiently large, the term $1 - c_i$ involved in (1.14) is approximately equal to 0. Finally, the posterior samples generated by Algorithm 1.1.1 can be used to obtain a t-by-t estimation of other functionals that convey meaningful information such as the posterior mode and median (together with the posterior mean), cf. Part 2 of Algorithm 1.1.2.

Algorithm 1.1.2 Bayesian inference

Part 1. Median survival time

1: use the weighted sample $(\omega_{\ell,t_i}, S_{\ell,t_i})_{\ell=1,\dots,\ell_{\max}}$ to estimate, for each $i = 1,\dots,q$, $c_i = \mathbb{P}(\tilde{S}(t_i) \leq 1/2|\boldsymbol{X})$

2: plug the c_i's in (1.15) to obtain $\hat{m}$

3: use the sequence $(c_i)_{i=1}^q$ as a proxy for the posterior distribution of m so to devise credible intervals for $\hat{m}$.

Part 2. t-by-t functionals

1: use the weighted sample $(\omega_{\ell,t_i}, S_{\ell,t_i})_{\ell=1,\dots,\ell_{\max}}$ to estimate, for each $i = 1,\dots,q$, $a_i = \inf_{x\in[0,1]}\{\mathbb{P}(\tilde{S}(t_i) \leq x|\boldsymbol{X}) \geq 1/2\}$ and $b_i = \mathrm{mode}\{\tilde{S}(t_i)|\boldsymbol{X}\}$

2: use the sequences $(a_i)_{i=1}^q$ and $(b_i)_{i=1}^q$ to approximately evaluate, t-by-t, posterior median and mode respectively

3: use the weighted sample $(\omega_{\ell,t_i}, S_{\ell,t_i})_{\ell=1,\dots,\ell_{\max}}$ to devise t-by-t credible intervals

---

The rest of this section is divided in two parts in which Algorithms 1.1.1 and 1.1.2 are applied to analyse simulated and real survival data. In Section 1.1.4 the focus is on the estimation of the median survival time for simulated samples of varying size. In Section 1.1.4 we analyse a real two-sample dataset and we estimate posterior

median and mode, together with credible intervals, of $\tilde{S}(t)$. In both illustrations our approximations are based on the first $N = 10$ moments.

Application to simulated survival data

Consider four samples of size $n = 25, 50, 100, 200$, from a mixture f of Weibull distributions, defined by

$$f = \frac{1}{2}\text{Weib}(2, 2) + \frac{1}{2}\text{Weib}(2, 1/2).$$

After observing that the largest observation in the samples is 4.21, we set $M = 5$ and $q = 100$ for the analysis of each sample. By applying Algorithms 1.1.1 and 1.1.2 we approximately evaluate, t-by-t, the posterior distribution of $\tilde{S}(t)$ together with the posterior distribution of the median survival time m. In Figure 1.4 the focus is on the sample corresponding to $n = 100$. On the left panel, true survival function and Kaplan–Meier estimate are plotted. By investigating the right panel it can be appreciated that the estimated HPD credible regions for $\tilde{S}(t)$ contain the true survival function. Moreover, the posterior distribution of m is nicely concentrated around the true value $m_0 = 0.724$.

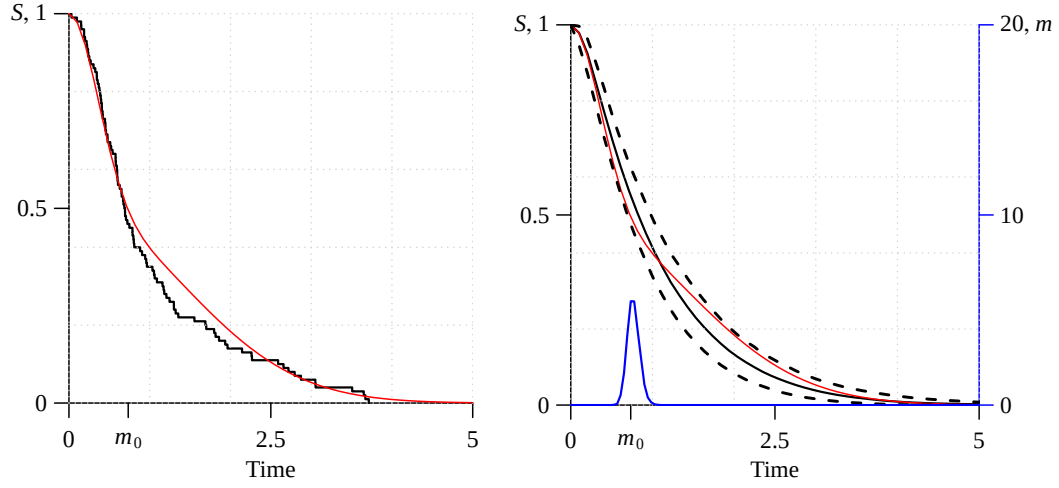


FIGURE 1.4: (Simulated dataset, $n = 100$.) Left: true survival function (red line) and Kaplan–Meier estimate (balk line). Right: true survival function (red line) and estimated posterior mean (black solid line) with 95% HPD credible intervals for $\tilde{S}(t)$ (black dashed lines); the blue plot appearing in the panel on the right is the posterior distribution of the median survival time m.

The performance of the introduced methodology is investigated as the sample size n grows. Table 1.1 summarizes the values obtained for $\hat{m}$ and the corresponding credible intervals. For all the sample sizes considered, credible intervals for $\hat{m}$ contain the true value. Moreover, as expected, as n grows, they shrink around m_0: for example the length of the interval reduces from 0.526 to 0.227 when the size n changes from 25 to 200. Finally, for all these samples, the estimated median survival time $\hat{m}$ is closer to m_0 than the empirical estimator $\hat{m}_e$.

TABLE 1.1: (Simulated datasets.) Comparison of the estimated median survival time ($\hat{m}$) obtained by means of our Bayesian nonparametric procedure (BNP) and the empirical median survival time $\hat{m}_e$, for different sample sizes. For BNP estimation we show $\hat{m}$, the absolute error $|\hat{m} - m_0|$ and the 95%-credible interval (CI); last two columns show the empirical estimate $\hat{m}_e$ and the corresponding absolute error $|\hat{m}_e - m_0|$. The true median survival time m_0 is 0.724.

	BNP			Empirical	
sample size	$\hat{m}$	error	CI	$\hat{m}_e$	error
25	0.803	0.079	(0.598, 1.124)	0.578	0.146
50	0.734	0.010	(0.577, 0.967)	0.605	0.119
100	0.750	0.026	(0.622, 0.912)	0.690	0.034
200	0.746	0.022	(0.669, 0.896)	0.701	0.023

Application to real survival data

The described methodology is now used to analyse a well known two-sample dataset involving leukemia remission times, in weeks, for two groups of patients, under active drug treatment and placebo respectively. The same dataset was studied, e.g., by [85]. Observed remission times for patients under treatment (T) are

$$\{6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*\},$$

where stars denote right-censored observations. On the other side, remission times of patients under placebo (P) are all exact and coincide with

$$\{1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23\}.$$

For this illustration we set $M = 2\max(\boldsymbol{X})$, that is $M = 70$, and $q = 50$. For both samples posterior mean, median and mode as well as 95% credible intervals, are estimated and compared. In the left panel of Figure 2.3 such estimates are plotted for sample T. By inspecting the plot, it is apparent that, for large values of t, posterior mean, median and mode show significantly different behaviors, with posterior mean being more optimistic than posterior median and mode. It is worth stressing that such differences, while very meaningful for clinicians, could not be captured by marginal methods for which only the posterior mean would be available. A fair analysis must take into account the fact that, up to $t = 23$, i.e. the value corresponding to the largest non-censored observation, the three curves are hardly distinguishable. The different patterns for larger t might therefore depend on the prior specification of the model. Nonetheless, this example is meaningful as it shows that a more complete posterior analysis is able to capture differences, if any, between posterior mean, median and mode.

When relying on marginal methods, the most natural choice for estimating the uncertainty of posterior estimates consists in considering the quantiles intervals corresponding to the output of the Gibbs sampler, that we refer to as *marginal intervals*. This leads to consider, for any fixed t, the interval whose lower and upper extremes

are the quantiles of order 0.025 and 0.975, respectively, of the sample of conditional moments $\{\mu_{1,t}^{(l_{\min}+1)}, \ldots, \mu_{1,t}^{(l_{\max})}\}$ defined in (1.13). In the middle panel of Figure 2.3 the estimated 95% HPD intervals for $\tilde{S}(t)$ and the marginal intervals corresponding to the output of the Gibbs sampler are compared. In this example, the marginal method clearly underestimates the uncertainty associated to the posterior estimates. This can be explained by observing that, since the underlying completely random measure has already been marginalized out, the intervals arising from the Gibbs sampler output, capture only the variability of the posterior mean that can be traced back to the latent variables $\boldsymbol{Y}$ and the parameters (c, β). As a result, the uncertainty detected by the marginal method leads to credible intervals that can be significantly narrower than the actual posterior credible intervals that we approximate through the moment-based approach. This suggests that the use of intervals produced by marginal methods as proxies for posterior credible intervals should be, in general, avoided.

The analysis is concluded by observing that the availability of credible intervals for survival functions can be of great help in comparing treatments. In the right panel of Figure 2.3 posterior means as well as corresponding 95% HPD intervals are plotted for both samples T and P. By inspecting the plot, for example, the effectiveness of the treatment seems clearly significant as, essentially, there is no overlap between credible intervals of the two groups.

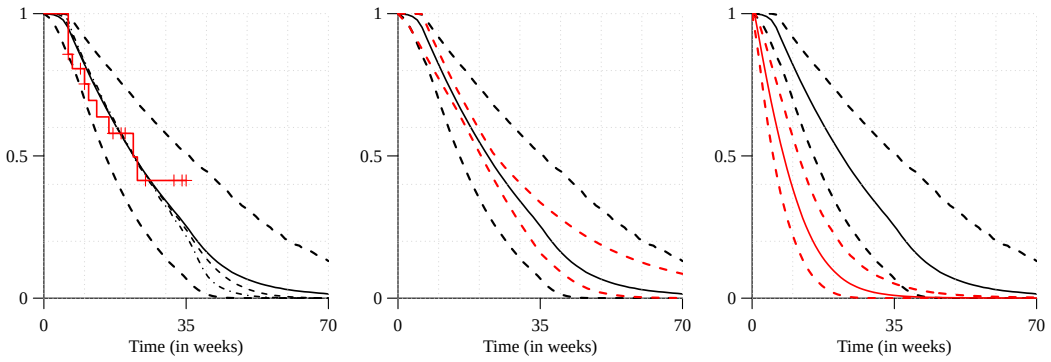


FIGURE 1.5: Left: comparison of posterior mean (solid line), median (dashed line) and mode (point dashed line) in dataset T, with 95% HPD credible intervals (dashed line). The Kaplan–Meier estimate is plotted in red. Middle: comparison of the 95% HPD credible interval (dashed black line) with the marginal interval (dashed red line). Right: comparison of samples T (black) and P (red), with posterior means (solid) and 95% HPD credible intervals (dashed).

[S2] H. Lü, J. Arbel, and F. Forbes. Bayesian Nonparametric Priors for Hidden Markov Random Fields. *Under major revision, Statistics and Computing*, 2019

1.2 Image segmentation: mixtures with hidden Markov random fields

1.2.1 Introduction

Hidden Markov random field (HMRF) models are widely used for clustering data under spatial constraints. Spatial dependencies are encoded by modelling the cluster labels as a discrete state Markov random field (MRF) such as Ising (two clusters or states) or Potts (more than two clusters) model [78, 307]. HMRF can be seen as spatial extensions of independent mixture models. As for standard mixtures, one concern is the automatic selection of the proper number of clusters in the data, or equivalently the number of states in the HMRF. In the independent data case, several criteria exist to select this number automatically based on penalized likelihoods (*e.g.*, AIC, BIC, ICL, etc.) and have been extended in the HMRF framework using variational approximation [120]. They require running several models with different cluster numbers so as to choose the best one, with a potential waste of computational effort as all the other models are usually discarded. Other techniques use a fully Bayesian setting including a prior on the number of components. The most celebrated method in this case is reversible jump Markov chain Monte Carlo [144]. Although simplifications in the inference have been proposed recently in [234], the computational cost of reversible jump techniques remains considerably high.

In the present work, we investigate alternatives based on Bayesian nonparametric (BNP) methods. In particular, Dirichlet process mixture (DPM) models have emerged as promising candidates for clustering applications where the number of clusters is unknown. Nevertheless, applications of DPMs involve observations which are assumed to be independent. For more complex tasks such as unsupervised image segmentation with spatial relationships or dependencies between the observations, DPMs are not satisfactory. Therefore, we propose to introduce MRF dependencies between data points in BNP models, and we term the resulting model BNP-MRF. This requires to extend finite state space MRF models to an infinite number of states. We show that this can be achieved by incorporating a stick-breaking scheme in an MRF formulation more general than the standard Potts model commonly used.

The addition of MRF dependencies between data points in BNP models raises the question of how they impact the natural clustering and rich-get-richer properties of BNP priors? We answer this question by providing theoretical results about two quantities of interest for BNP priors: the predictive distribution, that represents the distribution of one datum conditional on previous observations, and the number of clusters induced by a BNP-MRF prior.

The links to other similar attempts is reviewed in Section 1.2.1. The proposed BNP-MRF model is explained in Section 1.2.2 and theoretical properties are investigated

in Section 1.2.3. The model implementation using variational approximation is detailed in Section 1.2.4. An illustration of its performance on an image segmentation task is provided in Section 1.2.5 and a conclusion ends the section.

Related work

Attempts to build countably infinite state space MRF models using BNP priors have already appeared in the literature. In particular, we can distinguish attempts such as [79, 80, 89] from the work in [3, 256, 340, 305]. The approach in [79, 80, 89] differs in that it is not based on a generalization of the Potts model but on a transformation of an inference algorithm. More specifically in [79, 80], a standard mean field approximation is first considered and then transformed to account for an infinite number of states. In that sense it is closer to an Iterated Conditional Mode (ICM) algorithm [53], but does not provide a spatial generalization of DPMs. Typically, the simple Potts model considered in [79, 80] cannot be extended to an infinite number of states as it will become clear in our Section 1.2.2. Other attempts include the work in [179], but there the number of states is known to be three and the Dirichlet process (DP) prior is used instead to model intensity distributions non-parametrically. Segmentation with spatially dependent Pitman–Yor processes (PY) has also been considered in [308], but using Gaussian processes.

We build on the approach in [3] which differs from [256, 340, 305] which all use a partition model representation. In particular, [340] generalizes [256] and proposes a more efficient Markov chain Monte Carlo (MCMC) inference by means of the Swendsen–Wang algorithm, while [305] extends this idea to hierarchical DP priors for multiple image segmentation. In contrast to [256, 340, 305], we propose to use a stick-breaking-based scheme for the mixing weights, thus providing a more comprehensive representation than partition models which integrate out the process. In addition, stick-breaking representations lead naturally to variational approximations for performing inference [58]. The advantage is to reduce the computational cost in complex data clustering without suffering from label switching complications. In other words, in our approach the MRF is imposed internally in the BNP mechanics leading to well defined infinite state HMRF models. This construction is valid for any stick-breaking representation. We show how it can be implemented for the DP and PY priors, and provide references for extensions to larger classes of BNP priors.

1.2.2 BNP Markov random field mixture models

The clustering task is addressed through a missing data model that includes a set $\mathbf{y} = (y_1, \ldots, y_n)$ of observed variables from $\mathbb{R}^d$ and a set $\mathbf{z} = (z_1, \ldots, z_n)$ of missing (also called hidden) variables whose joint distribution $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta})$ is governed by a set of parameters denoted by $\boldsymbol{\Theta}$ and possibly by additional hyperparameters $\boldsymbol{\phi}$ not specified in the notation. The latter ones are usually fixed and not considered at first. Typically, the z_i's corresponding to group memberships (or labels), take their values in $\{1, \ldots, K\}$ where K is the number of clusters or groups. We shall denote by $\mathcal{Z} = \{1, \ldots, K\}^n$ the set in which $\mathbf{z}$ takes its values and by $\underline{\Theta}$ the parameter space. To account for dependencies between the z_i's, $\mathbf{z}$ can be modeled as a discrete MRF. If in addition, the y_j's are independent conditionally on $\mathbf{z}$, the joint distribution

$p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta})$ is referred to as an HMRF model. In this case, the conditional distribution $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\Theta})$ is also an MRF. For clustering dependent data into K groups, the most commonly used MRF is the so-called Potts model [78, 307].

As already mentioned, our goal is to bypass the issue of selecting the number K of clusters by considering a countably infinite number of them while allowing MRF dependencies between the $\mathbf{y}_i$'s. The construction of the proposed model is explained starting from the link between standard finite mixtures and Dirichlet process mixtures. Basic DP principles and notations are recalled in Section 1.2.2. The extension of finite state space MRF to a countably infinite number of states is given in Section 1.2.2 and the resulting BNP-MRF mixture models is summarized in Section 1.2.2.

From finite mixtures to DP mixture models

A generative approach to clustering consists of picking one of K clusters from a multinomial distribution with weights parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and then to generate a data point y from a cluster specific distribution $p(y \mid \theta_k^*)$ with cluster specific parameter θ_k^*. This yields a finite mixture model

$$p(y \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \, p(y \mid \theta_k^*) \tag{1.16}$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_K^*)$ and $\boldsymbol{\pi}$ are the parameters. For instance, for Gaussian mixtures, $\theta_k^* = (\mu_k, \Sigma_k)$ and $p(y \mid \theta_k^*)$ is a Gaussian distribution with mean μ_k and covariance matrix Σ_k, denoted by $\mathcal{N}(\mu_k, \Sigma_k)$ or $\mathcal{N}(y \mid \mu_k, \Sigma_k)$ when referring to the probability density function (pdf). The observations $(y_1, \ldots, y_n)$ are therefore *i.i.d.* and generated from the same mixture (1.16). It follows that the kth cluster is by definition the set of data points arising from the kth mixture component. This is usually expressed by introducing for each y_j an additional hidden variable Z_j that takes its values in $\{1, \ldots, K\}$, so that $p(z_j = k \mid \boldsymbol{\pi}) = \pi_k$. Another way to obtain a sample from a finite mixture model consists of defining a discrete measure $G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}$ and then of considering the following hierarchical representation, for all $j = 1, \ldots, n$,

$$\begin{aligned} \theta_j \mid G &\overset{\text{i.i.d.}}{\sim} G, \\ y_j \mid \theta_j &\overset{\text{ind}}{\sim} p(\,.\mid \theta_j). \end{aligned}$$

The subset of θ_j's that are equal to θ_k^* corresponds to the y_j's in the kth cluster.

In a Bayesian setting, in addition, a prior distribution is placed on $\boldsymbol{\theta}^*$ and $\boldsymbol{\pi}$. The most common choice for $\boldsymbol{\pi}$ is the Dirichlet distribution $\text{Dir}(\alpha_1, \ldots \alpha_K)$ depending on a vector of positive parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. The choice of the prior on $\boldsymbol{\theta}^*$ (denoted by G_0) is model-specific, usually following a conjugate prior such as a Normal inverse-Wishart distribution for Gaussian mixture models. Other cases are possible

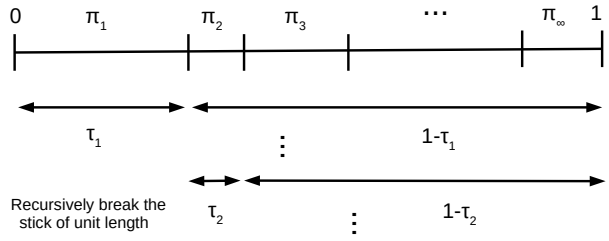


FIGURE 1.6: Illustration of the stick-breaking representation.

and tractable (*e.g.* [84]). It follows the hierarchical representation:

$$\theta_1^*, \dots, \theta_K^* \mid G_0 \sim G_0, \tag{1.17}$$
$$\boldsymbol{\pi} \mid \boldsymbol{\alpha} \sim \text{Dir}(\alpha_1, \dots \alpha_K), \tag{1.18}$$
$$G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}, \tag{1.19}$$
$$\theta_j \mid G \overset{\text{i.i.d.}}{\sim} G, \; j = 1, \dots, n, \tag{1.20}$$
$$y_j \mid \theta_j \overset{\text{ind}}{\sim} p(\,.\mid \theta_j)\; j = 1, \dots, n.$$

To become non-parametric, a first approach is to consider an infinite number of π_k's. Using an infinite number of random variables $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots)$ on $[0,1]$, we can construct an infinite number of π_k's that sum to one as follows:

$$\pi_1(\boldsymbol{\tau}) = \tau_1 \quad \text{and} \quad \pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1}(1-\tau_l), \quad k = 2, 3, \dots$$

The intuition behind this construction, referred to as *stick-breaking*, is that it consists of recursively breaking a unit-length stick as shown in Fig. 1.6. It follows an explicit formula for the π_k's. Hence, the τ_k's simulation replaces step (1.18), and G in (1.19) can be replaced by

$$G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\theta_k^*}.$$

We can also add after step (1.20) the fact that $z_j = k$ if $\theta_j = \theta_k^*$ and replace the last step by $y_j | z_j, \boldsymbol{\theta}^* \overset{\text{ind}}{\sim} p(\,.\mid \theta_{z_j}^*)$. Then the distributions of the τ_k's need to be specified. The Dirichlet process [117], denoted by $\text{DP}(G_0, \alpha)$, is characterized by a base distribution G_0 and a positive scaling parameter α. Its stick-breaking representation corresponds to *i.i.d* τ_k's that follow the same beta $\mathcal{B}(1, \alpha)$ distribution [164]. All together, using the same notation G_0 for the prior of each θ_k^* simulated as *i.i.d.* variables, it comes

the following hierarchical representation:

$$\theta_k^* \mid G_0 \overset{\text{i.i.d.}}{\sim} G_0, \quad k = 1, 2, \ldots, \tag{1.21}$$

$$\tau_k \mid \alpha \overset{\text{i.i.d.}}{\sim} \mathcal{B}(1, \alpha), \quad k = 1, 2, \ldots,$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \quad k = 1, 2, \ldots, \tag{1.22}$$

$$G = \sum_{k=1}^{\infty} \pi_k(\tau) \delta_{\theta_k^*}, \tag{1.23}$$

$$\theta_j \mid G \overset{\text{i.i.d.}}{\sim} G, \text{ and } z_j = k \text{ if } \theta_j = \theta_k^* \tag{1.24}$$

$$y_j \mid z_j, \boldsymbol{\theta}^* \overset{\text{ind}}{\sim} p(\,.\,\mid \theta_{z_j}^*). \tag{1.25}$$

The above hierarchical representation corresponds to a countably infinite mixture model referred to as a Dirichlet process mixture (DPM) model. It is an explicit characterization of the DP (Eq. (1.21) to (1.23)) and of the DPM (Eq. (1.21) to Eq. (1.25)) using a stick-breaking construction. The stick-breaking representation will be particularly useful in our study for both the definition of our model (Sections 1.2.2 and 1.2.2) and its estimation (Section 1.2.4).

Infinite MRF priors

The explicit use of the labels $\boldsymbol{z} = (z_1, \ldots, z_n)$ in the DPM construction above makes it closer to clustering generative models and opens the way to an HMRF extension. Such a generalization is only possible from Potts models with an external field parameter. In the finite state space case, an MRF model is defined using a dependence structure coded via a graph $\mathcal{G}$ whose nodes correspond to the variables. A K-state Potts model with an external field, defined over $\boldsymbol{z} = (z_1, \ldots, z_n)$ with for all $j = 1, \ldots, n$, $z_j \in \{1, \ldots, K\}$, corresponds to the following pdf,

$$p(\boldsymbol{z}; \beta, \boldsymbol{v}) \propto \exp\left(\sum_{j=1}^{n} v_{z_j} + \beta \sum_{i \sim j} \delta_{(z_i = z_j)} \right), \tag{1.26}$$

where $i \sim j$ means that i and j are neighbors, *i.e.* linked by an edge, in the considered dependence structure described by graph $\mathcal{G}$, $\delta_{(z_i = z_j)}$ is the indicator function which is 1 if $z_i = z_j$ and 0 otherwise, β is a positive scalar interaction parameter and $\boldsymbol{v} = (v_1, \ldots, v_K)$ represents an additional external field parameter where each v_k is a scalar. The distribution (1.26) is insensitive to an addition of the same constant to all the v_k's. Such non-identifiability can be overcome by an additional constraint on $\boldsymbol{v}$ such as requiring $\sum_{k=1}^{K} \pi_k = 1$ with $v_k = \log \pi_k$. The Potts model in (1.26) can then be rewritten as

$$p(\boldsymbol{z}; \beta, \boldsymbol{\pi}) \propto \left(\prod_{j=1}^{n} \pi_{z_j} \right) \exp\left(\beta \sum_{i \sim j} \delta_{(z_i = z_j)} \right). \tag{1.27}$$

In the finite state space case, we can equivalently use the Gibbs representation,

$$p(\boldsymbol{z}; \beta, \boldsymbol{\pi}) \propto \mathrm{e}^{V(\boldsymbol{z}; \beta, \boldsymbol{\pi})}, \tag{1.28}$$

where $V(\boldsymbol{z}; \beta, \boldsymbol{\pi}) := \sum_{j=1}^{n} \log \pi_{z_j} + \beta \sum_{i \sim j} \delta_{(z_i = z_j)}$ is often referred to as the *energy function*. The first sum in V represents the first order potentials while the second sum represents the second order potentials. In the finite state space case, the Hammersley–Clifford theorem [53] applied to the Gibbs representation (1.28) entails that the distribution in (1.26) is a Markov random field. What is interesting about formulas (1.26) and (1.27) is that they do not involve the number of states K. As long as a stick-breaking construction is available, we can consider a countably infinite number of probabilities π_k that sum to one, *i.e.*, $\sum_{k=1}^{\infty} \pi_k = 1$ and define the same energy function V as before but over an infinite countable set of states. Using the Gibbs representation (1.28), the Hammersley–Clifford theorem still holds if we can show that $\sum_{\boldsymbol{z}} \mathrm{e}^{V(\boldsymbol{z}; \boldsymbol{\pi}, \beta)} < \infty$, where the sum runs over all n-uples of positive integers $\boldsymbol{z} \in \{1, 2, \ldots\}^n$. Note that this latter condition that is automatically satisfied in the finite state space case (for reasonable potential functions), may not be satisfied in the infinite case. However, the stick-breaking representation of $\boldsymbol{\pi}$ ensures this property since:

$$\begin{aligned} \sum_{\boldsymbol{z}} \mathrm{e}^{V(\boldsymbol{z}; \beta, \boldsymbol{\pi})} &\overset{(a)}{\leq} \left(\sum_{\boldsymbol{z}} \prod_{j=1}^{n} \pi_{z_j} \right) \mathrm{e}^{\beta \frac{n(n-1)}{2}}, \\ &\overset{(b)}{=} \mathrm{e}^{\beta \frac{n(n-1)}{2}} < \infty \end{aligned}$$

where we used for (a) the fact that $n(n-1)/2$ is the maximum number of neighbors among n observations (complete dependence or graph), while (b) comes from $\sum_{\boldsymbol{z}} \prod_{j=1}^{n} \pi_{z_j} = \left(\sum_{k=1}^{\infty} \pi_k \right)^n = 1$. It follows that $p(\boldsymbol{z}; \beta, \boldsymbol{\pi})$, in the infinite state space case, is still a valid probability distribution and is an MRF by the Hammersley–Clifford theorem. Such a generalization is possible because of the presence of the external field parameters π_k that satisfy $\sum_{k=1}^{\infty} \pi_k = 1$ as ensured by the stick-breaking construction. A standard Potts model with equal or no external field parameters cannot be as simply extended to an infinite countable state space because in the K-state case this Potts model is equivalent to $\pi_k = 1/K$ for all k which possesses a degenerate limit when K tends to infinity.

BNP-MRF mixture models

The stick-breaking representation amounts to identifying a set of random variables $\boldsymbol{\tau} = (\tau_k)_{k=1}^{\infty}$ with each $\tau_k \in [0, 1]$ and so that the weights π_k are defined by (1.22). Then the Potts model construction (1.27) is valid for any set of parameters $\boldsymbol{\tau} = (\tau_k)_{k=1}^{\infty}$ with each $\tau_k \in [0, 1]$. Bayesian non-parametric priors specify a prior distribution on τ_k's. For instance, as already mentioned for the DP stick-breaking, all τ_k's are independent and identically distributed according to a $\mathcal{B}(1, \alpha)$ distribution. For the Pitman–Yor (PY) process [270], the τ_k's are independent but not identically distributed with

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \mathcal{B}(1 - \sigma, \alpha + k\sigma) \quad \text{for } k = 1, 2, \ldots, \tag{1.29}$$

where $\sigma \in (0,1)$ is a discount parameter and α a concentration parameter $\alpha > -\sigma$. The PY is a two-parameter generalisation of the DP which allows to control the tail behavior when modeling data with either exponential or power-law tails [164, 270]. When $\sigma = 0$, the PY reduces to a DP. More general stick-breaking representations are possible (*e.g.*, for Gibbs-type priors [93, 131] or homogeneous normalized random measures with independent increments (NRMIs) [114]) but the Pitman–Yor case provides a clear interpretation in terms of number of clusters. The rich-gets-richer property of the DP is preserved meaning that there are a small number of large clusters, but there is also a large number of small clusters with parameter σ decreasing the probability that observations join small clusters. The PY yields a power-law behavior which can make it more suitable for a number of applications. In other words, the number of clusters grows as $\mathcal{O}(n^\sigma)$ for the PY while it grows more slowly at $\mathcal{O}(\log n)$ for the DP.

The extension we propose is therefore to augment the original HMRF formulation with additional variables $(\tau_k)_{k=1}^{\infty}$. We refer to it as the BNP-MRF mixture model. It corresponds to the following hierarchical construction written here in the PY case:

$$\theta_k^* \mid G_0 \overset{\text{i.i.d.}}{\sim} G_0, \; k = 1, 2, \ldots, \tag{1.30}$$

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \mathcal{B}(1-\sigma, \alpha + k\sigma), \; k = 1, 2, \ldots, \tag{1.31}$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \tag{1.32}$$

$$p(\boldsymbol{z} \mid \boldsymbol{\tau}; \beta) \propto \left(\prod_{j=1}^{n} \pi_{z_j}(\boldsymbol{\tau}) \right) \exp\left(\beta \sum_{i \sim j} \delta_{(z_i = z_j)} \right), \tag{1.33}$$

$$y_j \mid z_j, \boldsymbol{\theta}^* \overset{\text{ind}}{\sim} p(y_j \mid \theta_{z_j}^*) \; . \tag{1.34}$$

The prior on τ_k's from (1.31) can be adapted to more general classes of BNP priors, see for example Theorem 14.23 of [131] for Gibbs-type priors, and [114] for NRMIs. Importantly, in the BNP-MRF model above, the θ_j's and z_j's are not *i.i.d* conditionally on G anymore. The joint distribution (1.33) on $\boldsymbol{z} = (z_1, \ldots, z_n)$ induces a joint distribution on $(\theta_1, \ldots, \theta_n)$ using that $\theta_j = \theta_{z_j}^*$. If we still denote for simplicity by G this joint distribution, we can define it in a similar manner as in the *i.i.d.* case, using its conditional specifications,

$$\theta_j | \theta_{\mathcal{N}_j}; G \sim \sum_{k=1}^{\infty} p(z_j = k | z_{\mathcal{N}_j}, \boldsymbol{\tau}; \beta) \; \delta_{\theta_k^*},$$

where $\mathcal{N}_j$ denotes the neighbors of j in the graph dependence structure $\mathcal{G}$ and the $p(z_j = k | z_{\mathcal{N}_j}, \boldsymbol{\tau}; \beta)$'s are the conditional specifications of (1.33).

In Section 1.2.4, we detail the case when cluster specific distributions are Gaussian, with $\theta_k^* = (\mu_k, \Sigma_k)$ and $p(y_j \mid \theta_k^*) = \mathcal{N}(y_j \mid \mu_k, \Sigma_k)$.

1.2.3 Predictive distribution and number of clusters for a BNP-MRF prior

In this section, we provide theoretical results about two quantities of interest for Bayesian nonparametric priors: the predictive distribution, that represents the distribution of one datum conditional on previous observations, and the number of clusters induced by a BNP-MRF prior. We consider data of varying sample size, and denote by $\mathcal{G}_n$ the subgraph of $\mathcal{G}$ induced by node $\{1, \ldots, n\}$.

We focus on the large class of Gibbs-type priors [93], of which the DP and PY are special cases. Consider n observations $(\theta_1, \ldots, \theta_n)$ sampled from a BNP-MRF prior Eq (1.30)-(1.33) but using a Gibbs-type prior instead of PY prior (1.31). We are interested in the predictive distribution of observation θ_{n+1} conditional on $(\theta_1, \ldots, \theta_n)$, but unconditional on G. With a BNP-MRF prior, this predictive distribution depends on the structure of the graph $\mathcal{G}$, more specifically on the neighbors of θ_{n+1}. Denote by K_n the number of clusters in $(\theta_1, \ldots, \theta_n)$, by $(\theta_1^\star, \ldots, \theta_{K_n}^\star)$ their K_n different values[1] and by $(n_1, \ldots, n_{K_n})$ their size. We first consider the Gibbs-type prior case without the addition of a Markov component. The predictive distribution [131] is given by,

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0 + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{\ell=1}^{k_n} (n_\ell - \sigma) \delta_{\theta_\ell^\star} \tag{1.35}$$

where the triangular array of nonnegative parameters $V_{n,k}$, $1 \leq k \leq n$, satisfy the backward recurrence relation

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1}, \tag{1.36}$$

with $V_{1,1} = 1$. This predictive can be specialized to the PY case with

$$V_{n,k} = \frac{\sigma^k (1 + \frac{\alpha}{\sigma})_{(k-1)}}{(1 + \alpha)_{(n-1)}},$$

where $(a)_{(x)} := \Gamma(a + x)/\Gamma(a)$ denotes the rising factorial. It follows

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{\alpha + \sigma k_n}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{\ell=1}^{k_n} (n_\ell - \sigma) \delta_{\theta_\ell^\star}, \tag{1.37}$$

while the case of the DP is obtained by setting $\sigma = 0$ above.

For the sake of simplicity, we propose to use the labels notation $z_{1:n} = (z_1, \ldots, z_n)$ defined so that $z_j = \ell$ when $\theta_j = \theta_\ell^\star$, we denote by $\{z_1, \ldots, z_n\}$ the set of label values which includes only K_n different labels. In the Gibbs-type prior case, it is clear from (1.35) that

$$\begin{aligned} p(z_{n+1} \mid z_{1:n}) &= \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \quad \text{if } z_{n+1} \notin \{z_1, \ldots, z_n\}, \\ p(z_{n+1} = \ell \mid z_{1:n}) &= \frac{V_{n+1,k_n}}{V_{n,k_n}} (n_\ell - \sigma) \quad \text{if } \ell \in \{z_1, \ldots, z_n\}\,. \end{aligned} \tag{1.38}$$

[1] Note that the notation introduced for the different $\theta_j^\star$ differs from that devoted to the stick-breaking variables, θ_j^*.

The next proposition indicates how the predictive is impacted by the addition of a Markov dependence. The neighbors of θ_{n+1} in $\mathcal{G}_n$ is denoted by $\mathcal{N}_{n+1}$ and $\tilde{n}_\ell$ is the number of neighbors of θ_{n+1} which belong to cluster ℓ, hence satisfying $\tilde{n}_\ell \leq n_\ell$. Also $z_{\mathcal{N}_{n+1}} = \{z_i, i \in \mathcal{N}_{n+1}\}$ denotes the labels in the neighborhood.

Proposition 1.2.1 (Predictive distribution of a Gibbs-MRF prior)**.** *The predictive distribution for a Gibbs-MRF prior is given by*

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{V_{n+1,k_n+1}}{V_{n,k_n} + V_{n+1,k_n}\boldsymbol{\eta}_{n+1}} G_0 + \frac{V_{n+1,k_n}}{V_{n,k_n} + V_{n+1,k_n}\boldsymbol{\eta}_{n+1}} \sum_{\ell=1}^{k_n} \boldsymbol{\lambda}_{n+1,\ell}\, \delta_{\theta_\ell^\star} \tag{1.39}$$

where

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_{n+1}(\sigma, \beta) = \sum_{\ell \in z_{\mathcal{N}_{n+1}}} (n_\ell - \sigma)(\mathrm{e}^{\beta\tilde{n}_\ell} - 1),$$

$$\boldsymbol{\lambda}_{n+1,\ell} = \boldsymbol{\lambda}_{n+1,\ell}(\sigma, \beta) = (n_\ell - \sigma)\, \mathrm{e}^{\beta\tilde{n}_\ell \delta_{\mathcal{N}_{n+1}}(\ell)}.$$

and $\delta_{\mathcal{N}_{n+1}}(\ell)$ is 1 when ℓ is a label present in the neighborhood of θ_{n+1} and 0 otherwise.

Refer to the paper [S2] for a proof.

Remark 1.2.1. *When $\beta = 0$, $\boldsymbol{\eta}_{n+1}(\sigma, 0) = 0$ and $\boldsymbol{\lambda}_{n+1,\ell}(\sigma, 0) = n_\ell - \sigma$ so that the Gibbs-type prior predictive (1.35) is recovered. In contrast, for $\beta > 0$, the above predictive specialized to the PY-MRF case is,*

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{\alpha + \sigma k_n}{\alpha + n + \boldsymbol{\eta}_{n+1}} G_0 + \frac{1}{\alpha + n + \boldsymbol{\eta}_{n+1}} \sum_{\ell=1}^{k_n} \boldsymbol{\lambda}_{n+1,\ell}\, \delta_{\theta_\ell^\star}, \tag{1.40}$$

while the case of the DP-MRF is obtained by setting $\sigma = 0$. Comparing the probability of a new draw for a Gibbs-type prior, $\frac{V_{n+1,k_n+1}}{V_{n,k_n}}$, with that of a new draw for a Gibbs-MRF prior, $\frac{V_{n+1,k_n+1}}{V_{n,k_n} + V_{n+1,k_n+1}\boldsymbol{\eta}_{n+1}}$, we see that the MRF has the effect of reducing this probability. In the PY case, this increase corresponds to increasing the sample size from n to $n + \boldsymbol{\eta}_{n+1}$ when $\beta > 0$, where $\boldsymbol{\eta}_{n+1}$ can be quite a large number. More specifically for a label ℓ in the neighborhood of z_{n+1}, the weight of each previous observations with label ℓ (in the neighborhood or not) is multiplied by a factor $(\mathrm{e}^{\beta\tilde{n}_\ell} - 1)$. The effect is then all the more important as β is large and as n_ℓ is large.

The predictive distribution (1.39) provides in turn the following lower bounds on the prior expectation of the number of clusters.

Proposition 1.2.2 (Lower bound for expected number of clusters)**.** *Assume that the graph $\mathcal{G}$ has maximal degree D. Then the expected prior number of clusters for a BNP-MRF distribution has the following lower bound*

$$\mathbb{E}[K_n] \gtrsim \frac{\alpha}{\mathrm{e}^{D\beta}} \log n \tag{1.41}$$

for the Dirichlet process and

$$\mathbb{E}[K_n] \gtrsim c n^{\sigma \mathrm{e}^{-D\beta}}, \tag{1.42}$$

for the Pitman–Yor process, with some positive constant c, and where $a_n \gtrsim b_n$ stands for $\limsup a_n / b_n \geq 1$.

Refer to the paper [S2] for a proof.

Remark 1.2.2. *We do not have a proof for the general case of Gibbs-type priors, but we conjecture that the same power-law lower bound (1.42) as for PY holds. Note that the MRF component of a BNP prior can only reduce the prior expected number of clusters. For instance, for the DP with a simple graph where the first two nodes are connected, we have*

$$\mathbb{E}[K_2] = 1 + \frac{\alpha}{\alpha + \mathrm{e}^{\beta}} \leq 1 + \frac{\alpha}{\alpha + 1} = \mathbb{E}[K_2; \beta = 0]$$

where the last two terms above correspond the expectation of K_2 for a DP, i.e. when $\beta = 0$. Thus natural upper bounds that complement the lower bounds of Proposition 1.2.2 are given by

$$\mathbb{E}[K_n] \lesssim \alpha \log n$$

for the Dirichlet process and

$$\mathbb{E}[K_n] \lesssim \frac{\Gamma(\alpha + 1)}{\sigma \Gamma(\alpha + \sigma)} n^{\sigma} \tag{1.43}$$

for the Pitman–Yor process (see [267]).

1.2.4 Inference using Variational approximation

Sampling based inference (MCMC) for a similar BNP-MRF model has been proposed in [256, 340] for the case of a DP prior. As an alternative, we propose a variational approximation that is facilitated by the stick-breaking representation. For that purpose, we shall briefly recall the variational principle.

Variational Bayesian Expectation Maximization

The clustering task consists primarily of estimating the unknown labels $\mathbf{z} = (z_1, \ldots, z_n)$ from observed $\mathbf{y} = (y_1, \ldots, y_n)$ assuming a joint distribution $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi})$ governed by a set of parameters denoted by $\boldsymbol{\Theta}$ and often by additional hyperparameters $\boldsymbol{\phi}$. However to perform good label estimation, the parameters $\boldsymbol{\Theta}$ values (and hyperparameters $\boldsymbol{\phi}$) have to be available. A natural approach for parameter estimation is based on maximum likelihood, where $\boldsymbol{\Theta}$ is estimated by $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta} \in \underline{\Theta}} p(\mathbf{y} \mid \boldsymbol{\Theta})$. Then an estimate of $\mathbf{z}$ can be obtained by maximizing $p(\mathbf{z} \mid \mathbf{y}, \hat{\boldsymbol{\Theta}})$. However, $p(\mathbf{y} \mid \boldsymbol{\Theta})$ is a marginal distribution over the unknown $\mathbf{z}$ variables, so that direct maximum likelihood is intractable in general. The Expectation-Maximization (EM) algorithm [230] is a general iterative technique for maximum likelihood estimation in the presence of unobserved latent variables or missing data. An EM iteration consists of two steps usually referred to as the E-step in which the expectation of the so-called complete log-likelihood is computed and the M-step in which this expectation is maximized over $\boldsymbol{\Theta}$. An equivalent way to define EM is the following. As discussed in [243], EM can be viewed as an alternating maximization procedure of a function

$\mathcal{F}_0$ defined, for any probability distribution q_Z on $\mathcal{Z}$ by

$$\begin{aligned}\mathcal{F}_0(q_Z, \boldsymbol{\Theta}, \boldsymbol{\phi}) &= \sum_{\mathbf{z}\in\mathcal{Z}} q_Z(\mathbf{z}) \ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi}) + I[q_Z] \\ &= \mathbb{E}_{q_Z}\left[\log \frac{p(\mathbf{y}, \mathbf{Z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi})}{q_Z(\mathbf{z})}\right] \end{aligned} \tag{1.44}$$

where $I[q_Z] = -\mathbb{E}_{q_Z}[\log q_Z(\mathbf{Z})]$ is the entropy of q_Z ($\mathbb{E}_q$ denotes the expectation with regard to q). The function $\mathcal{F}_0$ depends on observations $\mathbf{y}$ which are fixed throughout, hence are omitted from the notation.

Instead of considering only point estimation of $\boldsymbol{\Theta}$, a fully Bayesian approach can be carried out, for instance when prior knowledge on the parameters $\boldsymbol{\Theta}$ is available. In this case, we have to compute

$$p(\mathbf{z} \mid \mathbf{y}) = \int_\Theta p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\Theta}) \ p(\boldsymbol{\Theta} \mid \mathbf{y}) \mathrm{d}\boldsymbol{\Theta} \tag{1.45}$$

Integrating out $\boldsymbol{\Theta}$ in this way requires the computation of $p(\boldsymbol{\Theta} \mid \mathbf{y})$ which is not usually available in closed-form. As an alternative to costly simulation-based methods (MCMC), an EM-like procedure using variational approximation can provide approximations of the marginal posterior distributions $p(\boldsymbol{\Theta} \mid \mathbf{y})$ and $p(\mathbf{z} \mid \mathbf{y})$. This approach is referred to as VBEM for Variational Bayesian EM [48]. Let q_Z and q_Θ denote respectively distributions over $\mathbf{Z}$ and $\boldsymbol{\Theta}$ that will serve as approximations to the true posteriors. Similarly to standard EM, VBEM is maximizing the following *free energy* function defined for any q_Z and q_Θ distributions

$$\mathcal{F}(q_Z, q_\Theta, \boldsymbol{\phi}) = \mathbb{E}_{q_Z q_\Theta}\left[\log \frac{p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})}{q_Z(\mathbf{z}) q_\Theta(\boldsymbol{\Theta})}\right]$$

alternatively over q_Z, q_Θ and $\boldsymbol{\phi}$. Adding a prior on $\boldsymbol{\Theta}$ is formally the same as adding $\boldsymbol{\Theta}$ to the missing variables, while the hyperparameters $\boldsymbol{\phi}$ play the role of the parameters of interest in maximum likelihood estimation.

The alternate maximization of $\mathcal{F}$ yields the VBEM algorithm that decomposes into three steps. It is easy to show, using the Kullback–Leibler (KL) divergence properties, that the maximization over q_Z and q_Θ leads to the following E-steps (see Appendix A of [77]). At the rth iteration, using current values $\boldsymbol{\phi}^{(r-1)}$ and $q_\Theta^{(r-1)}$, we get the following updating,

$$\begin{aligned}
&\textbf{VB-E-Z:} \quad q_Z^{(r)}(\mathbf{z}) \propto \exp \mathbb{E}_{q_\Theta^{(r-1)}}[\log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)})], \\
&\textbf{VB-E-}\boldsymbol{\Theta}\textbf{:} \quad q_\Theta^{(r)}(\boldsymbol{\Theta}) \propto \exp \mathbb{E}_{q_Z^{(r)}}[\log p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)})], \\
&\textbf{VB-M-}\boldsymbol{\phi}\textbf{:} \quad \boldsymbol{\phi}^{(r)} = \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_Z^{(r)} q_\Theta^{(r)}}[\log p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})].
\end{aligned}$$

Also, it is worth noticing that if $\mathbf{Y}$ and $\mathbf{Z}$ are independent of $\boldsymbol{\phi}$ conditionally on $\boldsymbol{\Theta}$, as this is often the case when $\boldsymbol{\phi}$ gathers the parameters that describe the prior on $\boldsymbol{\Theta}$,

then the VB-M-step simplifies into

$$\boldsymbol{\phi}^{(r)} = \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_\Theta^{(r)}}[\log p(\boldsymbol{\Theta};\boldsymbol{\phi})] = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}(q_\Theta^{(r)} \| p(\boldsymbol{\Theta};\boldsymbol{\phi})). \tag{1.46}$$

Then $\boldsymbol{\phi}^{(r)}$ is the value that minimizes the KL distance between the prior $p(\boldsymbol{\Theta};\boldsymbol{\phi})$ and the variational posterior $q_\Theta^{(r)}(\boldsymbol{\Theta})$. In the conjugate exponential family case, it is known that both distributions belong to the same family [48]. If this family is identifiable it follows that $\boldsymbol{\phi}^{(r)} = \hat{\boldsymbol{\phi}}^{(r)}$ where $\hat{\boldsymbol{\phi}}^{(r)}$ are the variational parameters defining $q_\Theta^{(r)}(\boldsymbol{\Theta})$. A more detailed example is given in Section 1.2.4.

In practice, we can decide which parameters are treated as genuine parameters $\boldsymbol{\Theta}$ or as hyperparameters $\boldsymbol{\phi}$, depending on whether some prior knowledge is available only for a subset of the parameters or whether the model has hyperparameters $\boldsymbol{\phi}$ for which no prior information is available. Also for complex models, q_Θ and q_Z may need to be further restricted to simpler forms, such as factorized forms, in order to ensure tractable VB-E-steps. This is illustrated in the next section for the PY-MRF inference.

VBEM for a PY-MRF mixture model with Gaussian components

The VBEM steps are described for a PY-MRF mixture model as defined in Eq. (1.31) to (1.34), with Gaussian distributed observations $\mathbf{y}$. As hyperparameters α and σ may have a significant effect on the growth of the number of clusters with data sample size, it is possible to specify priors on them. For the DP case obtained with $\sigma = 0$, it is suggested in [58] to use a gamma prior over α with two hyperparameters s_1 and s_2, *i.e.* $\alpha \sim \mathcal{G}(s_1, s_2)$ where s_1 and s_2 can be estimated or fixed. A natural question that arises is then whether one can also find a tractable prior for the discount parameter σ. We propose to use the following prior that accounts for the constraints $\sigma \in (0,1)$ and $\alpha > -\sigma$,

$$p(\alpha, \sigma; s_1, s_2, a) = p(\alpha \mid \sigma; s_1, s_2)\; p(\sigma; a) \tag{1.47}$$

where $p(\alpha \mid \sigma; s_1, s_2)$ is a shifted gamma distribution $\mathcal{SG}(s_1, s_2, \sigma)$ and $p(\sigma; a)$ is a distribution depending on some parameter a not specified for the moment but that can typically be taken as the uniform distribution on the interval $(0,1)$. Such a shifted gamma distribution is the distribution of a variable $U - \sigma$ where σ is considered as fixed and U follows a gamma distribution $\mathcal{G}(s_1, s_2)$. The pdf of this shifted gamma distribution is obtained from the standard gamma distribution as $p(\alpha \mid \sigma; s_1, s_2) = \mathcal{G}(\alpha + \sigma; s_1, s_2)$. It follows that the joint distribution of the observed data $\mathbf{y}$ and all latent variables becomes

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\Theta}; \boldsymbol{\phi}) = p(\alpha, \sigma; s_1, s_2, a) \prod_{j=1}^{n} p(y_j | z_j, \boldsymbol{\theta}^*) p(\mathbf{z} | \boldsymbol{\tau}; \beta) \prod_{k=1}^{\infty} p(\tau_k | \alpha, \sigma) \prod_{k=1}^{\infty} p(\theta_k^*; \rho_k),$$

where the notation $\prod_{k=1}^{\infty}$ is a distributional notation, and in addition to the terms already defined in (1.31) and (1.33), we specify the likelihood term (1.34) as a Gaussian distribution $p(y_j | \theta_{z_j}^*) = \mathcal{N}(y_j | \mu_{z_j}, \Sigma_{z_j})$ and the G_0 prior on cluster specific parameters $\theta_k^* = (\mu_k, \Sigma_k)$ as a Normal-inverse-Wishart distribution parameterized by

$\rho_k = (m_k, \lambda_k, \Psi_k, \nu_k)$ with a pdf

$$p(\theta_k^*; \rho_k) = \mathcal{NIW}(\mu_k, \Sigma_k; \rho_k) = \mathcal{N}(\mu_k; m_k, \lambda_k^{-1}\Sigma_k)\, \mathcal{IW}(\Sigma_k; \Psi_k, \nu_k).$$

In the above notation, we consider as hyperparameters the set $\boldsymbol{\phi} = (s_1, s_2, a, \beta, (\rho_k)_{k=1}^{\infty})$ while $\boldsymbol{\Theta} = (\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*)$.

In most variational approximations, the posteriors are approximated in a factorized form (mean-field approximation). In particular, the intractable MRF posterior on $\boldsymbol{z}$ is approximated as $q_{\mathbf{z}}(\mathbf{z})$ that factorizes so as to handle intractability due to spatial dependencies, namely

$$q_{\mathbf{z}}(\mathbf{z}) = \prod_{j=1}^{n} q_{z_j}(z_j).$$

Then, the infinite state space for each z_i is dealt with by choosing a truncation of the state space to a maximum label K [58]. In practice, this consists of assuming that the variational distributions q_{z_j}, for $j = 1, \ldots, n$, satisfy $q_{z_j}(k) = 0$ for $k > K$ and that the variational distribution on $\boldsymbol{\tau}$ also factorizes as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k)$, with the additional condition that $\tau_K = 1$. Thus, the truncated variational posterior of parameters $\boldsymbol{\Theta}$ is given by

$$q_{\Theta}(\boldsymbol{\Theta}) = q_{\alpha,\sigma}(\alpha, \sigma) \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k) \prod_{k=1}^{K} q_{\theta_k^*}(\theta_k^*). \tag{1.48}$$

These forms of q_z and $q_{\boldsymbol{\Theta}}$ lead to four VB-E steps and three VB-M steps summarized below (refer to [2] for details). Set the initial value of $\boldsymbol{\phi}$ to $\boldsymbol{\phi}^{(0)}$. Then, repeat iteratively the following steps. The iteration index is omitted in the update formulas for simplicity.

VB-E-$\boldsymbol{\tau}$ step

The VB-E-$\boldsymbol{\tau}$ step corresponds to a variational approximation in the exponential family case and results in a posterior from the same family as the prior. It comes for $k = 1, \ldots, K$,

$$q_{\tau_k}(\tau_k) = \mathcal{B}(\tau_k; \hat{\gamma}_{k,1}, \hat{\gamma}_{k,2}) \tag{1.49}$$

with

$$\hat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_\sigma}[\sigma] + \bar{n}_k, \quad \hat{\gamma}_{k,2} = \mathbb{E}_{q_\alpha}[\alpha] + k\mathbb{E}_{q_\sigma}[\sigma] + \sum_{\ell=k+1}^{K} \bar{n}_\ell, \tag{1.50}$$

where

$$\text{for} \quad k = 1, \ldots, K, \quad \bar{n}_k = \sum_{j=1}^{n} q_{z_j}(k) \tag{1.51}$$

corresponds to the weight of cluster k.

VB-E-(α, σ) step

The (α, σ) variational posterior is more complex but has a simple gamma form in the DP ($\sigma = 0$) case. More specifically, we need to compute

$$\hat{s}_1 = s_1 + K - 1, \text{ and } \hat{s}_2 = s_2 - \sum_{k=1}^{K-1} \psi(\hat{\gamma}_{k,2}) - \psi(\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}) \tag{1.52}$$

where $\psi(\cdot)$ is the digamma function defined by $\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z} \log \Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$. When $\sigma = 0$ then q_α is a gamma distribution $\mathcal{G}(\hat{s}_1, \hat{s}_2)$ and $\mathbb{E}_{q_\alpha}[\alpha] = \frac{\hat{s}_1}{\hat{s}_2}$. Otherwise (PY case), $q_{\alpha,\sigma}$ is only identified up to a normalizing constant but the required $\mathbb{E}_{q_\alpha}[\alpha]$ and $\mathbb{E}_{q_\sigma}[\sigma]$ can be computed by importance sampling. See the appendix of [2] for details.

VB-E-Z step

Due to the mean field approximation and the truncation, this step consists in computing, for all $j = 1, \ldots, n$ and for $k \leq K$,

$$q_{z_j}(k) = \frac{\tilde{q}_j(k)}{\sum_{\ell=1}^{K} \tilde{q}_j(\ell)}, \tag{1.53}$$

where $\log \tilde{q}_j(k)$ is defined by

$$\begin{aligned} &-\frac{1}{2}\left\{\log\left|\frac{\hat{\Psi}_k}{2}\right| - \sum_{i=1}^{d} \psi\left(\frac{\hat{v}_k + (1-i)}{2}\right) + \hat{v}_k (y_j - \hat{m}_k)^T \hat{\Psi}_k^{-1} (y_j - \hat{m}_k) + \frac{d}{\hat{\lambda}_k}\right\} + \\ &\psi(\hat{\gamma}_{k,1}) - \psi(\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}) + \sum_{l=1}^{k-1} \psi(\hat{\gamma}_{l,2}) - \psi(\hat{\gamma}_{l,1} + \hat{\gamma}_{l,2}) + \beta \sum_{i \in \mathcal{N}_j} q_{z_i}(k), \end{aligned} \tag{1.54}$$

where in the last sum, $\mathcal{N}_j$ represents the neighbours of j. In the above formula, symbols $(\hat{m}_k, \hat{\lambda}_k, \hat{\mathbf{\Psi}}_k, \hat{v}_k)$ are the variational hyperparameters for $q_{\theta_k^*}$ more specifically defined in the following step and d is the dimension of the data. The advantage of Eq. (1.53) is that it provides assignment probabilities $q_{z_i}(k)$ and does not require intermediate commitments to hard assignments of the z_j's. The hard assignments can be postponed to the end if desired to get a segmentation through the following maximum a posteriori (MAP) estimation:

$$\hat{z}_j = \underset{k \in \{1,\ldots,K\}}{\arg\max}\, q_{z_j}(k). \tag{1.55}$$

VB-E-θ^* step

This step is divided into K parts where the computation is similar to that in standard Bayesian finite mixtures with a choice of conjugate prior, here for Gaussian distributions. Hence, for each $k \leq K$, the variational posterior is a Normal-inverse-Wishart distribution defined as

$$q_{\theta_k^*}(\mu_k, \Sigma_k) = \mathcal{NIW}(\mu_k, \Sigma_k; \hat{m}_k, \hat{\lambda}_k, \hat{\Psi}_k, \hat{v}_k), \tag{1.56}$$

where the hyperparameters are updated as follows (see for instance [238])

$$\begin{aligned}\hat{\lambda}_k &= \lambda_k + \bar{n}_k, \quad \hat{\nu}_k = \nu_k + \bar{n}_k, \\ \hat{\Psi}_k &= \Psi_k + S_k + \frac{\lambda_k \bar{n}_k}{\lambda_k + \bar{n}_k}(m_k - \bar{\mu}_k)(m_k - \bar{\mu}_k)^T, \\ \hat{m}_k &= \frac{\lambda_k m_k + \bar{n}_k \bar{\mu}_k}{\lambda_k + \bar{n}_k} = \frac{\lambda_k m_k + \bar{n}_k \bar{\mu}_k}{\hat{\lambda}_k},\end{aligned} \tag{1.57}$$

with $\bar{n}_k$ defined in (1.51) and

$$\begin{aligned}\bar{\mu}_k &= \frac{1}{\bar{n}_k} \sum_{j=1}^{n} q_{z_j}(k) y_j, \\ S_k &= \sum_{j=1}^{n} q_{z_j}(k)(y_j - \bar{\mu}_k)(y_j - \bar{\mu}_k)^T.\end{aligned} \tag{1.58}$$

VB-M steps

The maximization step consists of updating the hyperparameters $\boldsymbol{\phi} = (\beta, s_1, s_2, a, \boldsymbol{\rho})$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$, by maximizing the free energy, if they are not set heuristically:

$$\boldsymbol{\phi}^{(r)} = \arg\max_{\boldsymbol{\phi}} \; \mathbb{E}_{q_Z^{(r)} q_\tau^{(r)} q_{\alpha,\sigma}^{(r)} q_{\theta^*}^{(r)}} [\log p(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*; \boldsymbol{\phi})] \; . \tag{1.59}$$

The VB-M-step can therefore be divided into 3 independent sub-steps as listed below. From the conditional independence of $(s_1, s_2, a, \boldsymbol{\rho})$ and $(\boldsymbol{Y}, \boldsymbol{Z})$ given $(\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*)$, the VB-M-step writes as in (1.46) so that the solutions for the VB-M-(s_1, s_2) (in the DP case) and VB-M-$\boldsymbol{\rho}$ steps are straightforward. Only the β step and the M-(s_1, s_2, a) step (in the PY case) are more involved.

VB-M-β: The maximization of (1.59) with respect to β leads to

$$\beta^{(r)} = \arg\max_{\beta} \mathbb{E}_{q_Z^{(r)} q_\tau^{(r)}} [\log p(\boldsymbol{Z} | \boldsymbol{\tau}; \beta)] \; . \tag{1.60}$$

This step does not admit a closed-form solution but can be solved numerically.

VB-M-(s_1, s_2, a): This step is straightforward in the DP case ($\sigma = 0$). It can be expressed easily using the fact that both the prior and the variational posterior are Gamma distributions, and using the cross-entropy properties,

$$(s_1, s_2)^{(r)} = \arg\max_{(s_1, s_2)} \mathbb{E}_{q_\alpha^{(r)}} [\log p(\alpha; s_1, s_2)] \; = \; (\hat{s}_1^{(r)}, \hat{s}_2^{(r)}) \tag{1.61}$$

where $(\hat{s}_1^{(r)}, \hat{s}_2^{(r)})$ is given in (1.52). In the more general PY case, we can solve this step numerically using also importance sampling.

VB-M-ρ: This step divides into K sub-steps that involve again cross-entropies,

$$\rho_k^{(r)} = \arg\max_{\rho} \mathbb{E}_{q_{\theta_k^*}^{(r)}}[\log p(\theta_k^*; \rho_k)] \;=\; \hat{\rho}_k^{(r)} \tag{1.62}$$

where $\hat{\rho}_k^{(r)} = (\hat{\lambda}_k^{(r)}, \hat{\nu}_k^{(r)}, \hat{\Psi}_k^{(r)}, \hat{m}_k^{(r)})$ is given in Eq. (1.57).

1.2.5 Application to image segmentation

To validate the proposed approach, we consider its application to unsupervised image segmentation as a spatial clustering task. Image segmentation consists of partitioning a digital image into distinct regions that contain pixels with similar properties. Extensive research work has been done in this field using various clustering techniques. In practice, to be meaningful for image analysis and interpretation, the segmented regions should closely relate to depicted objects or features of interest. A number of tasks in image analysis often depends on the reliability of preliminary segments, but an accurate partitioning of an image is still quite challenging.

Feature extraction for image segmentation

The color and texture features in a natural image are often very complex. For our experiments, we mainly focus on two special types of features based on the HSV (Hue, Saturation, Value) color space and the maximum response (MR) filter bank. The HSV color space is often used in natural image analysis because it corresponds better to how people experience color than the RGB color space does. Regarding the texture information, we shall consider the MR8 filter bank [327], which consists of 38 filters but only 8 filter responses. More precisely, the MR8 filter bank contains filters at multiple orientations but their outputs are compressed by recording only the maximum filter response across all orientations. This achieves rotation invariance. Furthermore, the images are presegmented into superpixels that group pixels similar in color and other low-level properties [1]. In this respect, superpixels are regarded as more natural entities that allow reducing the number of observations drastically for running clustering algorithms. In all our experiments, each image is presegmented into approximately 1 000 superpixels using the SLIC algorithm proposed in [1]. Finally, we compute the feature vectors at superpixel level, *i.e.*, the average of features on the centroid of each superpixel.

Berkeley Segmentation Data Set

To quantify the performance of our segmentation algorithm, numerical experiments were conducted on a subset of images selected from the Berkeley Segmentation Data Set 500 (BSDS500) already studied by [33, 79], which provides multiple human annotated segments as many ground truths for each image. The considered subset consists of 154 images as listed in Tables 1 and 2 in [79].

In the literature, a standard measure for comparing a test segmentation to another is the rand index (RI) [279]. The RI is one when two segmentations are exactly the

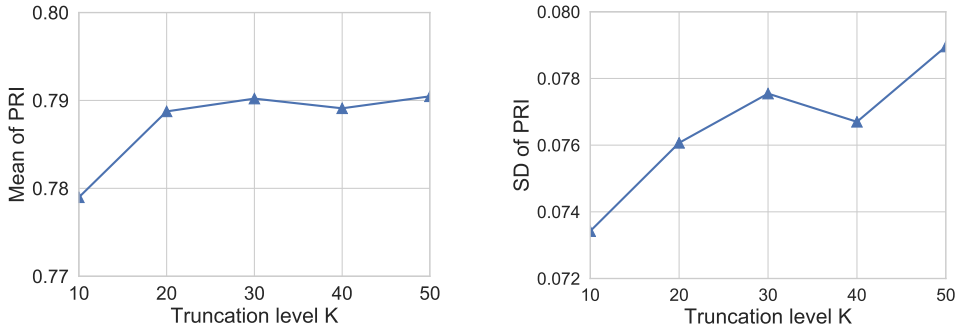


FIGURE 1.7: PY-MRF mixture model: Mean and standard deviation of the PRI score over the considered subset of the BSDS500 data set as a function of the truncation level K.

same. However, when having for one image a set of ground truths which do not completely agree, the probabilistic rand index (PRI) [325] is preferable. Given a set of ground truths $\mathcal{S} = (S_1, \ldots, S_T)$, the PRI is defined as follows:

$$\mathrm{PRI}(S_{\text{test}}, \mathcal{S}) = \frac{2}{n(n-1)} \sum_{i<j} [c_{ij} p_{ij} + (1 - c_{ij})(1 - p_{ij})] \tag{1.63}$$

where $c_{ij} = 1$ if pixels i and j belong to the same segment in S_{test} and $c_{ij} = 0$ otherwise, n is the number of image pixels and p_{ij} is the probability of two pixels i and j having the same label, *i.e.*, the fraction of all available ground truths in $\mathcal{S}$ where pixels i and j belong to the same segment. In fact, it can be shown that Eq. (1.63) is simply the mean of the RI computed between each pair (S_{test}, S_k), namely $\frac{1}{T} \sum_{k=1}^{T} \mathrm{RI}(S_{\text{test}}, S_k)$. By definition, the PRI always takes values in $[0, 1]$, where 0 means that S_{test} and $(S_1, \ldots, S_T)$ have no similarities and 1 means all segments are identical. The larger the PRI, the better. In practice, PRI values are often reported as percentages in $[0, 100]$.

Our approach has been tested on the considered subset of the BSDS500 and the summary statistics of the PRI score are shown in Figure 1.7 as a function of the truncation level K for the PY-MRF case. Similar results were observed for the DP. It appears that for $K \geq 30$, the global performance does not change much and is satisfying with respect to existing results in the literature. We compared our best results with those reported in [79]. Table 1.2 shows that our approach outperforms the existing results. The improvement in PRI may appear overall small but it can be assessed by visualizing original images and their segmentations. We show in Figure 1.8 segmentation results for four images. The main differences between the non spatial PY and PY-MRF mixture models can be visualized for the first image in the ground and water which are segmented in the latter case into a smaller number of regions whose shapes are in addition smoother. This is typical of more spatial interaction in the clustering process. Similarly, the same phenomenon is also visible in the peak part of the second image, in the sky and grass parts of the third image and in the plant parts of the fourth image.

We also examined, for the PY-MRF mixture model with $K = 50$, the values of the expected α, σ and β for each of the 154 segmented images presented in Figure 1.9

	Proposed model	Results given in [79]			
PRI (%)	**PY-MRF**	DPM	iHMRF	MRF-PYP	Graph Cuts
Mean	**79.05**	74.15	75.50	76.49	76.10
Median	**80.62**	75.49	76.89	78.08	77.59
St. Dev.	**7.9**	8.4	8.2	7.9	8.3

TABLE 1.2: Performance comparison: Summary statistics of the PRI score over the 154 images from BSDS500 studied by [80] for our PY-MRF mixture model and the approaches tested in [80, 79].

as scatter plots (one point per image). Recall from Section 1.2.4 that α and σ are elements of the parameters $\boldsymbol{\Theta}$ while β is considered as a hyperparameter from $\boldsymbol{\phi}$. Figure 1.9 shows also the correlations (across the 154 images) between the expected values of α, σ and β. It appears that the estimated σ values are most of the time smaller than 0.5 and sometimes closer to 0 with some anti-correlation with respect to α values. In contrast, β values appear quite independent from α or σ.

In terms of pure PRI performance, the BSDS500 data set is not an easy example because the ground truth segmentations are labeled manually by humans and are sometimes quite subjective and inconsistent across users. However, this example allows comparison of methods and visualization. Two interesting findings are that the choice of K does not seem to be too sensitive as soon as K is large enough, and there seems to be some correlation between α and σ while β is rather independent of the latest. Further analysis would be needed to confirm these properties but in practice, they could be used to guide the segmentations into more or less spatially smooth versions without risking to eliminate too small segments.

1.2.6 Discussion

In this paper, we proposed a general scheme to build BNP priors that can model dependencies through the addition of a Markov random field term. In contrast to other existing attempts that reduce to spatially constrained standard BNP priors such as [79, 80], our proposal leads to proper spatial priors. Our construction is based on the stick-breaking representation and was illustrated starting from the Dirichlet and Pitman–Yor processes, although this approach could be extended to other forms of BNP priors admitting a stick-breaking representation such as Gibbs-type priors. The stick-breaking representation was further exploited to derive clustering properties of the model and to provide a variational inference algorithm. In addition to the usual BNP parameters, an estimation of the Markov interaction parameter β was proposed. The variational approximation chosen was based on a standard truncation but it would be interesting to investigate other approximations, *e.g.* [337]. Also the variational algorithm is greatly simplified for standard stick-breaking representations (*e.g.* DP and PY) with independent weight variables. Nevertheless, it would be interesting to investigate more general stick-breaking representations possibly using some MCMC counterpart for estimation.

FIGURE 1.8: Segmentation results for four images from the BSDS500 data set. From left to right, columns show respectively, the original images, the segmentation results with the PY and PY-MRF mixture models.

The approach was illustrated on a challenging unsupervised image segmentation task with good results with respect to the literature, but the proposed scheme is quite flexible and can be used in more general settings including community detection or disease mapping in epidemiology.

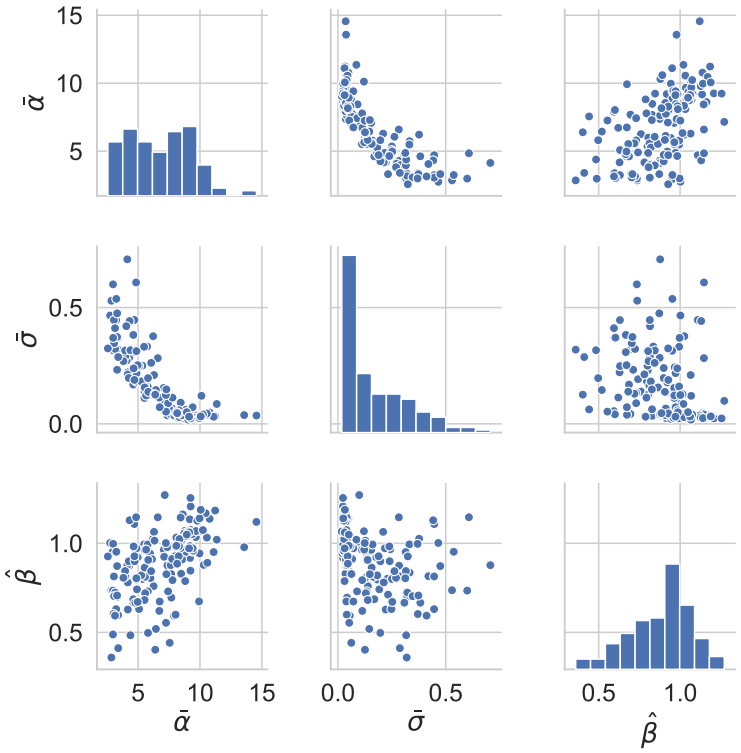


FIGURE 1.9: Estimated parameter values $(\bar{\alpha}, \bar{\sigma})$ obtained from **VB-E** steps and $\hat{\beta}$ obtained from a **VB-M** step using the PY-MRF model with truncation level $K = 50$, on the 154 images from the Berkeley benchmark.

[C3] G. Kon Kam King, J. Arbel, and I. Prünster. *Bayesian Statistics in Action*, chapter A Bayesian nonparametric approach to ecological risk assessment, pages 151–159. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017

[S3] J. Arbel, G. Kon Kam King, A. Lijoi, L. E. Nieto-Barajas, and I. Prünster. BNPdensity: Bayesian nonparametric mixture modeling in R. *Submitted*, 2019

1.3 Ecotoxicological application to species sensitivity distribution modeling

1.3.1 Introduction

Assessing the response of a community of species to an environmental stress is critical for ecological risk assessment. Methods for this purpose vary in levels of complexity and realism. Species Sensitivity Distribution (SSD) represents an intermediate tier, more refined than rudimentary assessment factors [272] but practical enough for routine use by environmental managers and regulators in most developed countries (Australia, Canada, China, EU, South Africa, USA...). The SSD approach is intended to provide, for a given contaminant, a description of the tolerance of all species possibly exposed using information collected on a sample of those species. This information consists of Critical Effect Concentrations (CECs), a concentration specific to a species which marks a limit over which the species suffers a critical level of effect. This is for instance the concentration at which 50% of the tested organisms died (Lethal Concentration 50%), or the concentration which inhibited growth or reproduction by 50% compared to the control experiment (Lethal Concentration 50%, LC_{50}). Each CEC is the summary of long and costly bioassay experiments for a single species, so they are rarely available in large number. Typical sample sizes range from 10 to 15 [108].

To describe the tolerance of all species to be protected, the distribution of the CECs is then estimated from the sample. In practice, a parametric distributional assumption is often adopted [121] where the CECs are assumed to follow a log-normal, log-logistic, triangular or BurrIII distributions.

Once the response of the community is characterised by the distribution, the goal of risk assessment is to define a safe concentration protecting all or most of the species. In the case of distributions without a lower threshold strictly above 0, a cut-off value is often chosen as the safe concentration. Typically, this is the Hazardous Concentration for 5% of the Species HC_5, which is the 5th percentile of the distribution. Reasonings behind this choice include: that the lowest bound of the confidence interval around the 5th percentile will be used instead of the estimate, that a safety factor will be subsequently applied to that value and that ecosystems have a certain resilience to perturbations.

The lack of justification for the choice of any given parametric distribution has sparked several research directions. Some authors [341, 153] have sought to find the best parametric distribution by model comparison using goodness-of-fit measures. The

general understanding is that no single distribution seems to provide a superior fit and that the answer is dataset dependent [121]. Therefore, the log-normal distribution has become the customary choice, notably because it readily provides confidence intervals on the HC_5, and because model comparison and goodness of fit tests have relatively low power on small datasets, precluding the emergence of a definite answer to the question. Another research direction consisted in seeking to avoid any reference to a distribution, using so-called nonparametric or distribution-free approaches. Those efforts included using the empirical distribution function [309, 180], bootstrap resampling [336] or nonparametric kernel density estimation [338]. All these approaches have in common that they require large sample sizes to be effectively applicable. Finally, authors have considered the possibility that the distribution of the CECs might not be a single distribution but rather a mixture of distributions [343], datasets being an assemblage of several log-normally distributed subgroups [186]. This is more realistic from an ecological point of view because several factors influence the tolerance of a species to a contaminant such as the taxonomic group or the mode of action, and contaminant such as pesticides might even target specific species groups. Therefore, there is strong evidence in favour of the presence of groups of CECs, although the CECs within a group might remain log-normally distributed.

Ignorance of the group structure is a strong motivation for a nonparametric approach. However, the method must remain applicable to small datasets, which suggests trying to improve on the existing frequentist nonparametric methods. Bayesian nonparametric mixture models offer an interesting solution for both large and small datasets, because the complexity of the mixture model adapts to the size of the dataset. It offers a good compromise between a simplistic one-component parametric model and a kernel density method which in a certain sense lacks flexibility and might cause overfitting. Moreover, the low amount of information available in small datasets to estimate the groups parameters can be complemented via the prior, as some a priori degree of information is generally available from other species or contaminants [36].

The rest of the section is organised as follows. In Section 1.3.2 we present the BNP model and existing frequentist models for SSD and explain how to obtain a density estimate. Then in Section 1.3.3 we compare the different methods on a real dataset, illustrating the benefits of the BNP SSD.

1.3.2 Models for SSD

Given that concentrations vary on a wide range, it is common practice to work on log-transformed concentrations. Consider a sample of n log-concentrations denoted by $\boldsymbol{X} = (X_1, \ldots, X_n)$. We propose to carry out density estimation for the SSD based on sample $\boldsymbol{X}$ by use of nonparametric mixtures. Bayesian nonparametric mixtures were introduced in [217] with DPM. Generalizations of the DPM correspond to allowing the mixing distribution to be any discrete nonparametric prior. A large class of such prior distributions is obtained by normalizing increasing additive processes [**?**]. The normalization step, under suitable conditions, gives rise to so-called NRMI as defined by [**?**], see also [43] for a recent review. An NRMI mixture model is

defined hierarchically as:

$$\begin{aligned} X_i|\mu_i,\sigma &\overset{\text{ind}}{\sim} k(\cdot|\mu_i,\sigma), \quad \mu_i|\tilde{P} \overset{\text{i.i.d.}}{\sim} \tilde{P}, \quad i=1,\dots,n, \\ \tilde{P} &\sim \text{NRMI}, \quad \sigma \sim \text{Ga}(a_\sigma, b_\sigma). \end{aligned} \tag{1.64}$$

where k is a kernel, which we assume parametrized by some $\theta = (\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_+$, and $\tilde{P}$ is a random probability on $\mathbb{R}$ whose distribution is an NRMI. In our model, all clusters have a common variance. This is easier to fit on a small dataset, because information about the variance is pooled across clusters. As described in the Introduction, concentrations are commonly fitted with a log-normal distribution. Our aim is to move from this parametric model to the nonparametric one in (1.64). In order to allow comparisons to be made, we stick to the normal specification for k on the log-concentrations $\boldsymbol{X}$ by letting: $k(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma)$. Under this framework, density estimation is carried out by evaluating the posterior predictive density along the lines of [43]:

$$\hat{f}(x|\tilde{P},\boldsymbol{X}) = \iint k(x|\mu,\sigma)\mathrm{d}\pi(\sigma)\mathrm{d}\tilde{P}(\mu) \tag{1.65}$$

for any x in $\mathbb{R}$, where π denotes the posterior distribution of σ.

To specify the prior, we choose as mixing random measure the normalized stable process [190] with:

i a stability parameter $\gamma = 0.4$, which controls the flatness of the prior on the number of clusters. The parameter γ can take values in $(0,1)$. Taking the limit $\gamma \to 0$ reduces the model to a Dirichlet process, larger values of γ lead to less informative priors on the number of clusters. The parameter γ was chosen as a good compromise between model flexibility and numerical stability. The total mass parameter is, without loss of generality, set equal to 1.

ii a base measure (which corresponds to the mean of the random probability measure) $P_0(\,\cdot\,) = \text{N}(\,\cdot\,|\,\varphi_1,\varphi_2)$ with mean φ_1 and standard deviation φ_2, hyperparameters fixed a priori to specify a certain knowledge in the degree of smoothness

iii a common variance for all the clusters with a vaguely informative prior distribution $Ga(0.5,0.5)$.

Recent years have witnessed the appearance of a wealth of softwares dedicated to implement Bayesian nonparametric models and sample from their posterior. To cite a few, the R package DPpackage, is a rather comprehensive bundle of functions for Bayesian nonparametric models, while Bayesian Regression [183] is a software for Bayesian nonparametric regression. For posterior sampling, we use the R package BNPdensity and the function MixNRMI1 which implements BNP density models under a general specification of normalized random measures based on the generalised gamma processes [see 43]. The package is available from the Comprehensive R Archive Network (CRAN).

To illustrate the interest of the Bayesian nonparametric SSD, we compare our proposed BNP model to two commonly used frequentist models: the normal distribution [4] and the nonparametric KDE recently proposed by [338]. For both frequentist

approaches, the data is assumed to be iid. Density estimates take on respectively the following form ($\hat{\mu}$ and $\hat{\sigma}$ are MLE)

$$\hat{f}_{\mathcal{N}}(x) = \mathcal{N}(x \mid \hat{\mu}, \hat{\sigma}) \quad \text{and} \quad \hat{f}_{KDE}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(x \mid X_i, 1.06\hat{\sigma}n^{-\frac{1}{5}}). \tag{1.66}$$

Model comparison and cross-validation

For the purpose of comparing the predictive performance of the model, we resort to Leave-One-Out (LOO) cross-validation. We compute the LOO for each of the methods as $\forall i, \text{LOO}_i = \hat{f}(X_i \mid \boldsymbol{X}_{-i})$ where $\hat{f}(x \mid \boldsymbol{X}_{-i})$ is the density for one of the three methods estimated from $\boldsymbol{X}$ with X_i left out. The LOO for the BNP model correspond to the conditional predictive ordinate (CPO) statistics which are commonly used in applications, see [125]. A CPO statistic is defined for each log-concentration X_i as follows:

$$\text{CPO}_i = \hat{f}(X_i|\boldsymbol{X}_{-i}) = \int k(X_i|\theta)\mathrm{d}\pi(\theta|\boldsymbol{X}_{-i}) \tag{1.67}$$

where $\boldsymbol{X}_{-i}$ denotes the whole sample $\boldsymbol{X}$ but X_i, $\mathrm{d}\pi(\theta|\boldsymbol{X}_{-i})$ is the posterior distribution associated to $\boldsymbol{X}_{-i}$ and $\hat{f}$ is the (cross-validated) posterior predictive distribution of Equation (1.65). As shown by [43], CPO can be easily approximated by Monte Carlo as

$$\widehat{\text{CPO}_i} = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{k(X_i|\theta^{(t)})}\right)^{-1} \tag{1.68}$$

where $\{\theta^{(t)}, t = 1, 2, \ldots, T\}$ is an MCMC sample from the posterior distribution.

Quantile estimation and HC_5

The quantity of interest for ecological risk assessment is the HC_5, which corresponds to the 5th percentile of the SSD distribution. We choose as an estimator the median of the posterior distribution of the 5th percentile, while the 95% credible bands are formed by the 2.5% and 97.5% quantiles of the posterior distribution of the 5th percentile. The 5th percentile of the KDE is obtained by numerical inversion of the cumulative distribution function, and the confidence intervals using nonparametric bootstrap. The 5th percentile of the normal SSD and its confidence intervals are obtained following the classical method of [4].

1.3.3 Application to real data

We applied this model to a selection of contaminants extracted from a large database collected by the National Institute for Public Health and the Environment (RIVM). We only considered non censored data, left or right censored data were discarded, while interval censored data were replaced by the centre of the interval. Using a continuous distribution for the CEC implies that the model does not support ties (or, in other words, observing ties has zero probability). However, ties may appear in the dataset due to the rounding of concentrations. Hence, we used a small jittering of the data.

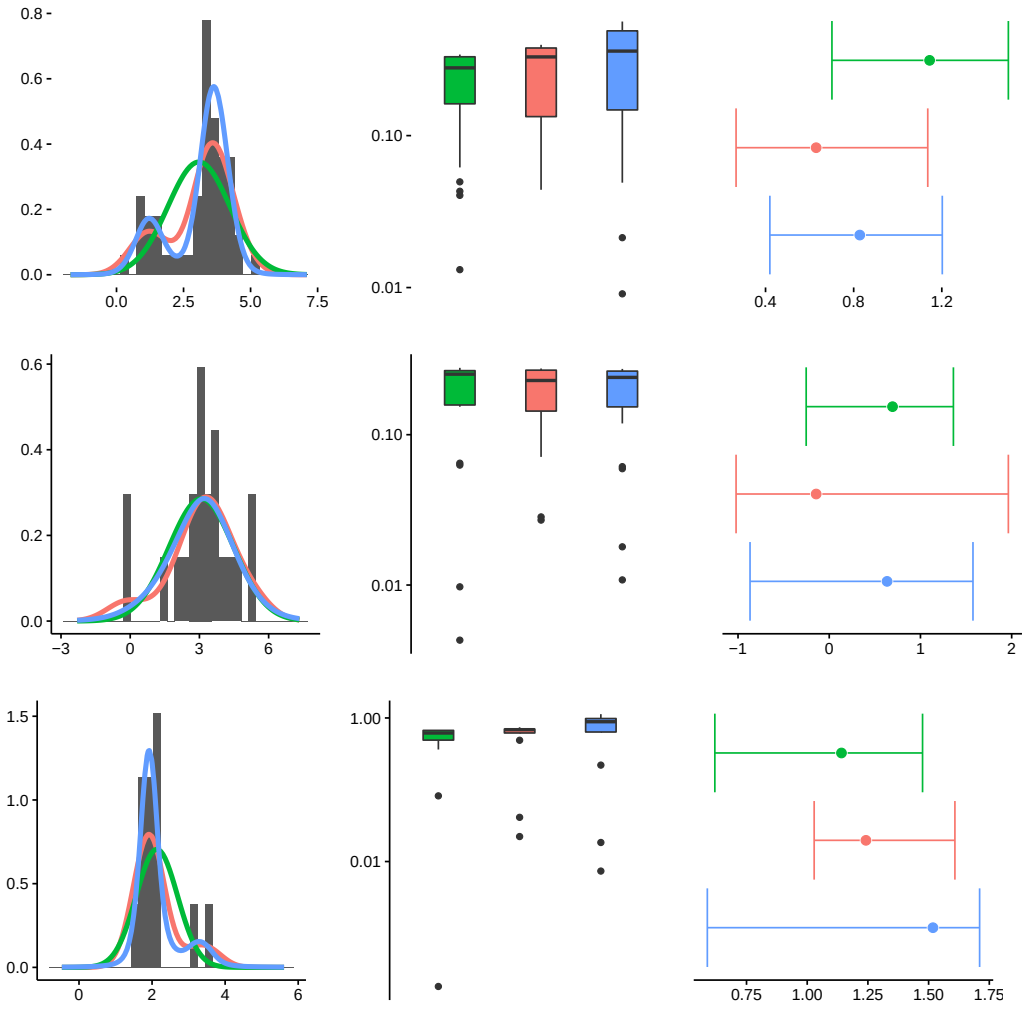
We selected three example datasets which feature three typical sample sizes: a relatively large `carbaryl` dataset (CAS: 63-25-2, insecticide, 55 species), a medium-sized `temephos` dataset (CAS: 3383-96-8, mosquito larvicide, 21 species), and a small `captan` dataset (CAS: 133-06-2, fungicide, 13 species). Datasets for new contaminants are always small, the minimum requirement set by the European Chemical Agency being 10 species. The datasets can be visualised on the histograms of Figure 1.10 (left panel).

These datasets illustrate different features of the three approaches: when there is a clear multimodality in the data, the BNP SSD is more flexible than the fixed bandwidth KDE SSD (Figure 1.10, `carbaryl` and `captan`). When the data do not exhibit strong multimodality, as for `temephos`, the BNP reduces to the normal SSD model, whereas the KDE remains by construction a mixture of many normal components.

One might think to increase the flexibility of the KDE by simply decreasing the bandwidth. However, that would also decrease the robustness of the method. On the middle panel of Figure 1.10, the LOO give an indication of the robustness to over-fitting of the three methods. For `carbaryl` and `captan`, they show that the superior flexibility of the BNP SSD compared to the KDE SSD does not come at the expense of robustness, because the median CPO of the BNP SSD is higher than the other two. In the case of `temephos`, the median LOO likelihood estimate of the normal model is very similar to the median CPO for the BNP SSD, sign that there is little over-fitting. This generally illustrates the fact that model complexity in a BNP model scales with the amount and structure of the data. On the right panel of Figure 1.10, the credible intervals of the HC_5 for the BNP SSD are generally larger than the confidence interval of the normal SSD, which is coherent with the model uncertainty of the non-parametric approach.

1.3.4 Discussion

The BNP SSD seems to perform well when the dataset deviates from a normal distribution. Its great flexibility is an asset to describe the variability of the data, while it does not seem prone to over-fitting. It can be thought of as an intermediate model between the normal SSD with a single component on the one hand, and the KDE which counts as many components as there are species on the other hand. We chose to base the BNP SSD on NRMI rather than on the more common Dirichlet Process, because it is more robust in case of misspecification of the number of clusters [211, 43]. The BNP SSD provides several benefits for risk assessment: it is an effective and robust standard model which adapts to many datasets. Moreover, it readily provides credible intervals. While it is always possible to obtain confidence intervals for a frequentist method using bootstrap, it can be difficult to stabilise the interval for small datasets even with a large number of bootstrap samples. As such, the BNP SSD represents a safe tool to remove one of the arbitrary parametric assumptions of SSD [121].



(A) Density (B) LOO and CPO (C) log HC_5

FIGURE 1.10: The top panel represents the large-size `carbaryl` dataset, the middle panel represents the medium-sized `temephos` dataset, the bottom panel represents the small-sized `captan` dataset. The Normal is in green, the KDE in red and the BNP in blue. Concentrations are log transformed.
Left: Histogram and density estimates.
Centre: Boxplot for the LOO (for Normal and KDE) and the CPO (for BNP) on logarithmic scale. The horizontal line corresponds to the median. The box hinges extend to the inner quartiles. The whiskers extend to cover points up to one and a half times the inter-quartile distance away from the hinges. For both frequentist methods, the n LOO are obtained by fitting the model n times, while an analytical expression is available for the BNP method (1.68).
Right: log HC_5 and associated confidence/credible intervals (for Normal, KDE and BNP).

Chapter 2

Approximate Bayesian Inference

Contents

This chapter is based on the following papers and preprints

Section 2.1

[A6] J. Arbel, P. De Blasi, and I. Prünster. Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*, 14(3):753–771, 2019

Section 2.2

[A8] J. Arbel and I. Prünster. A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17, 2017

[P9] J. Arbel and I. Prünster. Truncation error of a superposed gamma process in a decreasing order representation. *NeurIPS Advances in Approximate Bayesian Inference workshop*, 2016

[C4] J. Arbel and I. Prünster. *Bayesian Statistics in Action*, chapter On the truncation error of a superposed gamma process, pages 11–19. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017

Section 2.3

[A1] J. Arbel and S. Favaro. Approximating predictive probabilities of Gibbs-type priors. *Sankhyā, forthcoming*, 2019

[A9] J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858, 2017

Section 2.4

[S4] H. D. Nguyen, J. Arbel, H. Lü, and F. Forbes. Approximate Bayesian computation via the energy statistic. *Submitted*, 2019

[A6] J. Arbel, P. De Blasi, and I. Prünster. Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*, 14(3):753–771, 2019

2.1 Truncation-based approximations: Pitman–Yor

2.1.1 Introduction

The Pitman–Yor process defines a rich and flexible class of random probability measures used as prior distribution in Bayesian nonparametric inference. It originates from the work of [261], further investigated in [265, 264], and its use in nonparametric inference was initiated by [164]. Thanks to its analytical tractability and flexibility, it has found applications in a variety of inferential problems which include species sampling [210, 113, 240], survival analysis and gene networks [175, 246], linguistics and image segmentation [318, 308], curve estimation [71] and time-series and econometrics [74, 44]. The Pitman–Yor process is a discrete probability measure

$$P(\mathrm{d}x) = \sum_{i\geq 1} p_i \delta_{\xi_i}(\mathrm{d}x) \tag{2.1}$$

where $(\xi_i)_{i\geq 1}$ are iid random variables with common distribution P_0 on a Polish space $\mathcal{X}$, and $(p_i)_{i\geq 1}$ are random frequencies, i.e. $p_i \geq 0$ and $\sum_{i\geq 1} p_i = 1$, independent of $(\xi_i)_{i\geq 1}$. The distribution of the frequencies of the Pitman–Yor process is known in the literature as the two-parameter Poisson–Dirichlet distribution. Its distinctive property is that the frequencies in *size-biased order*, that is the random arrangement in the order of appearance in a simple random sampling without replacement, admit the *stick-breaking representation*, or residual allocation model,

$$p_i \stackrel{d}{=} V_i \prod_{j=1}^{i-1}(1-V_j), \quad V_j \stackrel{\text{ind}}{\sim} \text{beta}(1-\alpha, \theta + j\alpha) \tag{2.2}$$

for $0 \leq \alpha < 1$ and $\theta > -\alpha$, see [270]. By setting $\alpha = 0$ one recovers the Dirichlet process of [117]. Representation (2.2) turns out very useful in devising finite support approximation to the Pitman–Yor process obtained by truncating the summation in (2.1). A general method consists in setting the truncation level n by replacing p_{n+1} with $1-(p_1+\cdots+p_n)$ in (2.1). The key quantity is the *truncation error* of the infinite summation (2.1),

$$R_n = \sum_{i>n} p_i = \prod_{j\leq n}(1-V_j), \tag{2.3}$$

since the resulting truncated process, say $P_n(\cdot)$, will be close to $P(\cdot)$ according to $|P(A) - P_n(A)| \leq R_n$ for any measurable $A \subset \mathcal{X}$. It is then important to study the distribution of the truncation error R_n as n gets large in order to control the approximation error. [164] proposes to determine the truncation level based on the moments of R_n. Cf. also [166, 126]. In this paper we propose and investigate a random truncation by setting n such that R_n is smaller than a predetermined value

$\epsilon \in (0,1)$ with probability one. Specifically, we define

$$\tau(\epsilon) = \min\{n \geq 1 : \ R_n < \epsilon\} \tag{2.4}$$

as the stopping time of the multiplicative process $(R_n)_{n\geq 1}$ and, following Section 4.3.3 of [131], we call *ϵ-Pitman–Yor (ϵ-PY) process* the Pitman–Yor process truncated at $n = \tau(\epsilon)$, namely

$$P_\epsilon(\mathrm{d}x) = \sum_{i=1}^{\tau(\epsilon)} p_i \delta_{\xi_i}(\mathrm{d}x) + R_{\tau(\epsilon)}\delta_{\xi_0}(\mathrm{d}x), \tag{2.5}$$

where ξ_0 has distribution P_0, independent of the sequences $(p_i)_{i\geq 1}$ and $(\xi_i)_{i\geq 1}$. By construction, P_ϵ is the finite stick-breaking approximation to P with the smallest number of support points given a predetermined approximation level. In fact $\tau(\epsilon)$ controls the error of approximation according to the total variation bound

$$d_{TV}(P_\epsilon, P) = \sup_{A\subset\mathcal{X}} |P(A) - P_\epsilon(A)| \leq \epsilon \tag{2.6}$$

almost surely (a.s.). As such, it also guarantees the almost sure convergence of measurable functionals of P by the corresponding functionals of P_ϵ as $\epsilon \to 0$, cf. Proposition 4.20 of [131]. A typical application is in Bayesian nonparametric inference on mixture models where the Pitman–Yor process is used as prior distribution on the mixing measure. The approximation P_ϵ can be applied to the posterior distribution given the latent variables, cf. Section 2.1.2 for details. In the Dirichlet process case, P_ϵ has been studied by [236]. In this setting $\tau(\epsilon) - 1$ is Poisson distributed with parameter $\theta \log 1/\epsilon$, which makes an exact sampling of the ϵ-approximation (2.5) feasible. This has been implemented in the highly popular R software DPpackage, see [173], to draw posterior inference on the random effect distribution of linear and generalized linear mixed effect model. Finally, in [2] a different type of finite dimensional truncation of the Pitman–Yor process based on decreasing frequencies has been proposed, see Section 2.1.5 for a discussion.

The main theoretical contribution of this paper is the derivation of the asymptotic distribution of $\tau(\epsilon)$ as $\epsilon \to 0$ for $\alpha > 0$. As (2.4) suggests, the asymptotic distribution of $\tau(\epsilon)$ is related to that of R_n in (2.3) as $n \to \infty$. According to [268, Lemma 3.11], the latter involves a polynomially tilted stable random variable $T_{\alpha,\theta}$, see Section 2.1.2 for a formal definition. The main idea is to work with $T_n = -\log R_n$ so to deal with sums of the independent random variables $Y_i = -\log(1 - V_i)$. The distribution of $\tau(\epsilon)$ can be then studied in terms of the allied *renewal counting process* $N(t) = \max\{n : \ T_n \leq t\}$, according to the relation $\tau(\epsilon) = N(\log 1/\epsilon) + 1$. The problem boils down to the derivation of an appropriate a.s. convergence of $N(t)$ as $t \to \infty$, which, in turn, is obtained from the asymptotic distribution of T_n by showing that $N(t) \to \infty$ a.s. as $t \to \infty$ together with a (non standard) application of the law of large numbers for randomly indexed sequences. This strategy proves successful in establishing the almost sure convergence of $\tau(\epsilon) - 1$ to $(\epsilon T_{\alpha,\theta}/\alpha)^{-\alpha/(1-\alpha)}$ as $\epsilon \to 0$. The form of the asymptotic distribution reveals how large the truncation point $\tau(\epsilon)$ is as ϵ gets small in terms of the model parameters α and θ. In particular, it highlights the power law behavior of $\tau(\epsilon)$ as $\epsilon \to 0$, namely the growth at the polynomial rate $1/\epsilon^{\alpha/(1-\alpha)}$ compared to the slower logarithmic rate $\theta \log 1/\epsilon$ in the Dirichlet process

case. This is further illustrated by a simulation study in which we generate from the asymptotic distribution of $\tau(\epsilon)$ by using Zolotarev's integral representation of the positive stable distribution as in [98]. As far as the simulation of the ϵ-PY process is concerned, exact sampling is feasible by implementing the stopping rule in (2.4), that is by simulating the stick breaking frequencies p_j until the error R_n crosses the approximation level ϵ. As this can be computationally expensive when ϵ is small, as an alternative we propose to use the asymptotic distribution of $\tau(\epsilon)$ by simulating the truncation point first, then run the stick breaking procedure up to that point. It results in an approximate sampler of the ϵ-PY process that we compare with the exact sampler in a simulation study involving moments and mean functionals.

The rest of the paper is organized as follows. In Section 2.1.2, we derive the asymptotic distribution of $\tau(\epsilon)$ and explain how to use it to simulate from the ϵ-PY process. Section 2.1.3 reports a simulation study on the distribution of $\tau(\epsilon)$ and on functionals of the ϵ-PY process. In Section 2.1.4, to help the understanding and gain additional insight on the asymptotic distribution, we highlight the connections of $\tau(\epsilon)$ with Pitman's theory on random partition structures. We conclude with a discussion of open problems in Section 2.1.5. The details of Devroye's algorithm for generating from a polynomially tilted positive stable random variable are given in Section 2.1.6.

2.1.2 Theory and algorithms

Asymptotic distribution of $\tau(\epsilon)$

In this section we derive the asymptotic distribution of the stopping time $\tau(\epsilon)$ and show how to simulate from it. We start by introducing the renewal process interpretation which is crucial for the asymptotic results. As explained in the previous section, in order to study the distribution of $\tau(\epsilon)$ it is convenient to work with the log transformation of the truncation error R_n in (2.3), that is

$$T_n = \sum_{i=1}^{n} Y_i, \quad Y_i = -\log(1 - V_i), \tag{2.7}$$

with $V_j \overset{\text{ind}}{\sim} \text{beta}(1-\alpha, \theta + j\alpha)$ as in (2.2). Being a sum of independent and nonnegative random variables, $(T_n)_{n\geq 1}$ takes the interpretation of a (generalized) renewal process with independent waiting times Y_i. For $t \geq 0$ define

$$N(t) = \max\{n : \ T_n \leq t\}, \tag{2.8}$$

to be the *renewal counting process* associated to $(T_n)_{n\geq 1}$, which is related to $\tau(\epsilon)$ via $\tau(\epsilon) = N(\log 1/\epsilon) + 1$. Classical renewal theory pertains to iid waiting times while here there is no identity in distribution unless $\alpha = 0$, i.e. the Dirichlet process case. In the latter setting, one gets $Y_i \overset{\text{iid}}{\sim} \text{Exp}(\theta)$ so that T_n has gamma distribution with scale parameter n. We immediately get from the relation $\{T_n \leq t\} = \{N(t) \geq n\}$ that $N(t) \sim \text{Pois}(\theta t)$ and, in turn, that $\tau(\epsilon) - 1$ has $\text{Pois}(\theta \log(1/\epsilon))$ distribution. As far as asymptotics is concerned, T_n satisfies the CLT with $(T_n - n/\theta)/(\sqrt{n}/\theta) \to_d Z$ where $Z \sim \text{N}(0,1)$. The asymptotic distribution of $N(t)$ can be obtained via Ascombe theorem, cf. [151, Theorem 7.4.1], to get $(N(t) - \theta t)/(\sqrt{\theta t}) \to_d Z$, as $t \to \infty$,

in accordance with the standard normal approximation of the Poisson distribution with large rate parameter.

In the general Pitman–Yor case $\alpha > 0$, the waiting times Y_i are no more identically distributed. More importantly, generalizations of the CLT such as the Lindeberg–Feller theorem do not apply for T_n, hence we cannot resort to Anscombe's theorem to derive the asymptotic distribution of $N(t)$ and, in turn, of $\tau(\epsilon)$. Nevertheless, the limit exists but is not normal as stated in Theorem 2.1.1 below. To this aim, let T_α be a positive stable random variable with exponent α, that is $\mathbb{E}(\mathrm{e}^{-sT_\alpha}) = \mathrm{e}^{-s^\alpha}$, and denote its density by $f_\alpha(t)$. A polynomially tilted version of T_α is defined as the random variable $T_{\alpha,\theta}$ with density proportional to $t^{-\theta} f_\alpha(t)$, that is

$$f_{\alpha,\theta}(t) = \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} t^{-\theta} f_\alpha(t), \quad t > 0. \tag{2.9}$$

The random variable $T_{\alpha,\theta}$ is of paramount importance in the theory of random partition structures associated to the frequency distribution of the Pitman–Yor process, see Section 2.1.4 for details. In particular, the convergence of R_n can be expressed in terms of $T_{\alpha,\theta}$. In Theorem 2.1.1 the a.s. limit of $\log N(t)$ as $t \to \infty$ is obtained from that of $T_n = -\log R_n$ as $n \to \infty$ by showing that $N(t) \to \infty$ a.s. as $t \to \infty$ and by an application of the law of large numbers for randomly indexed sequences.

Theorem 2.1.1. *Let $N(t)$ be defined in (2.7)–(2.8) and $T_{\alpha,\theta}$ be the random variable with density in (2.9). Then $t - (1/\alpha - 1)\log N(t) + \log\alpha \to_{a.s.} \log T_{\alpha,\theta}$ as $t \to \infty$.*

Proof. By definition (2.8), the renewal process $N(t)$ is related to the sequence of renewal epochs T_n through

$$\{T_n \leq t\} = \{N(t) \geq n\}. \tag{2.10}$$

Since $N(T_n) = n$, we have $T_{N(t)} = T_n$ when $t = T_n$, thus $0 = t - T_{N(t)}$ for $t = T_n$. Moreover, since $N(t)$ is increasing, when $T_n < t < T_{n+1}$, $N(T_n) < N(t) < N(t)+1$, hence $T_{N(t)} < t < T_{N(t)+1}$, i.e. $0 < t - T_{N(t)} < T_{N(t)+1} - T_{N(t)} = Y_{N(t)+1}$. Together the two relations above yield

$$0 \leq t - T_{N(t)} < Y_{N(t)+1}. \tag{2.11}$$

From Lemma 3.11 of [268] and an application of the continuous mapping theorem [see Theorem 10.1 in 151] the asymptotic distribution of T_n is obtained as

$$T_n - (1/\alpha - 1)\log n + \log\alpha \to_{a.s.} \log T_{\alpha,\theta} \quad \text{as } n \to \infty.$$

Now we would like to take the limit with respect to $n = N(t)$ as $t \to \infty$, that is apply the law of large numbers for randomly indexed sequence [see Theorem 6.8.1 in 151]. To this aim, we first need to prove that $N(t) \to_{a.s.} \infty$ as $t \to \infty$. Since $N(t)$ is non decreasing, by an application of Theorem 5.3.5 in [151], it is sufficient to prove that $N(t) \to \infty$ in probability as $t \to \infty$, that is $\mathbb{P}(N(t) \geq n) \to 1$ as $t \to \infty$ for any $n \in \mathbb{N}$. But this is an immediate consequence of the inversion formula (2.10). We have then established that

$$T_{N(t)} - (1/\alpha - 1)\log N(t) + \log\alpha \to_{a.s.} \log T_{\alpha,\theta} \quad \text{as } t \to \infty.$$

To conclude the proof, we need to replace $T_{N(t)}$ with t in the limit above. Note that, from (2.11), $|t - T_{N(t)}| \leq Y_{N(t)+1}$ so it is sufficient to show that the upper bound goes to zero a.s.. Actually, by a second application of Theorem 6.8.1 in [151] it is sufficient to show that $Y_n \to_{a.s.} 0$ as $n \to \infty$. This last result is established as follows. Recall that $Y_j = -\log(1 - V_j)$ for $V_j \overset{\text{ind}}{\sim} \text{beta}(1-\alpha, \theta + j\alpha)$. For $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(1 - V_n < \mathrm{e}^{-\epsilon}) &= \int_0^{\mathrm{e}^{-\epsilon}} \frac{\Gamma(\theta + n\alpha + 1 - \alpha)}{\Gamma(\theta + n\alpha)\Gamma(1-\alpha)} v^{\theta + n\alpha - 1}(1-v)^{-\alpha}\mathrm{d}x \\
&\leq \frac{(1 - \mathrm{e}^{-\epsilon})^{-\alpha}}{\Gamma(1-\alpha)} \frac{\Gamma(\theta + n\alpha + 1 - \alpha)}{\Gamma(\theta + n\alpha)} \frac{\mathrm{e}^{-\epsilon(\theta + n\alpha)}}{\theta + n\alpha} \\
&= \frac{(1 - \mathrm{e}^{-\epsilon})^{-\alpha}}{\Gamma(1-\alpha)} (\theta + n\alpha)^{-\alpha} \mathrm{e}^{-\epsilon(\theta + n\alpha)} \Big(1 + O\Big(\frac{1}{\theta + n\alpha}\Big)\Big) \qquad (2.12)
\end{aligned}$$

where in equality (2.12) we have used Euler's formula

$$\Gamma(z + \alpha)/\Gamma(z + \beta) = z^{\alpha - \beta}\left[1 + \frac{(\alpha - \beta)(\alpha + \beta - 1)}{2z} + O(z^{-2})\right]$$

for $z \to \infty$, see [322]. Since $\mathbb{P}(Y_n > \epsilon) = \mathbb{P}(1 - V_n < \mathrm{e}^{-\epsilon})$, (2.12) implies that $\mathbb{P}(Y_n > \epsilon)$ is exponentially decreasing in n and, in turn, that $\sum_{n \geq 1} \mathbb{P}(Y_n > \epsilon) < \infty$. An application of Borel–Cantelli Lemma yields $Y_n \to_{a.s.} 0$ and the proof is complete. ■

The asymptotic distribution of $\tau(\epsilon)$ is readily derived from Theorem 2.1.1 via the formula $\tau(\epsilon) = N(\log 1/\epsilon) + 1$ and an application of the continuous mapping theorem. The proof is omitted.

Theorem 2.1.2. *Let $\tau(\epsilon)$ be defined in (2.4) and $T_{\alpha,\theta}$ be the random variable with density in (2.9). Then $\tau(\epsilon) - 1 \sim_{a.s.} (\epsilon T_{\alpha,\theta}/\alpha)^{-\alpha/(1-\alpha)}$ as $\epsilon \to 0$.*

In order to sample from the asymptotic distribution of $\tau(\epsilon)$, the key ingredient is random generation from the polynomially tilted stable random variable $T_{\alpha,\theta}$. Following [98], we resort to Zolotarev's integral representation, so let $A(u)$ be the Zolotarev function

$$A(x) = \left(\frac{\sin(\alpha x)^{\alpha} \sin((1-\alpha)x)^{1-\alpha}}{\sin(x)}\right)^{\frac{1}{1-\alpha}}, \quad x \in [0, \pi]$$

and $Z_{\alpha,b}$, $\alpha \in (0,1)$ and $b > -1$ be a Zolotarev random variable with density given by

$$f(x) = \frac{\Gamma(1 + b\alpha)\Gamma(1 + b(1-\alpha))}{\pi\Gamma(1+b)A(x)^{b(1-\alpha)}}, \quad x \in [0, \pi].$$

According to Theorem 1 of [98], for G_a a gamma distributed random variable with shape $a > 0$ and unit rate,

$$T_{\alpha,\theta} \overset{d}{=} \left(\frac{A(Z_{\alpha,\theta/\alpha})}{G_{1+\theta(1-\alpha)/\alpha}}\right)^{\frac{1-\alpha}{\alpha}}$$

so that random variate generation simply requires one gamma random variable and one Zolotarev random variable. For the latter, rejection sampler can be used as detailed in [98]. See ALGORITHM 3 in Appendix 2.1.6.

Simulation of the ϵ-PY process

Given α, θ, ϵ and a probability measure P_0 on $\mathcal{X}$, an ϵ-PY process can be generated by implementing the stopping rule in the definition of $\tau(\epsilon)$, cf. (2.4). The algorithm consists in a while loop as follows:

ALGORITHM 1 (Exact sampler of ϵ-PY)

1. `set` $i = 1,\ R = 1$
2. `while` $R \geq \epsilon$: `generate` V `from` $\text{beta}(1-\alpha, \theta + i\alpha)$. `set` $p_i = VR,\ R = R(1-V),\ i = i+1$
3. `set` $\tau = i,\ R_\tau = R$
4. `generate` $\tau + 1$ `random variates` $\xi_0, \xi_1, \ldots, \xi_\tau$ `from` P_0
5. `set` $P_\epsilon(\mathrm{d}x) = \sum_{i=1}^{\tau} p_i \delta_{\xi_i}(\mathrm{d}x) + R_\tau \delta_{\xi_0}(\mathrm{d}x)$

When ϵ is small, the while loop happens to be computationally expensive since conditional evaluations at each iteration slow down computation, and memory allocation for the frequency and location vectors cannot be decided beforehand. In order to avoid these pitfalls and make the algorithm faster, one should generate the stopping time $\tau(\epsilon)$ first, and the frequencies up to that point later. We propose to exploit the asymptotic distribution of $\tau(\epsilon)$ in Theorem 2.1.2 as follows:

ALGORITHM 2 (Approximate sampler of ϵ-PY)

1: `generate` $T \overset{d}{=} T_{\alpha,\theta}$
2: `set` $\tau \leftarrow 1 + \lfloor (\epsilon T/\alpha)^{-\alpha/(1-\alpha)} \rfloor$
3. `for` $i = 1, \ldots, \tau$: `generate` V_i `from` $\text{beta}(1-\alpha, \theta + i\alpha)$. `set` $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$
4. `set` $R_\tau = 1 - \sum_{i=1}^{\tau} p_i = \prod_{i=1}^{\tau}(1 - V_j)$
5: `generate` $\tau + 1$ `random variates` $\xi_0, \xi_1, \ldots, \xi_\tau$ `from` P_0
6: `set` $P_\epsilon(\mathrm{d}x) = \sum_{i=1}^{\tau} p_i \delta_{\xi_i}(\mathrm{d}x) + R \delta_{\xi_0}(\mathrm{d}x)$

ALGORITHM 2 is an *approximate* sampler of the ϵ-PY process (while ALGORITHM 1 is an *exact* one) since it introduces two sources of approximations. First, through the use of the asymptotic distribution of $\tau(\epsilon)$. Second, through `Step 3` since the V_i's are not generated according to the conditional distribution given $\tau(\epsilon)$, rather unconditionally. Finding the conditional distribution of V_i, or an asymptotic approximation thereof, is not an easy task and is object of current research. In terms of the renewal process interpretation in (2.7)–(2.8), the problem is to generate the waiting times $Y_i = -\log(1 - V_i)$, $i = 1, \ldots, n$, from the conditional distribution of the renewal epochs $(T_1, \ldots, T_n)$ given $N(t) = n$ for $t = -\log 1/\epsilon$.

A typical use of samples from the Pitman–Yor process we have in mind is in infinite mixture models. In fact, the discrete nature of the Pitman–Yor process makes it a

suitable prior on the *mixing distribution*. ALGORITHM 1 or ALGORITHM 2 can be then applied to approximate a functional of the posterior distribution of the mixing distribution. In such models, the process components can be seen as latent features exhibited by the data. Let P denote such a process, n denote the sample size and $X_{1:n} = (X_1, \dots, X_n)$ be an exchangeable sequence from P, that is $X_{1:n}|P \overset{\text{iid}}{\sim} P$. Variables $X_{1:n}$ are latent variables in a model conditionally on which observed data $Y_{1:n}$ come from: $Y_j|X_j \overset{\text{ind}}{\sim} f(\,\cdot\,|X_j)$ where f denotes a kernel density. Actually, independence is not necessary here and applications also encompass dependent models such as Markov chain transition density estimation. In order to deal with the infinite dimensionality of the process, a strategy is to marginalize it and to draw posterior inference with a *marginal* sampler. Since draws from a marginal sampler allows to make inference only on posterior expectations of the process, for more general functionals of P, in the form of $\psi(P)$, one typically needs to resort to an additional sampling step. Exploiting the composition rule $\mathcal{L}(\psi(P)|Y_{1:n}) = \mathcal{L}(\psi(P)|X_{1:n}) \times \mathcal{L}(X_{1:n}|Y_{1:n})$ this additional step boils down to sampling P conditional on latent variables $X_{1:n}$. At this stage, recalling the conditional conjugacy of the Pitman–Yor process is useful. Among $X_{1:n}$, there are a number $k \leq n$ of unique values that we denote by $X^*_{1:k}$. Let $n^*_{1:k}$ denote their frequencies. Then the following identity in distribution holds

$$P|X_{1:n} = \sum_{j=1}^{k} q_j \delta_{X^*_j} + q_{k+1} P^*,$$

where, independently, $(q_1, \dots, q_k, q_{k+1}) \sim \text{Dirichlet}(n^*_1 - \alpha, \dots, n^*_k - \alpha, \theta + \alpha k)$ and P^* is a Pitman–Yor process of parameter $(\alpha, \theta + \alpha k)$, see Corollary 20 of [**?**]. Thus sampling from $\mathcal{L}(P|X_{1:n})$, hence from $\mathcal{L}(\psi(P)|X_{1:n})$, requires sampling the infinite dimensional P^*. Cf. [164, Section 4.4]. For the sake of comparison, the conjugacy of the Dirichlet process similarly leads to the need of sampling an infinite dimensional process, where $P|X_{1:n}$ takes the form of a Dirichlet process. As already noticed, the truncation of the Dirichlet process is very well understood, both theoretically and practically. The popular R package DPpackage [173] makes use of the *posterior* truncation point $\tau^*(\epsilon)$, as defined in (2.5), but here with respect to the posterior distribution of the process. Thus, it satisfies $\tau^*(\epsilon) - 1 \sim \text{Pois}((\theta + n)\log(1/\epsilon))$, where $\theta + n$ is the precision of the posterior Dirichlet process. Adopting here similar lines for the Pitman–Yor process, we replace P^* by the truncated process P^*_ϵ

$$P^*_\epsilon(\mathrm{d}x) = \sum_{i=1}^{\tau^*(\epsilon)} p^*_i \delta_{\xi_i}(\mathrm{d}x) + R_{\tau^*(\epsilon)} \delta_{\xi_0}(\mathrm{d}x),$$

cf. equation (2.5). Here $(p^*_i)_{i \geq 1}$ are defined according to (2.2) with $\theta + \alpha k$ in place of θ, i.e. $V_j \overset{\text{ind}}{\sim} \text{beta}(1 - \alpha, \theta + \alpha(k + j))$. Hence, according to Theorem 2.1.2 we have

$$\tau^*(\epsilon) - 1 \sim_{a.s.} (\epsilon T_{\alpha,\theta+\alpha k}/\alpha)^{-\alpha/(1-\alpha)}, \quad \text{as } \epsilon \to 0$$

hence ALGORITHM 2 can be applied here.

2.1.3 Simulation study

Stopping time $\tau(\epsilon)$

According to Theorem 2.1.2, the asymptotic distribution of $\tau(\epsilon)$ changes with ϵ, α and θ. For illustration, we simulate $\tau(\epsilon)$ from `Steps 1.-2.` in ALGORITHM 2 using Devroye's sampler, cf. ALGORITHM 3 in Appendix 2.1.6. In Figure 2.1 we compare density plots obtained with 10^4 iterations with respect to different combinations of ϵ, α and θ. The plot in the left panel shows how smaller values of ϵ result in larger values of $\tau(\epsilon)$. In fact, as $\epsilon \to 0$, $\tau(\epsilon)$ increases proportional to $1/\epsilon^{\alpha/(1-\alpha)}$. Note also that $(\epsilon T_{\alpha,\theta}/\alpha)^{-\alpha/(1-\alpha)}$ is nonnegative for $T_{\alpha,\epsilon} < \alpha/\epsilon$, which happens with high probability when ϵ is small. As for α, the plot in the central panel shows how $\tau(\epsilon)$ increases as α gets large. In fact, it is easy to see that $(\epsilon T_{\alpha,\theta}/\alpha)^{-\alpha/(1-\alpha)}$ is increasing in α when $T_{\alpha,\epsilon} < e^{1-\alpha}\alpha/\epsilon$, which also happens with high probability when ϵ is small, so the larger α, the more stick-breaking frequencies are needed in order to account for a prescribed approximation error ϵ. Finally, the plot in the right panel shows that the larger θ, the larger $\tau(\epsilon)$. In fact, by definition, the polynomial tilting makes $T_{\alpha,\theta}$ stochastically decreasing in θ.

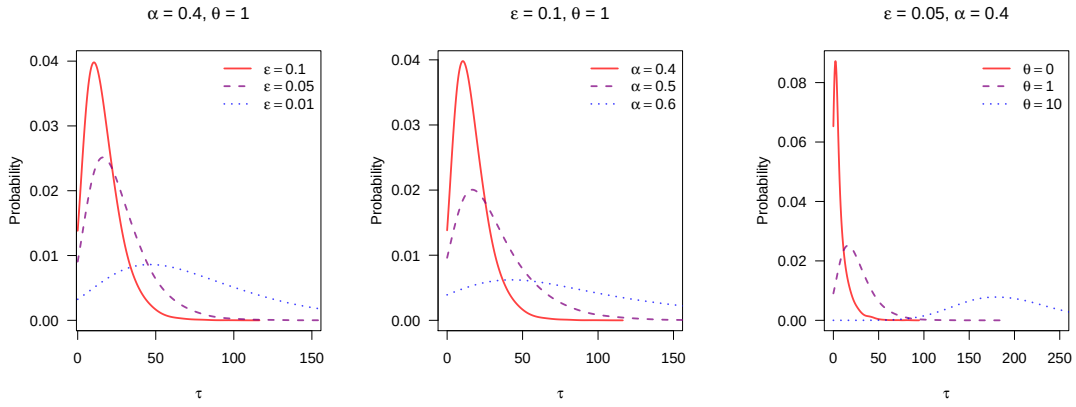


FIGURE 2.1: Density plot for the asymptotic approximation of $\tau(\epsilon)$ based on 10^4 values under the following parameter configurations. Left: $\epsilon \in \{0.10, 0.05, 0.01\}$, $\alpha = 0.4$, $\theta = 1$. Center: $\alpha \in \{0.4, 0.5, 0.6\}$, $\theta = 1$, $\epsilon = 0.1$. Right: $\theta \in \{0, 1, 10\}$, $\alpha = 0.25$, $\epsilon = 0.05$.

In order to illustrate the rate of convergence in Theorem 2.1.2, we compare next the exact distribution of $\tau(\epsilon)$ with the asymptotic one. To do so, we repeat the following experiment several times: we simulate $\tau(\epsilon)$ from `Steps 1.-3.` in ALGORITHM 1, then we compare the empirical distribution of $(\epsilon/\alpha)^{\alpha}(\tau(\epsilon) - 1)^{1-\alpha}$ with $T_{\alpha,\theta}^{-\alpha}$, the latter corresponding to the α-diversity of the PY process, see Section 2.1.4 for a formal definition. In Table 2.1 we report the Kolmogorov distance together with expected value, median, first and third quartiles for both the exact and the asymptotic distribution obtained with 10^4 iterations. This is repeated for $\alpha = 0.5$, $\theta = \{0, 1, 10\}$ and $\epsilon = \{0.10, 0.05, 0.01\}$. As expected, as we decrease ϵ, the Kolmogorov distance gets smaller to somehow different rates according to the parameter choice. The derivation of convergence rates is left for future research.

		d_K	Mean		25%		Median		75%	
θ	ϵ		As	Ex	As	Ex	As	Ex	As	Ex
0	0.10	3.42	1.06	1.05	0.45	0.45	0.89	0.89	1.61	1.55
0	0.05	2.17	1.10	1.08	0.45	0.45	0.95	0.95	1.64	1.58
0	0.01	1.73	1.14	1.11	0.45	0.45	0.97	0.95	1.64	1.60
1	0.10	4.79	2.24	2.14	1.55	1.48	2.14	2.10	2.86	2.76
1	0.05	2.38	2.25	2.20	1.55	1.52	2.17	2.14	2.86	2.79
1	0.01	1.40	2.26	2.25	1.57	1.54	2.19	2.19	2.87	2.85
10	0.10	11.93	6.39	6.07	5.69	5.40	6.34	6.06	7.04	6.72
10	0.05	6.12	6.39	6.24	5.70	5.56	6.34	6.22	7.05	6.88
10	0.01	1.93	6.40	6.37	5.71	5.70	6.34	6.34	7.05	7.00

TABLE 2.1: Summary statistics for the asymptotic distribution (As) and exact distribution (Ex) of $\tau(\epsilon)$ at the scale of the α-diversity based on 10^4 values. The Kolmogorov distance (d_K) is between the empirical cumulative distribution function of the sample from the exact distribution and the asymptotic one (multiplied by a factor of 100). The parameter values are $\alpha = 0.5$, $\theta \in \{0, 1, 10\}$ and $\epsilon \in \{0.10, 0.05, 0.01\}$.

Functionals of the ϵ-PY process

In the case that P is defined on $\mathcal{X} \subseteq \mathbb{R}$, the total variation bound (2.6) implies that $|F(x) - F_\epsilon(x)| < \epsilon$ almost surely for any $x \in \mathbb{R}$, where F_ϵ and F are the cumulative distribution functions of P_ϵ and P. Also, measurable functionals $\psi(P)$ such as the mean $\mu = \int xP(\mathrm{d}x)$ can be approximated in distribution by the corresponding functionals $\psi(P_\epsilon)$. For illustration, we set $\mathcal{X} = [0, 1]$ and P_0 the uniform distribution on $[0, 1]$. For given α and θ, we then compare the distribution under P with that under the ϵ-PY process P_ϵ for $F(1/2)$, $F(1/3)$ and $\mu = \int xP(\mathrm{d}x)$. As for the distribution of the finite dimensional distributions $F(1/2)$ and $F(1/3)$ under the full process P, we set $\alpha = 0.5$ so to exploit results in [169]. According to their Proposition 4.7, the finite dimensional distributions of P when $\alpha = 0.5$ are given by

$$f(w_1, \ldots, w_{n-1}) = \frac{(\prod_{i=1}^n p_i)\Gamma(\theta + n/2)}{\pi^{(n-1)/2}\Gamma(\theta + 1/2)} \frac{w_1^{-3/2} \cdots w_{n-1}^{-3/2}(1 - \sum_{i=1}^{n-1} w_i)^{-3/2}}{\mathcal{A}_n(w_1, \ldots, w_{n-1})^{\theta+n/2}}$$

for any partition $A_1, \ldots, A_n$ of $\mathcal{X}$ with $p_i = P_0(A_i)$ and $\mathcal{A}_n(w_1, \ldots, w_{n-1}) = \sum_{i=1}^{n-1} p_i^2 w_i^{-1} + p_n^2(1 - \sum_{i=1}^{n-1} w_i)^{-1}$. Direct calculation shows that $F(1/2)$ has beta distribution with parameters $(\theta + 1/2, \theta + 1/2)$ while $F(1/3)$ has density

$$f(w) = \frac{2}{\sqrt{\pi}} 9^\theta \frac{\Gamma(\theta + 1)}{\Gamma(\theta + 1/2)} \frac{(w(1 - w))^{\theta - 1/2}}{(1 + 3w)^{\theta+1}}.$$

As for the mean functional $\mu = \int xP(\mathrm{d}x)$, the distribution under the full process P is approximated by simulations by setting a deterministic truncation point sufficiently large. As for the distribution under P_ϵ, we use both ALGORITHM 1 and ALGORITHM 2.

In Figure 2.2 we compare the density plots of $F(1/2)$ for $\epsilon = \{0, 1.0.05, 0.001\}$ and $\theta = \{0, 10\}$ under P_ϵ with the beta density under P so to illustrate that the two distributions get close as ϵ gets small. As for $F(1/3)$ and $\mu = \int xP(\mathrm{d}x)$, in Tables 2.2

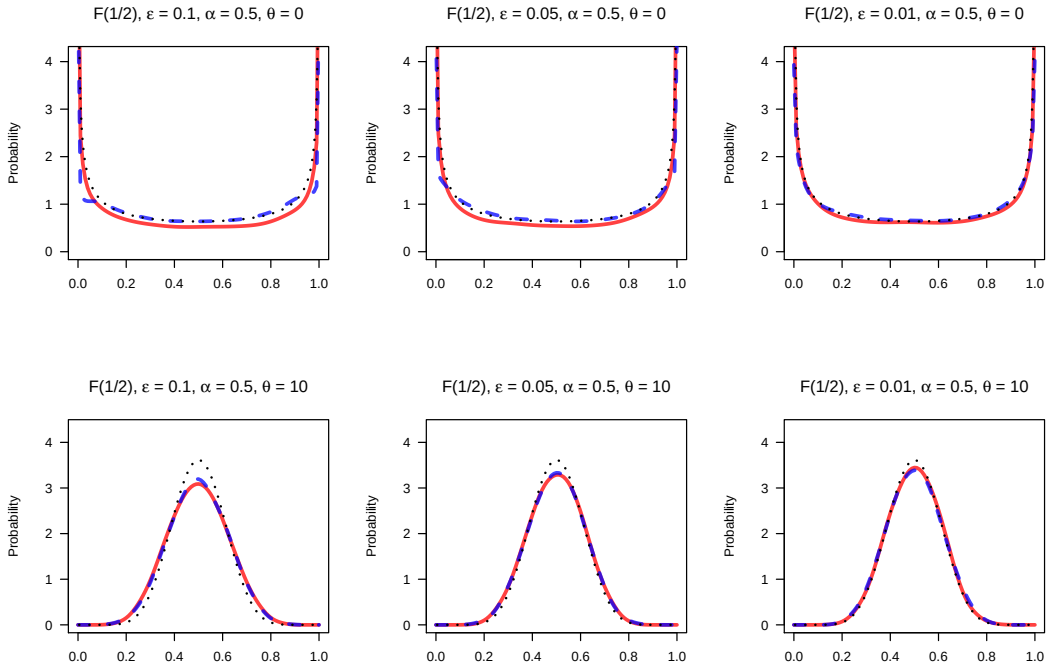


FIGURE 2.2: Density plots for the random probability $F(1/2)$ using the ALGORITHM 2 (in red solid curve) and ALGORITHM 1 (in blue dashed curve) to sample from the ϵ-PY process. The density under the Pitman–Yor process is the black dotted curve. The parameter α is fixed equal to 0.5, θ is equal to 0 on the first row and 10 on the second row, while ϵ is respectively equal to $\{0.10, 0.05, 0.01\}$ in the left, center and right columns.

and 2.3 we report the Kolmogorov distance between P and P_ϵ for the two sampling algorithms, together with expected value, median, first and third quartiles. For each case and each parameter configuration, we have sampled 10^4 trajectories from the ϵ-PY process and 10^4 trajectories from the Pitman–Yor process in the case of $\mu = \int xP(\mathrm{d}x)$. As expected, the Kolmogorov distances are generally larger, still close, when using ALGORITHM 2 versus ALGORITHM 1 due to the approximate nature of the former.

Computation time

In this section, we provide a concrete justification of the computational advantage of using ALGORITHM 2 versus ALGORITHM 1. We simulate 10^4 ϵ-PY iterations by using ALGORITHM 1 and ALGORITHM 2 for different combinations of the α and θ parameters and of the ϵ error threshold. In Table 2.4 (resp. Table 2.5), we report the average computing time[1] *per iteration* (resp. *per support point*) for ALGORITHM

[1]The experiments were conducted on an Intel Core i5 processor (3.1 GHz) computer.

TABLE 2.2: Simulation study on $F(1/3)$

		d_K		Mean			25%			Median			75%		
θ	ϵ	AL1	AL2	AL1	AL2	PY	AL1	AL2	PY	AL1	AL2	PY	AL1	AL2	PY
0	0.10	16.29	16.48	0.33	0.33	0.33	0.04	0.01	0.04	0.20	0.16	0.20	0.60	0.64	0.59
0	0.05	11.53	12.52	0.33	0.33	0.33	0.05	0.01	0.04	0.20	0.17	0.20	0.58	0.63	0.59
0	0.01	5.49	5.60	0.34	0.33	0.33	0.04	0.03	0.04	0.21	0.19	0.20	0.59	0.61	0.59
1	0.10	3.08	5.65	0.33	0.33	0.33	0.14	0.12	0.14	0.29	0.28	0.28	0.49	0.50	0.49
1	0.05	1.34	3.11	0.33	0.33	0.33	0.14	0.13	0.14	0.28	0.28	0.28	0.48	0.50	0.49
1	0.01	0.56	0.89	0.33	0.34	0.33	0.14	0.14	0.14	0.28	0.29	0.28	0.49	0.49	0.49
10	0.10	3.10	3.81	0.33	0.33	0.33	0.25	0.25	0.26	0.32	0.32	0.32	0.40	0.41	0.40
10	0.05	1.41	1.38	0.33	0.33	0.33	0.26	0.26	0.26	0.32	0.32	0.32	0.40	0.40	0.40
10	0.01	0.75	0.65	0.33	0.33	0.33	0.26	0.26	0.26	0.33	0.32	0.32	0.40	0.40	0.40

TABLE 2.3: Simulation study on $\mu = \int xP(\mathrm{d}x)$

		d_K		Mean			25%			Median			75%		
θ	ϵ	AL1	AL2	AL1	AL2	PY	AL1	AL2	PY	AL1	AL2	PY	AL1	AL2	PY
0	0.10	1.60	3.57	0.50	0.50	0.50	0.36	0.34	0.36	0.50	0.50	0.50	0.64	0.67	0.65
0	0.05	0.94	2.72	0.50	0.50	0.50	0.35	0.34	0.36	0.50	0.50	0.50	0.64	0.66	0.65
0	0.01	1.18	2.10	0.50	0.50	0.50	0.36	0.35	0.36	0.50	0.50	0.50	0.64	0.65	0.65
1	0.10	1.61	3.18	0.50	0.50	0.50	0.40	0.39	0.40	0.50	0.50	0.50	0.60	0.61	0.60
1	0.05	1.28	2.32	0.50	0.50	0.50	0.40	0.40	0.40	0.50	0.50	0.50	0.59	0.60	0.60
1	0.01	1.12	0.57	0.50	0.50	0.50	0.40	0.41	0.40	0.50	0.50	0.50	0.60	0.60	0.60
10	0.10	2.81	4.18	0.50	0.50	0.50	0.45	0.46	0.46	0.50	0.50	0.50	0.55	0.55	0.54
10	0.05	1.78	1.28	0.50	0.50	0.50	0.46	0.46	0.46	0.50	0.50	0.50	0.54	0.54	0.54
10	0.01	2.01	1.09	0.50	0.50	0.50	0.46	0.46	0.46	0.50	0.50	0.50	0.54	0.54	0.54

Summary statistics for $F(1/3)$ (Table 2.2) and $\mu = \int xP(\mathrm{d}x)$ (Table 2.3) using ALGORITHM 1 (AL1) and ALGORITHM 2 (AL2) to sample from the ϵ-PY process. The Kolmogorov distance (d_K) is between the cumulative distribution functions with respect to the Pitman–Yor (PY) process (multiplied by a factor of 100). The parameter values are $\alpha = 0.5$, $\theta \in \{0, 1, 10\}$ and $\epsilon \in \{0.10, 0.05, 0.01\}$.

1 and ALGORITHM 2. By *iteration*, we mean a full realization of the ϵ-PY process including frequencies and locations, while by *support point*, we mean that we divide the total time by the number of support points $\tau(\epsilon) + 1$. In order to account for the computational task required per iteration, the expected stopping time $\mathbb{E}[\tau(\epsilon)]$ is also reported. Both tables illustrate that our proposed approach is faster than ALGORITHM 1 when the ϵ-PY is composed of about 20 support points or more. The more support points, the faster ALGORITHM 2 is compared to ALGORITHM 1. This disadvantage of the former for small numbers of support points comes from the

fixed cost of initially generating a random variable with the same distribution as $T_{\alpha,\theta}$. Conversely, as the number of support points increases, this fixed cost is largely counterbalanced by the fast vector-sampling of a prescribed size, which is in contrast with ALGORITHM 1 `while` loop whose cost increases with the number of support points. This can be seen in Table 2.5 where the actual sampling time per support point is essentially increasing for ALGORITHM 1 and decreasing for ALGORITHM 2. With the parameter configurations tested, ALGORITHM 2 can be up to 90 times faster ALGORITHM 1 for $\alpha = 0.6$, $\theta = 10$ and $\epsilon = 0.01$.

TABLE 2.4: Computing time (ms) per **iteration**

		$\alpha = 0.4$			$\alpha = 0.5$			$\alpha = 0.6$		
θ	ϵ	AL1	AL2	n	AL1	AL2	n	AL1	AL2	n
0	0.10	0.01	0.20	5	0.02	0.04	11	0.11	0.05	38
0	0.05	0.01	0.04	8	0.04	0.04	21	0.20	0.06	105
0	0.01	0.04	0.04	20	0.36	0.05	101	15.10	0.26	1163
1	0.10	0.03	0.19	17	0.07	0.05	31	0.23	0.07	92
1	0.05	0.06	0.06	26	0.13	0.06	61	0.80	0.12	258
1	0.01	0.18	0.09	73	0.80	0.12	301	27.75	0.57	2877
10	0.10	0.22	0.15	121	0.61	0.10	211	2.11	0.18	567
10	0.05	0.45	0.10	191	1.52	0.15	421	9.22	0.37	1603
10	0.01	1.93	0.20	558	13.24	0.48	2101	760.68	4.01	17911

TABLE 2.5: Computing time (μs) per **support point**

		$\alpha = 0.4$			$\alpha = 0.5$			$\alpha = 0.6$		
θ	ϵ	AL1	AL2	n	AL1	AL2	n	AL1	AL2	n
0	0.10	1.92	38.46	5	1.82	3.64	11	2.91	1.32	38
0	0.05	1.30	5.22	8	1.90	1.90	21	1.91	0.57	105
0	0.01	1.95	1.95	20	3.56	0.50	101	12.98	0.22	1163
1	0.10	1.81	11.45	17	2.26	1.61	31	2.50	0.76	92
1	0.05	2.33	2.33	26	2.13	0.98	61	3.10	0.46	258
1	0.01	2.45	1.23	73	2.66	0.40	301	9.65	0.20	2877
10	0.10	1.82	1.24	121	2.89	0.47	211	3.72	0.32	567
10	0.05	2.35	0.52	191	3.61	0.36	421	5.75	0.23	1603
10	0.01	3.46	0.36	558	6.30	0.23	2101	42.47	0.22	17911

Average computing time per iteration (in *milli*second in Table 2.4) and per support point (in *micro*second in Table 2.5) for ALGORITHM 1 (AL1) and ALGORITHM 2 (AL2) based on 10^4 iterations, and expected stopping time $n = \mathbb{E}[\tau(\epsilon)]$. The parameter values are $\alpha \in \{0.4, 0.5, 0.6\}$, $\theta \in \{0, 1, 10\}$ and $\epsilon \in \{0.10, 0.05, 0.01\}$.

2.1.4 Connections with random partition structures

α-diversity and asymptotic distribution of R_n

The random variable $T_{\alpha,\theta}$ in Theorem 2.1.1 plays a key role in the Pitman–Yor process, in particular for its link with the *α-diversity* of the process. The α-diversity is defined as the almost sure limit of $n^{-\alpha}K_n$ where K_n denotes the (random) number of unique values in the first n terms of an exchangeable sequence from P in (2.1). According to Theorem 3.8 in [268], $n^{-\alpha}K_n \sim_{a.s.} (T_{\alpha,\theta})^{-\alpha}$, in particular, for $\theta = 0$, $T_\alpha^{-\alpha}$ has a Mittag-Leffler distribution with p-th moment $\Gamma(p+1)/\Gamma(p\alpha+1)$, $p > -1$. According to [268, Lemma 3.11, eqn (3.36)], the asymptotic distribution of the truncation error R_n can be derived from that of K_n to get $R_n \sim_{a.s.} \alpha(T_{\alpha,\theta})^{-1} n^{1-1/\alpha}$ as $n \to \infty$. The proof relies on Kingman's representation of random partitions [191] together with techniques set forth by [135]. In the proof of Theorem 2.1.1 the asymptotic distribution of $T_n = -\log R_n$ is a direct consequence of the above by an application of the continuous mapping theorem.

When $\theta = 0$ it is possible to give an interpretation of the asymptotic distribution of R_n in terms of the jumps of a stable subordinator. In this case the weights of P can be represented as the renormalized jumps of a stable subordinator, with T_α denoting the total mass. Denote the (unormalized) jumps as $(J_i)_{i\geq 1}$ in decreasing order and as $(\tilde{J}_i)_{i\geq 1}$ when in size-biased order,

$$T_\alpha = \sum_{i\geq 1} J_i = \sum_{i\geq 1} \tilde{J}_i, \quad \text{and } T_\alpha R_n = \sum_{i>n} \tilde{J}_i.$$

By the asymptotic distribution of R_n, $n^{1/\alpha-1}\sum_{i>n}\tilde{J}_i \to_{a.s.} \alpha$ as $n \to \infty$. That is, once properly scaled, the small jumps of the stable subordinator (in size-biased random order), interpreted as the "dust", converge to the "proportion" α. This is reminiscent to the number of singletons which is asymptotically $(n \to \infty)$ a α proportion of the number of groups in a sample of size n, see Lemma 3.11, eqn (3.39), of [268].

Regenerative random compositions and Anscombe's theorem

We review next the connections of the counting renewal process $N(t)$ defined in (2.7)–(2.8) and the theory of regenerative random compositions. The reader is referred to the survey of [134] for a review. Recall that, when $\alpha = 0$ (Dirichlet process case), $V_i \overset{\text{iid}}{\sim} \text{beta}(1,\theta)$ in the stick-breaking representation (2.2), and in turns $Y_i = -\log(1-V_i) \overset{\text{iid}}{\sim} \text{Exp}(\theta)$ and $T_n = -\log R_n \sim \text{Gamma}(n,\theta)$. By direct calculus, $N(t) \sim \text{Pois}(\theta t)$ so that $\tau(\epsilon) - 1 = N(\log 1/\epsilon)$ has $\text{Pois}(\theta \log 1/\epsilon)$ distribution. More generally, the stick-breaking frequencies $(p_i)_{i\geq 1}$ correspond to the gaps in $[0,1]$ identified by the *multiplicative regenerative* set $\mathcal{R} \subset (0,1)$ consisting of the random partial sums $1 - R_k = \sum_{i\leq k} p_i$. The complement open set $\mathcal{R}^c = (0,1)/\mathcal{R}$ can be represented as a disjoint union of countably many open intervals or gaps, $\mathcal{R}^c = \bigcup_{k=0}^{\infty}(1-R_k, 1-R_{k+1})$, $R_0 = 1$. A random composition of the integer n into an ordered sequence $\varkappa_n = (n_1, n_2, \ldots, n_k)$ of positive integers with $\sum_j n_j = n$ can be generated as follows: independently of $\mathcal{R}$, sample $U_1, U_2, \ldots$ from the uniform distribution on $[0,1]$ and group them in clusters by the rule: U_i, U_j belong to the same

cluster if they hit the same interval. The random composition of $\varkappa_n$ corresponds then to the record of positive counts in the left-to-right order of the intervals. The composition structure $(\varkappa_n)$ is called regenerative since for all $n > m \geq 1$, conditionally given the first part of $\varkappa_n$ is m, if the part is deleted then the remaining composition of $n - m$ is distributed like $\varkappa_{n-m}$. The regenerative set $\mathcal{R}$ corresponds to the closed range of the multiplicative subordinator $\{1 - \exp(-S_t), t \geq 0\}$, where S_t is the compound Poisson process with Lévy intensity $\tilde{\nu}(\mathrm{d}y) = \theta \mathrm{e}^{-\theta y}\mathrm{d}y$. Since the range of S_t is a homogeneous Poisson point process on $\mathbb{R}_+$ with rate θ, $\mathcal{R}$ is an inhomogeneous Poisson point process $\mathcal{N}(\mathrm{d}x)$ on $[0,1]$ with Lévy intensity $\nu(\mathrm{d}x) = \theta/(1-x)\mathrm{d}x$ so that, for $t = \log 1/\epsilon$,

$$N(\log 1/\epsilon) = \mathcal{N}[0, 1-\epsilon] \sim \mathrm{Pois}(\lambda), \quad \lambda = \int_0^{1-\epsilon} \frac{\theta}{1-x}\mathrm{d}x = \theta \log 1/\epsilon$$

as expected. Suppose now that $(V_i)_{i\geq 1}$ are independent copies of some random variable V on $[0,1]$, not necessarily beta$(1,\theta)$ distributed. The corresponding random composition structure has been studied in [137, 138] as the outcome of a *Bernoulli sieve* procedure. We recall here the relevant asymptotic analysis. Let $\mu = \mathbb{E}(-\log(1-V))$ and $\sigma^2 = \mathrm{Var}(-\log(1-V))$, equal respectively to $1/\theta$ and $1/\theta^2$ in the DP case, respectively. If those moments are finite, by the CLT,

$$\frac{T_n - n\mu}{\sqrt{n}\sigma} \to_d Z, \quad \text{as } n \to \infty,$$

where $Z \sim \mathrm{N}(0,1)$, and, by means of Anscobe's Theorem, one obtains that

$$\frac{N(t) - t/\mu}{\sqrt{\sigma^2 t/\mu^3}} \to_d Z, \quad \text{as } t \to \infty.$$

It turns out that the normal limit of $N(\log n)$ corresponds to the normal limit of K_n,

$$\frac{K_n - \log n/\mu}{\sqrt{\sigma^2 \log n/\mu^3}} \to_d Z, \quad \text{as } n \to \infty$$

provided that $\mathbb{E}(-\log V) < \infty$. To see why, consider iid random variables $X_1, X_2, \ldots$ with values in $\mathbb{N}$ such that $\{X_i = k\} = \{U_i \in (1-R_{k-1}, 1-R_k)\}$. Hence $\mathbb{P}(X_1 = k|\mathcal{R}) = p_k$. We then have that $K_n = \#\{k : X_i = k \text{ for at least one } i \text{ among } 1, \ldots, n\}$. Define $M_n = \max\{X_1, \ldots, X_n\}$. For $U_{1,n} \leq U_{2,n} \leq \ldots \leq U_{n,n}$ denoting the order statistics corresponding to the uniform variates $U_1, \ldots, U_n$, we have $M_n = \min\{j : 1 - R_j \geq U_{n,n}\} = \min\{j : T_j \geq E_{n,n}\}$ upon transformation $x \to -\log(1-x)$, where $E_{n,n}$ is the maximum of an iid sample of size n from the standard exponential distribution. Since $N(t) = \max\{n : T_n \leq t\} = \min\{n : T_n \geq t\} - 1$ we have $M_n - 1 = N(E_{n,n})$. [138] proves the equivalence

$$\frac{M_n - b_n}{a_n} \to_d X \iff \frac{N(\log n) - b_n}{a_n} \to_d X$$

where X is a random variable with a proper and non degenerate distribution with $a_n > 0$, $a_n \to \infty$ and $b_n \in \mathbb{R}$. A key fact exploited in the proof is that, from extreme-value theory, $E_{n,n} - \log n$ has an asymptotic distribution of Gumbel type. That M_n can be replaced by K_n in the equivalence relation above follows from the fact that

$M_n - K_n$, the number of integers $k < M_n$ not appearing in the sample $X_1, \ldots, X_n$, is bounded in probability when $\mathbb{E}(-\log V) < \infty$, see Proposition 5.1 in [138].

Back to the Pitman–Yor process case, by Theorem 2.1.1 we have $n^{-\alpha/(1-\alpha)} N(\log n) \to_d (T_{\alpha,\theta}/\alpha)^{-\alpha/(1-\alpha)}$ while $n^{-\alpha} K_n \to_{a.s.} (T_{\alpha,\theta})^{-\alpha}$. So we see that $N(\log n)$ and K_n do not have the same asymptotic behavior as in the $\alpha = 0$ case. By using the fact that

$$\mathbb{P}(X_1 > n | (p_i)) = R_n, \quad R_n \sim_{a.s.} \alpha n^{-(1-\alpha)/\alpha} T_{\alpha,\theta}^{-1},$$

and the fact that, conditional on $(p_i)_{i \geq 1}$, M_n belongs to the domain of attraction of Fréchet distribution, [269, Theorem 6.1] establishes that

$$\mathbb{P}(M_n \leq x n^{\alpha/(1-\alpha)}) \to \mathbb{E}\big[\exp\big(-\alpha T_{\alpha,\theta}^{-1} x^{-(1-\alpha)/\alpha}\big)\big]$$

so we see that $N(\log n)$ and M_n do not have the same asymptotic behavior as in the $\alpha = 0$ case, although they share the same growth rate $n^{\alpha/(1-\alpha)}$. Finally, the non correspondence of the asymptotic distribution of M_n and K_n suggests that the behavior of $M_n - K_n$ is radically different with respect to the $\alpha = 0$ case.

2.1.5 Discussion

In this paper we have studied stochastic approximations of the Pitman–Yor process consisting in the truncation of the sequence of stick-breaking frequencies at a random stopping time $\tau(\epsilon)$ that controls the accuracy of the approximation in the total variation distance by ϵ. We name this finite dimensional approximation the ϵ-Pitman–Yor process. We have derived the asymptotic distribution of $\tau(\epsilon)$ as ϵ goes to zero and we have advanced its use to devise a sampling scheme that generates the stopping time first, and then the frequencies up to that point. The simulations in Section 2.1.3 show that the proposed sampler proves computationally very efficient in the moderate to large stopping time regime (for approximately $\tau(\epsilon) \geq 20$). The asymptotic distribution illustrates how large the stopping time is as the approximation error gets small in terms of the prior parameters θ and α. In particular, it shows that the distribution of $\tau(\epsilon)$ in the Dirichlet process case is not recovered in the limit $\alpha \to 0$ in Theorem 2.1.2. In fact, in the Dirichlet process case $\tau(\epsilon)$ grows at a logarithmic rate in $1/\epsilon$ while in Pitman–Yor case it grows at the polynomial rate $\epsilon^{\alpha/(1-\alpha)}$ and the first regime is not recovered by letting α approach 0 in the second regime. We have also drawn important connections with the theory of random partition structures developed by Jim Pitman and coauthors which highlight the relationship of the the stopping time $\tau(\epsilon)$ with the number K_n of unique values in a sample of size n from the Pitman–Yor process.

We have left as open problem for future research the study of the conditional distribution of the stick-breaking frequencies given the stopping time. In the Dirichlet process case one can exploit the renewal process interpretation to generate exactly from this conditional distribution. In fact, when $\alpha = 0$, the sequence $(-\log R_n)_{n \geq 1}$ corresponds to the jump times of a Poisson process and the conditional distribution of the jumps given the number of jumps at time t can be described in terms of the ordered statistics of i.i.d. uniform random variates on $(0, t)$. The case $\alpha > 0$ does not

seem to be easily tractable, as it would be if the counting process associated to $\tau(\epsilon)$ were a mixed sample process or, equivalently, a Cox process, cf. Section 6.3 in [141].

It would be also interesting to compare the accuracy of our finite dimensional approximation of the Pitman–Yor process to the one proposed in [2]. The latter is based on a representation of the frequencies in decreasing order, cf. [Proposition 22 270]. [2] compare the accuracy of their approximation scheme to a stick-breaking truncation at a number n of stick-breaking frequencies that matches the number of frequencies used in their scheme. Not surprisingly, their approximation is superior since it generates weights in decreasing order, specially when α is large. In contrast, Theorem 2.1.2 describes precisely how large the truncation threshold n should be as α gets large for a given approximation level ϵ, cf. the center panel of Figure 2.1. It also underlines that the approximation deteriorates for fixed n and increasing α, which is coherent with the findings in [2]. A fair comparison with their approach can only be done for a given nominal approximation error, but unfortunately the authors did not provide a precise assessment of it. The number of stick-breaking frequencies needed to match the approximation accuracy of [2] would be *de facto* larger due to the non monotonicity. However, since the stopping rule (2.4) adapts to the size of α, we do not expect the accuracy of our approximation scheme to deteriorate for α large. As for computation time, the techniques used by [2] in order to obtain decreasing frequencies are computational heavy. Their average computing time for $\alpha = 0.5$ is about 2.30 seconds/iteration with 10^4 locations. This amounts to 0.23 *milli*seconds/support point, which is 1000 times slower than the computing time for our Algorithm 2 in the parameter configuration $\alpha = 0.5$, $\theta = 10$ and $\epsilon = 0.01$, equal to 0.23 *micro*second/support point. It would be interesting to investigate what are the consequences in terms of computation time per iteration for a given approximation error.

2.1.6 Appendix: Random generation of $T_{\alpha,\theta}$

Let Y be a standard $\mathrm{Exp}(1)$ random variable. Note that

$$\begin{aligned}\mathbb{P}((Y/T_\alpha)^\alpha > x) &= \int_0^\infty \mathbb{P}(Y > x^{1/\alpha}t) f_\alpha(t)\mathrm{d}t = \int_0^\infty \exp[-x^{1/\alpha}t] f_\alpha(t)\mathrm{d}t \\ &= \mathbb{E}\left[\mathrm{e}^{-x^{1/\alpha}T_\alpha}\right] = \mathrm{e}^{-x} = \mathbb{P}(Y > x)\end{aligned}$$

so we have $Y =_d (Y/T_\alpha)^\alpha$. For $r < \alpha$, $\mathbb{E}(Y^{-r/\alpha}) < \infty$, so we find that $\mathbb{E}(Y^{-r/\alpha}) = \mathbb{E}(T_\alpha^r)\mathbb{E}(Y^{-r})$ and

$$\mathbb{E}(T_\alpha^r) = \frac{\mathbb{E}(Y^{-r/\alpha})}{\mathbb{E}(Y^{-r})} = \frac{\Gamma(1-r/\alpha)}{\Gamma(1-r)}. \tag{2.13}$$

The normalizing constant in $f_{\alpha,\theta}(t)$ is $\int_0^\infty t^{-\theta} f_\alpha(t)\mathrm{d}t = \mathbb{E}(T_\alpha^{-\theta})$, so set $r = -\theta$ and note that $-\theta < \alpha$. Let G_a be a gamma random variable with shape $a > 0$ and unit rate. Simple moment comparisons using (2.13) yield the distributional equality $G_{1+\theta/\alpha} \overset{d}{=} (G_{1+\theta}/T_{\alpha,\theta})^\alpha$, which, however, does not provide a way to generate from $T_{\alpha,\theta}$. For this we resort to [98]. First we recall how to generate a Zolotarev random

variable $Z_{\alpha,b}$ for $\alpha \in (0,1)$ and $b = \theta/\alpha > -1$. Let

$$C = \frac{\Gamma(1+b\alpha)\Gamma(1+b(1-\alpha))}{\pi\Gamma(1+b)}$$

and

$$B(u) = A(u)^{-(1-\alpha)} = \frac{\sin(u)}{\sin(\alpha u)^{\alpha}\,\sin((1-\alpha)u)^{1-\alpha}}.$$

A simple asymptotic argument yields the value $B(0) = \alpha^{-\alpha}(1-\alpha)^{-(1-\alpha)}$. Then $f(x) = C\,B(x)^b, \quad 0 \le x \le \pi$. The following bound holds

$$f(x) \le CB(0)^b e^{-\frac{x^2}{2\sigma^2}}, \quad \text{with } \sigma^2 = \frac{1}{b\alpha(1-\alpha)}.$$

This Gaussian upper bound suggests a simple rejection sampler for sampling Zolotarev random variates. Following [98], it is most efficient to adapt the sampler to the value of σ. If $\sigma \ge \sqrt{2\pi}$, rejection from a uniform random variate is best. Otherwise, use a normal dominating curve as suggested in the bound above. The details are given below.

ALGORITHM 3 (Sampler of $T_{\alpha,\theta}$)

1. set $b = \theta/\alpha$ and $\sigma = \sqrt{b\alpha(1-\alpha)}$

2. if $\sigma \ge \sqrt{2\pi}$:
   then repeat: generate $U \sim \text{Unif}(0,\pi)$ and $V \sim \text{Unif}(0,1)$.
   set $X \leftarrow U$, $W \leftarrow B(X)$.
   until $V \le (W/B(0))^b$
   else repeat: generate $N \sim \text{N}(0,1)$ and $V \sim \text{Unif}(,1)$.
   set $X \leftarrow \sigma|N|$, $W \leftarrow B(X)$.
   until $X \le \pi$ and $Ve^{-N^2/2} \le (W/B(0))^b$

3. generate $G \stackrel{d}{=} G_{1+b(1-\alpha)/\alpha}$

4. set $T \leftarrow 1/(WG^{1-\alpha})^{1/\alpha}$

5. return T

[A8] J. Arbel and I. Prünster. A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17, 2017

[P9] J. Arbel and I. Prünster. Truncation error of a superposed gamma process in a decreasing order representation. *NeurIPS Advances in Approximate Bayesian Inference workshop*, 2016

[C4] J. Arbel and I. Prünster. *Bayesian Statistics in Action*, chapter On the truncation error of a superposed gamma process, pages 11–19. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017

2.2 Truncation-based approximations: normalized random measures

2.2.1 Introduction

Independent increment processes or, more generally, completely random measures (CRMs) are ubiquitous in modern stochastic modeling and inference. They form the basic building block of countless popular models in, e.g., Finance, Biology, Reliability, Survival Analysis. Within Bayesian nonparametric statistics they play a pivotal role. The Dirichlet process, the cornerstone of the discipline introduced in [117], can be obtained as normalization or exponentiation of suitable CRMs [see 118]. Moreover, as shown in [213], CRMs can be seen as the unifying concept of a wide variety of Bayesian nonparametric models. See also [182]. The concrete implementation of models based on CRMs often requires to simulate their realizations. Given they are discrete infinite objects, $\sum_{i\geq 1} J_i \delta_{Z_i}$, some kind of truncation is required, producing an approximation error $\sum_{i\geq M+1} J_i \delta_{Z_i}$. Among the various representations useful for simulating realizations of CRMs the method due to [119] and popularized by [334] stands out in that, for each realization, the weights J_i's are sampled in decreasing order. This clearly implies that for a given truncation level M the approximation error over the whole sample space is minimized. The appealing feature of decreasing jumps has lead to a huge literature exploiting the Ferguson & Klass algorithm. Limiting ourselves to recall contributions within Bayesian Nonparametrics we mention, among others, [34, 43, 94, 111, 149, 148, 251, 249, 252, 248, 247]. General references dealing with the simulation of Lévy processes include [289] and [83], who review the Ferguson & Klass algorithm and the compound Poisson process approximation to a Lévy process.

However, the assessment of the quality of the approximation due to the truncation for general CRMs is limited to some heuristic criteria. For instance, [43] implement the Ferguson & Klass algorithm for mixture models by using the so called *relative error* index. The corresponding stopping rule prescribes to truncate when the relative size of an additional jump is below a pre-specified fraction of the sum of sampled jumps. The inherent drawbacks of such a procedure and related heuristic threshold-type procedures employed in the several of the above references is two-fold. On the one hand the threshold is clearly arbitrary without quantifying the total mass of the ignored jumps. On the other hand the total mass of the jumps beyond the

threshold, i.e. the approximation error, can be very different for different CRMs or, even, for the same CRM with different parameter values; this implies that the same threshold can produce very different approximation errors in different situations. Starting from similar concerns about the quality of the approximation, the recent paper by [148] adopts an algorithmic approach and proposes an adaptive truncation sampler based on sequential Monte Carlo for infinite mixture models based on normalized random measures and on stick-breaking priors. The measure of discrepancy that is used in order to assess the convergence of the sampler is based on the effective sample size (ESS) calculated over the set of particles: the algorithm is run until the absolute value of the difference between two consecutive ESS gets under a pre-specified threshold. Also motivated by the same concerns, [34, 35] adopt an interesting approach to circumvent the problem of truncation by changing the model in the sense of replacing the CRM part of their model with a Poisson process approximation, which having an (almost surely) finite number of jumps can be sampled exactly. However, this leaves the question of the determination of the quality of approximation for truncated CRMs open. Another line of research, originated by [164], is dedicated to validating the trajectories from the point of view of the marginal density of the observations in mixture models. In this context, the quality of the approximation is measured by the L_1 distance between the marginal densities under truncated and non-truncated priors. Recent interesting contributions in this direction include bounds for a Ferguson & Klass representation of the beta process [101] and bounds for the beta process, the Dirichlet process as well as for arbitrary CRMs in a *size biased representation* [258, 69].

This paper faces the problem by a simple yet effective idea. In contrast to the above strategies, our approach takes all jumps of the CRMs into account and hence leads to select truncation levels in a principled way, which vary according to the type of CRM and its parameters. The idea is as follows: given moments of CRMs are simple to compute, one can quantify the quality of the approximation by evaluating some measure of discrepancy between the actual moments of the CRM at issue (which involve all its jumps) and the "empirical" moments, i.e. the moments computed based on the truncated sampled realizations of the CRM. By imposing such a measure of discrepancy not to exceed a given threshold and selecting the truncation level M large enough to achieve the desired bound, one then obtains a set of "validated" realizations of the CRM, or, in other terms, satisfying a moment-matching criterion. An important point to stress is that our validation criterion is all-purpose in spirit since it aims at validating the CRM samples themselves rather than samples of a transformation of the CRM. Clearly the latter type of validation would be ad hoc, since it would depend on the specific model. For instance, with the very same set of moment-matching realizations of a gamma process, one could obtain a set of realizations of the Dirichlet process via normalization and a set gamma mixture hazards by combination with a suitable kernel. Moreover, given moments of transformed CRMs are typically challenging to derive, a moment-matching strategy would not be possible in most cases. Hence, while the quantification of the approximation error does not automatically translate to transformed CRMs, one can still be confident that the moment-matching output at the CRM level produces good approximations. That this is indeed the case is explicitly shown in some practical examples both for prior and posterior quantities in Section 2.2.3.

2.2.2 Ferguson and Klass algorithm for completely random measures

Definition and main properties

Let $\mathscr{M}_{\mathbb{X}}$ be the set of boundedly finite measures on $\mathbb{X}$, which means that if $\mu \in \mathscr{M}_{\mathbb{X}}$ then $\mu(A) < \infty$ for any bounded set A. $\mathbb{X}$ is assumed to be a complete and separable metric space and both $\mathbb{X}$ and $\mathscr{M}_{\mathbb{X}}$ are equipped with the corresponding Borel σ-algebras. See [90] for details.

Definition 2.2.1. *A random element $\tilde{\mu}$, defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $\mathscr{M}_{\mathbb{X}}$, is called a* completely random measure *(CRM) if, for any collection of pairwise disjoint sets $A_1, \ldots, A_n$ in $\mathbb{X}$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent.*

An important feature is that a CRM $\tilde{\mu}$ selects (almost surely) discrete measures and hence can be represented as

$$\tilde{\mu} = \sum_{i \geq 1} J_i \delta_{Z_i} \tag{2.14}$$

where the jumps J_i's and locations Z_i's are random and independent. In (2.14) and throughout we assume there are no fixed points of discontinuity a priori. The main technical tool for dealing with CRMs is given by their Laplace transform, which admits a simple structural form known as Lévy–Khintchine representation. In fact, the Laplace transform of $\tilde{\mu}(A)$, for any A in $\mathbb{X}$, is given by

$$L_A(u) = \mathbb{E}\left[\mathrm{e}^{-\lambda \tilde{\mu}(A)}\right] = \exp\left\{ - \int_{\mathbb{R}^+ \times A} \left[1 - \mathrm{e}^{-\lambda v}\right] \nu(\mathrm{d}v, \mathrm{d}x) \right\} \tag{2.15}$$

for any $\lambda > 0$. The measure ν is known as *Lévy intensity* and uniquely characterizes $\tilde{\mu}$. In particular, there corresponds a unique CRM $\tilde{\mu}$ to any measure ν on $\mathbb{R}^+ \times \mathbb{X}$ satisfying the integrability condition

$$\int_B \int_{\mathbb{R}^+} \min\{v, 1\} \nu(\mathrm{d}v, \mathrm{d}x) < \infty \tag{2.16}$$

for any bounded B in $\mathbb{X}$. From an operational point of view this is extremely useful, since a single measure ν encodes all the information about the jumps J_i's and the locations Z_i's. The measure ν will be conveniently rewritten as

$$\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v|x)\alpha(\mathrm{d}x), \tag{2.17}$$

where ρ is a transition kernel on $\mathbb{R}^+ \times \mathbb{X}$ controlling the jump intensity and α is a measure on $\mathbb{X}$ determining the locations of the jumps. If ρ does not depend on x, the CRM is said homogeneous, otherwise it is non-homogeneous.

We now introduce two popular examples of CRMs that we will serve as illustrations throughout the paper.

Example 2.2.1. *The generalized gamma process introduced by [63] is characterized by a Lévy intensity of the form*

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\mathrm{e}^{-\theta v}}{\Gamma(1-\gamma) v^{1+\gamma}} \mathrm{d}v\, \alpha(\mathrm{d}x), \tag{2.18}$$

whose parameters $\theta \geq 0$ and $\gamma \in [0, 1)$ are such that at least one of them is strictly positive. Notable special cases are: (i) the gamma *CRM which is obtained by setting $\gamma = 0$; (ii) the*

inverse-Gaussian CRM, which arises by fixing $\gamma = 0.5$; (iii) the stable *CRM which corresponds to $\theta = 0$. Moreover, such a CRM stands out for its analytical tractability. In the following we work with $\theta = 1$, a choice which excludes the stable CRM. This is justified in our setting because the moments of the stable process do not exist. See Remark 2.2.1.*

Example 2.2.2. *The stable-beta process, or three-parameter beta process, was defined by [319] as an extension of the beta process [156]. Its jump sizes are upper-bounded by 1 and its Lévy intensity on $[0,1] \times \mathbb{X}$ is given by*

$$\nu(\dot{v}, \dot{x}) = \frac{\Gamma(c+1)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} v^{-\sigma-1}(1-v)^{c+\sigma-1} \dot{v}\, \alpha(\dot{x}), \tag{2.19}$$

where $\sigma \in [0,1)$ is termed discount parameter and $c > -\sigma$ concentration parameter. When $\sigma = 0$, the stable-beta process reduces to the beta CRM of [156]. Moreover, if $c = 1 - \sigma$, it boils down to a stable CRM where the jumps larger than 1 are discarded.

Moments of a CRM

For any measurable set A of $\mathbb{X}$, the n-th (raw) moment of $\tilde{\mu}(A)$ is defined by

$$m_n(A) = \mathbb{E}\left[\tilde{\mu}^n(A)\right].$$

In the sequel the multinomial coefficient is denoted by $\binom{n}{k_1 \cdots k_n} = \frac{n!}{k_1! \ldots k_n!}$. In the next proposition we collect known results about moments of CRMs which are crucial for our methodology.

Proposition 2.2.1. *Let $\tilde{\mu}$ be a CRM with Lévy intensity $\nu(\dot{v}, \dot{x})$. Then the i-th cumulant of $\tilde{\mu}(A)$, denoted by $\kappa_i(A)$, is given by*

$$\kappa_i(A) = \int_{\mathbb{R}^+ \times A} v^i \nu(\dot{v}, \dot{x}),$$

which, in the homogeneous case $\nu(\dot{v}, \dot{x}) = \rho(\dot{v})\alpha(\dot{x})$, simplifies to

$$\kappa_i(A) = \alpha(A) \int_0^\infty v^i \rho(\dot{v}).$$

The n-th moment of $\tilde{\mu}(A)$ is given by

$$m_n(A) = \sum_{(*)} \binom{n}{k_1 \cdots k_n} \prod_{i=1}^{n} \left(\kappa_i(A)/i!\right)^{k_i},$$

where the sum $(*)$ is over all n-tuples of nonnegative integers $(k_1, \ldots, k_n)$ satisfying the constraint $k_1 + 2k_2 + \cdots + nk_n = n$.

In the following we focus on (almost surely) finite CRMs i.e. $\tilde{\mu}(\mathbb{X}) < \infty$. This is motivated by the fact that most Bayesian nonparametric models, but also models in other application areas, involve finite CRMs. Hence, we assume that the measure α in (2.16) is finite i.e. $\alpha(\mathbb{X}) := a \in (0, \infty)$. This is a sufficient condition for $\tilde{\mu}(\mathbb{X}) < \infty$ in the non-homogeneous case and also necessary in the homogeneous case [see e.g. 282]. A common useful parametrization of α is then given as aP^* with P^* a probability measure and a a finite constant. Note that, if $\tilde{\mu}(\mathbb{X}) = \infty$, one could still

CRM	Cumulants	Moments			
	κ_i	m_1	m_2	m_3	m_4
G	$a(i-1)!$	a	$a_{(2)}$	$a_{(3)}$	$a_{(4)}$
IG	$a(1/2)_{(i-1)}$	a	$a^2+\frac{1}{2}a$	$a^3+\frac{3}{2}a^2$ $+\frac{3}{4}a$	a^4+3a^3 $+\frac{15}{4}a^2+\frac{15}{8}a$
GG	$a(1-\gamma)_{(i-1)}$	a	$a^2+a(1-\gamma)$	$a^3+3a^2(1-\gamma)$ $+a(1-\gamma)_{(2)}$	$a^4+6a^3(1-\gamma)$ $+a^2(1-\gamma)(11-7\gamma)+a(1-\gamma)_{(3)}$
B	$a\frac{(i-1)!}{(c+1)_{(i-1)}}$	a	$a^2+\frac{a}{c+1}$	$a^3+\frac{3a^2}{c+1}$ $+\frac{2a}{(c+1)_{(2)}}$	$a^4+\frac{6a^3}{c+1}+\frac{8a^2}{(c+1)_{(2)}}$ $+\frac{3a^2}{(c+1)^2}+\frac{6a}{(c+1)_{(3)}}$
SB	$a\frac{(1-\sigma)_{(i-1)}}{(c+1)_{(i-1)}}$	a	$a^2+a\frac{1-\sigma}{c+1}$	$a^3+3a^2\frac{1-\sigma}{c+1}$ $+a\frac{(1-\sigma)_{(2)}}{(c+1)_{(2)}}$	$a^4+6a^3\frac{1-\sigma}{c+1}+4a^2\frac{(1-\sigma)_{(2)}}{(c+1)_{(2)}}$ $+3a^2\frac{(1-\sigma)^2}{(c+1)^2}+a\frac{(1-\sigma)_{(3)}}{(c+1)_{(3)}}$

TABLE 2.6: Cumulants and first four moments of the random total mass $\tilde{\mu}(\mathbb{X})$ for the gamma (G), inverse-Gaussian (IG), generalized gamma (GG), beta (B) and stable-beta (SB) CRMs.

identify a bounded set of interest A and the whole following analysis carries over by replacing $\tilde{\mu}(\mathbb{X})$ with $\tilde{\mu}(A)$.

As we shall see in Section 2.2.2, the key quantity for evaluating the truncation error is given by the random total mass of the CRM, $\tilde{\mu}(\mathbb{X})$. Proposition 2.2.1 shows how the moments $m_n = m_n(\mathbb{X})$ can be obtained from the cumulants $\kappa_i = \kappa_i(\mathbb{X})$ and, in particular, the relations between the first four moments and the cumulants are

$$m_1 = \kappa_1,\ m_2 = \kappa_1^2+\kappa_2,\ m_3 = \kappa_1^3+3\kappa_1\kappa_2+\kappa_3,\ m_4 = \kappa_1^4+6\kappa_1^2\kappa_2+4\kappa_1\kappa_3+3\kappa_2^2+\kappa_4.$$

With reference to the two examples considered in Section 2.2.2, in both cases the expected value of $\tilde{\mu}(\mathbb{X})$ is a, which explains the typical terminology *total mass parameter* attributed to a. For the generalized gamma CRM the variance is given by $\text{Var}(\tilde{\mu}(\mathbb{X})) = a(1-\gamma)$, which shows how the parameter γ affects the variability. Moreover, $\kappa_i = a(1-\gamma)_{(i-1)}$ with $x_{(k)} = x(x+1)\dots(x+k-1)$ denoting the ascending factorial. As for the stable-beta CRM, we have $\text{Var}(\tilde{\mu}(\mathbb{X})) = a\frac{1-\sigma}{c+1}$ with both discount and concentration parameter affecting the variability, and also $\kappa_i = a\frac{(1-\sigma)_{(i-1)}}{(1+c)_{(i-1)}}$. Table 2.6 summarizes the cumulants κ_i and moments m_n for the random total mass $\tilde{\mu}(\mathbb{X})$ for the generalized gamma (assuming as in Example 2.2.1 $\theta = 1$), stable-beta CRMs and some of their special cases.

Remark 2.2.1. The *stable* CRM, which can be derived from the generalized gamma CRM by setting $\theta = 0$, does not admit moments. Hence, it cannot be included in our moment-matching methodology. However, the *stable* CRM with jumps larger than 1 discarded, derived from the stable-beta process by setting $c = 1-\sigma$, has all moments. Moreover, even when working with the standard stable CRM, posterior

quantities typically involve an exponential updating of the Lévy intensity [see 213], which makes the corresponding moments finite. This then allows to apply the moment matching methodology to the posterior.

Ferguson & Klass algorithm

For notational simplicity we present the Ferguson & Klass algorithm for the case $\mathbb{X} = \mathbb{R}$. However, note that it can be readily extended to more general Euclidean spaces [see e.g. 257]. Given a CRM

$$\tilde{\mu} = \sum_{i=1}^{\infty} J_i \delta_{Z_i}, \tag{2.20}$$

the Ferguson & Klass representation consists in expressing random jumps J_i occurring at random locations Z_i in terms of the underlying Lévy intensity.

In particular, the random locations Z_i, conditional on the jump sizes J_i, are obtained from the distribution function $F_{Z_i|J_i}$ given by

$$F_{Z_i|J_i}(s) = \frac{\nu(\mathrm{d}J_i, (-\infty, s])}{\nu(\mathrm{d}J_i, \mathbb{R})}.$$

In the case of a homogeneous CRM with Lévy intensity $\nu(\dot{\mathrm{v}}, \dot{\mathrm{x}}) = \rho(\dot{\mathrm{v}})\, aP^*(\dot{\mathrm{x}})$, the jumps are independent of the locations and, therefore $F_{Z_i|J_i} = F_{Z_i}$ implying that the locations are i.i.d. samples from P^*.

As far as the random jumps are concerned, the representation produces them in decreasing order, that is, $J_1 \geq J_2 \geq \cdots$. Indeed, they are obtained as $\xi_i = N(J_i)$, where $N(v) = \nu([v, \infty), \mathbb{R})$ is a decreasing function, and $\xi_1, \xi_2, \ldots$ are jump times of a standard Poisson process (PP) of unit rate i.e. $\xi_1, \xi_2 - \xi_1, \ldots \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$. Therefore, the J_i's are obtained by solving the equations $\xi_i = N(J_i)$. In general, this is achieved by numerical integration, e.g., relying on quadrature methods [see, e.g. 68]. For specific choices of the CRM, it is possible to make the equations explicit or at least straightforward to evaluate. For instance, if $\tilde{\mu}$ is a generalized gamma process (see Example 2.2.1), the function N takes the form

$$N(v) = \frac{a}{\Gamma(1-\gamma)} \int_v^{\infty} \mathrm{e}^{-u} u^{-(1+\gamma)}\, \dot{\mathrm{u}} = \frac{a}{\Gamma(1-\gamma)} \Gamma(v; -\gamma), \tag{2.21}$$

with $\Gamma(\,\cdot\,;\,\cdot\,)$ indicating an incomplete gamma function. If $\tilde{\mu}$ is the stable-beta process, one has

$$N(v) = a \frac{\Gamma(c+1)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \int_v^1 u^{-\sigma-1}(1-u)^{c+\sigma-1}\, \dot{\mathrm{u}} = a \frac{\Gamma(c+1)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} B(1-v; c+\sigma, -\sigma), \tag{2.22}$$

where $B(\,\cdot\,;\,\cdot\,,\,\cdot\,)$ denotes the incomplete beta function.

Hence, the Ferguson & Klass algorithm can be summarized as follows.

Algorithm 2.2.1 Ferguson & Klass algorithm

1: Sample $\xi_i \sim \text{PP}$ for $i = 1, \dots, M$
2: Define $J_i = N^{-1}(\xi_i)$ for $i = 1, \dots, M$
3: Sample $Z_i \sim P^*$ for $i = 1, \dots, M$
4: Approximate $\tilde{\mu}$ by $\sum_{i=1}^{M} J_i \delta_{Z_i}$

Since it is impossible to sample an infinite number of jumps, approximate simulation of $\tilde{\mu}$ is in order. This becomes a question of determining the number M of jumps to sample in (2.20) leading to the truncation

$$\tilde{\mu} \approx \tilde{\mu}_M = \sum_{i=1}^{M} J_i \delta_{Z_i}, \tag{2.23}$$

with approximation error in terms of the un-sampled jumps equal to $\sum_{i=M+1}^{\infty} J_i$. The Ferguson & Klass representation has the key advantage of generating the jumps in decreasing order implicitly minimizing such an approximation error. Then, the natural path to determining the truncation level M would be the evaluation of the Ferguson & Klass tail sum

$$\sum_{i=M+1}^{\infty} N^{-1}(\xi_i). \tag{2.24}$$

[63, Theorem A.1] provided an upper bound for (2.24) in the generalized gamma case.

Moment-matching criterion

Our methodology for assessing the quality of approximation of the Ferguson & Klass algorithm consists in comparing the actual distribution of the random total mass $\tilde{\mu}(\mathbb{X})$ with its empirical counterpart, where by empirical distribution we mean the distribution obtained by the sampled trajectories, i.e. by replacing random quantities by Monte Carlo averages of their sampled trajectories. In particular, based on the fact that the first K moments carry much information about a distribution, theoretical and empirical moments of $\tilde{\mu}(\mathbb{X})$ are compared.

The infinite vector of jumps is denoted by $\boldsymbol{J} = (J_i)_{i=1}^{\infty}$ and a vector of jumps sampled by the Ferguson & Klass algorithm by $\boldsymbol{J}^{(l)} = (J_1^{(l)}, \dots, J_M^{(l)})$. Here, $l = 1, \dots, N_{\text{FK}}$ stands for the l-th iteration of the algorithm, i.e. for the l-th sampled realization. We then approximate the expectation $\mathbb{E}$ of a statistic of the jumps, say $S(\boldsymbol{J})$, by the following empirical counterpart, denoted by $\mathbb{E}_{\text{FK}}$,

$$\mathbb{E}\big[S(\boldsymbol{J})\big] \approx \mathbb{E}_{\text{FK}}\big[S(\boldsymbol{J})\big] := \frac{1}{N_{\text{FK}}} \sum_{l=1}^{N_{\text{FK}}} S\big(\boldsymbol{J}^{(l)}\big). \tag{2.25}$$

Note that there are two layers of approximation involved in (2.25): first, only a finite number of jumps M is used; second, the actual expected value is estimated through an empirical average which typically conveys on Monte Carlo error. The latter is not

the focus of the paper, so we take a large enough number of trajectories, $N_{\text{FK}} = 10^4$, in order to insure a limited Monte Carlo error of the order of 0.01. We focus on the first approximation inherent to the Ferguson & Klass algorithm.

More specifically, as far as moments are concerned, $\boldsymbol{m}_K = (m_1, \ldots, m_K)$ denotes the first K moments of the random total mass $\tilde{\mu}(\mathbb{X}) = \sum_{i=1}^{\infty} J_i$ provided in Section 2.2.2 and $\hat{\boldsymbol{m}}_K = (\hat{m}_1, \ldots, \hat{m}_K)$ indicates the first K empirical moments given by

$$\hat{m}_n = \mathbb{E}_{\text{FK}}\left[\left(\sum_{i=1}^{M} J_i\right)^n\right]. \tag{2.26}$$

As measure of discrepancy between theoretical and empirical moments, a natural choice is given by the mean squared error between the vectors of moments or, more precisely, between the n-th roots of theoretical and empirical moments

$$\ell = \ell(\boldsymbol{m}_K, \hat{\boldsymbol{m}}_K) = \left(\frac{1}{K}\sum_{n=1}^{K}\left(m_n^{1/n} - \hat{m}_n^{1/n}\right)^2\right)^{1/2}. \tag{2.27}$$

When using the Ferguson & Klass representation for computing the empirical moments the index ℓ depends on the truncation level M and we highlight such a dependence by using the notation ℓ_M. Of great importance is also a related quantity, namely the number of jumps necessary for achieving a given level of precision, which essentially consists in inverting ℓ_M and is consequently denoted by $M(\ell)$.

The index of discrepancy (2.27) clearly also depends on K, the number of moments used to compute it and $1/K$ in (2.27) normalizes the indices in order to make them comparable as K varies. A natural question is then about the sensitivity of (2.27) w.r.t. K. It is desirable for ℓ_M to capture fine variations between the theoretical and empirical distributions, which is assured for large K. In extensive simulation studies not reported here we noted that increasing K in the range $\{1, \ldots, 10\}$ makes the index increase and then plateau and this holds for all processes and parameter specifications used in the paper. Recalling also the whole body of work by Pearson on eponymous curves, which shows that the knowledge of four moments suffices to cover a large number of known distributions, we adhere to his rule of thumb and choose $K = 4$ in our analyses. On the one hand it is a good compromise between targeted precision of the approximation and speed of the algorithm. On the other hand it is straightforward to check the results as K varies in specific applications; for the ones considered in the following sections the differences are negligible.

In the literature several heuristic indices based on the empirical jump sizes around the level of truncation have been discussed [cf Remark 3 in 43]. Here, in order to compare such procedures with our moment criterion, we consider the relative error index which is based on the jumps themselves. It is defined as the expected value of the relative error between two consecutive partial sums of jumps. Its empirical counterpart is denoted by e_M and given by

$$e_M = \mathbb{E}_{\text{FK}}\left[\frac{J_M}{\sum_{i=1}^{M} J_i}\right]. \tag{2.28}$$

2.2.3 Applications to Bayesian Nonparametrics

In this section we concretely implement the proposed moment-matching Ferguson & Klass algorithm to several Bayesian nonparametric models. The performance in terms of both a priori and a posteriori approximation is evaluated. A comparison of the quality of approximation resulting from using (2.28) as benchmark index is provided.

A priori simulation study

We start by investigating the performance of the proposed moment-matching version of the Ferguson & Klass algorithm w.r.t. the CRMs defined in Examples 2.2.1 and 2.2.2, namely the generalized gamma and stable-beta processes. Figure 2.3 displays the behaviour of both the moment-matching distance ℓ_M (left panel) and the relative jumps' size index e_M (right panel) as the truncation level M increases. The plots, from top to bottom, correspond to: the generalized gamma process with varying γ and $a = 1$ fixed; the inverse-Gaussian process with varying total mass a (which is a generalized gamma process process with $\gamma = 0.5$); the stable-beta process with varying discount parameter σ and $a = 1$ fixed.

First consider the behaviour of the indices as the parameter specifications vary. It is apparent that, for any fixed truncation level M, the indices ℓ_M and e_M increase as each of the parameters a, γ or σ increases. For instance, roughly speaking, a total mass parameter a corresponds to sampling trajectories defined on the interval $[0, a]$ [see 282], and a larger interval worsens the quality of approximation for any given truncation level. Also it is natural that γ and σ impact in similar way ℓ_M and e_M given they stand for the "stable" part of the Lévy intensity. See first and third rows of Figure 2.3.

As far as the comparison between ℓ_M and e_M is concerned, it is important to note that e_M consistently downplays the error of approximation related to the truncation. This can be seen by comparing the two columns of Figure 2.3. ℓ_M is significantly more conservative than e_M for both the generalized gamma and the stable-beta processes, especially for increasing values of the parameters γ, a or σ. This indicates quite a serious issue related to e_M as a measure for the quality of approximation and one should be cautious when using it. In contrast, the moment-matching index ℓ_M matches more accurately the known behaviour of these processes as the parameters vary.

By reversing the viewpoint and looking at the truncation level $M(\ell)$ needed for achieving a certain error of approximation ℓ in terms of moment-match, the results become even more intuitive. We set $\ell = 0.1$ and computed $M(\ell)$ on a grid of size 20×20 with equally-spaced points for the parameters $(a, \gamma) \in (0, 2) \times (0, 0.8)$ for the generalized gamma process and $(a, c) \in (0, 2) \times (0, 30)$ for the beta process. Figure 2.4 displays the corresponding plots. In general, it is interesting to note that a limited number of jumps is sufficient to achieve good precision levels. Analogously to Figure 2.3, larger values of the parameters require a larger number of jumps to achieve a given precision level. In particular, when $\gamma > 0.5$, one needs to sample a

significantly larger number of jumps. For instance, in the generalized gamma process case, with $a = 1$, the required number of jumps increases from 28 to 53 when passing from $\gamma = 0.5$ to $\gamma = 0.75$. It is worth noting that for the normalized version of the generalized gamma process, to be discussed in Section 2.2.3 and quite popular in applications, the estimated value of γ rarely exceeds 0.75 in species sampling, whereas it is typically in the range $[0.2, 0.4]$ in mixture modeling.

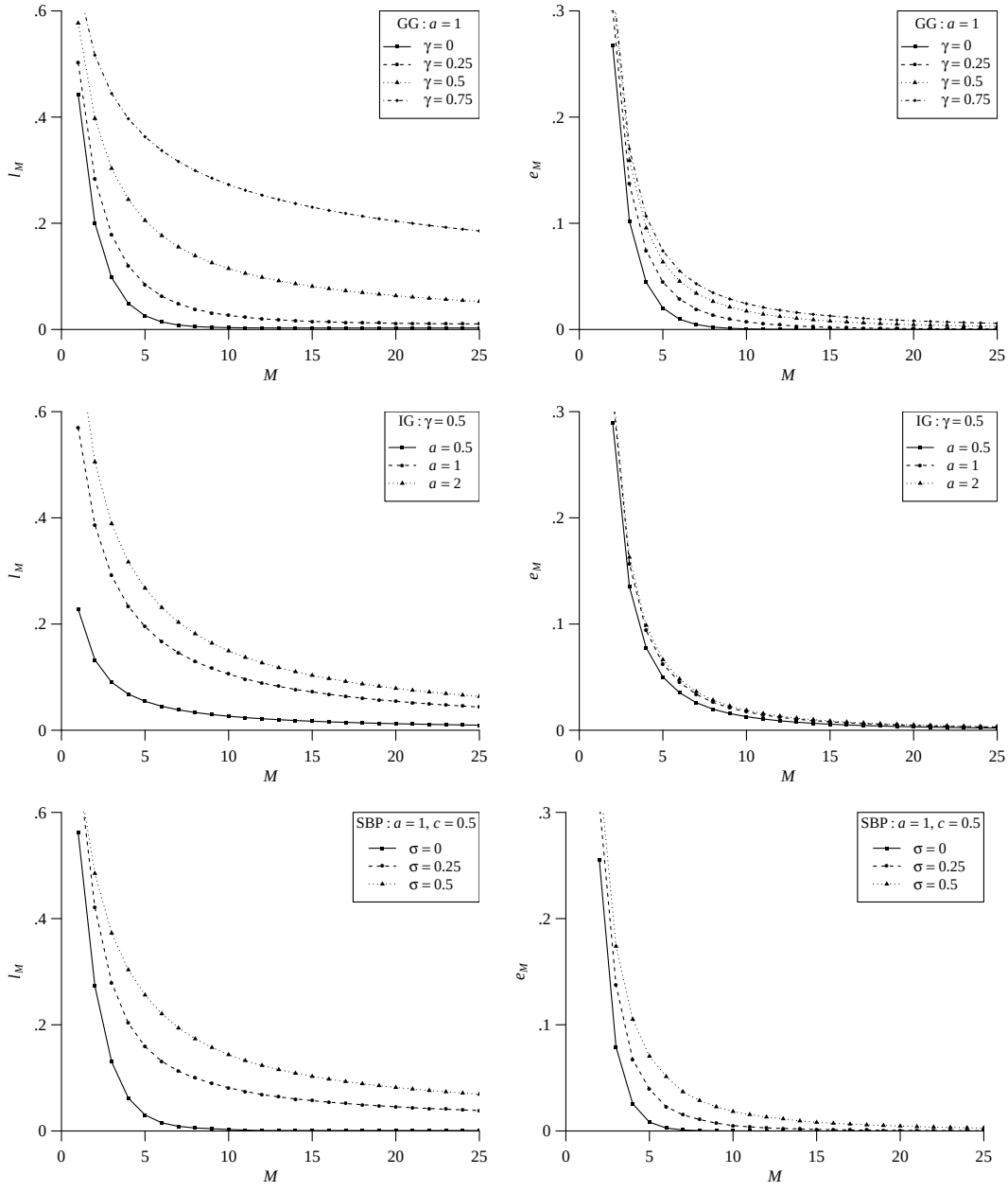
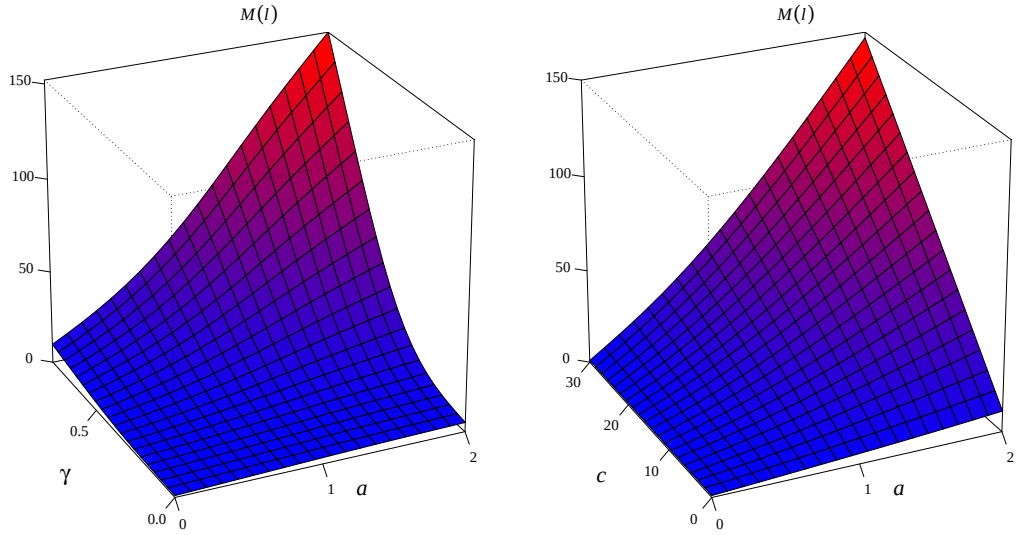


FIGURE 2.3: Left panel: ℓ_M as M varies; right panel: e_M as M varies. Top row: generalized gamma process (GG) with varying γ and $a = 1$ fixed; middle row: inverse-Gaussian process (IG), $\gamma = 0.5$, with varying total mass a; bottom row: stable-beta process (SBP) with $a = 1$, $c = 0.5$ fixed and varying discount parameter σ. The points are connected by straight lines only for visual simplification.



(A) generalized gamma process (B) beta process

FIGURE 2.4: Number of jumps $M(\ell)$ required to achieve a precision level of $\ell = 0.1$ for ℓ_M. Left panel: generalized gamma process for $a \in (0, 2)$ and $\gamma \in (0, 0.8)$. Right panel: beta process for $a \in (0, 2)$ and $c \in (0, 30)$.

Normalized random measures with independent increments

Having illustrated the behaviour of the moment-matching methodology for plain CRMs we now investigate it on specific classes of nonparametric priors, which typically involve a transformation of the CRM. Moreover, given their posterior distributions involve updated CRMs it is important to test the moment-matching Ferguson & Klass algorithm also on posterior quantities. The first class of models we consider are normalized random measures with independent increments (NRMI) introduced by [282]. Such nonparametric priors have been used as ingredients of a variety of models and in several application contexts. Recent reviews can be found in [213, 43].

If $\tilde{\mu}$ is a CRM with Lévy intensity (2.17) such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ (almost surely), then an NRMI is defined as

$$\tilde{P} = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}. \tag{2.29}$$

Particular cases of NRMI are then obtained by specifying the CRM in (2.29). For instance, by picking the generalized gamma process defined in Example 2.2.1 one obtains the normalized generalied gamma process, first used in a Bayesian context by [211].

Posterior Distribution of an NRMI The basis of any Bayesian inferential procedure is represented by the posterior distribution. In the case of NRMIs, the determination of the posterior distribution is a challenging task since one cannot rely directly on Bayes' theorem (the model is not dominated) and, with the exception of the Dirichlet process, NRMIs are not conjugate as shown in [171]. Nonetheless, a

posterior characterization has been established in [172] and it turns out that, even though NRMIs are not conjugate, they still enjoy a sort of "conditional conjugacy." This means that, conditionally on a suitable latent random variable, the posterior distribution of an NRMI coincides with the distribution of an NRMI having fixed points of discontinuity located at the observations. Such a simple structure suggests that when working with a general NRMI, instead of the Dirichlet process, one faces only one additional layer of difficulty represented by the marginalization with respect to the conditioning latent variable.

Before stating the posterior characterization to be used with our algorithm, we need to introduce some notation and basic facts. Let $(Y_n)_{n\geq 1}$ be an exchangeable sequence directed by an NRMI, i.e.

$$\begin{aligned} &Y_i|\tilde{P} \overset{\text{i.i.d.}}{\sim} \tilde{P}, \quad \text{for } i = 1, \dots, n, \\ &\tilde{P} \sim Q, \end{aligned} \tag{2.30}$$

with Q the law of NRMI, and set $\mathbf{Y} = (Y_1, \dots, Y_n)$. Due to the discreteness of NRMIs, ties will appear with positive probability in $\mathbf{Y}$ and, therefore, the sample information can be encoded by the $K_n = k$ distinct observations $(Y_1^*, \dots, Y_k^*)$ with frequencies $(n_1, \dots, n_k)$ such that $\sum_{j=1}^k n_j = n$. Moreover, introduce the nonnegative random variable U such that the distribution of $[U|\mathbf{Y}]$ has density, w.r.t. the Lebesgue measure, given by

$$f_{U|\mathbf{Y}}(u) \propto u^{n-1} \exp\{-\psi(u)\} \prod_{j=1}^k \tau_{n_j}(u|Y_j^*), \tag{2.31}$$

where $\tau_{n_j}(u|Y_j^*) = \int_0^\infty v^{n_j} \mathrm{e}^{-uv} \rho(v|Y_j^*)$ and ψ is the Laplace exponent of $\tilde{\mu}$ defined by $\psi(u) = -\log\big(L_{\mathbb{X}}(u)\big)$, cf (2.15). Finally, assume $P^* = \mathbb{E}[\tilde{P}]$ to be nonatomic.

Proposition 2.2.2 ([172])**.** *Let $(Y_n)_{n\geq 1}$ be as in (2.30) where $\tilde{P}$ is an NRMI defined in (2.29) with Lévy intensity as in (2.17). Then the posterior distribution of the unnormalized CRM $\tilde{\mu}$, given a sample $\mathbf{Y}$, is a mixture of the distribution of $[\tilde{\mu}|U, \mathbf{Y}]$ with respect to the distribution of $[U|\mathbf{Y}]$. The latter is identified by (2.31), whereas $[\tilde{\mu}|U = u, \mathbf{Y}]$ is equal in distribution to a CRM with fixed points of discontinuity at the distinct observations Y_j^*,*

$$\tilde{\mu}^* + \sum_{j=1}^k J_j^* \delta_{Y_j^*} \tag{2.32}$$

such that:

(a) $\tilde{\mu}^*$ is a CRM characterized by the Lévy intensity

$$\nu^*(v, x) = \mathrm{e}^{-uv} \nu(v, x), \tag{2.33}$$

(b) the jump height J_j^* corresponding to Y_j^* has density, w.r.t. the Lebesgue measure, given by

$$f_j^*(v) \propto v^{n_j} \mathrm{e}^{-uv} \rho(v|Y_j^*), \tag{2.34}$$

(c) $\tilde{\mu}^*$ and J_j^*, $j = 1, \dots, k$, are independent.

Moreover, the posterior distribution of the NRMI $\tilde{P}$, conditional on U, is given by

$$[\tilde{P}|U,\mathbf{Y}] \stackrel{d}{=} w\frac{\tilde{\mu}^*}{\tilde{\mu}^*(\mathbb{X})} + (1-w)\frac{\sum_{k=1}^{k} J_j^* \delta_{Y_j^*}}{\sum_{l=1}^{k} J_l^*}, \tag{2.35}$$

where $w = \tilde{\mu}^*(\mathbb{X})/(\tilde{\mu}^*(\mathbb{X}) + \sum_{l=1}^{k} J_l^*)$.

In order to simplify the notation, in the statement we have omitted explicit reference to the dependence on $[U|\mathbf{Y}]$ of both $\tilde{\mu}^*$ and $\{J_j^* : j = 1,\ldots,k\}$, which is apparent from (2.33) and (2.34). A nice feature of the posterior representation of Proposition 2.2.2 is that the only quantity needed for deriving explicit expressions for particular cases of NRMI is the Lévy intensity (2.17). For instance, in the case of the generalized gamma process, the CRM part $\tilde{\mu}^*$ in (2.32) is still a generalized gamma process characterized by a Lévy intensity of the form of (2.18)

$$\nu^*(\mathrm{v},\mathrm{y}) = \frac{\mathrm{e}^{-(1+u)v}}{\Gamma(1-\gamma)v^{1+\gamma}}\,\mathrm{v}\,aP^*(\mathrm{y}). \tag{2.36}$$

Moreover, the distribution of the jumps (2.34) corresponding to the fixed points of discontinuity Y_j^*'s in (2.32) reduces to a gamma distribution with density

$$f_j^*(v) = \frac{(1+u)^{n_j-\gamma}}{\Gamma(n_j-\gamma)} v^{n_j-\gamma-1}\mathrm{e}^{-(1+u)v}. \tag{2.37}$$

Finally, the conditional distribution of the non-negative latent variable U given $\mathbf{Y}$ (2.31) is given by

$$f_{U|\mathbf{Y}}(u) \propto u^{n-1}(u+1)^{k\gamma-n}\exp\left\{-\frac{a}{\gamma}(u+1)^{\gamma}\right\}. \tag{2.38}$$

The availability of this posterior characterization makes it then possible to determine several important quantities such as the predictive distributions and the induced partition distribution. See [172] for general NRMI and [211] for the subclass of the normalized generalized gamma process. See also [34] for another approach to approximate the normalized generalized gamma process with a finite number of jumps.

Moment-matching for posterior NRMI From (2.32) it is apparent that the posterior of the unnormalized CRM $\tilde{\mu}$, conditional on the latent variable U, is composed of the independent sum of a CRM $\tilde{\mu}^*$ and fixed points of discontinuity at the distinct observations Y_j^*. The part which is at stake here is obviously $\tilde{\mu}^*$ for which only approximate sampling is possible. As for the fixed points of discontinuities, they are independent from $\tilde{\mu}^*$ and can be sampled exactly, at least in special cases.

We focus on the case of the normalized generalized gamma process. By (2.33) the Lévy intensity of $\tilde{\mu}^*$ is obtained by exponentially tilting the Lévy intensity of the prior $\tilde{\mu}$. Hence, the Ferguson & Klass algorithm applies in the same way as for the prior. The sampling of the fixed points jumps is straightforward from the gamma distributions (2.37). As far as the moments are concerned, key ingredient of our

algorithm, the cumulants of $\tilde{\mu}^*$ are equal to $\kappa_i^* = a\frac{(1-\gamma)_{(i-1)}}{(u+1)^{i-\gamma}}$ and the corresponding moments are then obtained via Proposition 2.2.1.

Our simulation study is based on a sample of size $n = 10$. Such a small sample size is challenging in the sense that the data provide rather few information and the CRM part of the model is still prevalent. We examine three possible clustering configurations of the observations Y_i^*s: (i) $k = 1$ group, with $n_1 = 10$, (ii) $k = 3$ groups, with $n_1 = 1$, $n_2 = 3$, $n_3 = 6$, and (iii) $k = 10$ groups, with $n_j = 1$ for $j = 1, \ldots, 10$. First let us consider the behaviour of $f_{U|\mathbf{Y}}$, which is illustrated in Figure 2.5 for $n = 10$ and $k \in \{1, 2, \ldots, 10\}$. It is clear that the smaller the number of clusters, the more $f_{U|\mathbf{Y}}$ is concentrated on small values, and vice versa.

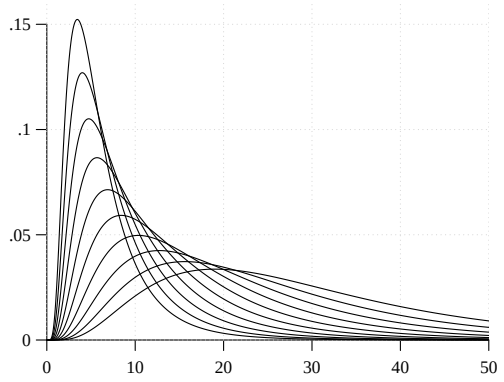


FIGURE 2.5: NGG posterior: density $f_{U|\mathbf{Y}}$ with $n = 10$ observations, $a = 1$, $\gamma = 0.5$, and number of clusters $k \in \{1, \ldots, 10\}$; $k = 1$ corresponds to the most peaked density and $k = 10$ to the flattest.

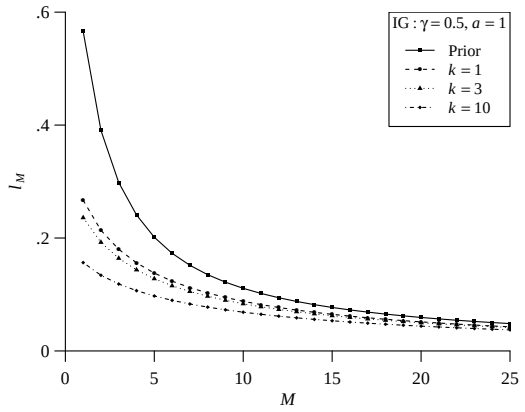
Now we consider $\tilde{\mu}^*(\mathbb{X})$, the random total mass corresponding to the CRM part of the posterior only given in (2.35). Such a quantity depends on U whose distribution is driven by the data $\mathbf{Y}$. In order to keep the presentation as neat as possible, and in the same time to remain consistent with the data, we choose to condition on $U = u$ for u equal to the mean of $f_{U|\mathbf{Y}}$, the most natural representative value. Given this, it is possible to run the Ferguson & Klass algorithm on the CRM part $\tilde{\mu}^*$ of the posterior and compute moment-matching index ℓ_M as the number of jumps varies. Figure 2.6 shows these results for the inverse-Gaussian CRM, a special case of the generalized gamma process corresponding to $\gamma = 0.5$. Such posteriors were sampled under the above mentioned $\mathbf{Y}$ clustering configuration scenarios (i)-(iii), which led to mean values of $U|\mathbf{Y}$ of, respectively, 6.3, 8.9 and 25.1. The plot also displays a comparison to the prior values of ℓ_M and indicates that for a given number of jumps the approximation error, measured in terms of ℓ_M, is smaller for the posterior CRM part $\tilde{\mu}^*$ w.r.t. to the prior CRM $\tilde{\mu}$.

Additionally, instead of considering only the CRM part $\tilde{\mu}^*$ of the posterior, one may be interested in the quality of the full posterior which includes also the fixed discontinuities. For this purpose we consider an index which is actually of interest in its own. In particular, we evaluate the relative importance of the CRM part w.r.t. the part corresponding to the fixed points of discontinuity in terms of the ratio $\mathbb{E}\big(\sum_{j=1}^k J_j^*\big) / \mathbb{E}\big(\tilde{\mu}^*(\mathbb{X})\big)$. Loosely speaking one can think of the numerator as the expected weight of the data and the denominator as the expected weight of the prior.

Recall that in the normalized generalized gamma process case, for a given pair (n,k) and conditional on $U=u$, the sum of fixed location jumps is a gamma$(n-k\gamma, u+1)$. Hence, the index becomes

$$\frac{\mathbb{E}\left(\sum_{j=1}^{k} J_j^* | U=u\right)}{\mathbb{E}\left(\tilde{\mu}^*(\mathbb{X})|U=u\right)} = \frac{(n-k\gamma)/(u+1)}{a/(u+1)^{1-\gamma}} = \frac{n-k\gamma}{a(u+1)^{\gamma}}. \tag{2.39}$$

By separately mixing the conditional expected values in (2.39) over $f_{U|\mathbf{Y}}$ (we use an adaptive rejection algorithm to sample from $f_{U|\mathbf{Y}}$) we obtained the results summarized in the table of Figure 2.6. We can appreciate that the fixed part typically overcomes (or is at least of the same order than) the CRM part, a phenomenon which uniformly accentuates as the sample size n increases. Returning to the original problem of measuring the quality of approximation in terms of moment matching, these findings make it apparent that the comparative results of Figure 2.6 between prior and posterior are conservative. In fact, if performing the moment-match on the whole posterior, i.e. including the fixed jumps which can be sampled exactly, the corresponding moment-matching index would, for any given truncation level M, indicate a better quality of approximation w.r.t. the index based solely on $\tilde{\mu}^*$. Note that computing the moments of $\tilde{\mu}^*(\mathbb{X}) + \sum_{i=1}^{k} J_i$ straightforward given the independence between $\tilde{\mu}^*$ and the fixed jumps J_i's and also among the jumps themselves. From a practical point of view the findings of this section suggest that a given quality of approximation ℓ in terms of moment-match for the prior represents an upper bound for the quality of approximation in the posterior.



	$\mathbb{E}\left(\sum_{j=1}^{k} J_j^*\right)/\mathbb{E}\left(\tilde{\mu}^*(\mathbb{X})\right)$		
$k \setminus n$	10	30	100
1	3.34	7.30	13.50
$\sqrt{n}$	2.65	4.68	6.05
n	0.89	0.98	0.99

FIGURE 2.6: Inverse-Gaussian process ($\gamma = 0.5$) with $a = 1$. Left: Moment-matching errors ℓ_M as the number of jumps M varies. ℓ_M corresponding to prior $\tilde{\mu}$ (continuous line) and posterior $\tilde{\mu}^*$ under $\mathbf{Y}$ clustering scenarios (i) (dashed line), (ii) (dotted line), (iii) (dotted-dashed line). Right: Index of relative importance $\mathbb{E}\left(\sum_{j=1}^{k} J_j^*\right)/\mathbb{E}\left(\tilde{\mu}^*(\mathbb{X})\right)$ for varying (n,k).

A note on the inconsistency for diffuse distributions In the context of Gibbs-type priors, of which the normalized generalized gamma process is a special case, [95]

showed that, if the data are generated from a "true" P_0, the posterior of $\tilde{P}$ concentrates at a point mass which is the linear combination

$$bP^*(\cdot) + (1-b)P_0(\cdot)$$

of the prior guess $P^* = \mathbb{E}(\tilde{P})$ and P_0. The weight b depends on the prior and, indirectly, on P_0, since P_0 dictates the rate at which the distinct observations k are generated. For a diffuse P_0, all observations are distinct and $k = n$ (almost surely). In the normalized generalized gamma process case this implies that $b = \gamma$ and hence the posterior is inconsistent since it does not converge to P_0. For the inverse-Gaussian process, i.e. with $\gamma = 0.5$, the posterior distribution gives asymptotically the same weight to P^* and P_0. The last row of the table of Figure 2.6, which displays the ratio $\mathbb{E}\big(\sum_{j=1}^{k} J_j^*\big)/\mathbb{E}\big(\tilde{\mu}^*(\mathbb{X})\big)$ for $k = n$, is an illustration of this inconsistency result since the ratio gets close to 1 as n grows. In contrast, when P_0 is discrete, which implies that k increases at a slower rate than n, one always has consistency. This is illustrated by the first two rows of the table of Figure 2.6, where one can appreciate that the ratio $\mathbb{E}\big(\sum_{j=1}^{k} J_j^*\big)/\mathbb{E}\big(\tilde{\mu}^*(\mathbb{X})\big)$ increases as n increases, giving more and more weight to the data. These findings suggest that consistency issues for general NRMI could be explored from new perspectives based on the study of the asymptotic behavior of $f_{U|\mathbf{Y}}$, which will be subject to future work.

Stable-beta Indian buffet process

The Indian buffet process (IBP), introduced in [130], is one of the most popular models for feature allocation and is closely connected to the beta process discussed in Example 2.2.2. In fact, when marginalizing out the Dirichlet process and considering the resulting partition distribution one obtains the well known Chinese restaurant process. Likewise, as shown in [320], when integrating out a beta process in a Bernoulli process (BeP) model one obtains the IBP. Recall that a Bernoulli process, with an atomic base measure $\tilde{\mu}$, is a stochastic process whose realizations are collections of atoms of mass 1, with possible locations given by the atoms of the base measure $\tilde{\mu}$. Such an atom is element of the collection with probability given by the jump size in $\tilde{\mu}$. Later, [319] generalized the construction and defined the stable-beta Indian buffet process as

$$\begin{aligned} &Y_i|\tilde{\mu} \overset{\text{i.i.d.}}{\sim} \text{BeP}(\tilde{\mu}) \quad \text{for } i = 1, \dots, n, \\ &\tilde{\mu}|c, \sigma, aP^* \sim \text{SBP}(c, \sigma, aP^*). \end{aligned} \tag{2.40}$$

Given the construction involves a CRM, it is clear that any conditional simulation algorithm will need to rely on some truncation for which we use our moment-matching Ferguson & Klass algorithm.

Posterior distribution in the IBP Let us consider a conditional iid sample $\mathbf{Y} = (Y_1, \dots, Y_n)$ as in (2.40). Note that due to the discreteness of $\tilde{\mu}$, ties appear with positive probability. We adopt the same notations for the ties Y_j^* and frequencies n_j as in Section 2.2.3. Then we can state the following result which highlights the posterior structure of the stable-beta process in the Indian buffet process.

Proposition 2.2.3 ([319]). *Let $(Y_n)_{n\geq 1}$ be as in (2.40). Then the posterior distribution of $\tilde{\mu}$ conditional on $\mathbf{Y}$ is given by the distribution of*

$$\tilde{\mu}^* + \sum_{j=1}^{k} J_j^* \delta_{Y_j^*}$$

where

(a) $\tilde{\mu}^*$ is a stable-beta process characterized by the Lévy intensity

$$\nu^*(v, x) = (1-v)^n \nu(v, x),$$

(b) the jump height J_j^* corresponding to Y_j^* is beta distributed

$$J_j^* \sim \text{beta}(n_j - \sigma, c + \sigma + n - n_j),$$

(c) $\tilde{\mu}^*$ and J_j^*, $j = 1, \ldots, k$, are independent.

Note that due to the polynomial tilting of ν by $(1-u)^n$ in *(a)* above, the CRM part $\tilde{\mu}^*$ is still a stable-beta process with updated parameters

$$c^* = c + n \text{ and } a^* = a\frac{(c+\sigma)_{(n)}}{(c+1)_{(n)}},$$

while the discount parameter σ remains unchanged.

Moment-matching for the IBP In order to implement the moment-matching methodology we first need to evaluate the posterior moments of the random total mass. For this purpose, we rely on the moments characterization in terms of the cumulants provided in Proposition 2.2.1. The cumulants κ_i^* of the CRM part $\tilde{\mu}^*(\mathbb{X})$ are obtained from Table 2.6 with the appropriate parameter updates which leads to

$$\kappa_i^* = a^* \frac{(1-\sigma)_{(i-1)}}{(1+c^*)_{(i-1)}} = a\frac{(1-\sigma)_{(i-1)}(c+\sigma)_{(n)}}{(1+c)_{(n+i-1)}}.$$

We consider two stable-beta processes: the beta process prior $\tilde{\mu} \sim \text{SBP}(c = 1, \sigma = 0, a = 1)$ and the stable-beta process prior $\tilde{\mu} \sim \text{SBP}(c = 1, \sigma = 0.5, a = 1)$. We let n vary in $\{5, 10, 20\}$. In contrast to the NRMI case, there is no need to work under different scenarios for the clustering profile of the data, since the posterior CRM $\tilde{\mu}^*$ is not affected by them with only the sample size entering the updating scheme. We compare the prior moment-match for $\tilde{\mu}$ with the posterior moment-match for $\tilde{\mu}^*$ in terms of our discrepancy index ℓ_M and the results are displayed in Figure 2.7. The comparison shows that there is a gain in precision between prior and posterior distributions in terms of ℓ_M suggesting that the a priori error level ℓ represents an upper bound for the posterior approximation error.

As in Section 2.2.3, we also evaluate the relative weights of fixed jumps and posterior CRM or, roughly, of the data w.r.t. the prior. Recalling that fixed location jumps J_j^* are independent and $\text{beta}(n_j - \sigma, c + \sigma + n - n_j)$ and some algebra allow to re-write

the ratio of interest as

$$\frac{\mathbb{E}\left(\sum_{j=1}^{k} J_j^*\right)}{\mathbb{E}\left(\tilde{\mu}^*(\mathbb{X})\right)} = \frac{(n-k\sigma)(c+1)_{(n-1)}}{a(c+\sigma)_{(n)}}.$$

Table 2.7 displays the corresponding values for different choices of n and k. As in the NRMI case, the fixed part overcomes the CRM part, which means that the data dominate the prior, and, moreover, their relative weight increases as n increases. In terms of moment-matching this shows that, if one looks at the overall posterior structure, the approximation error connected to the truncation is further dampened.

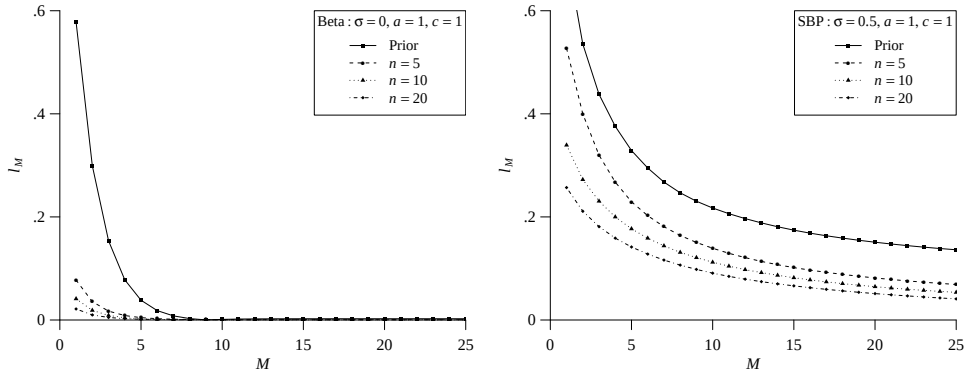


FIGURE 2.7: Moment-matching errors ℓ_M as the number of jumps M varies for the stable-beta process with $c = 1$, $a = 1$, and, respectively, $\sigma = 0$ (left panel) and $\sigma = 0.5$ (right panel). ℓ_M corresponding to prior $\tilde{\mu}$ (continuous line) and the posterior $\tilde{\mu}^*$ given with $n = 5$ (dashed line) and $n = 10$ (dotted line) and $n = 20$ (dashed-dotted line) observations.

	$\mathbb{E}\left(\sum_{j=1}^{k} J_j^*\right)/\mathbb{E}\left(\mu^*(\mathbb{X})\right)$		
$k \quad \backslash \quad n$	10	30	100
1	2.57	4.71	8.79
n^σ	2.28	4.36	8.39
n	1.35	2.40	4.41

TABLE 2.7: Stable-beta process with $\sigma = 0.5$, $c = 1$ and $a = 1$: Index of relative importance $\mathbb{E}\left(\sum_{j=1}^{k} J_j^*\right)/\mathbb{E}\left(\mu^*(\mathbb{X})\right)$ for varying (n, k).

2.2.4 Moment-matching criterion implementation for mixtures

We illustrate the use of the moment-matching strategy by implementing it within location-scale NRMI mixture models, which can be represented in hierarchical form

as

$$\begin{aligned} Y_i|\mu_i,\sigma_i &\overset{\text{ind}}{\sim} k(\cdot|\mu_i,\sigma_i), \quad i=1,\dots,n,\\ (\mu_i,\sigma_i)|\tilde{P} &\overset{\text{i.i.d.}}{\sim} \tilde{P}, \quad i=1,\dots,n,\\ \tilde{P} &\sim \text{NRMI}, \end{aligned}$$

where k is a kernel parametrized by $(\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+$ and the NRMI $\tilde{P}$ is defined in (2.29). Under this framework, density estimation is carried out by evaluating the posterior predictive density. Specifically, we consider the Gaussian kernel $k(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma)$ and normalized generalized gamma process on locations and scales with a normal base measure P_0, parameter $\theta = 1$ in Equation (2.18), and varying stability parameter $\gamma \in \{0, 0.25, 0.5, 0.75\}$.

The dataset we consider is the popular `Galaxy` dataset, which consists of velocities of 82 distant galaxies diverging from our own galaxy. Since the data are clearly away from zero (range from 9.2 to 34), Gaussian kernels, although having the whole real line as support, are typically employed in its analysis.

As far as the simulation algorithm is concerned, based on Sections 2.2.3 to 2.2.3, the following moment-matching Ferguson & Klass posterior sampling strategy is implemented: (1) evaluate the threshold $M(\ell)$ which validates trajectories of the CRM using Algorithm 2.2.1 on the prior distribution; (2) implement Algorithm 2.2.1 on the posterior distribution using the threshold $M(\ell)$. More elaborate and suitably tailored moment-matching strategies can be devised for specific models. However, to showcase the generality and simplicity of our proposal we do not pursue this here.

In particular, we set $\ell_M = 0.01$. We compare the output to the Ferguson & Klass algorithm with heuristic relative error e_M criterion, which consists of step (2) only with truncation dictated by the relative error for which we set $e_M \in \{0.1, 0.05, 0.01\}$. For both algorithms the Gibbs sampler is run for $20,000$ iterations with a burn-in of $4,000$, thinned by a factor of 5.

In order to compare the results, we compute the Kolmogorov–Smirnov distance $d_{KS}(\hat{F}_{\ell_M}, \hat{F}_{e_M})$ between associated estimated cumulative distribution functions (cdf) $\hat{F}_{\ell_M}$ and $\hat{F}_{e_M}$ under, respectively, the moment-match and the relative error criteria. The results are displayed in Table 2.8. The estimated cdf $\hat{F}_{\ell_M}$ with $\ell_M = 0.01$ can be seen as a reference estimate since the truncation error is controlled uniformly across the different values of γ by the moment-match at the CRM level. First, one immediately notes that the smaller e_M, the closer the two estimates become (in the d_{KS} distance). Second, and more importantly, the numerical values of the distances heavily depend on the particular choice of the parameter γ for any given e_M. In fact, $\hat{F}_{\ell_M}$ and $\hat{F}_{e_M}$ are significantly further apart for large values of γ than for small ones. This clearly shows that the quality of approximation with the heuristic criterion of the relative index is highly variable in terms of a single parameter; in passing from $\gamma = 0$ to $\gamma = 0.75$ the distance increases by at least a factor of 2. This means that for comparing correctly CRM based models with different parameters one would need to pick different relative indices for each value of the parameter. However, there is no way to guess such thresholds without the guidance of an analytic criterion. And, this already happens by varying a single parameter, let alone when changing CRMs

for which the same e_M could imply drastically different truncation errors. This seems quite convincing evidence supporting the abandonment of heuristic criteria for determining the truncation threshold and the adoption of principled approaches such as the moment-matching criterion proposed in this paper.

γ	$e_M = 0.1$	$e_M = 0.05$	$e_M = 0.01$
0	19.4	15.5	9.2
0.25	31.3	23.7	15.1
0.5	42.4	28.9	18.3
0.75	64.8	41.0	23.2

TABLE 2.8: `Galaxy` dataset. Kolmogorov–Smirnov distance $d_{KS}(\hat{F}_{\ell_M}, \hat{F}_{e_M})$ between estimated cdfs $\hat{F}_{\ell_M}$ and $\hat{F}_{e_M}$ under, respectively, the moment-match (with $\ell_M = 0.01$) and the relative error (with $e_M = 0.1, 0.05, 0.01$) criteria. The mixing measure of normal mixture is the normalized generalized gamma process with varying $\gamma \in \{0, 0.25, 0.5, 0.75\}$.

[A1] J. Arbel and S. Favaro. Approximating predictive probabilities of Gibbs-type priors. *Sankhyā, forthcoming*, 2019

[A9] J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858, 2017

2.3 Approximating the predictive weights of Gibbs-type priors

2.3.1 Introduction and main result

Gibbs-type random probability measures, or Gibbs-type priors, are arguably the most “natural" generalization of the Dirichlet process prior by [117]. They have been first introduced in the seminal works of [266] and [136], and their importance in Bayesian nonparametrics have been extensively discussed in [213], [93] and [38]. Gibbs-type priors have been widely used in the context of Bayesian nonparametric inference for species sampling problems, where their mathematical tractably allowed to obtain explicit expressions for the posterior distributions of various population's features, and to predict features of additional unobservable samples. See, e.g., [210], [214], [113], [115], [40] and [17]. The class of Gibbs-type priors has been also applied in the context of nonparametric mixture modeling, thus generalizing the celebrated Dirichlet process mixture model of Lo [217]. In particular, nonparametric mixture models based in Gibbs-type priors are characterized by a more flexible parameterization than Dirichlet process mixture model, thus allowing for a better control of the clustering behaviour. See, e.g., [164], [209], [211], [116]. Most recently, Gibbs-type priors have been used in Bayesian nonparametric inference for ranked data [75], sparse exchangeable random graphs and networks [74, 155], exchangeable feature allocations [e.g. 319, 64, 154, 290, 45], reversible Markov chains [39], dynamic textual data [81], and bipartite graphs [73].

The definition of Gibbs-type random probability measures relies on the notion of α-stable Poisson–Kingman model, first introduced by [266]. Specifically, let $(J_i)_{i\geq 1}$ be the decreasing ordered jumps of an α-stable subordinator, i.e. subordinator with Lévy measure $\rho(\mathrm{d}x) = C_\alpha x^{-\alpha-1}\mathrm{d}x$ for some constant C_α, and let $P_i = J_i/T_\alpha$ with $T_\alpha = \sum_{i\geq 1} J_i < +\infty$ almost surely; in particular T_α is a positive α-stable random variable, and we denote its density function by f_α. If $\mathrm{PK}(\alpha; t)$ denotes the conditional distribution of $(P_i)_{i\geq 1}$ given $T_\alpha = t$, and if $T_{\alpha,h}$ is a random variable with density function $f_{T_{\alpha,h}}(t) = h(t)f_\alpha(t)$, for any nonnegative function h, then an α-stable Poisson–Kingman model is defined as the discrete random probability measure $P_{\alpha,h} = \sum_{i\geq 1} P_{i,h}\delta_{X_i^*}$, where $(P_{i,h})_{i\geq 1}$ is distributed as $\int_{(0,+\infty)} \mathrm{PK}(\alpha; t) f_{T_{\alpha,h}}(t)\mathrm{d}t$ and $(X_i^*)_{i\geq 1}$ are random variables, independent of $(P_{i,h})_{i\geq 1}$, and independent and identically distributed according to a nonatomic probability measure ν_0. An α-stable Poisson–Kingman model thus provides with a generalization of the normalized α-stable process in [190], which is recovered by setting $h = 1$. According to the work of [136], Gibbs-type random probability measures are defined as a class of (almost sure)

discrete random probability measures indexed by a parameter $\alpha < 1$ such that: i) for any $\alpha < 0$ they are M-dimensional symmetric Dirichlet distribution, with M being a nonnegative random variable on the set $\mathbb{N}$; ii) for $\alpha = 0$ they coincide with the Dirichlet process; iii) for any $\alpha \in (0,1)$ they are α-stable Poisson–Kingman models.

In this paper we focus on the predictive probabilities of Gibbs-type priors with $\alpha \in (0,1)$, i.e. the posterior expectation $\mathbb{E}[P_{\alpha,h}(\cdot) \mid \boldsymbol{X}_n]$, with $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ being a random sample from $P_{\alpha,h}$. Due to the (almost sure) discreteness of the Gibbs-type random probability measure $P_{\alpha,h}$, we expect ties in a sample $\boldsymbol{X}_n$ from $P_{\alpha,h}$, that is $\boldsymbol{X}_n$ features $K_n = k_n \leq n$ distinct types, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $(N_1, \ldots, N_{K_n}) = (n_1, \ldots, n_{k_n})$ such that $\sum_{1 \leq i \leq k_n} n_i = n$. That is, the sample $\boldsymbol{X}_n$ induces a random partition of the set $\{1, \ldots, n\}$; see [268] for details on Gibbs-type random partitions. According to [266], the predictive probabilities of $P_{\alpha,h}$ are

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \nu_0(\cdot) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{i=1}^{k_n} (n_i - \alpha) \delta_{X_i^*}(\cdot) \tag{2.41}$$

for $n \geq 1$, where

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n - k_n\alpha)} \int_0^{+\infty} \int_0^1 t^{-k_n\alpha} p^{n-k_n\alpha-1} h(t) f_\alpha((1-p)t) \mathrm{d}t \mathrm{d}p, \tag{2.42}$$

with $\Gamma(\cdot)$ being the Gamma function. See, e.g., [266] and [136] for a detailed account on (2.41) and (2.42). Hereafter we briefly recall two noteworthy examples of Gibbs-type random probability measures: the Pitman–Yor process and the normalized generalized gamma process.

Example 2.3.1. *Let $(a)_n$ be the rising factorial of a of order n, i.e. $(a)_n = \prod_{0 \leq i \leq n-1}(a+i)$, for $a > 0$. For any $\alpha \in (0,1)$ and $\theta > -\alpha$ the Pitman–Yor process, say $P_{\alpha,\theta}$, is a Gibbs-type random probability measure with*

$$h(t) = \frac{\alpha\Gamma(\theta)}{\Gamma(\theta/\alpha)} t^{-\theta} \tag{2.43}$$

such that

$$V_{n,k_n} = \frac{\prod_{i=0}^{k_n-1}(\theta + i\alpha)}{(\theta)_n}. \tag{2.44}$$

The normalized α-stable process is $P_{\alpha,0}$, whereas the Dirichlet process may be recovered as a limiting special case for $\alpha \to 0$. See, e.g., [261, 270, 169, 266, 170] for detailed accounts on $P_{\alpha,\theta}$.

Example 2.3.2. *Let $\Gamma(\cdot,\cdot)$ be the incomplete Gamma function, i.e., $\Gamma(a,b) = \int_b^\infty x^{a-1}\exp\{-x\}\mathrm{d}x$ for $(a,b) \in \mathbb{R} \times \mathbb{R}^+$. For any $\alpha \in (0,1)$ and $\tau \geq 0$ the normalized generalized gamma process, say $G_{\alpha,\tau}$, is a Gibbs-type random probability measure with*

$$h(t) = \mathrm{e}^{\tau^\alpha - \tau t} \tag{2.45}$$

such that

$$V_{n,k_n} = \frac{\alpha^{k_n}\mathrm{e}^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right). \tag{2.46}$$

The normalized α-stable process coincides with $G_{\alpha,0}$, whereas $G_{1/2,\tau}$ is the normalized inverse Gaussian process. See [169, 266, 209, 210, 214, 170] for detailed accounts on $G_{\alpha,\tau}$ and

applications.

Within the large class of predictive probabilities of the form (2.41), those of the Pitman–Yor process $P_{\alpha,\theta}$ certainly stand out for their mathematical tractability, and for having an intuitive interpretability with respect to the parameter $\alpha \in (0,1)$ and $\theta > -\alpha$. See [38] for a description of the predictive probabilities of $P_{\alpha,\theta}$ in terms of a simple generalized Pólya like urn scheme. These desirable features of $P_{\alpha,\theta}$ arise from the product form of the V_{n,k_n}'s in (2.44), which makes the ratio $V_{n+1,k_n+1}/V_{n,k_n}$ a simple linear function of k_n, and the ratio $V_{n+1,k_n}/V_{n,k_n}$ independent of k_n. Specifically, the predictive probabilities of $P_{\alpha,\theta}$ reduce to the following

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] = \frac{\theta + k_n\alpha}{\theta + n}\nu_0(\cdot) + \frac{1}{\theta + n}\sum_{i=1}^{k_n}(n_i - \alpha)\delta_{X_i^*}(\cdot), \tag{2.47}$$

for $n \geq 1$. The weight attached to ν_0 in (2.47) can be read as a sum of two terms with distinct asymptotic orders of magnitude: i) αk_n, referred to as the first order term, and θ, referred to as the second order term. An analogous two-term decomposition holds for the weight attached to the empirical part of (2.47). Our distinction and phrasing is formally captured by writing the weights as follows

$$\frac{\theta + k_n\alpha}{\theta + n} = \frac{k_n\alpha}{n} + \frac{\theta}{n} + o\left(\frac{1}{n}\right) \tag{2.48}$$

and

$$\frac{1}{\theta + n} = \frac{1}{n} - \frac{\theta}{n^2} + o\left(\frac{1}{n^2}\right), \tag{2.49}$$

where o is almost sure, recovering both contributions in a two-term asymptotic decomposition. Equations (2.48) and (2.49) lead to two large n approximations of the predictive distribution displayed in (2.47): i) a first order approximation of (2.47), denoted by $\sim$, is obtained by combining (2.47) with the first term on the right-hand side of (2.48) and (2.49); ii) a second order approximation of (2.47), denoted by $\approx$, is obtained by combining (2.47) with the first two terms on the right-hand side of (2.48) and (2.49).

[292] and [17] extended the decompositions displayed in (2.48) and (2.49) to the normalized inverse Gaussian process and the normalized generalized gamma process, respectively, thus covering the setting described in Example 2.3.2. In the next theorem we generalize (2.48) and (2.49) to the entire class of Gibbs-type priors, that is, for any continuously differentiable function h and any $\alpha \in (0,1)$ we provide a two-term asymptotic decomposition for the weights $V_{n+1,k_n+1}/V_{n,k_n}$ and $V_{n+1,k_n}/V_{n,k_n}$ of the predictive probabilities (2.41).

Theorem 2.3.1. *Let $\boldsymbol{X}_n$ be a sample from $P_{\alpha,h}$ featuring $K_n = k_n \leq n$ distinct types, labelled by $X_1^*, \ldots, X_{K_n}^*$, with frequencies $(N_1, \ldots, N_{K_n}) = (n_1, \ldots, n_{k_n})$. Assume that function h is continuously differentiable and denote its derivative by h'. Then*

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{k_n\alpha}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right) \tag{2.50}$$

and

$$\frac{V_{n+1,k_n}}{V_{n,k_n}} = \frac{1}{n} - \frac{\beta_n}{n^2} + o\left(\frac{1}{n^2}\right) \tag{2.51}$$

for any $n \geq 1$, *where* $\beta_n = \varphi_h(nk_n^{-1/\alpha})$ *with* φ_h *being defined as* $\varphi_h(t) = -th'(t)/h(t)$.

Theorem 2.3.1 may be applied to obtain a first and a second order approximations of the predictive probabilities of an arbitrary Gibbs-type prior $P_{\alpha,h}$. This result then contributes to a remarkable simplification in the evaluation of (2.41) for any choice of the function h. Besides that, Theorem 2.3.1 highlights, for large n, the role of h from a purely predictive perspective. In particular, according to Theorem 2.3.1, the function h does not affect the first order term in the asymptotic decompositions (2.50) and (2.51), and it is sufficient to consider a second order term in order to take into account h. This leads to two meaningful approximations of the predictive probabilities (2.41). In particular, by considering the sole first order term in (2.50) and (2.51), one obtains the first order approximation

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] \sim \frac{k_n \alpha}{n} \nu_0(\cdot) + \frac{1}{n} \sum_{i=1}^{k_n} (n_i - \alpha) \delta_{X_i^*}(\cdot), \tag{2.52}$$

which is the predictive of the normalized α-stable process, i.e. $h = 1$. By including the second order term in (2.50) and (2.51), one obtains the second order approximation

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] \approx \frac{\beta_n + k_n \alpha}{\beta_n + n} \nu_0(\cdot) + \frac{1}{\beta_n + n} \sum_{i=1}^{k_n} (n_i - \alpha) \delta_{X_i^*}(\cdot), \tag{2.53}$$

which resembles the predictive probabilities (2.47) of the Pitman–Yor process $P_{\alpha,\theta}$, with the parameter θ replaced by a suitable function of h, α and the number k_n of distinct types in the sample $\mathbf{X}_n$. Note that (2.53) is obtained by normalizing the weights (2.50) and (2.51) which lead to a proper predictive distribution (the weights of (2.53) sum up to one) while preserving the second order approximation since

$$\frac{\beta_n + k_n \alpha}{\beta_n + n} = \frac{k_n \alpha}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right) \quad \text{and} \quad \frac{1}{\beta_n + n} = \frac{1}{n} - \frac{\beta_n}{n^2} + o\left(\frac{1}{n^2}\right).$$

The predictive probabilities of any Gibbs-type prior thus admit a second order approximation, for large n, with an error term vanishing as $o(1/n)$. More importantly, such a second order approximation maintains the same mathematical tractability and interpretability as the predictive probability of the Pitman–Yor process.

The section is structured as follows. In Section 2.3.2 we prove Theorem 2.3.1 and the approximate predictive probabilities displayed in Equation (2.52) and Equation (2.53). In Section 2.3.3 we present a numerical illustration of our approximate predictive probabilities, thus showing their usefulness from a practical point of view. Section 2.3.4 describes a marginal Blackwell–MacQueen Pólya urn posterior sampling scheme based on the proposed first order and second approximations. Section 2.3.5 contains a brief discussion of our results.

2.3.2 Proofs

Throughout this section, we will use the notation $a_n \asymp b_n$ when $a_n / b_n \to 1$ as $n \to \infty$, almost surely. The main argument of the proof consists in a Laplace approximation of the integral form for V_{n,k_n} in (2.42) as $n \to \infty$. This approximation basically replaces an exponentially large term in an integrand by a Gaussian kernel which

matches both mean and variance of the integrand. From evaluating the Gibbs-type predictive probabilities (2.41) on the whole space it is clear that we have

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = 1 - (n - \alpha k_n)\frac{V_{n+1,k_n}}{V_{n,k_n}}. \tag{2.54}$$

Denote the integrand function of (2.42) by $f_n(p,t) = t^{-\alpha k_n} p^{n-1-k_n\alpha} h(t) f_\alpha((1-p)t)$, and denote integration over its domain $(0,1)\times\mathbb{R}_+^*$ by $\iint$. Then we can write

$$\frac{V_{n+1,k_n}}{V_{n,k_n}} = \frac{1}{n-\alpha k_n}\frac{\iint p f_n}{\iint f_n}. \tag{2.55}$$

Note that this ratio of integrals coincides with $\mathbb{E}_n(P)$, that is the expectation under the probability distribution with density proportional to f_n. This, combined with (2.54) provides $V_{n+1,k_n+1}/V_{n,k_n} = \mathbb{E}_n(1-P)$. In order to apply the Laplace approximation method, write the nonnegative integrand f_n in exponential form $f_n = \mathrm{e}^{nl_n}$, and further define functions $g(p,t) = 1-p$ and $\tilde{g}(p,t) = 1$. Then

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\iint g\mathrm{e}^{nl_n}}{\iint \tilde{g}\mathrm{e}^{nl_n}}. \tag{2.56}$$

The mode (t_n, p_n) of f_n (or equivalently of l_n) is determined by the root of the partial derivatives

$$n\frac{\partial l_n(p,t)}{\partial p} = \frac{n-\alpha k_n - 1}{p} - t\frac{f'_\alpha(t(1-p))}{f_\alpha(t(1-p))} \tag{2.57}$$

and

$$n\frac{\partial l_n(p,t)}{\partial t} = \frac{-\alpha k_n}{t} + \frac{h'(t)}{h(t)} + (1-p)\frac{f'_\alpha(t(1-p))}{f_\alpha(t(1-p))}, \tag{2.58}$$

where f'_α and h' denote respectively the derivatives of the α-stable density f_α and of the function h. Now consider the Laplace approximations to the numerator and the denominator of the ratio (2.56) with the notations set forth in Section 6.9 of Small [304]. The exponential term is identical in both integrands of the ratio (2.56), hence the term involving $\det f_n$, the Hessian of f_n, is also identical and equal to

$$C_n = (2\pi/n)^{2/2}(-\det f_n)^{-1/2}\mathrm{e}^{nl_n(t_n,p_n)}.$$

Thus it simplifies in the ratio. One needs only to consider the asymptotic series expansions, where we require a second order term $a(t_n,p_n)$ for the numerator, that is

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{C_n\times\left(g(t_n,p_n) + \frac{1}{n}a(t_n,p_n) + \mathcal{O}\left(\frac{1}{n^2}\right)\right)}{C_n\times\left(\tilde{g}(t_n,p_n) + \mathcal{O}\left(\frac{1}{n}\right)\right)}.$$

The expression of $a(t_n, p_n)$ is provided in Equation (6.14) of Small [304]. In our case, $a(t_n, p_n) = o(1/n)$, hence with $\tilde{g} = 1$, the previous display simplifies to the following

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = g(t_n, p_n) + o\left(\frac{1}{n}\right). \tag{2.59}$$

Let $\varphi_h(t) = -th'(t)/h(t)$. Note that, adding $(1-p_n)\times$(2.57) and $t_n\times$(2.58) we can write

$$g(t_n, p_n) = 1 - p_n = \frac{\alpha k_n + \varphi_h(t_n)}{n + \varphi_h(t_n) - 1} \tag{2.60}$$

so, in view of (2.59),

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\alpha k_n + \varphi_h(t_n)}{n + \varphi_h(t_n) - 1} + o\left(\frac{1}{n}\right). \tag{2.61}$$

Let $\psi(x) = (xf'_\alpha(x))/(\alpha f_\alpha(x))$. By (2.57), $\psi((1-p_n)t_n) = (1-p_n)(n-\alpha k_n - 1)/\alpha p_n$. By Theorem 2 in Arbel et al. [17], $1 - p_n \asymp \alpha k_n/n$. Hence, $\psi((1-p_n)t_n)$ grows to infinity when $n \to \infty$ at the same rate as k_n. But studying the variations of the α-stable density f_α, Nolan [254] shows that the only infinite limit of ψ is in 0^+ according to

$$\psi(x) \overset{[}{0}{}^{+}]{} \asymp (\alpha/x)^{\frac{\alpha}{1-\alpha}}.$$

In order that $\psi((1-p_n)t_n)$ matches with its infinite limit when $n \to \infty$, its argument $(1-p_n)t_n$ needs go to 0^+, which yields to the following asymptotic equivalence

$$k_n \asymp \psi((1-p_n)t_n) \asymp \left(\frac{\alpha}{(1-p_n)t_n}\right)^{\frac{\alpha}{1-\alpha}},$$

which in turn gives

$$t_n \asymp \alpha \frac{k_n^{1-1/\alpha}}{1-p_n} \asymp \alpha \frac{k_n^{1-1/\alpha}}{\alpha k_n/n} \asymp \frac{n}{k_n^{1/\alpha}} \asymp T_{\alpha,h},$$

where the last equivalence is from [266]. Since function h is assumed to be positive and continuous differentiable, $\varphi_h(T_{\alpha,h})$ is a.s. well defined (and finite) and $\varphi_h(t_n) \asymp \varphi_h(nk_n^{-1/\alpha}) \asymp \varphi_h(T_{\alpha,h})$ a.s., so (2.61) can be rewritten

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\alpha k_n}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right),$$

where we set $\beta_n = \varphi_h(nk_n^{-1/\alpha})$. In other terms, to match the expression of the second order approximate predictive probability displayed in Equation (2.53), we have

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\beta_n + k_n\alpha}{\beta_n + n} + o\left(\frac{1}{n}\right).$$

The expression of the second weight in the predictive of the theorem follows from (2.54), i.e.,

$$\begin{aligned}\frac{V_{n+1,k_n}}{V_{n,k_n}} &= \frac{1 - V_{n+1,k_n+1}/V_{n,k_n}}{n - \alpha k_n} \\ &= \left(1 - \frac{\alpha k_n}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right)\right)\left(\frac{1}{n} + \frac{\alpha k_n}{n^2} + o\left(\frac{k_n}{n^2}\right)\right), \\ &= \frac{1}{n} - \frac{\alpha k_n}{n^2} - \frac{\beta_n}{n^2} + \frac{\alpha k_n}{n^2} + o\left(\frac{1}{n^2}\right) = \frac{1}{n} - \frac{\beta_n}{n^2} + o\left(\frac{1}{n^2}\right),\end{aligned}$$

or, to match the expression of the second order approximate predictive of equation (2.53),

$$\frac{V_{n+1,k_n}}{V_{n,k_n}} = \frac{1}{\beta_n + n} + o\left(\frac{1}{n^2}\right).$$

2.3.3 Numerical illustrations

As we recalled in Example 2.3.1, the Pitman–Yor process $P_{\alpha,\theta}$ is a Gibbs-type random probability measure with $\alpha \in (0,1)$ and $h(t) = t^{-\theta}\Gamma(\theta+1)/\Gamma(\theta/\alpha+1)$, for any $\theta > -\alpha$. By an application of Theorem 2.3.1, the predictive probabilities of $P_{\alpha,\theta}$ admit a first order approximation of the form (2.52) and a second order approximation of the form (2.53) with $\varphi_h(t) = \theta$, and such that $\beta_n = \theta$. Among Gibbs-type random probability measures with $\alpha \in (0,1)$, the Pitman–Yor process certainly stands out for a predictive structure which admits a simple numerical evaluation. This made the Pitman–Yor process a natural candidate in several applications within the large class of Gibbs-type priors. Hereafter we present a brief numerical illustration to compare the predictive probabilities of $P_{\alpha,\theta}$ with their first and second order approximations given in terms of Equation (2.48) and Equation (2.49). While there is no practical reason to make use our approximate predictive probabilities, because of the simple expression of (2.47), the illustration is useful to show the accuracy of our approximations. We then present the same numerical illustration for the normalized generalized gamma process $G_{\alpha,\tau}$ of Example 2.3.2. We will see that, differently from the Pitman–Yor process, the predictive probabilities of the normalized generalized gamma process do not admits a simple numerical evaluation. This motivates the use of Theorem 2.3.1.

We consider 500 data points sampled independently and identically distributed from the ubiquitous Zeta distribution. For any $\sigma > 1$ this is a distribution with probability mass function $\Pr(Z = z) \propto z^{-\sigma}$, for $z \in \mathbb{N}$. Here we choose $\sigma = 1.5$. For each $n = 1, \ldots, 500$ we record the number k_n of distinct types at the n-th draw, and we evaluate the predictive weight $V_{n+1,k_n+1}/V_{n,k_n}$ for the Pitman–Yor process, i.e. the left-hand side of (2.48). We consider the following pairs of parameters (α,θ): $(0.25,1)$, $(0.25,3)$, $(0.25,10)$, $(0.5,1)$, $(0.5,3)$, $(0.5,10)$, $(0.75,1)$, $(0.75,3)$ and $(0.75,10)$. For each of these pairs we compare the left-hand side of Equation (2.48) with the first term of the right-hand side of Equation (2.48) (first oder approximation) and with the first two terms of the right-hand side of Equation (2.48) (second

order approximation), that are

$$\frac{\theta + k_n \alpha}{\theta + n}, \tag{2.62}$$

$$\frac{k_n \alpha}{n} \tag{2.63}$$

and

$$\frac{k_n \alpha}{n} + \frac{\theta}{n}, \tag{2.64}$$

respectively. Figure 2.8 shows the curve, as functions of n, of the "exact" predictive weight (2.62) and its first order approximation (2.63) and second order approximation (2.64). The first order approximation consistently underestimates the "exact" predictive weight, while the second order approximation consistently overestimates it. This is due to the fact that the parameter θ is positive. The discrepancy between the first order approximation and (2.62) stays substantial even for large values of n, all the more for large θ. On the contrary, the second order approximation consistently outperforms the first order approximation, closely following (2.62). For $n = 500$, the "exact" predictive weight and its second order approximation are barely distinguishable in all the considered pairs of parameters.

The normalized generalized gamma process

As we recalled in Example 2.3.2, the normalized generalized gamma process is a Gibbs-type random probability measure with $\alpha \in (0, 1)$ and $h(t) = \exp\{\tau^\alpha - \tau t\}$, for any $\tau \geq 0$. From Theorem 2.3.1, the predictive probabilities of the normalized generalized gamma process admit a first order approximation of the form (2.52) and a second order approximation of the form (2.53) with $\varphi_h(t) = \tau t$, and

$$\beta_n = \frac{\tau n}{k_n^{1/\alpha}}.$$

The predictive probabilities of the normalized generalized gamma process are of the form (2.41), with the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ and $V_{n+1,k_n}/V_{n,k_n}$ admitting an explicit (closed-form) expression in terms of (2.46). However, differently from the Pitman–Yor process, the evaluation of the predictive weights is cumbersome, thus preventing their practical implementation. In particular, as pointed out in Lijoi et al. [211] in the context of mixture models with a normalized generalized Gamma prior, the evaluation of (2.46) gives rise to severe numerical issues, even for not too large values of n. These issues are mainly due to the evaluation of the incomplete gamma function, as well as with handling very small terms and very large terms within the summation (2.46). Because of these numerical issues in evaluating (2.46), we propose an alternative approach to evaluate the V_{n,k_n}'s of the normalized generalized gamma process. This is a Monte Carlo approach, and it relies on the fact that V_{n,k_n} in (2.46) can be written as the expectation of a suitable ratio of independent random variables. Recall that f_α denotes the density function of a positive α-stable random variable.

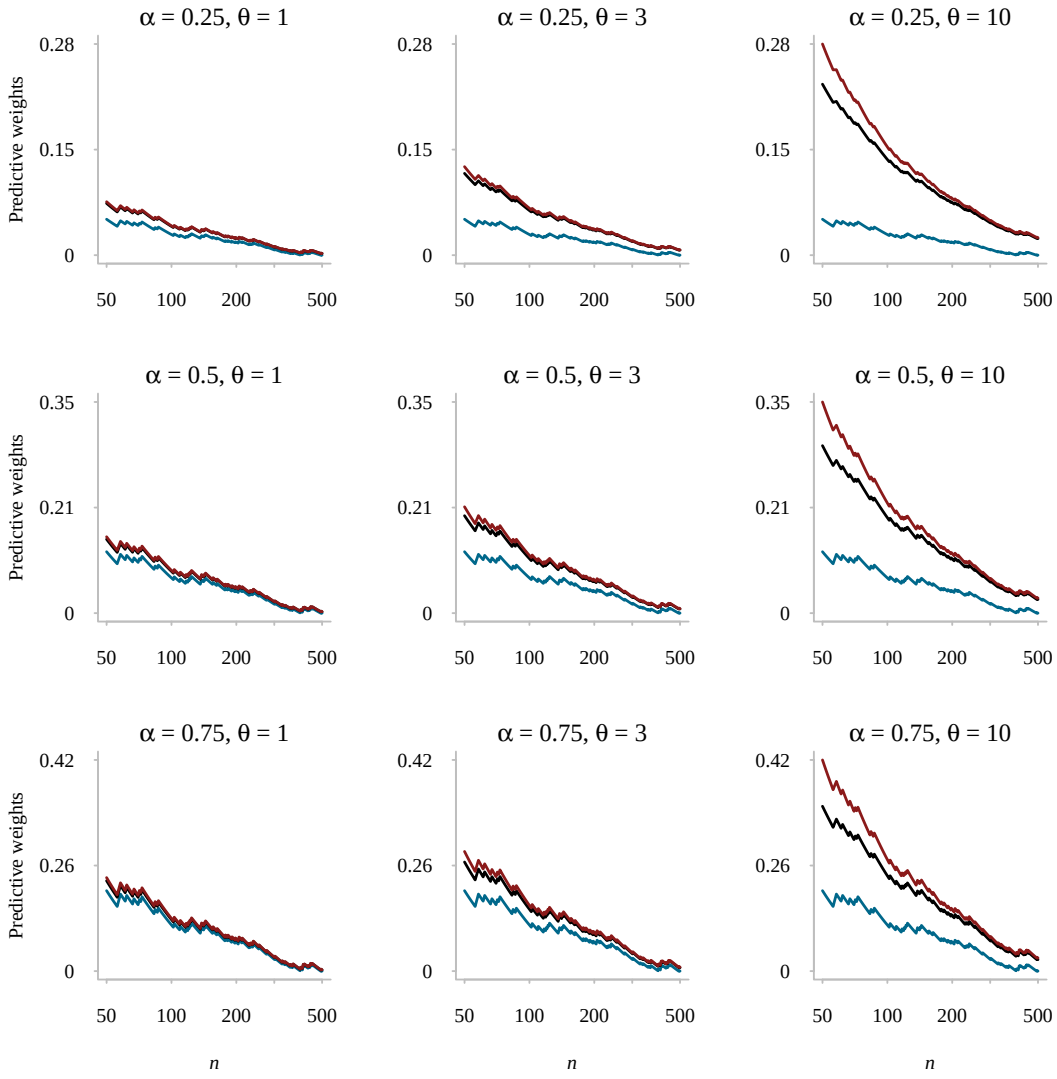


FIGURE 2.8: Predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the Pitman–Yor process. In black: the "exact" value (2.62). In blue: the first order approximation (2.63). In red: the second order approximation (2.64). The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\theta = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.

Then, using (2.42) with $h(t) = \exp\{\tau^\alpha - \tau t\}$, we can write

$$\begin{aligned} V_{n,k_n} &= \frac{\alpha^{k_n}}{\Gamma(n-k_n\alpha)} \int_0^{+\infty} \int_0^1 p^{n-1-k_n\alpha} t^{-k_n\alpha} \exp\{\tau^\alpha - \tau t\} f_\alpha(t(1-p)) \mathrm{d}p\mathrm{d}t \\ &= \frac{\alpha^{k_n-1}\Gamma(k_n)}{\Gamma(n)} \int_0^{+\infty} \exp\{\tau^\alpha - \tau t\} \frac{\alpha\Gamma(n)}{\Gamma(k_n)\Gamma(n-k_n\alpha)} t^{-k_n\alpha} \\ &\qquad \int_0^1 (1-p)^{n-k_n\alpha-1} f_\alpha(tp) \mathrm{d}p\mathrm{d}t \\ &= \frac{\alpha^{k_n-1}\Gamma(k_n)}{\Gamma(n)} \mathbb{E}\left[\exp\left\{\tau^\alpha - \frac{\tau X}{Y}\right\}\right], \end{aligned} \tag{2.65}$$

where X and Y are two independent random variables such that Y is distributed according to a Beta distribution with parameter $(k_n\alpha, n - k_n\alpha)$, and X is distributed according to a polynomially tilted positive α-stable random variable, i.e.,

$$\Pr[X \in \mathrm{d}x] = \frac{\Gamma(k_n\alpha + 1)}{\Gamma(k_n + 1)} x^{-k_n\alpha} f_\alpha(x) \mathrm{d}x. \tag{2.66}$$

We refer to [266], [268] and Devroye [98] for a detailed account on the polynomially tilted α-stable random variable X. Given the representation (2.65) we can perform a Monte Carlo evaluation of V_{n,k_n} by simply sampling from the Beta random variable Y and from the random variable X with distribution (2.66).

Sampling form the Beta random variable Y is straightforward. The random variable X can be sampled by using an augmentation argument that reduces the problem of sampling X to the problem of sampling a Gamma random variable and, given that, an exponentially tilted α-stable random variable, i.e. a random variable with density function $\exp\{c^\alpha - cx\} f_\alpha(x)$, for some constant $c > 0$. The problem of sampling exponentially tilted α-stable random variables has been considered in Devroye [98] and Hofert [160]. Specifically, we can write (2.66) as follows

$$\begin{aligned} \frac{\Gamma(k_n\alpha + 1)}{\Gamma(k_n + 1)} x^{-k_n\alpha} f_\alpha(x) &= \frac{\alpha}{\Gamma(k_n)} \int_0^{+\infty} c^{k_n\alpha - 1} \exp\{-c^\alpha\} \frac{\exp\{-cx\} f_\alpha(x)}{\exp\{-c^\alpha\}} \mathrm{d}c \\ &= \int_0^{+\infty} f_C(c) f_{X|C=c}(x) \mathrm{d}c, \end{aligned}$$

where $f_{X|C=c}$ is the density function of an exponentially tilted positive α-stable random variable, and f_C is the density function of the random variable $C = G^{1/\alpha}$, where G being a Gamma random variable with parameter $(k_n, 1)$. We apply Hofert [160] for sampling the exponentially tilted positive α-stable random variable with density function $f_{X|C=c}$. Note that, as k_n grows, the tilting parameter $C = G^{1/\alpha}$ gets larger in distribution. As a result, the acceptance probability decreases and the Monte Carlo algorithm slows down. Let Be, Ga and tSt respectively denote Beta, Gamma and exponentially tilted positive α-stable distributions, and let Γ_l represents the logarithm of the Γ function. Hereafter is the step-by-step pseudocode for the Monte Carlo evaluation of the V_{n,k_n}'s:

(i) Set $M = 10^4$, n, k_n, α, τ;

(ii) Sample $Y \sim \mathrm{Be}(\alpha k_n, n - \alpha k_n)$ of size M;

(iii) Sample $G \sim \text{Ga}(k_n, 1)$ of size M;

(iv) Sample $X \sim \text{tSt}(\alpha, G^{1/\alpha})$ of size M

(v) Set $v = (k_n - 1)\log\alpha + \Gamma_l(k_n) - \Gamma_l(n) + \tau^\alpha - \tau X/Y$;

(vi) Set $V = \exp(v)$.

In the same setting described for the Pitman–Yor process, we perform a numerical study for the normalized generalized gamma process. More specifically, 500 data points are sampled independently and identically distributed from the Zeta distribution with parameter $\sigma = 1.5$. We consider the following pairs of parameters (α, τ): $(0.25, 1)$, $(0.25, 3)$, $(0.25, 10)$, $(0.5, 1)$, $(0.5, 3)$, $(0.5, 10)$, $(0.75, 1)$, $(0.75, 3)$ and $(0.75, 10)$. For these pairs of parameters the predictive weight $V_{n+1,k_n+1}/V_{n,k_n}$ is evaluated by means of the above steps 1-6, and this evaluation is compared with the first order approximation and with the second order approximation of $V_{n+1,k_n+1}/V_{n,k_n}$ given by Theorem 2.3.1, i.e.

$$\frac{k_n\alpha}{n} \tag{2.67}$$

and

$$\frac{k_n\alpha}{n} + \frac{\tau}{k_n^{1/\alpha}}, \tag{2.68}$$

respectively. Figure 2.9 shows that the Monte Carlo evaluation of $V_{n+1,k_n+1}/V_{n,k_n}$ lays between the first order approximation and the second order approximation of $V_{n+1,k_n+1}/V_{n,k_n}$. As n moves, the difference between the resulting Monte Carlo curve and the approximate curves is imperceptible for $\alpha = 0.25$; such a difference is also very small for $\tau = 1$. Larger values of α and/or τ lead to larger discrepancies between the Monte Carlo curve and the approximate curves. The second order approximation is consistently closer to the Monte Carlo value than the first order approximation. In particular we observe that for $n = 500$ the second order approximation and the Monte Carlo value are indistinguishable, whereas the first order approximation may still be far from the Monte Carlo value for several choices of the parameters, e.g. $(\alpha, \tau) = (0.75, 3)$ and $(\alpha, \tau) = (0.75, 10)$.

We conclude by motivating the use of the second order approximation instead of the Monte Carlo evaluation. First of all, for pairs of parameters with large α and large τ, e.g. $(\alpha, \tau) = (0.75, 10)$ in our numerical study, the Monte Carlo evaluation is extremely noisy, although we have used a large number of iterations, i.e 10^4. In particular, as shown in Figure 2.9, the noise does not vanish as n grows. On the contrary, the second order approximation has a more stable behavior, and for $(\alpha, \tau) = (0.75, 10)$ it converges to the bulk of the Monte Carlo curve, which makes it more reliable than the latter for large values of n. Furthermore, evaluating the second order approximation is fast. On the other hand, the computational burden of the Monte Carlo evaluation is very heavy, e.g. 35 hours were required for the nine configurations of Figure 2.9, with 10^4 iterations for each weight. This is because of the sampling of the exponentially tilted α stable random variable. Indeed the rejection sampler originally proposed by Hofert [160] has an acceptance probability that decreases as n grows, making this approach prohibitive for large sample sizes. Although our Monte Carlo code could certainly be fastened, our empirical study suggests that the computing time increases exponentially with the sample size n. See the average Monte Carlo running time in Figure 2.10, as well as the running

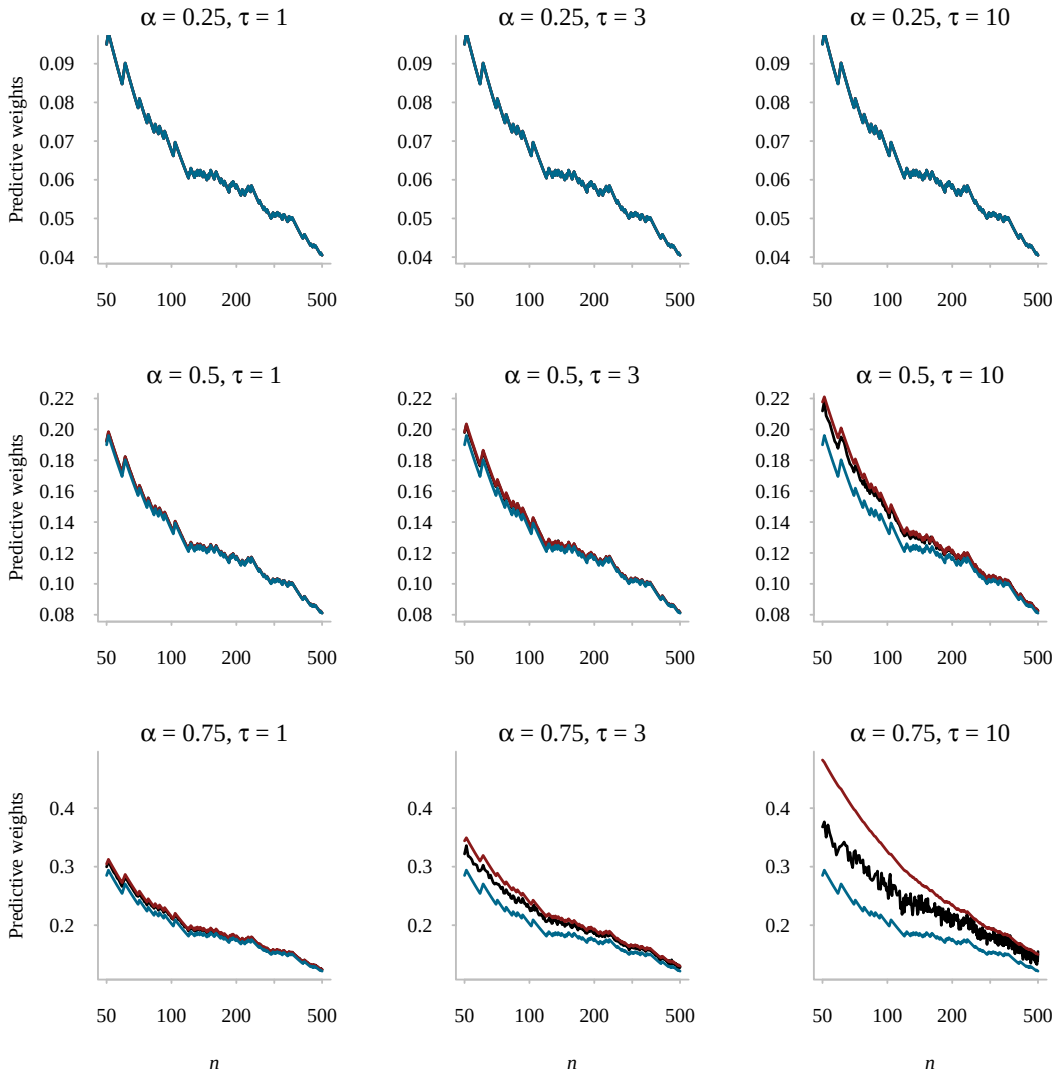


FIGURE 2.9: Predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized gamma process. In black: the "exact" value evaluated by the Monte Carlo approach. In blue: the first order approximation (2.67). In red: the second order approximation (2.68). The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.

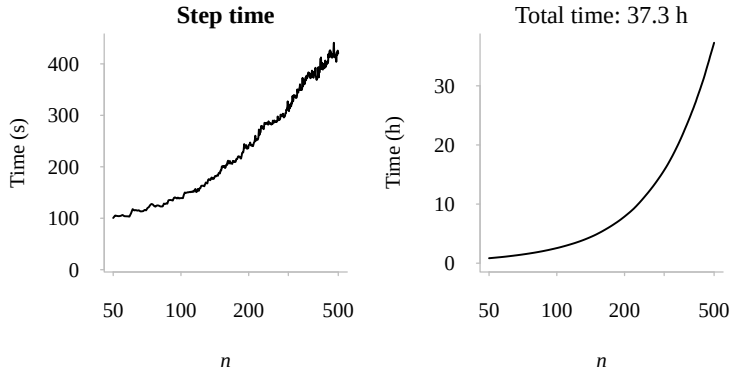


FIGURE 2.10: Left panel: running time (in seconds) averaged over all nine parameter configurations, and right panel: cumulated running time (in hours) averaged over all nine parameter configurations, for the Monte Carlo approach applied to the evaluation of the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized gamma process case. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.

time and cumulated running time for each of the nine parameter configurations in Figure 2.11 and Figure 2.12.

2.3.4 Posterior sampling

In this section we present an application of Theorem 2.3.1 in the context of Bayesian nonparametric mixture modeling. Among various posterior sampling schemes for Bayesian nonparametric mixture modeling, the so-called Blackwell–MacQueen Pólya urn scheme certainly stands out. It is a Markov chain Monte Carlo sampling scheme belonging to the class of "marginal" schemes, since it relies on the predictive distributions. See MacEachern [220] and Escobar and West [112] for a description of the Blackwell–MacQueen Pólya urn scheme in the context of mixture modeling based on Dirichlet process priors, and Ishwaran and James [164] for mixture modeling based on general stick-breaking priors, e.g., the Pitman–Yor process prior. We compare the Blackwell–MacQueen Pólya urn scheme based on the exact predictive distributions with the Blackwell–MacQueen Pólya urn scheme based on our approximated predictive distributions. The performance is evaluated by computing the Kolmogorov–Smirnov (KS) distance between the estimated distribution function and the cumulative distribution function (cdf) of the true data generating process.

As an illustrative example, we considered simulated data of varying size $n = 50, 100, 200, 500$ sampled from a mixture of two Gaussian distributions, say $w_1\mathcal{N}(\mu_1, \sigma_1^2) + (1 - w_1)\mathcal{N}(\mu_2, \sigma_2^2)$. Precisely, we set $(\mu_1, \sigma_1^2) = (1, 0.2)$, $(\mu_2, \sigma_2^2) = (10, 0.2)$ and $w_1 = 0.5$. The Bayesian nonparametric mixture model can be defined as

$$\begin{aligned} &Y_i \mid X_i \overset{\text{ind}}{\sim} \mathcal{N}(Y_i \mid X_i, \sigma^2), \quad i = 1, \ldots, n, \\ &X_i \mid P_{\alpha,h} \overset{\text{i.i.d.}}{\sim} P_{\alpha,h}, \quad i = 1, \ldots, n, \\ &P_{\alpha,h} \sim \mathcal{P}_{\alpha,h}, \\ &\sigma^2 \sim \mathcal{IG}(a, b), \end{aligned} \tag{2.69}$$

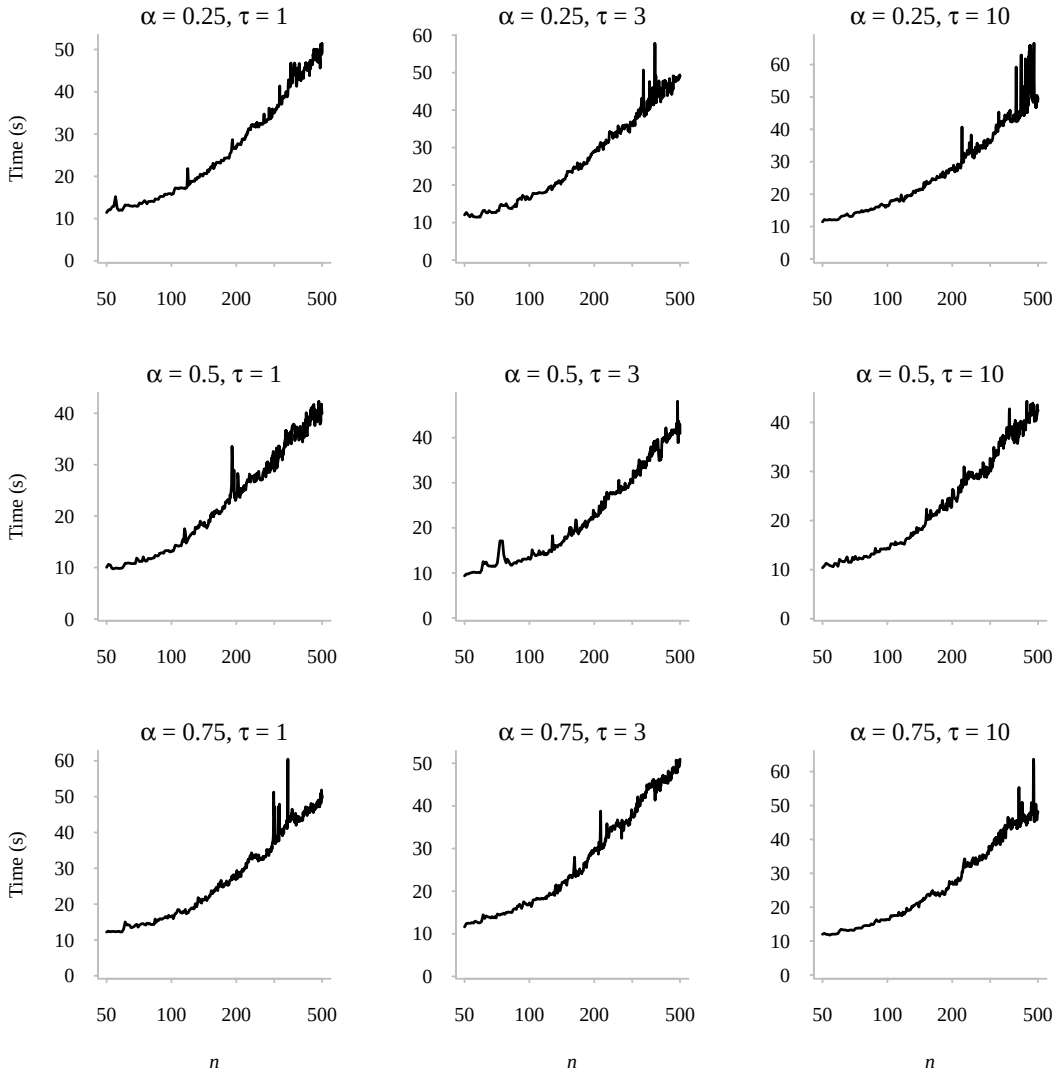


FIGURE 2.11: Running time (in seconds) for the Monte Carlo approach for evaluating the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized gamma process case. The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.

where $\mathcal{P}_{\alpha,h}$ denotes a Gibbs-type prior, and $\mathcal{IG}(a, b)$ stands for an inverse-gamma distribution with parameters a and b. Following Section 2.3.3, we focus on the two common choices for the random probability measure $P_{\alpha,h}$, namely the Pitman–Yor process and the normalized generalized gamma process. In both cases we assume that the nonatomic probability measure ν_0 is the standard Gaussian distribution. In the model (2.69) we assume that $a = b = 1$.

Under the assumption of the Pitman–Yor process prior and the assumption of the normalized generalized gamma process prior, we apply the Blackwell–MacQueen Pólya urn scheme with the exact predictive distributions and with the corresponding approximated predictive distributions given by Theorem 2.3.1. We used 10^4 iterations after a burn-in of 2 000. In Figure 2.13, we show the KS distance between the

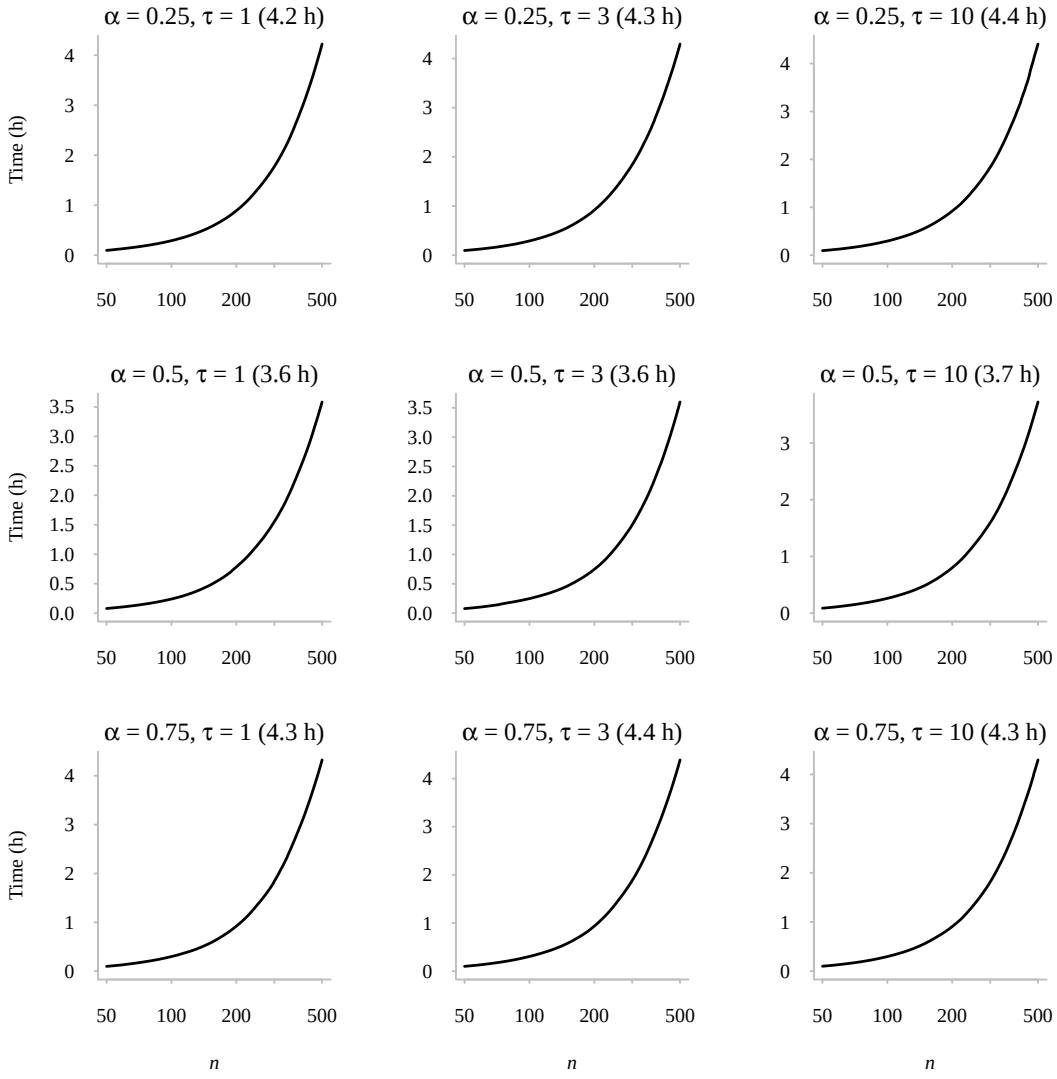


FIGURE 2.12: Cumulated running time (in hours) for the Monte Carlo approach for evaluating the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized gamma process case. The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.

true distribution function and the estimated distribution function obtained by using the Blackwell–MacQueen Pólya urn scheme with

- the exact predictive distributions (2.47) of the Pitman–Yor process; the second order approximation (2.53) of the predictive distribution of the Pitman–Yor process coincides with this exact predictive distribution.
- the first order approximation (2.52) of the predictive distributions of the Pitman–Yor process, which coincides with the first order approximation of the predictive distributions of the normalized generalized gamma process.
- the second order approximation (2.53) of the predictive distributions of the normalized generalized gamma process, which is different from the second order approximation of the predictive distribution of the Pitman–Yor process.

The values of the hyperparameters α, θ and τ correspond to those used in the numerical illustrations of Section 2.3.3. Results in Figure 2.13 show that both first and second order approximations of predictive distributions produce posterior estimates with comparable performance to that the exact predictive distribution of the Pitman–Yor process. Also, the sampling scheme based on the first order approximation outperforms the sampling scheme based on the exact predictive distributions of the Pitman–Yor process, and of the second order approximation of the predictive distribution of the normalized generalized gamma process. A reason for this superiority of the first order approximation is the following: this first order approximation, both for the Pitman–Yor process and for the normalized generalized gamma process, boils down to the normalized α-stable process. For a given parameter α, such normalized α-stable process has a lower prior expected number of clusters than the Pitman–Yor process and the normalized generalized gamma process counterparts. Thus the normalized α-stable process is a better specified prior than the latter two processes for the true data generating process which is only made of two components, leading to an overall better performance.

2.3.5 Discussion

Gibbs-type priors form a flexible class of nonparametric priors, which is parameterized by an index $\alpha \in (0,1)$ and a function h. According to the definition of Gibbs-type random probability measures in terms of α-stable Poisson–Kingman models, the function h has the primary role of enriching the parameterization of the normalized α-stable process by introducing additional parameters other than α. See, e.g., Example 2.3.1 and Example 2.3.2. In this paper we introduced a first order approximation (2.52) and a second order approximation (2.53) for the predictive probabilities of Gibbs-type priors, for any $\alpha \in (0,1)$ and any function h. In particular, we have proved that at the level of the first order approximation the function h has no impact on the predictive probabilities. Indeed Equation (2.52) coincides with the predictive probability of the normalized α-stable process, i.e. a Gibbs-type random probability measure with $\alpha \in (0,1)$ and $h(t) = 1$. However, it is sufficient to consider a second order approximation in order to take into account the function h. Indeed, Equation (2.53) coincides with the predictive probability of the two parameter Poisson–Dirichlet process in which the parameter θ is replaced by a suitable function of h. The proposed approximations thus highlight the role of the function h from a

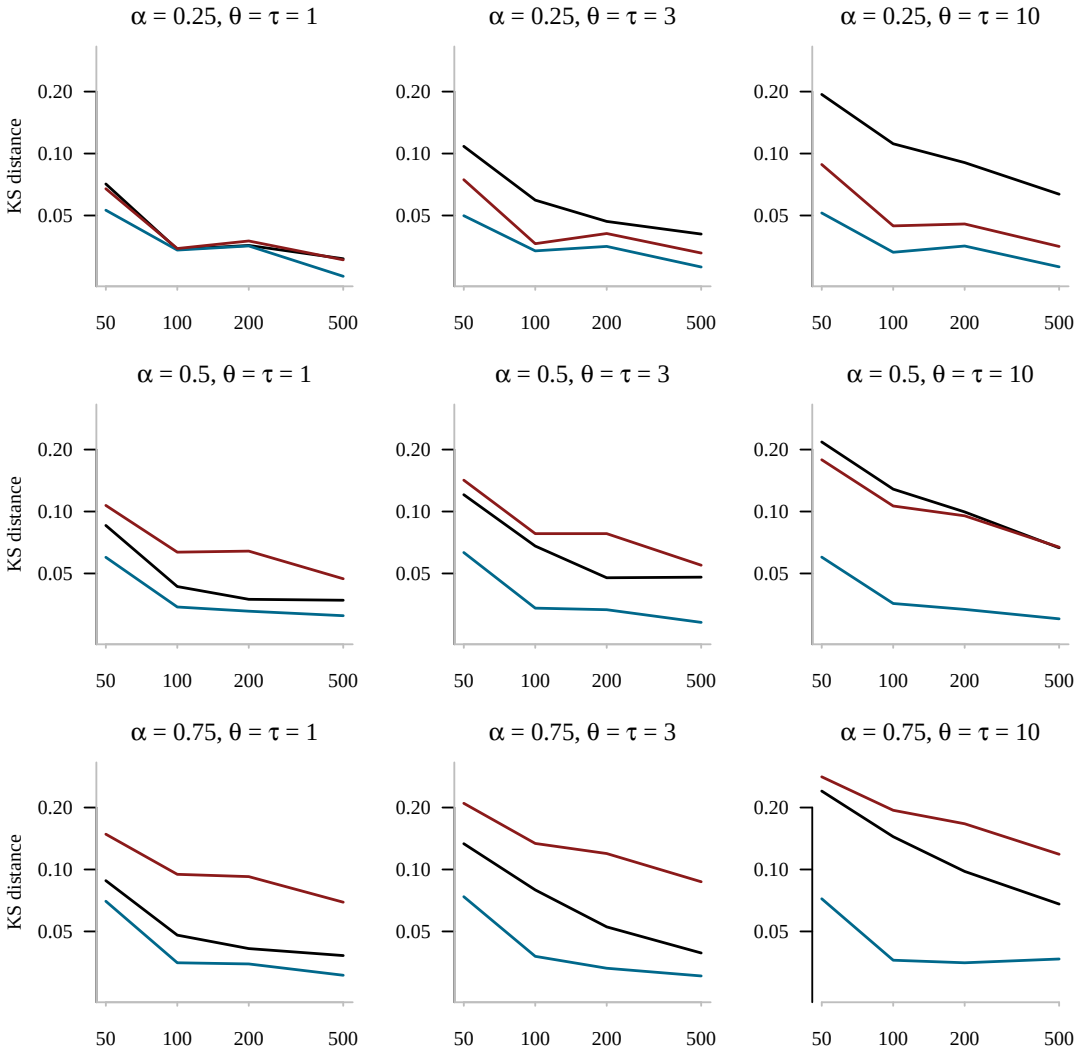


FIGURE 2.13: Kolmogorov–Smirnov distance between the true cdf and the cdf obtained by using the mixture model (2.69) with the following color code. In black: exact predictive distributions (2.47) of the Pitman–Yor process. In blue: first order approximation (2.52) of the predictive distributions of the Pitman–Yor process and the normalized generalized gamma process. In red: second order approximation (2.53) of the predictive distributions of the normalized generalized gamma process. The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\theta = \tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.

purely predictive perspective, and at the same time they provide practitioners with a way to easily handle the predictive probabilities of any Gibbs-type prior.

[S4] H. D. Nguyen, J. Arbel, H. Lü, and F. Forbes. Approximate Bayesian computation via the energy statistic. *Submitted*, 2019

2.4 Approximate Bayesian computation based on the energy distance

2.4.1 Introduction

In recent years, Bayesian inference has become a popular paradigm for machine learning and statistical analysis. Good introductions and references to the primary methods and philosophies of Bayesian inference can be found in texts such as [273], [132], [193], [195], [284], [41], and [239].

In this article, we are concerned with the problem of parametric, or classical Bayesian inference. For details regarding nonparametric Bayesian inference, the reader is referred to the expositions of [133], [158], and [131].

When conducting parametric Bayesian inference, we observe some realizations $\boldsymbol{x}$ of the data $\boldsymbol{X} \in \mathbb{X}$ that are generated from some data generating process (DGP), which can be characterized by a parametric likelihood, given by a probability density function (PDF) $f(\boldsymbol{x}|\boldsymbol{\theta})$, determined entirely via the parameter vector $\boldsymbol{\theta}$. Using the information that the parameter vector $\boldsymbol{\theta}$ is a realization of a random variable $\boldsymbol{\Theta} \in \mathbb{T}$, which arises from a DGP that can be characterized by some known prior PDF $\pi(\boldsymbol{\theta})$, we wish to characterize the posterior distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{c(\boldsymbol{x})}, \tag{2.70}$$

where

$$c(\boldsymbol{x}) = \int_{\mathbb{T}} f(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,.$$

In very simple cases, such as cases when the prior PDF is a conjugate of the likelihood (cf. [284, Sec. 3.3]), the posterior distribution (2.70) can be expressed explicitly. In the case of more complex but still tractable pairs of likelihood and prior PDFs, one can sample from (2.70) via a variety of Monte Carlo methods, such as those reported in [273, Ch. 6].

In cases where the likelihood function is known but not tractable, or when the likelihood function has entirely unknown form, one cannot exactly sample from (2.70) in an inexpensive manner, or at all. In such situations, a sample from an approximation of (2.70) may suffice in order to conduct the user's desired inference. Such a sample can be drawn via the method of approximate Bayesian computation (ABC).

It is generally agreed that the ABC paradigm originated from the works of [291], [316], and [274]; see [315] for details. Stemming from the initial listed works, there are now numerous variants of ABC methods. Some good reviews of the current ABC literature can be found in the expositions of [224], [332, Sec. 5.1], [216], and [184]. The volume of [302] provides a comprehensive treatment regarding ABC methodologies.

The core philosophy of ABC is to define a quasi-posterior by comparing data with plausibly simulated replicates. The comparison is traditionally based on some summary statistics, the choice of which being regarded as a key challenge of the approach.

In recent years, data discrepancy measures bypassing the construction of summary statistics have been proposed by viewing data sets as empirical measures. Examples of such an approach is via the use of the Kullback–Leibler divergence, the Wasserstein distance, or a maximum mean discrepancy (MMD) variant.

In this section, we develop upon the discrepancy measurement approach of [176], via the importance sampling ABC (IS-ABC) approach which makes use of a weight function [see e.g., 184]. In particular, we report on a class of ABC algorithms that utilize the two-sample energy statistic (ES) of [312] (see also [42], [313], and [314]). Our approach is related to the maximum MMD ABC algorithms that were implemented in [260], [176], and [52]. The MMD is a discrepancy measurement that is closely related to the ES (cf. [297]).

We establish new asymptotic results that have not been proved in these previous papers. In the IS-ABC setting and in the regime where both the observation sample size and the simulated data sample size increase to infinity, our theoretical result highlights how the data discrepancy measure impacts the asymptotic pseudo-posterior. More specifically, under the assumption that the data discrepancy measure converges to some asymptotic value $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, we show that the pseudo-posterior distribution converges almost surely to a distribution proportional to $\pi(\boldsymbol{\theta}) w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}))$: the prior distribution times the IS weight w function evaluated at $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, where $\boldsymbol{\theta}_0$ stands for the 'true' parameter value associated to the DGP that generates observations $\boldsymbol{X}$. Although devised in settings where likelihoods are assumed intractable, ABC can also be cast in the setting of robucitestness with respect to misspecification, where the ABC posterior distribution can be viewed as a special case of a coarsened posterior distribution [cf. 233].

The remainder of the article proceeds as follows. In Section 2.4.2, we introduce the general IS-ABC framework. In Section 2.4.3, we introduce the two-sample ES and demonstrate how it can be incorporated into the IS-ABC framework. Theoretical results regarding the IS-ABC framework and the two-sample ES are presented in Section 2.4.4. Illustrations of the IS-ABC framework are presented in Section 2.4.5. Conclusions are drawn in Section 2.4.6.

2.4.2 Importance sampling ABC

Assume that we observe n independent and identically distributed (IID) replicates of $\boldsymbol{X}$ from some DGP, which we put into $\mathbf{X}_n = \{\boldsymbol{X}_i\}_{i=1}^n$. We suppose that the DGP that generates $\boldsymbol{X}$ is dependent on some parameter vector $\boldsymbol{\theta}$, a realization of $\boldsymbol{\Theta}$ from space $\mathbb{T}$, which is random and has prior PDF $\pi(\boldsymbol{\theta})$.

Denote $f(\boldsymbol{x}|\boldsymbol{\theta})$ to be the PDF of $\boldsymbol{X}$, given $\boldsymbol{\theta}$, and write

$$f(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\theta}),$$

where $\mathbf{x}_n$ is a realization of $\mathbf{X}_n$, and each $\boldsymbol{x}_i$ is a realization of $\boldsymbol{X}_i$ ($i \in [n] = \{1, \ldots, n\}$).

If $f(\mathbf{x}_n|\boldsymbol{\theta})$ were known, then we could use (2.70) to write the posterior PDF

$$\pi(\boldsymbol{\theta}|\mathbf{x}_n) = \frac{f(\mathbf{x}_n|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{c(\mathbf{x}_n)}, \tag{2.71}$$

where $c(\mathbf{x}_n) = \int_{\mathbb{T}} f(\mathbf{x}_n|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is a constant that makes $\int_{\mathbb{T}} \pi(\boldsymbol{\theta}|\mathbf{x}_n)\,\mathrm{d}\boldsymbol{\theta} = 1$. When evaluating $f(\boldsymbol{x}|\boldsymbol{\theta})$ is prohibitive and ABC is required, then operating with $f(\mathbf{x}_n|\boldsymbol{\theta})$ is similarly difficult. We suppose that given any $\boldsymbol{\theta}_0 \in \mathbb{T}$, we at least have the capability of sampling from the DGP with PDF $f(\boldsymbol{x}|\boldsymbol{\theta}_0)$. That is, we have a simulation method that allows us to feasibly sample the IID vector $\mathbf{Y}_m = \{\boldsymbol{Y}_i\}_{i=1}^m$, for any $m \in \mathbb{N}$, for a DGP with PDF

$$f(\mathbf{y}_n|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(\boldsymbol{y}_i|\boldsymbol{\theta}).$$

Using the simulation mechanism that generates samples $\mathbf{Y}_m$ and the prior distribution that generates parameters $\boldsymbol{\Theta}$, we can simulate a set of $N \in \mathbb{N}$ simulations $\mathbf{Z}_N = \{\boldsymbol{Z}_{m,k}\}_{k=1}^N$, where $\boldsymbol{Z}_{m,k}^\top = \left(\mathbf{Y}_{m,k}^\top, \boldsymbol{\Theta}_k^\top\right)$ and $(\cdot)^\top$ is the transposition operator. Here, for each $k \in [N]$, $\boldsymbol{Z}_{m,k}$ is an observation from the DGP with joint PDF $f(\mathbf{y}_m|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$, hence each $\boldsymbol{Z}_{m,k}$ is composed of a parameter value and a datum conditional on the parameter value. We now consider how $\mathbf{X}_n$ and $\mathbf{Z}_N$ can be combined in order to construct an approximation of (2.71).

Following the approach of [176], we define $\mathcal{D}(\mathbf{x}_n, \mathbf{y}_m)$ to be some non-negative real-valued function that outputs a small value if $\mathbf{x}_n$ and $\mathbf{y}_m$ are similar, and outputs a large value if $\mathbf{x}_n$ and $\mathbf{y}_m$ are different, in some sense. We call $\mathcal{D}(\mathbf{x}_n, \mathbf{y}_m)$ the data discrepancy measurement between $\mathbf{x}_n$ and $\mathbf{y}_m$, and we say that $\mathcal{D}(\cdot,\cdot)$ is the data discrepancy function.

Next, we let $w(d, \epsilon)$ be a non-negative, decreasing (in d), and bounded (importance sampling) weight function (cf. Section 3 of [184]), which takes as inputs a data discrepancy measurement $d = \mathcal{D}(\mathbf{x}_n, \mathbf{y}_m) \geq 0$ and a calibration parameter $\epsilon > 0$. Using the weight and discrepancy functions, we can propose the following approximation for (2.71).

In the language of [176], we call

$$\pi_{m,\epsilon}(\boldsymbol{\theta}|\mathbf{x}_n) = \frac{\pi(\boldsymbol{\theta})\,L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta})}{c_{m,\epsilon}(\mathbf{x}_n)} \tag{2.72}$$

the quasi-posterior PDF, where

$$L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta}) = \int_{\mathbb{X}^m} w(\mathcal{D}(\mathbf{x}_n, \mathbf{y}_m), \epsilon)\, f(\mathbf{y}_m|\boldsymbol{\theta})\,\mathrm{d}\mathbf{y}_m$$

is the approximate likelihood function, and

$$c_{m,\epsilon}(\mathbf{x}_n) = \int_{\mathbb{T}} \pi(\boldsymbol{\theta})\,L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

is a normalization constant. We can use (2.72) to approximate (2.71) in the following way. For any functional of the parameter vector $\boldsymbol{\Theta}$ of interest, $g(\boldsymbol{\Theta})$ say, we may approximate the posterior Bayes estimator of $g(\boldsymbol{\Theta})$ via the expression

$$\mathbb{E}\left[g(\boldsymbol{\Theta})\,|\mathbf{x}_n\right] \approx \frac{\int_{\mathbb{T}} g(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{c_{m,\epsilon}(\mathbf{x}_n)}, \tag{2.73}$$

where the right-hand side of (2.73) can be unbiasedly estimated using $\mathbf{Z}_N$ via

$$\mathbb{M}\left[g(\boldsymbol{\Theta})\,|\mathbf{x}_n\right] = \frac{\sum_{k=1}^{N} g(\boldsymbol{\Theta}_k)\,w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_{m,k}), \epsilon)}{\sum_{k=1}^{N} w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_{m,k}), \epsilon)}. \tag{2.74}$$

We call the process of constructing (2.74), to approximate (2.73), the IS-ABC procedure. The general form of the IS-ABC procedure is provided in Algorithm 2.4.1.

Algorithm 2.4.1 IS-ABC procedure for approximating $\mathbb{E}\left[g(\boldsymbol{\Theta})\,|\mathbf{x}_n\right]$

Input: a data discrepancy function $\mathcal{D}$, a weight function w, and a calibration parameter $\epsilon > 0$.
For $k \in [N]$;
sample $\boldsymbol{\Theta}_k$ from the DGP with PDF $\pi(\boldsymbol{\theta})$;
generate $\mathbf{Y}_{m,k}$ from the DGP with PDF $f(\mathbf{y}_m|\boldsymbol{\Theta}_k)$;
put $\boldsymbol{Z}_k = (\mathbf{Y}_{m,k}, \boldsymbol{\Theta}_k)$ into $\mathbf{Z}_N$.
Output: $\mathbf{Z}_N$ and construct the estimator $\mathbb{M}\left[g(\boldsymbol{\Theta})\,|\mathbf{x}_n\right]$.

2.4.3 The energy statistic (ES)

Let δ define a metric and let $\boldsymbol{X} \in \mathbb{X} \subseteq \mathbb{R}^d$ and $\boldsymbol{Y} \in \mathbb{X}$ be two random variables that are in a metric space endowed with δ, where $d \in \mathbb{N}$. Furthermore, let $\boldsymbol{X}'$ and $\boldsymbol{Y}'$ be two random variables that have the same distributions as $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Here, $\boldsymbol{X}$, $\boldsymbol{X}'$, $\boldsymbol{Y}$, and $\boldsymbol{Y}'$ are all independent of one another.

Upon writing

$$\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y}) = 2\mathbb{E}\left[\delta(\boldsymbol{X}, \boldsymbol{Y})\right] - \mathbb{E}\left[\delta(\boldsymbol{X}, \boldsymbol{X}')\right] - \mathbb{E}\left[\delta(\boldsymbol{Y}, \boldsymbol{Y}')\right],$$

we can define the original ES of [42] and [312], as a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$, via the expression $\mathcal{E}_{\delta_2}(\boldsymbol{X}, \boldsymbol{Y})$, where $\delta_p(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p$ is the metric corresponding to the ℓ_p-norm ($p \in [1, \infty]$). Thus, the original ES statistic, which we shall also denote as $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y})$, is defined using the Euclidean norm δ_2.

The original ES has numerous useful mathematic properties. For instance, under the assumption that $\mathbb{E}\|\boldsymbol{X}\|_2 + \mathbb{E}\|\boldsymbol{Y}\|_2 < \infty$, it was shown that

$$\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi_X(\boldsymbol{t}) - \varphi_Y(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t}, \tag{2.75}$$

in Proposition 1 of [313], where $\Gamma(\cdot)$ is the gamma function and φ_X (respectively, φ_Y) is the characteristic function of $\boldsymbol{X}$ (respectively, $\boldsymbol{Y}$). Thus, we have the fact that

$\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ for any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{X}$, and $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

The result above is generalized in Proposition 3 of [313], where we have the following statement. If $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x} - \boldsymbol{y})$ is a continuous function and $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$ are independent random variables, then it is necessary and sufficient that $\delta(\cdot)$ is strictly negative definite (see [313] for the precise definition) for the following conclusion to hold: $\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ for any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{X}$, and $\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

We observe that there is thus an infinite variety of functions δ from which we can construct energy statistics. We shall concentrate on the use of the original ES, based on δ_2, since it is the most well known and popular of the varieties.

The V-statistic estimator

Suppose that we observe $\mathbf{X}_n = \{\boldsymbol{X}_i\}_{i=1}^n$ and $\mathbf{Y}_m = \{\boldsymbol{Y}_i\}_{i=1}^m$, where the former is a sample containing n IID replicates of $\boldsymbol{X}$, and the latter is a sample containing m IID replicates of $\boldsymbol{Y}$, respectively, with $\mathbf{X}_n$ and $\mathbf{Y}_m$ being independent. In [146], it was shown that for any δ, upon assuming that $\delta(\boldsymbol{x}, \boldsymbol{y}) < \infty$, the so-called V-statistic estimator (cf. [299, Ch. 5] and [196])

$$\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) = \frac{2}{mn} \sum_{i=1}^n \sum_{j=1}^m \delta(\boldsymbol{X}_i, \boldsymbol{Y}_j) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \delta(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \delta(\boldsymbol{Y}_i, \boldsymbol{Y}_j), \tag{2.76}$$

can be proved to converge in probability to $\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y})$, as $n \to \infty$ and $m \to \infty$, under the condition that $m/n \to \alpha < \infty$, for some constant α (see also [145]).

We note that the assumption of this result is rather restrictive, since it either requires the bounding of the space $\mathbb{X}$ or the function δ. In the sequel, we will present a result for the almost sure convergence of the V-statistic that depends on the satisfaction of a more realistic hypothesis.

It is noteworthy that if the ES is non-negative, then the V-statistic retains the non-negativity property of its corresponding ES (cf. [146]). That is, for any continuous and negative definite function $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x} - \boldsymbol{y})$, we have $\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) \geq 0$.

The ES-based IS-ABC algorithm

From Algorithm 2.4.1, we observe that an IS-ABC algorithm requires three components. A data discrepancy measurement $d = \mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) \geq 0$, a weighting function $w(d, \epsilon) \geq 0$, and a tuning parameter $\epsilon > 0$. We propose the use of the ES in the place of the data discrepancy measurement d, in combination with various weight functions that have been used in the literature. That is we set

$$\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) = \mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m),$$

in Algorithm 2.4.1.

In particular, we consider original ES, where $\delta = \delta_2$. We name our framework the ES-ABC algorithm. In Section 2.4.4, we shall demonstrate that the proposed algorithm possesses desirable large sample qualities that guarantees its performance in practice, as illustrated in Section 2.4.5.

Related methods

The ES-ABC algorithm that we have presented here is closely related to ABC algorithms based on the maximum mean discrepancy (MMD) that were implemented in [260], [176], and [52]. For each positive definite Mercer kernel function $\chi(\boldsymbol{x}, \boldsymbol{y})$ $(\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X})$, the corresponding MMD is defined via the equation

$$\mathrm{MMD}_{\chi}^{2}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\chi\left(\boldsymbol{X}, \boldsymbol{X}'\right)\right] + \mathbb{E}\left[\chi\left(\boldsymbol{Y}, \boldsymbol{Y}'\right)\right] - 2\mathbb{E}\left[\chi\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right],$$

where $\boldsymbol{X}, \boldsymbol{X}', \boldsymbol{Y}, \boldsymbol{Y}'$ are random variable such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed to $\boldsymbol{X}'$ and $\boldsymbol{Y}'$, respectively.

The MMD as a statistic for testing goodness-of-fit was studied prominently in articles such as [145], [147], and [146]. It is clear that if $\delta = -\chi$, the forms of the ES and the squared MMD are identical. More details regarding the relationship between the two classes of statistics can be found in [297].

We note two shortcomings with respect to the applications of the MMD as a basis for an ABC algorithm in the previous literature. Firstly, no theoretical results regarding the consistency of the MMD-based methods have been proved. And secondly, in the application by [260] and [176], the MMD was implemented using the unbiased U-statistic estimator, rather than the biased V-statistic estimator. Although both estimators are consistent, in the sense that they can be proved to be convergent to the desired limiting MMD value, the U-statistic estimator has the unfortunate property of not being bounded from below by zero (cf. [146]). As such, it does not meet the criteria for a data discrepancy measurement.

2.4.4 Theoretical results

General asymptotic analysis

We now establish a consistency result for the quasi-posterior density (2.72), when n and m approach infinity. Our result generalizes the main result of [176] (i.e., Theorem 1), which is the specific case when the weight function is restricted to the form

$$w(d, \epsilon) = [\![d < \epsilon]\!], \tag{2.77}$$

where $[\![\cdot]\!]$ is the Iverson bracket notation, which equals 1 when the internal statement is true, and 0, otherwise (cf. [140]).

The weighting function of form (2.77), when implemented within the IS-ABC framework, produces the common rejection ABC algorithms, that were suggested by [316], and [274]. We extended upon the result of [176] so that we may provide theoretical

guarantees for more exotic ABC procedures, such as the kernel-smoothed ABC procedure of [260], which implements weights of the form

$$w(d, \epsilon) = \exp\left(-d^q/\epsilon\right), \tag{2.78}$$

for $q > 0$. See [184] for further discussion and examples.

In order to prove our consistency result, we require Hunt's lemma, which is reported in [97], as Theorem 45 of Section V.5. For convenience to the reader, we present the result, below.

Theorem 2.4.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with increasing σ-fields $\{\mathcal{F}_n\}$ and let $\mathcal{F}_\infty = \cup_n \mathcal{F}_n$. Suppose that $\{U_n\}$ is a sequence of random variables that is bounded from above in absolute value by some integrable random variable V, and further suppose that U_n converges almost surely to the random variable U. Then, $\lim_{n\to\infty} \mathbb{E}(U_n|\mathcal{F}_n) = \mathbb{E}(U|\mathcal{F}_\infty)$ almost surely, and in $\mathcal{L}_1$ mean, as $n \to \infty$.*

Define the continuity set of a function $d \mapsto w(d)$ as

$$C(w) = \{d : w \text{ is continuous at } d\}.$$

Using Theorem 2.4.1, we can now prove the following result regarding the asymptotic behavior of the quasi-posterior density function (2.72).

Theorem 2.4.2. *Let $\mathbf{X}_n$ and $\mathbf{Y}_m$ be IID samples from DGPs that can be characterized by PDFs $f(\mathbf{x}_n|\boldsymbol{\theta}_0) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}_m|\boldsymbol{\theta}) = \prod_{i=1}^m f(\boldsymbol{y}_i|\boldsymbol{\theta})$, respectively, with corresponding parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. Suppose that the data discrepancy $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m)$ converges to some $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, which is a function of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, almost surely as $n \to \infty$, for some $m = m(n) \to \infty$. If $w(d, \epsilon)$ is piecewise continuous and decreasing in d and $w(d, \epsilon) \leq a < \infty$ for all $d \geq 0$ and any $\epsilon > 0$, and if*

$$\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \in C(w(\cdot, \epsilon)),$$

then we have

$$\pi_{m,\epsilon}(\boldsymbol{\theta}|\mathbf{x}_n) \to \frac{\pi(\boldsymbol{\theta}) w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)}{\int \pi(\boldsymbol{\theta}) w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon) \, \mathrm{d}\boldsymbol{\theta}}, \tag{2.79}$$

almost surely, as $n \to \infty$.

Proof. Using the notation of Theorem 2.4.1, we set $U_n = w(d(\mathbf{X}_n, \mathbf{Y}_m), \epsilon)$. Since $w(d, \epsilon) \leq a < \infty$, for any d, we have the existence of a $|U_n| \leq V < \infty$ such that V is integrable, since we can take $V = a$. Since $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, and $w(\cdot, \epsilon)$ is continuous at $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, we have $U_n \to U = w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$ with probability one by the extended continuous mapping theorem (cf. [92, Thm. 7.10]).

Now, let $\mathcal{F}_n$ be the σ-field generated by the sequence $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. Thus, $\mathcal{F}_n$ is an increasing σ-field, which approaches $\mathcal{F}_\infty = \cup_n \mathcal{F}_n$. We are in a position to directly apply Theorem 2.4.1. This yields

$$\mathbb{E}[w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m), \epsilon)|\mathbf{X}_n] \to \mathbb{E}[w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)|\mathbf{X}_\infty],$$

almost surely, as $n \to \infty$, where the right-hand side equals $w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$.

Notice that the left-hand side has the form

$$\mathbb{E}\left[w\left(\mathcal{D}\left(\mathbf{X}_n, \mathbf{Y}_m\right), \epsilon\right) | \mathbf{X}_n\right] = L_{m,\epsilon}\left(\mathbf{X}_n | \boldsymbol{\theta}\right)$$

and therefore $L_{m,\epsilon}\left(\mathbf{X}_n|\boldsymbol{\theta}\right) \to w\left(\mathcal{D}_\infty\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}\right), \epsilon\right)$, almost surely, as $n \to \infty$. Thus, the numerator of (2.72) converges to

$$\pi\left(\boldsymbol{\theta}\right) w\left(\mathcal{D}_\infty\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}\right), \epsilon\right), \tag{2.80}$$

almost surely.

To complete the proof, it suffices to show that the denominator of (2.72) converges almost surely to

$$\int_{\mathbb{T}} \pi\left(\boldsymbol{\theta}\right) w\left(\mathcal{D}_\infty\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}\right), \epsilon\right) \mathrm{d}\boldsymbol{\theta}. \tag{2.81}$$

Since $L_{m,\epsilon}\left(\mathbf{X}_n|\boldsymbol{\theta}\right) \to w\left(\mathcal{D}_\infty\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}\right), \epsilon\right)$ and $c_{m,\epsilon}\left(\mathbf{x}_n\right) = \int_{\mathbb{T}} \pi\left(\boldsymbol{\theta}\right) L_{m,\epsilon}\left(\mathbf{x}_n|\boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta}$, we obtain our desired convergence via the dominated convergence theorem, because $w\left(d, \epsilon\right) \leq a < \infty$. An application of a Slutsky-type theorem yields the almost sure convergence of the ratio between (2.80) and (2.81) to the right-hand side of (2.79), as $n \to \infty$. ■

The following result and proof guarantees the applicability of Theorem 2.4.2 to rejection ABC procedures, and to kernel-smoothed ABC procedures, as used in [176] and [260], respectively.

Proposition 2.4.1. *The result of Theorem 2.4.2 applies to rejection ABC and importance sampling ABC, with weight functions of respective forms (2.77) and (2.78).*

Proof. For weights of form (2.77), we note that $w\left(d, \epsilon\right) = [\![d < \epsilon]\!]$ is continuous in d at all points, other than when $d = \epsilon$. Furthermore, $w\left(d, \epsilon\right) \in \{0, 1\}$ and is hence non-negative and bounded. Thus, under the condition that $\mathcal{D}_\infty\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}\right) \neq \epsilon$, we have the desired conclusion of Theorem 2.4.2.

For weights of form (2.78), we note that for fixed ϵ, $w\left(d, \epsilon\right)$ is continuous and positive in d. Since w is uniformly bounded by 1, differentiating with respect to d, we obtain $\mathrm{d}w/\mathrm{d}d = -\left(q/\epsilon\right) d^{q-1} \exp\left(-d^q/\epsilon\right)$, which is negative for any $d \geq 0$ and $q > 0$. Thus, (2.78) constitutes a weight function and satisfies the conditions of Theorem 2.4.2.

■

Asymptotic of the energy statistic

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be arbitrary elements of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively. That is $\boldsymbol{X}$ and $\boldsymbol{Y}$ arise from DGPs that can be characterized by PDFs $f\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)$ and $f\left(\boldsymbol{y}; \boldsymbol{\theta}\right)$, respectively. Under the assumption $\mathbb{E}\left\|\boldsymbol{X}\right\|_2 + \mathbb{E}\left\|\boldsymbol{Y}\right\|_2 < \infty$, Proposition 1 of [313] states that we can write the ES as

$$\mathcal{E}\left(\boldsymbol{X}, \boldsymbol{Y}\right) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{\left|\varphi\left(\boldsymbol{t}; \boldsymbol{\theta}_0\right) - \varphi\left(\boldsymbol{t}; \boldsymbol{\theta}\right)\right|^2}{\left\|\boldsymbol{t}\right\|_2^{d+1}} \mathrm{d}\boldsymbol{t}, \tag{2.82}$$

where $\varphi(\boldsymbol{t};\boldsymbol{\theta})$ is the characteristic function corresponding to the PDF $f(\boldsymbol{y};\boldsymbol{\theta})$.

We write $\log^+ x = \log(\max\{1, x\})$. From [312] we have the fact that for arbitrary δ,

$$\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) = \frac{1}{n^2 m^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} \kappa_\delta(\boldsymbol{X}_{i_1}, \boldsymbol{X}_{i_2}; \boldsymbol{Y}_{j_1}, \boldsymbol{Y}_{j_2}),$$

where

$$\kappa_\delta(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}; \boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2}) = \delta(\boldsymbol{x}_{i_1}, \boldsymbol{y}_{j_1}) + \delta(\boldsymbol{x}_{i_2}, \boldsymbol{y}_{j_2}) - \delta(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) - \delta(\boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2})$$

is the kernel of the V-statistic that is based on the function δ. The following result is a direct consequence of Theorem 1 of [298], when applied to V-statistics constructed from functionals δ that satisfy the hypothesis of [313, Prop. 3].

Lemma 2.4.1. *Make the same assumptions regarding $\mathbf{X}_n$ and $\mathbf{Y}_m$ as in Theorem 2.4.2. Let $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x} - \boldsymbol{y})$ be a continuous and strictly negative definite function. If*

$$\mathbb{E}\left(|\kappa_\delta(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{Y}_1, \boldsymbol{Y}_2)| \log^+ |\kappa_\delta(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{Y}_1, \boldsymbol{Y}_2)|\right) < \infty, \tag{2.83}$$

then $\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{E}_\delta(\boldsymbol{X}_1, \boldsymbol{Y}_1) \geq 0$, as $\min\{n, m\} \to \infty$, where $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{X}$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{X}$ are arbitrary elements of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively. Furthermore, $\mathcal{E}_\delta(\boldsymbol{X}_1, \boldsymbol{Y}_1) = 0$ if and only if $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ are identically distributed.

We may apply the result of Lemma 2.4.1 directly to the case of $\delta = \delta_2$ in order to provide an almost sure convergence result regarding the V-statistic $\mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$.

Corollary 2.4.1. *Make the same assumptions regarding $\mathbf{X}_n$ and $\mathbf{Y}_m$ as in Theorem 2.4.2. If $\boldsymbol{X}_1 \in \mathbb{X}$ and $\boldsymbol{Y}_1 \in \mathbb{X}$ are arbitrary elements of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively, and*

$$\mathbb{E}\left(\|\boldsymbol{X}_1\|_2^2\right) + \mathbb{E}\left(\|\boldsymbol{Y}_1\|_2^2\right) < \infty, \tag{2.84}$$

and if $\min\{n, m\} \to \infty$, then $\mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{E}(\boldsymbol{X}_1, \boldsymbol{Y}_1)$, of form (2.82).

Proof. By the law of total expectation, we apply Lemma 2.4.1 by considering the two cases of (2.83): when $|\kappa_{\delta_2}| \leq 1$ and when $|\kappa_{\delta_2}| > 1$, separately, to write

$$\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}|\right) = p_0 \mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}| \,|\, |\kappa_{\delta_2}| \leq 1\right) + p_1 \mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}| \,|\, |\kappa_{\delta_2}| > 1\right), \tag{2.85}$$

where $p_0 = \mathbb{P}(|\kappa_{\delta_2}| \leq 1)$ and $p_1 = \mathbb{P}(|\kappa_{\delta_2}| > 1)$. The first term on the right-hand side of (2.85) is equal to zero, since $\log^+ |\kappa_{\delta_2}| = \log(1) = 0$, whenever $|\kappa_{\delta_2}| \leq 1$. Thus, we need only be concerned with bounding the second term.

For $|\kappa_{\delta_2}| > 1$, $|\kappa_{\delta_2}| \log |\kappa_{\delta_2}| \leq |\kappa_{\delta_2}|^2$, thus

$$\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}| \,|\, |\kappa_{\delta_2}| > 1\right) \leq \mathbb{E}\left(|\kappa_{\delta_2}|^2 \,|\, |\kappa_{\delta_2}| > 1\right)$$

The condition that $\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}|\right) < \infty$ is thus fulfilled if $\mathbb{E}\left(|\kappa_{\delta_2}|^2 \,|\, |\kappa_{\delta_2}| > 1\right) < \infty$, which is equivalent to

$$\mathbb{E}\left(|\kappa_{\delta_2}|^2\right) = p_0 \mathbb{E}\left(|\kappa_{\delta_2}|^2 \,|\, |\kappa_{\delta_2}| \leq 1\right) + p_1 \mathbb{E}\left(|\kappa_{\delta_2}|^2 \,|\, |\kappa_{\delta_2}| > 1\right) < \infty,$$

by virtue of the integrability of $\left\{ |\kappa_{\delta_2}|^2 \,|\, |\kappa_{\delta_2}| \leq 1 \right\}$ implying the existence of

$$\mathbb{E}\left(|\kappa_{\delta_2}|^2 \,|\, |\kappa_{\delta_2}| \leq 1 \right),$$

since it is defined on a bounded support.

Next, by the triangle inequality, $|\kappa_{\delta_2}| \leq 2\left(\|\boldsymbol{X}_1\|_2 + \|\boldsymbol{X}_2\|_2 + \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{Y}_2\|_2\right)$, and hence

$$\begin{aligned} |\kappa_{\delta_2}|^2 \leq 4 & \left(\|\boldsymbol{X}_1\|_2^2 + \|\boldsymbol{X}_2\|_2^2 + \|\boldsymbol{Y}_1\|_2^2 + \|\boldsymbol{Y}_2\|_2^2 \right) \\ & + 8(\|\boldsymbol{X}_1\|_2 \|\boldsymbol{X}_2\|_2 + \|\boldsymbol{X}_1\|_2 \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{X}_1\|_2 \|\boldsymbol{Y}_2\|_2 \\ & + \|\boldsymbol{X}_2\|_2 \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{X}_2\|_2 \|\boldsymbol{Y}_2\|_2 + \|\boldsymbol{Y}_1\|_2 \|\boldsymbol{Y}_2\|_2). \end{aligned}$$

Since $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2$ are all pairwise independent, and $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ are identically distributed to $\boldsymbol{X}_2$ and $\boldsymbol{Y}_2$, respectively, we have

$$\begin{aligned} \mathbb{E}\left(|\kappa_{\delta_2}|^2 \right) \leq & \, 8\left[\mathbb{E}\left(\|\boldsymbol{X}_1\|_2^2 \right) + \mathbb{E}\left(\|\boldsymbol{Y}_1\|_2^2 \right) \right] + 8\left[\left(\mathbb{E}\|\boldsymbol{X}_1\|_2 \right)^2 + \left(\mathbb{E}\|\boldsymbol{Y}_1\|_2 \right)^2 \right] \\ & + 32\left[\mathbb{E}\|\boldsymbol{X}_1\|_2 \, \mathbb{E}\|\boldsymbol{Y}_1\|_2 \right], \end{aligned}$$

which concludes the proof since $\mathbb{E}\|\boldsymbol{X}_1\|_2^2 + \mathbb{E}\|\boldsymbol{Y}_1\|_2^2 < \infty$ is satisfied by the hypothesis and implies $\mathbb{E}\|\boldsymbol{X}_1\|_2 + \mathbb{E}\|\boldsymbol{Y}_1\|_2 < \infty$.

■

We note that condition (2.84) is stronger than a direct application of condition (2.83), which may be preferable in some situations. However, condition (2.84) is somewhat more intuitive and verifiable since it is concerned with the polynomial moments of norms and does not involve the piecewise function $\log^+ x$. It is also suggested in [345] that one may replace $\log^+ x$ by $\log(2 + x)$ if it is more convenient to do so.

Combining the result of Theorem 2.4.2 with Corollary 2.4.1 and the conclusion from Proposition 1 of [313] provided in Equation (2.82) yields the key result below. This result justifies the use of the V-statistic estimator $\mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$ for the energy distance $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y})$ within the IS-ABC framework.

Corollary 2.4.2. *Under the assumptions of Corollary 2.4.1. If* $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) = \mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$, *then the conclusion of Theorem 2.4.2 follows with*

$$\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) \to \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi(\boldsymbol{t}; \boldsymbol{\theta}_0) - \varphi(\boldsymbol{t}; \boldsymbol{\theta})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t} = \mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}),$$

almost surely, as $n \to \infty$, *where* $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq 0$ *and* $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0$, *if and only if* $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$.

2.4.5 Illustrations

We illustrate the use of the ES on some standard models. The standard rejection ABC algorithm is employed (that is, we use Algorithm 2.4.1 with weight function w of form (2.77)) for constructing estimators (2.74). The proposed ES is compared to

the Kullback–Leibler divergence (KL), the Wasserstein distance (WA), and the maximum mean discrepancy (MMD). Here, the ES is applied using the Euclidean metric δ_2, the Wasserstein distance using the exponent $p = 2$ [cf. 52] and the MMD using a Gaussian kernel $\chi(\boldsymbol{x}, \boldsymbol{y}) = \exp\left[-(\boldsymbol{x} - \boldsymbol{y})^2\right]$. The Gaussian kernel is commonly used in the MMD literature, and was also considered for ABC in [260] and [176]. Details regarding the use of the Kullback–Leibler divergence as a discrepancy function for ABC algorithms can be found in [176, Sec. 2].

We use $\boldsymbol{X} \sim \mathcal{L}$ to denote that the random variable $\boldsymbol{X}$ has probability law $\mathcal{L}$. Furthermore, we denote the normal law by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ states that the DGP of $\boldsymbol{X}$ is multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We further denote the uniform law, in the interval (a, b), for $a < b$, by $\text{Unif}(a, b)$.

We consider examples explored in [176, Sec. 4.1]. For each illustration below, we sample synthetic data of the same size m as the observed data size, n, whose value is specified for each model below. We consider only the rejection weight function, and the number of ABC iterations in Algorithm 2.4.1 is set to $N = 10^5$. The tuning parameter ϵ is set so that only the 0.05% smallest discrepancies are kept to form ABC posterior sample. We postpone the discussion of the results of our simulation experiments to Section 2.4.5

The experiments were implemented in R, using in particular the winference package [52] and the FNN package [54]. The Kullback–Leibler divergence between two PDFs is computed within the 1-nearest neighbor framework [59]. Moreover, the k-d trees is adopted for implementing the nearest neighbor search, which is the same as the method of [176]. For estimating the 2-Wasserstein distance between two multivariate empirical measures, we propose to employ the swapping algorithm [276], which is simple to implement, and is more accurate and less computationally expensive than other algorithms commonly used in the literature [52]. Regarding the MMD, the same unbiased U-statistic estimator is adopted as given in [176] and [260]. For reproduction of the the experimental results, the original source code can be accessed at `https://github.com/hiendn/Energy_Statistics_ABC`.

Bivariate Gaussian mixture model

Let $\mathbf{X}_n$ be a sequence of IID random variables, such that each $\boldsymbol{X}_i$ has mixture of Gaussian probability law

$$\boldsymbol{X}_i \sim p\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + (1 - p)\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \tag{2.86}$$

with known covariance matrices

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}.$$

We aim to estimate the generative parameters $\boldsymbol{\theta}^\top = (p, \boldsymbol{\mu}_0^\top, \boldsymbol{\mu}_1^\top)$ consisting of the mixing probability p and the population means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. To this end, we perform ABC using $n = 500$ observations, sampled from model (2.86) with $p = 0.3$,

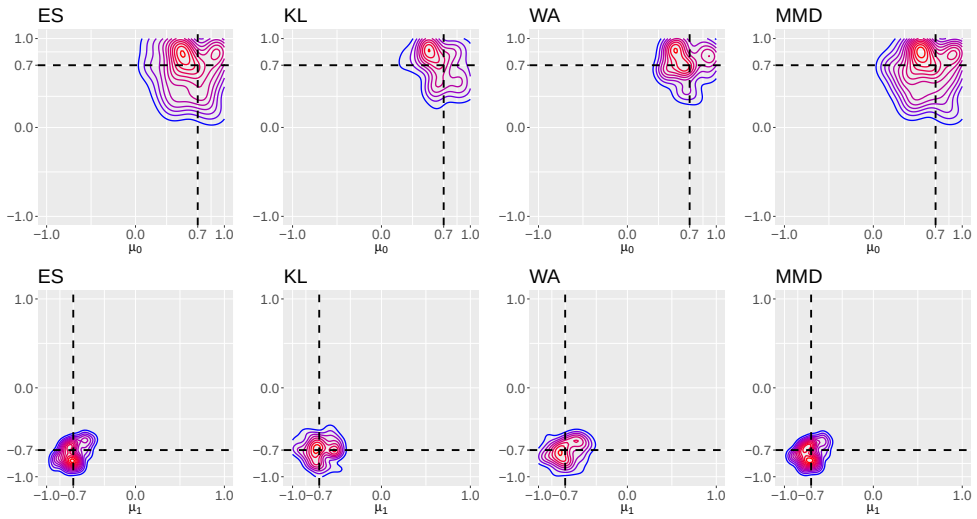


FIGURE 2.14: Marginal KDEs of the ABC posterior for the mean parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ of the bivariate Gaussian mixture model (2.86). The intersections of black dashed lines indicate the positions of the population means.

$\boldsymbol{\mu}_0^\top = (0.7, 0.7)$ and $\boldsymbol{\mu}_1^\top = (-0.7, -0.7)$. A kernel density estimate (KDE) of the ABC posterior distribution is presented in Figure 2.14.

Moving-average model of order 2

The moving-average model of order q, MA(q), is a stochastic process $\{Y_t\}_{t\in\mathbb{N}^*}$ defined as

$$Y_t = Z_t + \sum_{i=1}^{q} \theta_i Z_{t-i},$$

with $\{Z_t\}_{t\in\mathbb{Z}}$ being a sequence of unobserved noise error terms. [176] used a MA(2) model for their benchmarking; namely $Y_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$, $t \in [d]$. Each observation $\boldsymbol{Y}$ corresponds to a time series of length d. Here, we use the same model as that proposed in [176], where Z_t follows the Student-t distribution with 5 degrees of freedom, and $d = 10$. The priors on the model parameters θ_1 and θ_2 are taken to be uniform, that is, $\theta_1 \sim \text{Unif}(-2, 2)$ and $\theta_2 \sim \text{Unif}(-1, 1)$. We performed ABC using $n = 200$ samples generated from a model with the true parameter values $(\theta_1, \theta_2) = (0.6, 0.2)$. A KDE of the ABC posterior distribution is displayed in Figure 2.15.

Bivariate beta model

The bivariate beta model proposed by [86] is defined with five positive parameters $\theta_1, \ldots, \theta_5$ by letting

$$V_1 = \frac{U_1 + U_3}{U_5 + U_4}, \text{ and } V_2 = \frac{U_2 + U_4}{U_5 + U_3}, \tag{2.87}$$

where $U_i \sim \text{Gamma}(\theta_i, 1)$, for $i \in [5]$, and setting $Z_1 = V_1/(1 + V_1)$ and $Z_2 = V_2/(1 + V_2)$. The bivariate random variable $\boldsymbol{Z}^\top = (Z_1, Z_2)$ has marginal laws $Z_1 \sim \text{Beta}(\theta_1 + \theta_3, \theta_5 + \theta_4)$ and $Z_2 \sim \text{Beta}(\theta_2 + \theta_4, \theta_5 + \theta_3)$. We performed ABC using

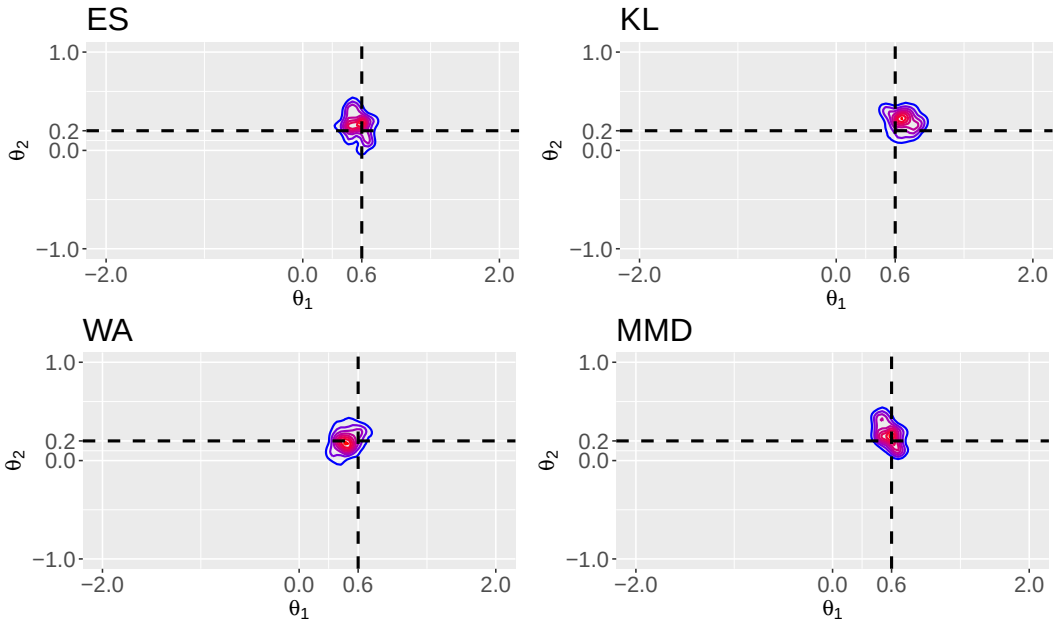


FIGURE 2.15: KDE of the ABC posterior for the parameters θ_1 and θ_2 of the MA(2) model experiment. The intersections of black dashed lines indicate the true parameter values.

samples of size $n = 500$, which are generated from a DGP with true parameter values $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1, 1, 1, 1, 1)$. The prior on each of the model parameters is taken to be independent Unif$(0, 5)$. A KDE of the ABC posterior distribution is displayed in Figure 2.16.

Multivariate g-and-k distribution

A univariate g-and-k distribution can be defined via its quantile function [102]:

$$F^{-1}(x) = A + B\left[1 + 0.8\frac{1 - \exp(-g \times z_x)}{1 + \exp(-g \times z_x)}\right](1 + z_x^2)^k z_x, \tag{2.88}$$

where parameters (A, B, g, k) respectively relate to location, scale, skewness, and kurtosis. Here, z_x is the xth quantile of the standard normal distribution. Given a set of parameters (A, B, g, k), it is easy to simulate d observations of a DGP with quantile function (2.88), by generating a sequence of IID sample $\{Z_i\}_{i=1}^d$, where $Z_i \sim \mathcal{N}(0, 1)$, for $i \in [d]$.

A so-called d-dimensional g-and-k DGP can instead be defined by applying the quantile function (2.88) to each of the d elements of a multivariate normal vector $\mathbf{Z}^\top = (Z_1, ..., Z_d) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a covariance matrix. In our experiment, we use a 5-dimensional g-and-k model with the same covariance matrix and parameter values for (A, B, g, k) as that considered by [176]. That is, we generate samples of size $n = 200$ from a g-and-k DGP with the true parameter values

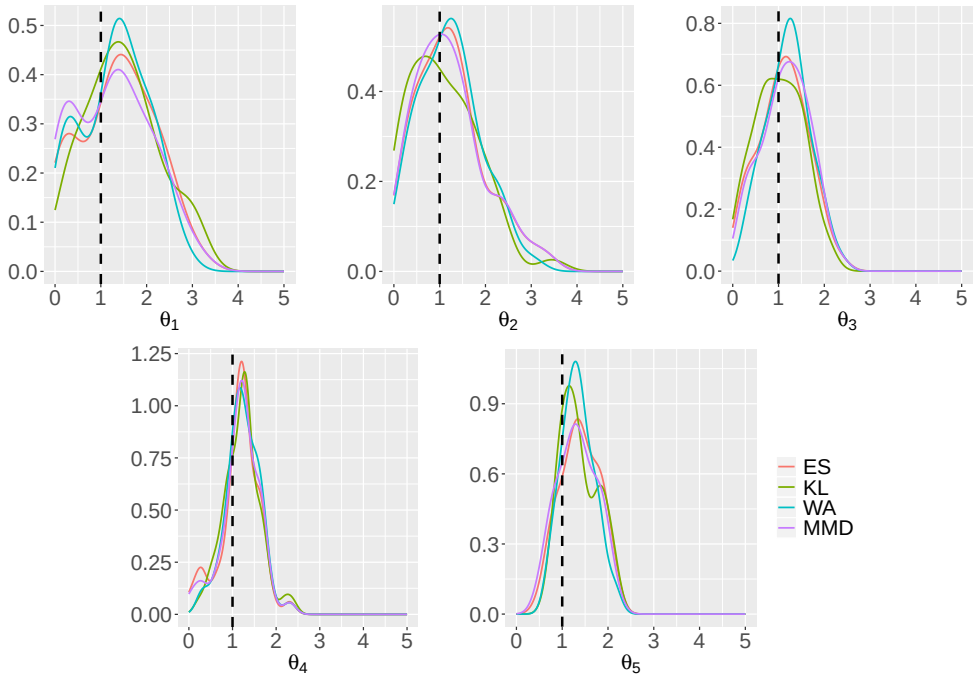


FIGURE 2.16: Marginal KDEs of the ABC posterior for the parameters $\theta_1, \ldots, \theta_5$ for the bivariate beta model. The black dashed lines indicate the true parameter values.

$(A, B, g, k) = (3, 1, 2, 0.5)$ and the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & \rho & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{bmatrix},$$

where $\rho = -0.3$. Marginal KDEs of the ABC posterior distributions is presented in Figure 2.17.

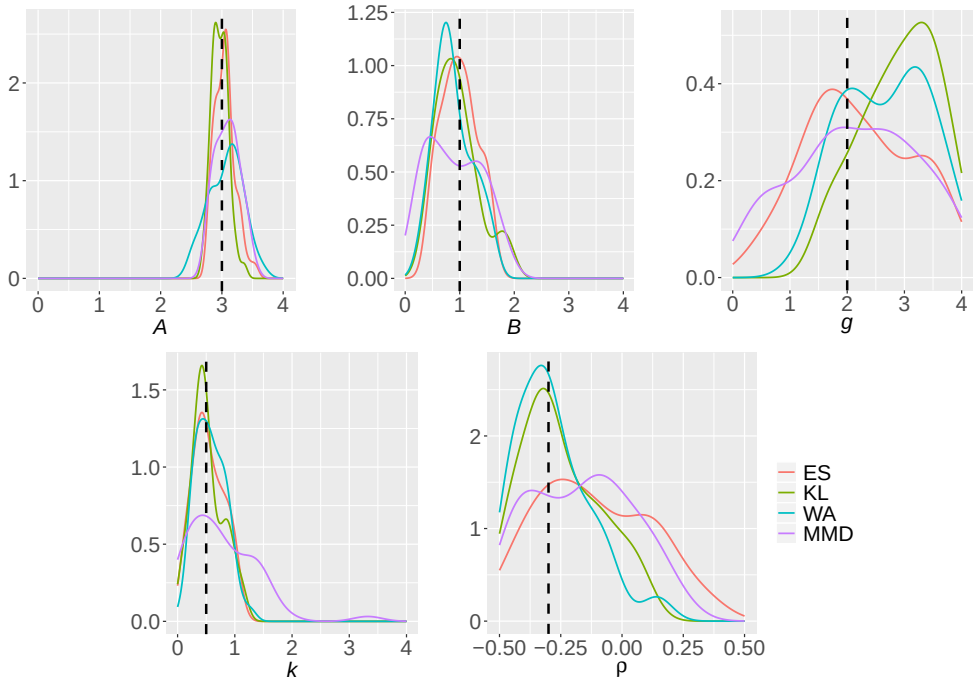


FIGURE 2.17: Marginal KDEs of the ABC posterior for the parameters A, B, g, k and ρ of the g-and-k model. The black dashed lines indicate the true parameter values.

Discussion of the results and performance

For each of the four experiments and each parameter, we computed the posterior mean $\hat{\theta}_{\text{mean}}$, posterior median $\hat{\theta}_{\text{med}}$, mean absolute error and mean squared error defined by

$$\text{MAE} = \frac{1}{M}\sum_{k=1}^{M}|\theta_k - \theta_0|, \text{ and } \quad \text{MSE} = \frac{1}{M}\sum_{k=1}^{M}|\theta_k - \theta_0|^2,$$

where $\{\theta_k\}_{k=1}^{M}$ denotes the pseudo-posterior sample and θ_0 denotes the true parameter. Here $M = 50$ since $N = 10^5$ and ϵ is chosen as to retain 0.05% of the samples. Each experiment was replicated ten times by keeping the same fixed (true) values for the parameters and by sampling new observed data each of the ten times. The estimated quantities $\hat{\theta}_{\text{mean}}$, $\hat{\theta}_{\text{med}}$, and errors MAE and RMSE $= \text{MSE}^{1/2}$ were then averaged over the ten replications, and are reported along with standard deviations $\sigma(\cdot)$ in columns associated with each estimator and true values θ_0 for each parameter in Tables 2.9, 2.10, 2.11 and 2.12.

Upon inspection, Tables 2.9, 2.10, 2.11 and 2.12 showed some advantage in performance from WA on the bivariate Gaussian mixtures, some advantage from the MMD on the bivariate beta model, and some advantage from the ES on the g-and-k model, while multiple methods are required to make the best inference in the case of the MA(2) experiment. When we further take into account the standard deviations of the estimators, we observe that all four data discrepancy measures essentially perform comparatively well across the four experimental models. Thus, we may conclude that there is no universally best performing discrepancy measure, and one must choose the right method for each problem of interest. Alternatively, one may also consider some kind of averaging over the results of the different discrepancy measures. We have not committed to an investigation of such methodologies and leave it as a future research direction.

TABLE 2.9: Estimation performance for bivariate Gaussian mixtures (Section 2.4.5). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$\mu_{00} = 0.7$	ES	0.594	0.045	0.607	0.063	0.215	0.030	0.283	0.055
	KL	0.648	0.039	0.666	0.048	0.165	0.016	0.205	0.026
	WA	**0.675**	0.035	**0.682**	0.043	**0.152**	0.020	**0.181**	0.021
	MMD	0.564	0.079	0.582	0.076	0.234	0.054	0.311	0.101
$\mu_{01} = 0.7$	ES	0.587	0.063	0.613	0.059	0.215	0.038	0.282	0.069
	KL	0.651	0.042	0.667	0.061	0.169	0.022	0.210	0.027
	WA	**0.655**	0.050	**0.669**	0.047	**0.152**	0.015	**0.187**	0.019
	MMD	0.559	0.076	0.598	0.075	0.235	0.049	0.313	0.092
$\mu_{10} = -0.7$	**ES**	**-0.699**	0.046	-0.716	0.040	1.401	0.043	1.412	0.039
	KL	-0.709	0.029	-0.712	0.035	1.409	0.029	1.415	0.029
	WA	**-0.699**	0.030	**-0.704**	0.037	**1.399**	0.030	**1.404**	0.030
	MMD	-0.709	0.054	-0.731	0.036	1.411	0.051	1.422	0.038
$\mu_{11} = -0.7$	**ES**	**-0.696**	0.058	-0.712	0.043	1.396	0.058	1.407	0.049
	KL	-0.711	0.047	**-0.704**	0.057	1.411	0.047	1.416	0.047
	WA	-0.695	0.043	-0.695	0.053	**1.395**	0.043	**1.401**	0.043
	MMD	-0.711	0.066	-0.726	0.046	1.411	0.066	1.424	0.052

TABLE 2.10: Estimation performance for the MA(2) model (Section 2.4.5). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$\theta_1 = 0.6$	ES	0.569	0.042	0.570	0.045	0.083	0.015	0.100	0.017
	KL	0.664	0.028	0.658	0.031	0.106	0.017	0.132	0.019
	WA	0.509	0.033	0.505	0.038	0.112	0.022	0.133	0.026
	MMD	**0.583**	0.044	**0.586**	0.048	**0.079**	0.013	**0.096**	0.015
$\theta_2 = 0.2$	ES	0.215	0.035	0.219	0.035	0.111	0.015	0.135	0.019
	KL	0.274	0.023	0.280	0.027	0.110	0.014	0.134	0.014
	WA	**0.205**	0.025	**0.207**	0.030	**0.090**	0.029	**0.112**	0.034
	MMD	0.220	0.037	0.220	0.036	0.108	0.010	0.132	0.012

TABLE 2.11: Estimation performance for the bivariate beta model (Section 2.4.5). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$\theta_1 = 1.0$	ES	1.299	0.223	1.189	0.264	0.713	0.130	0.885	0.165
	KL	1.389	0.190	1.333	0.165	0.696	0.151	0.877	0.205
	WA	**1.286**	0.220	1.193	0.265	0.672	0.128	**0.828**	0.153
	MMD	1.229	0.188	**1.143**	0.241	**0.676**	0.092	0.836	0.121
$\theta_2 = 1.0$	ES	1.362	0.185	1.290	0.237	0.716	0.118	0.904	0.131
	KL	**1.235**	0.152	**1.153**	0.170	**0.588**	0.070	**0.745**	0.097
	WA	1.292	0.196	1.240	0.241	0.657	0.114	0.817	0.139
	MMD	1.268	0.173	1.170	0.171	0.669	0.103	0.841	0.131
$\theta_3 = 1.0$	ES	1.170	0.132	1.183	0.157	0.459	0.045	0.552	0.049
	KL	**1.083**	0.100	**1.077**	0.088	**0.394**	0.034	**0.496**	0.045
	WA	1.229	0.118	1.216	0.132	0.426	0.054	0.521	0.059
	MMD	1.181	0.116	1.182	0.143	0.456	0.051	0.548	0.061
$\theta_4 = 1.0$	**ES**	**1.128**	0.112	1.113	0.138	0.435	0.032	0.534	0.045
	KL	1.133	0.111	**1.086**	0.135	**0.390**	0.038	**0.498**	0.051
	WA	1.218	0.110	1.196	0.108	0.409	0.049	0.514	0.066
	MMD	1.150	0.098	1.133	0.130	0.423	0.041	0.518	0.049
$\theta_5 = 1.0$	ES	1.343	0.096	1.360	0.104	0.428	0.052	0.514	0.059
	KL	1.300	0.087	1.250	0.065	0.384	0.040	0.491	0.061
	WA	1.300	0.101	1.298	0.105	**0.370**	0.058	**0.446**	0.066
	MMD	**1.258**	0.115	**1.232**	0.120	0.375	0.055	0.454	0.063

TABLE 2.12: Estimation performance for the g-and-k distribution (Section 2.4.5). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$A = 3.0$	**ES**	**3.024**	0.044	**3.009**	0.047	0.133	0.016	0.170	0.018
	KL	2.955	0.030	2.948	0.033	**0.105**	0.013	**0.128**	0.013
	WA	3.043	0.045	3.052	0.067	0.232	0.020	0.277	0.020
	MMD	3.081	0.061	3.062	0.065	0.177	0.029	0.221	0.036
$B = 1.0$	**ES**	**1.046**	0.062	**1.027**	0.079	**0.268**	0.024	**0.322**	0.029
	KL	0.918	0.071	0.885	0.068	0.313	0.026	0.375	0.029
	WA	0.894	0.127	0.869	0.136	0.277	0.044	0.334	0.045
	MMD	0.899	0.069	0.855	0.079	0.374	0.029	0.440	0.030
$g = 2.0$	ES	2.289	0.101	2.264	0.210	0.872	0.098	1.026	0.091
	KL	2.993	0.080	3.046	0.121	1.043	0.070	1.193	0.066
	WA	2.581	0.101	2.599	0.147	**0.858**	0.078	**1.025**	0.075
	MMD	**2.184**	0.128	**2.227**	0.190	0.904	0.103	1.052	0.100
$k = 0.5$	**ES**	**0.476**	0.046	0.444	0.067	0.225	0.014	0.270	0.015
	KL	0.550	0.059	**0.498**	0.064	0.252	0.029	0.317	0.045
	WA	0.544	0.095	0.526	0.094	**0.189**	0.035	**0.238**	0.046
	MMD	0.691	0.056	0.621	0.072	0.380	0.041	0.502	0.070
$\rho = -0.3$	ES	-0.163	0.047	-0.178	0.069	0.197	0.032	0.246	0.034
	KL	**-0.291**	0.034	-0.324	0.037	**0.117**	0.014	**0.144**	0.020
	WA	-0.288	0.026	**-0.314**	0.035	0.125	0.016	0.152	0.020
	MMD	-0.194	0.047	-0.210	0.063	0.174	0.030	0.218	0.035

2.4.6 Discussion

We have introduced a novel importance-sampling ABC algorithm that is based on the so-called *two-sample energy statistic*. Along with other data discrepancy measures that view data sets as empirical measures, such as the Kullback–Leibler divergence, the Wasserstein distance and maximum mean discrepancies, our proposed approach bypasses the cumbersome use of summary statistics.

We have shown that the V-statistic estimator of the ES is consistent under mild moment conditions. Furthermore, we have established a new asymptotic result for cases when the observed sample and simulated sample sizes increasing to infinity, that shows a kind of consistency of the pseudo-posterior in the infinite data scenario. This is in concordance with previous results in such cases [see for instance 176, 52] and extends upon existing theory for the application in the general IS-ABC framework.

Illustrations of the proposed ES-ABC algorithm on four experimental models have shown that it performs comparatively well to alternative discrepancy measures. Considering computing costs, KL should be preferred over the other three discrepancy measures, with a *linearithmic* computational time of $\mathcal{O}((n+m)\log(n+m))$. This can be contrasted against the quadratic time $\mathcal{O}((n+m)^2)$ for a single computation of $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m)$ when we consider the Wasserstein distance, instead. Both the ES and MMD estimators require quadratic computational time, like the Wasserstein distance. We note that linear time estimators are also available for the MMD and the ES, although these are unbiased and cannot be guaranteed to be positive [see Lemma 14 in 146].

In the rejection ABC setting, Proposition 2 of [52] shows that under some regularity assumptions on the DGP and if the data discrepancy measure satisfies the condition:

$$\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_n) = 0 \text{ if and only if } \mathbf{X}_n = \mathbf{Y}_n, \tag{2.89}$$

then the ABC pseudo-posterior contracts to the posterior distribution as the rejection threshold ϵ converges to zero. It can be shown that the V-statistic estimator of the ES only satisfies the *only if* direction of (2.89) and thus does not necessarily enjoy the conclusions of Proposition 2 of [52]. The condition is not known to be necessary and thus we do not know if the conclusion can be satisfied in another way. We observe, from our simulation experiments, that ES did not perform differently to the Wasserstein distance, which can be shown to satisfy Proposition 2 of [52].

Chapter 3

Distributional Properties of Statistical and Machine Learning Models

Contents

This chapter is based on the following papers and preprints

Section 3.1

[A7] O. Marchal and J. Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017

[A2] J. Arbel, O. Marchal, and H. D. Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability & Statistics, forthcoming*, 2019

[S5] M. Vladimirova and J. Arbel. Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Submitted*, 2019

Section 3.2

[A4] M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel. Understanding Priors in Bayesian Neural Networks at the Unit Level. *ICML*, 2019

[P5] M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural networks become heavier-tailed with depth. *NeurIPS Bayesian Deep Learning Workshop*, 2018

[P6] M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural network priors at the level of units. *1st Symposium on Advances in Approximate Bayesian Inference*, 2018

Section 3.3

[A3] J. Arbel, M. Crispino, and S. Girard. Dependence properties and Bayesian inference for asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 174, 2019

[A7] O. Marchal and J. Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017

[A2] J. Arbel, O. Marchal, and H. D. Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability & Statistics, forthcoming*, 2019

[S5] M. Vladimirova and J. Arbel. Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Submitted*, 2019

3.1 Sub-Gaussian and sub-Weibull properties

3.1.1 Introduction

The sub-Gaussian property [66, 65, 263] and related concentration inequalities [60, 278] have attracted a lot of attention in the last couple of decades due to their applications in various areas such as pure mathematics, physics, information theory and computer sciences. Recent interest focused on deriving the optimal proxy variance for discrete random variables like the Bernoulli distribution [67, 185, 50] and the missing mass [229, 228, 50, 49]. Our focus is instead on two continuous random variables, the Beta and Dirichlet distributions, for which the optimal proxy variance was not known to the best of our knowledge. Some upper bounds were recently conjectured by [109] that we prove in the present article by providing the optimal proxy variance for both Beta and Dirichlet distributions. Similar concentration properties of the Beta distribution have been recently used in many contexts including Bayesian adaptive data analysis [109], Bayesian nonparametrics [76] and spectral properties of random matrices [262].

We start by reminding the definition of sub-Gaussian property for random variables:
Definition 3.1.1 (Sub-Gaussian variables)**.** *A random variable X with finite mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ such that:*

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}. \tag{3.1}$$

Such a constant σ^2 is called a proxy variance (or sub-Gaussian norm), and we say that X is σ^2-sub-Gaussian. If X is sub-Gaussian, one is usually interested in the optimal proxy variance:

$$\sigma^2_{opt}(X) = \min\{\sigma^2 \geq 0 \text{ such that } X \text{ is } \sigma^2\text{-sub-Gaussian}\}.$$

Note that the variance always gives a lower bound on the optimal proxy variance: $Var[X] \leq \sigma^2_{opt}(X)$. In particular, when $\sigma^2_{opt}(X) = Var[X]$, X is said to be strictly sub-Gaussian.

Every compactly supported distribution, as is the Beta(α, β) distribution, is sub-Gaussian. This can be seen by Hoeffding's classic inequality: any random variable

X supported on $[0,1]$ with mean μ satisfies

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2}{8}},$$

thus exhibiting $\frac{1}{4}$ as an upper bound to the proxy variance. This bound can be improved by taking into account the location of the mean μ within the interval $[0,1]$. An early step in this direction is the second inequality in [159] paper, indexed (2.2). It states that if $\mu < 1/2$, then for any positive ϵ, $\mathbb{P}(X - \mu > \epsilon) \leq e^{-\epsilon^2 g(\mu)}$, where

$$g(\mu) = \frac{1}{1-2\mu} \ln \frac{1-\mu}{\mu} \tag{3.2}$$

thus indicating that X has a right tail lighter than a Gaussian tail of variance $\frac{1}{2g(\mu)}$. Hoeffding's result was strengthened by [185] to comply with Definition 3.1.1 of sub-Gaussianity[1] as follows

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp\left(\frac{\lambda^2}{4g(\mu)}\right) \text{ for all } \lambda \in \mathbb{R}, \tag{3.3}$$

thus indicating that $\frac{1}{2g(\mu)}$ is a distribution-sensitive proxy variance for any $[0,1]$-supported random variable with mean μ [see also 50, for a detailed proof of this result]. If this is the optimal proxy variance for the Bernoulli distribution [see Theorem 2.1 and Theorem 3.1 of 67], it is clear from our result that it does not hold true for the Beta distribution. However, fixing $\frac{\alpha}{\alpha+\beta} = \mu$ and letting $\alpha \to 0$, $\beta \to 0$, the Beta(α,β) distribution concentrates to the Bern(μ) distribution, and we show that we recover the optimal proxy variance for the Bernoulli distribution (Theorem 3.1.2).

3.1.2 Optimal proxy variance for the Beta distribution

The Beta(α,β) distribution, with $\alpha,\beta > 0$, is characterized by a density on the segment $[0,1]$ given by:

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $B(\alpha,\beta) = \int_0^\infty x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function. The moment-generating function of a Beta(α,β) distribution is given by a confluent hypergeometric function (also known as Kummer's function):

$$\mathbb{E}[\exp(\lambda X)] \;=\; {}_1F_1(\alpha;\alpha+\beta;\lambda) = \sum_{j=0}^{\infty} \frac{\Gamma(\alpha+j)\Gamma(\alpha+\beta)}{(j!)\Gamma(\alpha)\Gamma(\alpha+\beta+j)}\lambda^j. \tag{3.4}$$

This is equivalent to say that the j^{th} raw moment of a Beta(α,β) random variable X is given by:

$$\mathbb{E}[X^j] = \frac{(\alpha)_j}{(\alpha+\beta)_j}, \tag{3.5}$$

[1] Note indeed that Equation (3.1), together with Markov inequality, imply $\mathbb{P}(X - \mu > \epsilon) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$.

where $(x)_j = x(x+1)\cdots(x+j-1) = \frac{\Gamma(x+j)}{\Gamma(x)}$ is the *Pochhammer symbol*, also known in the literature as a *rising factorial*. In particular, the mean and variance are given by:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The Beta distribution is ubiquitous in statistics. It plays a central role in the binomial model in Bayesian statistics where it is a conjugate prior distribution (the associated posterior distribution is also Beta): if $X \sim \mathrm{Binomial}(\theta, N)$ and $\theta \sim \mathrm{Beta}(\alpha, \beta)$, then $\theta|X \sim \mathrm{Beta}(\alpha+X, \beta+N-X)$. It is also key to Bayesian nonparametrics where it embodies, among others, the distribution of the breaks in the stick-breaking representation of the Dirichlet process and the Pitman–Yor process; marginal distributions of Polya trees [76]; the posterior distribution of discovery probabilities under a Bayesian nonparametrics model [17]. Our main result opens new research avenues for instance about asymptotic (frequentist) assessments of these procedures.

Theorem 3.1.1 (Optimal proxy variance for the Beta distribution)**.** *For any $\alpha, \beta > 0$, the Beta distribution $\mathrm{Beta}(\alpha, \beta)$ is $\sigma^2_{opt}(\alpha,\beta)$-sub-Gaussian with optimal proxy variance $\sigma^2_{opt}(\alpha,\beta)$ given by:*

$$\begin{cases} \sigma^2_{opt}(\alpha,\beta) = \frac{\alpha}{(\alpha+\beta)x_0}\left(\frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)} - 1\right) \\ \qquad \textit{where } x_0 \textit{ is the unique solution of the equation} \\ \ln({}_1F_1(\alpha;\alpha+\beta;x_0)) = \frac{\alpha x_0}{2(\alpha+\beta)}\left(1 + \frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)}\right). \end{cases} \tag{3.6}$$

A simple and explicit upper bound to $\sigma^2_{opt}(\alpha,\beta)$ is given by $\sigma^2_0(\alpha,\beta) = \frac{1}{4(\alpha+\beta+1)}$:
- for $\alpha \neq \beta$ we have $\mathrm{Var}[\mathrm{Beta}(\alpha,\beta)] < \sigma^2_{opt}(\alpha,\beta) < \frac{1}{4(\alpha+\beta+1)}$
- for $\alpha = \beta$ we have $\mathrm{Var}[\mathrm{Beta}(\alpha,\alpha)] = \sigma^2_{opt}(\alpha,\alpha) = \frac{1}{4(2\alpha+1)}$.

refer to [223] for a proof.

Equation (3.6) defining x_0 is a transcendental equation, the solution of which is not available in closed form. However, it is simple to evaluate numerically. The values of the variance, optimal proxy variance and its simple upper bound are illustrated on Figure 3.1. Note that for a fixed value of the sum of the parameters, $\alpha+\beta = S$, the optimal proxy variance deteriorates when α, or equivalently β, gets close to 0 or to S. This is reminiscent of the Bernoulli optimal proxy variance behavior which deteriorates when the success probability moves away from $\frac{1}{2}$ [67].

Corollary 3.1.1. *The Beta distribution $\mathrm{Beta}(\alpha,\beta)$ is strictly sub-Gaussian if and only if $\alpha = \beta$.*

As a direct consequence, we obtain the strict sub-Gaussianity of the uniform, the arcsine and the Wigner semicircle distributions, as special cases up to a trivial rescaling of the $\mathrm{Beta}(\alpha,\alpha)$ distribution respectively with α equal to 1, $\frac{1}{2}$ and $\frac{3}{2}$.

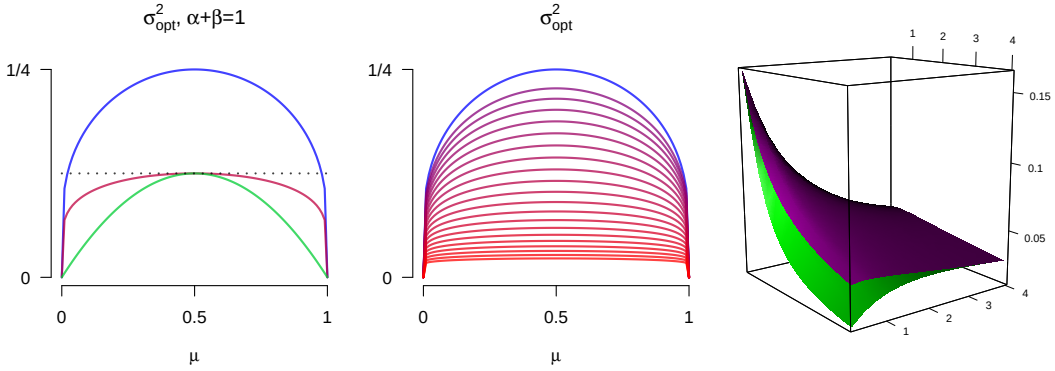


FIGURE 3.1: *Left*: curves of $\mathrm{Var}[\mathrm{Beta}(\alpha,\beta)]$ (green), $\sigma^2_{\mathrm{opt}}(\alpha,\beta)$ (purple) and $\frac{1}{4(\alpha+\beta+1)}$ (dotted black) for the $\mathrm{Beta}(\alpha,\beta)$ distribution with $\alpha+\beta$ set to 1, $\sigma^2_{\mathrm{opt}}(\mu)$ for the $\mathrm{Bern}(\mu)$ distribution (blue); varying mean μ on the x-axis. *Center*: curves of $\sigma^2_{\mathrm{opt}}(\mu)$ for the $\mathrm{Bern}(\mu)$ distribution (blue), and of $\sigma^2_{\mathrm{opt}}(\alpha,\beta)$ for the $\mathrm{Beta}(\alpha,\beta)$ distribution with $\alpha+\beta$ varying on a log scale from 0.1 (purple) to 10 (red); varying mean μ on the x-axis. *Right*: surfaces of $\mathrm{Var}[\mathrm{Beta}(\alpha,\beta)]$ (green) and $\sigma^2_{\mathrm{opt}}(\alpha,\beta)$ (purple), for values of α and β varying in $[0.2,4]$.

Optimal proxy variance for the Bernoulli distribution

The proof technique can be used to recover the optimal proxy variance for the Bernoulli distribution, known since [185]. This is illustrated by the center panel of Figure 3.1.

Theorem 3.1.2 (Optimal proxy variance for the Bernoulli distribution)**.** *For any $\mu \in (0,1)$, the Bernoulli distribution with mean μ is sub-Gaussian with optimal proxy variance $\sigma^2_{opt}(\mu)$ given by:*

$$\sigma^2_{opt}(\mu) = \frac{(1-2\mu)}{2\ln\frac{1-\mu}{\mu}}. \tag{3.7}$$

3.1.3 On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables

We focus here on the study of almost surely bounded random variables, where Bernoulli, beta, binomial, Kumaraswamy [181] or triangular [197] distributions are taken as standard and common examples. If sub-Gaussianity *per se* is *de facto* ensured because the support of said random variables is bounded, then exciting research avenues remain open in the area. Among these questions are (a) how to obtain the optimal sub-Gaussian proxy variance, and (b) how to characterize *strict* sub-Gaussianity?

Regarding question (a), we propose general conditions characterizing the optimal sub-Gaussian proxy variance, thus generalizing previous work [223] that was tailored to the beta and Dirichlet distributions. Several techniques based on studying variations of functions are proposed.

As for question (b), it turns out that the *symmetry* of the distribution plays a crucial role. By symmetry, we mean symmetry with respect to the mean $\mu = \mathbb{E}[X]$. That is,

we say that X is symmetrically distributed if X and $2\mu - X$ have the same distribution. Thus, if X has a density, this means that the density is symmetric with respect to μ. A simple, and remarkable, equivalence holds for most of the standard bounded random variables.

Proposition 3.1.1. *Let X be a Bernoulli, beta, binomial, Kumaraswamy or triangular random variable. Then,*

$$X \text{ is symmetric} \iff X \text{ is strictly sub-Gaussian.}$$

The result is known for the beta distribution [223]. In this article, we provide proofs for the Bernoulli, binomial, Kumaraswamy and triangular distributions.

From Proposition 3.1.1, it may be tempting to conjecture that the equivalence holds true for *any* random variable having a bounded support. However, we establish that this is not the case. This was actually one of the starting points for the present work. More precisely, we shall provide a proof of the following result.

Proposition 3.1.2. *Symmetry of X is neither*

(i) a sufficient condition, nor

(ii) a necessary condition,

for the strict sub-Gaussian property.

The proof of this result is presented in [22], where we demonstrate that *(i)* there exists simple symmetric mixtures of distributions (e.g., a two-components mixture of beta distribution and a three-components mixture distribution of Dirac masses) which are not strictly sub-Gaussian, and that *(ii)* there exists an asymmetric three-components mixture of Dirac masses which is strictly sub-Gaussian.

Before delving into detailing the strict sub-Gaussianity property in Section 3.1.3, we first investigate some conditions that characterize the optimal proxy variance σ^2_{opt}, in Section 3.1.3.

Characterizations of the optimal proxy variance σ^2_{opt}

Let X be an almost surely bounded random variable with mean $\mu = \mathbb{E}[X]$. Then, X is sub-Gaussian and satisfies Definition 3.1 for some $\sigma^2 > 0$.

An equivalent definition is that

$$\forall\, \lambda \in \mathbb{R} \,:\, \sigma^2 \geq \frac{2}{\lambda^2}\mathcal{K}(\lambda),$$

where the function $\mathcal{K}$, defined on $\mathbb{R}$ by: $\mathcal{K}(\lambda) = \ln \mathbb{E}[\exp(\lambda[X - \mu])]$, corresponds to the cumulants generating function of $X - \mu$. Thus the optimal proxy variance σ^2_{opt} can be defined as the supremum

$$\sigma^2_{\text{opt}} = \sup_{\lambda \in \mathbb{R}} \frac{2}{\lambda^2}\mathcal{K}(\lambda). \tag{3.8}$$

If X is almost surely bounded, then this supremum is attained.

Note that the function h, defined on $\mathbb{R}$ by

$$h(\lambda) = \frac{2}{\lambda^2}\mathcal{K}(\lambda), \tag{3.9}$$

is continuous at $\lambda = 0$, since a standard series expansion demonstrates that:

$$h(\lambda) \overset{\lambda\to 0}{=} \mathrm{Var}[X] + o(1). \tag{3.10}$$

Moreover, h may never vanish. In fact, since the logarithm function is strictly concave, Jensen's inequality implies that for any $\lambda \in \mathbb{R}$,

$$h(\lambda) = \frac{2}{\lambda^2}\ln \mathbb{E}[\mathrm{e}^{\lambda(X-\mu)}] > \frac{2}{\lambda^2}\mathbb{E}[\ln \mathrm{e}^{\lambda(X-\mu)}] = 0. \tag{3.11}$$

Equation (3.10) also explains directly why $\sigma^2_{\mathrm{opt}} \geq \mathrm{Var}[X]$, since the variance is the value of the right-hand side (r.h.s.) function at $\lambda = 0$ and thus the maximum is always greater or equal to it. We therefore have the following result.

Proposition 3.1.3 (Characterization of σ^2_{opt} by h)**.** *The optimal proxy variance is given by:*

$$\sigma^2_{\mathrm{opt}} = \max_{\lambda\in\mathbb{R}} h(\lambda) = \max_{\lambda\in\mathbb{R}} \frac{2}{\lambda^2}\mathcal{K}(\lambda). \tag{3.12}$$

We may now present a necessary (but not always sufficient) system of equations for σ^2_{opt}. Indeed, since the maximum is achieved at a finite point, then this point must necessarily be a zero of the derivative of h, if h is differentiable (we will denote by $\mathcal{D}^k$ the space of functions that are k times differentiable on $\mathbb{R}$ and by $\mathcal{C}^k$ the space of functions that are k times differentiable on $\mathbb{R}$ and for which the k^{th} derivative is continuous on $\mathbb{R}$).

Thus, we obtain the following corollary.

Corollary 3.1.2 (Necessary condition for σ^2_{opt}, with respect to h)**.** *Let σ^2_{opt} be the optimal proxy variance, and assume that h and $\mathcal{K}$ are $\mathcal{D}^1$. Then there exists a finite λ_0, such that*

$$\sigma^2_{\mathrm{opt}} = h(\lambda_0) \text{ and } h'(\lambda_0) = 0, \tag{3.13}$$

which is equivalent to

$$\sigma^2_{\mathrm{opt}} = \frac{2}{\lambda_0^2}\mathcal{K}(\lambda_0) \text{ and } \lambda_0\mathcal{K}'(\lambda_0) = 2\mathcal{K}(\lambda_0), \tag{3.14}$$

using only the centered cumulants generating function $\mathcal{K}$.

In practice, the previous set of equations has to used with caution, since there may be more than one solution to the second equation involving the derivative of h (or that of $\mathcal{K}$), and a global maximizer is required to be picked among the stationary points, instead of a minimizer or a local maximizer. On a case-by-case basis, the following approach based on ordinary differential equations (ODEs), satisfied by h, can be used to demonstrate that it has a unique global maximum.

Proposition 3.1.4. *If the function h is $\mathcal{C}^2$, then it is the unique solution of the ordinary differential equations:*

$$h'(\lambda) + \frac{2}{\lambda}h(\lambda) = \frac{2}{\lambda^2}\mathcal{K}'(\lambda) \text{ with } h(0) = \text{Var}[X], \tag{3.15}$$

or

$$h''(\lambda) + \frac{3}{\lambda}h'(\lambda) = \frac{2}{\lambda}\left(\frac{\mathcal{K}'(\lambda)}{\lambda}\right)' \text{ with } h(0) = \text{Var}[X] \text{ and } h'(0) = \frac{1}{3}\mathbb{E}[(X-\mu)^3]. \tag{3.16}$$

Proof. The result is directly obtained by differentiating h and via standard analysis theorems. ■

Remark 3.1.1. *For cases such as the Bernoulli and uniform distributions, we may prove that the r.h.s. of (3.16) is strictly negative on $\mathbb{R}^* := \mathbb{R} \setminus \{0\}$. This implies that if λ_0 is extremal (i.e., $h'(\lambda_0) = 0$), then it satisfies $h''(\lambda_0) < 0$ so that it is a local maximum. This implies that h has no local minimum and thus may only have one critical point which is necessarily the unique global maximum.*

We conclude this section with another possible methodology for deriving a necessary and sufficient condition for σ^2_{opt}. To this end, the problem needs to be addressed from a different point of view, by studying the difference of the terms of Definition 3.1:

$$\Delta : (\sigma^2, \lambda) \in \mathbb{R}^*_+ \times \mathbb{R} \mapsto \exp\left(\frac{\lambda^2\sigma^2}{2}\right) - \mathbb{E}[\exp(\lambda[X-\mu])]. \tag{3.17}$$

Proposition 3.1.5 (Characterization of σ^2_{opt}, with respect to Δ)**.** *If Δ is $\mathcal{C}^1$, then the optimal proxy variance is characterized by:*

$$\lambda \mapsto \Delta(\sigma^2_{\text{opt}}, \lambda) \geq 0 \text{ and } \exists\, \lambda_0 \in \mathbb{R}, \text{ such that } \Delta(\sigma^2_{\text{opt}}, \lambda_0) = 0 \text{ and } \partial_\lambda \Delta(\sigma^2_{\text{opt}}, \lambda_0) = 0. \tag{3.18}$$

Proof. See [22]. ■

This proof technique was used by [223] for obtaining the optimal proxy variance of the beta and Dirichlet distributions. However we find more convenient to use the conditions stated in Proposition 3.1.4 using the function h to address the issues presented in this article.

Remark 3.1.2. *In general, we would like to remove the condition: $\lambda \mapsto \Delta(\sigma^2, \lambda) \geq 0$ on the r.h.s. of Proposition 3.1.5, in order to have a simpler (and local) characterization of the optimal proxy variance, as a solution of (3.18). However, this is not possible, since we may not exclude that there exists a value $\sigma^2 < \sigma^2_{\text{opt}}$ for which $\Delta(\sigma^2, \lambda)$ presents a double zero λ_0 where locally it remains non-negative but at the same time a whole interval far from λ_0 where it would be strictly negative.*

On strict sub-Gaussianity

Conditions based on the cumulants Strict sub-Gaussianity is fulfilled when the optimal proxy variance equals the variance. In view of Equation (3.10), Proposition 3.1.3 can be rewritten as the following corollary in order to characterize the strict sub-Gaussianity property.

Corollary 3.1.3 (Corollary of Proposition 3.1.3)**.** *A distribution is strictly sub-Gaussian if and only if the maximum of function h, defined in (3.9), is attained in zero (and is automatically equal to $\mathrm{Var}[X]$). That is:*

$$\max_{\lambda \in \mathbb{R}} h(\lambda) = h(0) = \mathrm{Var}[X]. \tag{3.19}$$

This characterization provides necessary conditions, based on cumulants, that are required for strict sub-Gaussianity to hold.

Proposition 3.1.6 (Necessary conditions based on cumulants)**.** *If X is strictly sub-Gaussian, then the 3^{rd} and 4^{th} cumulants of X must satisfy*

$$\kappa_3 = \mathbb{E}[(X - \mathbb{E}[X])^3] = 0, \text{ and} \tag{3.20}$$

$$\kappa_4 = \mathbb{E}[(X - \mathbb{E}[X])^4] - 3\,\mathrm{Var}[X]^2 \leq 0. \tag{3.21}$$

Proof. By definition of the cumulant generating function $\mathcal{K}(\lambda)$ of $X - \mu$,

$$\mathcal{K}(\lambda) = \sum_{i=1}^{\infty} \kappa_i \frac{\lambda^i}{i!}, \tag{3.22}$$

where κ_i are the cumulants of $X - \mu$. Since $\kappa_1 = \mu - \mu = 0$ and $\kappa_2 = \mathrm{Var}[X]$, and using values for the third and fourth cumulants given in (3.20) and (3.21), we may write (locally around $\lambda \to 0$):

$$h(\lambda) = \mathrm{Var}[X] + \mathbb{E}[(X-\mu)^3]\frac{\lambda}{3} + \left(\mathbb{E}[(X-\mu)^4] - 3\,\mathrm{Var}[X]^2\right)\frac{\lambda^2}{12} + O(\lambda^3). \tag{3.23}$$

Therefore if $\mathbb{E}[(X-\mu)^3] \neq 0$, the maximum of $h(\lambda)$ cannot be $h(0)$ and thus strict sub-Gaussianity cannot be achieved. We conclude the proof by noting that if $\mathbb{E}[(X-\mu)^3] = 0$, we have the fact that $\lambda = 0$ can be a local maximum, only if $\mathbb{E}[(X-\mu)^4] \leq 3\,\mathrm{Var}[X]^2$. ■

Condition (3.20) requires that the third centered moment is zero and Condition (3.21) imposes a relation between the second and fourth centered moments. Note that the latter condition can be compactly formulated via an alternative condition on the kurtosis of X:

$$\mathrm{Kurt}[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^2} \leq 3.$$

More specifically, sub-Gaussianity requires that the random variable has kurtosis less than or equal to three, which is the kurtosis of a standard Gaussian random

variable. Such distributions are referred to as *platycurtic*. The fourth cumulant defined in (3.21) is also termed *excess kurtosis*. Thus, strict sub-Gaussianity requires negative excess kurtosis.

When the above necessary conditions (3.20) and (3.21) hold, we are not able to obtain simple additional necessary conditions on the next cumulants. In particular, note that strict sub-Gaussianity *does not* imply symmetry (i.e., $\mathbb{E}[(X - \mathbb{E}[X])^{2j+1}] = 0$, for any $j \geq 0$), as will be discussed in the next section.

In contrast, more can be said when the distribution is symmetric. In fact, in the symmetric case, the moments of odd order are zero, and a simple sufficient condition can be readily obtained by comparing the Taylor expansions at $\lambda = 0$ of both terms of inequality (3.1), as stated in the following proposition.

Proposition 3.1.7 (Sufficient condition based on moments)**.** *If X is symmetric with respect to its mean $\mu = \mathbb{E}[X]$, then a sufficient condition for X to be strictly sub-Gaussian can be stated in terms of all its even moments. That is, for X to be strictly sub-Gaussian, it is sufficient that*

$$\forall j \geq 2, \quad \frac{\mathbb{E}[(X-\mu)^{2j}]}{(2j)!} \leq \frac{(\mathrm{Var}[X])^j}{2^j j!} \tag{3.24}$$

holds.

Proof. The proof is based on series expansions at $\lambda = 0$ of both terms of inequality (3.1), when the proxy variance σ^2 is set to the variance $\mathrm{Var}[X]$. Namely:

$$\mathbb{E}[\exp(\lambda X)] = \sum_{j=0}^{\infty} \mathbb{E}\left[X^{2j}\right] \frac{\lambda^{2j}}{(2j)!}, \quad \text{and} \quad \exp\left(\frac{\lambda^2 \,\mathrm{Var}[X]}{2}\right) = \sum_{j=0}^{\infty} \frac{(\mathrm{Var}[X])^j}{2^j} \frac{\lambda^{2j}}{j!}, \tag{3.25}$$

when compared term-by-term, leads to inequality (3.1), under assumption (3.24). Note that inequality (3.24) needs be checked only for $j \geq 2$, as it trivially holds for $j = 0, 1$. ■

This technique was used by [223] (Section 2.2) for showing that a (symmetric) Beta(α, α) random variable is strictly sub-Gaussian. We also use it to address the cases of Bernoulli and binomial, and triangular distributions in Section 2.4.5.

Link with symmetry The relationship between strict sub-Gaussianity and symmetry was discussed in the Introduction. Here, we provide a proof of Proposition 3.1.2, while the proof of Proposition 3.1.1 is deferred to Section 2.4.5.

Symmetry is neither a sufficient condition...

Simple symmetric distributions which break the necessary condition of negative excess kurtosis can easily be constructed by hand. One such construction is by means of mixture of Dirac masses. First, consider the discrete random variable

$$X \sim \frac{\eta}{2}(\delta_{-1} + \delta_1) + (1 - \eta)\delta_0, \tag{3.26}$$

which is a three-component mixture of Dirac masses at locations -1, 0 and 1, with $\eta \in [0,1]$. It is symmetric, by construction, and its excess kurtosis equals

$$\kappa_4 = \mathbb{E}[X^4] - 3\,\mathrm{Var}[X]^2 = \eta - 3\eta^2 = \eta(1-3\eta), \tag{3.27}$$

which is *strictly positive* for all values $\eta \in \left(0, \frac{1}{3}\right)$, hence X is not strictly sub-Gaussian for these values by virtue of Proposition 3.1.6. On the other hand when $\eta \to 1$, the distribution of X degenerates to that of the so-called Rademacher random variable, which leads to the least possible excess kurtosis of -2.

Similar counter-examples to the sufficientness of symmetry can be built in the form of mixtures of two symmetric beta variables:

$$X \sim \eta\,\mathrm{Beta}(\alpha,\alpha) + (1-\eta)\,\mathrm{Beta}(\beta,\beta),$$

for $\eta \in (0,1)$ and $\alpha, \beta > 0$. For any value of $\eta \in (0,1)$, values for α, β leading to positive excess kurtosis can be obtained. For instance, we may set $(\eta, \alpha, \beta) = (0.1, 1.5, 9)$, to obtain the excess kurtosis $\kappa_4 \approx 1.1 \times 10^{-4}$.

. . . nor a necessary condition for strict sub-Gaussianity

Although most typical bounded random variables that are strictly sub-Gaussian are symmetric (see, e.g., Proposition 3.1.1), the symmetry of the distributions of such variables is not a necessary condition for strict sub-Gaussianity. Examples of such distributions include mixtures of Dirac masses. For example,

$$X \sim \sum_{i=1}^{3} p_i \delta_{x_i} \text{ with } \sum_{i=1}^{3} p_i = 1 \tag{3.28}$$

with $(x_1, x_2, x_3) = \left(-2, -\frac{1}{2}, \frac{5}{4}\right)$ and $(p_1, p_2, p_3) = \left(\frac{1}{13}, \frac{4}{7}, \frac{32}{91}\right)$. The function h for the random variable characterized by (3.28) is plotted in Figure 3.2b. Note that it attains its maximum in $\lambda = 0$.

3.1.4 Sub-Weibull property

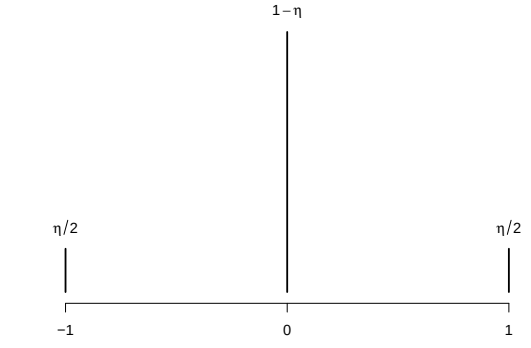
It is tempting to generalise the sub-Gaussian property by considering the class of distributions satisfying

$$\mathbb{P}(|X| \geq x) \leq a \exp\left(-bx^{1/\theta}\right), \quad \text{for all } x > 0, \quad \text{for some } \theta > 0, \tag{3.29}$$

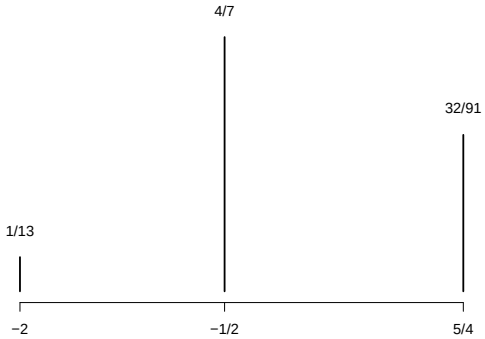
which is the goal of the present section. Since a *Weibull* random variable X on $\mathbb{R}_+$ is defined by a survival function, for $x > 0$,

$$\bar{F}(x) = \mathbb{P}(X \geq x) = \exp\left(-bx^{1/\theta}\right), \quad \text{for some } \; b > 0, \theta > 0, \tag{3.30}$$

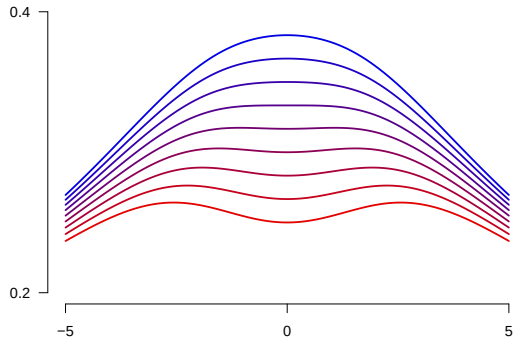
we term a distribution satisfying (3.37) a *sub-Weibull* distribution in the following definition.



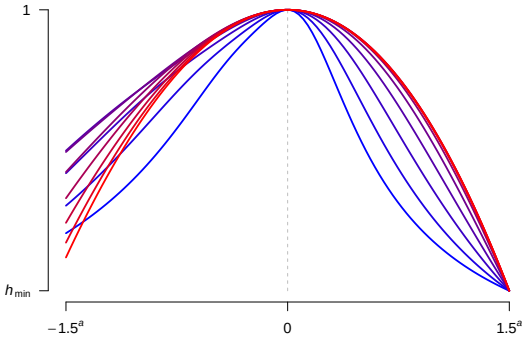
(A) Distribution (3.26): symmetric but not strictly sub-Gaussian when $\eta \in \left(0, \frac{1}{3}\right)$



(B) Distribution (3.28): asymmetric but strictly sub-Gaussian



(C) Function h for (3.26) with $\eta \in [0.25, 0.4]$



(D) Function h for (3.28), zoomed-in for red curves

FIGURE 3.2: Illustration of the mixtures of Dirac masses, proving Proposition 3.1.2 described in Sections 3.1.3 (a,c) and 3.1.3 (b,d). In (c), η varies from 0.25 (red curve, maximum not at zero) to 0.4 (blue curve, maximum at zero). In (d), we illustrate the function h with different zooming scales (around $\lambda = 0$): from $[-1.5^5, 1.5^5]$ (blue curve, maximum zoom-out) to $[-1.5^{-5}, 1.5^{-5}]$ (red curve, maximum zoom-in), with an adapted y-scale, showing that the maximum is attained in zero.

Definition 3.1.2 (Sub-Weibull random variable)**.** *A random variable X, satisfying (3.37) for some positive a, b and θ, is called a sub-Weibull random variable with tail parameter θ, which is denoted by $X \sim \mathsf{subW}(\theta)$.*

Interest in such heavier-tailed distributions than Gaussian or Exponential arises in our experience from their emergence in the field Bayesian deep neural networks [331]. While writing this note, we came across the preprint [200], independent of our work, which also introduces sub-Weibull distributions but from a different perspective. The definition proposed by [200] is based on Orlitz norm and is equivalent to Definition 3.2.3. While [200] focus on establishing tail bounds and rates of convergence for problems in high dimensional statistics, including covariance estimation and linear regression, under the sole sub-Weibull assumption, we prove here sub-Weibull characterization properties. In addition, we illustrate their link with deep neural networks, not in the form of a model assumption as in [200], but as a characterisation of the prior distribution of deep neural networks units.

Sub-Weibull distributions: characteristic properties

Let X be a random variable. When the kth moment of X exist, $k > 0$, we denote $\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k}$. The following theorem states different equivalent distribution properties, such as tail decay and growth of moments. The proof of this result shows how to transform one type of information about the random variable into another. See [**?**] for similar characteristic properties of sub-Gaussian and sub-Exponential distributions.

Theorem 3.1.3 (Sub-Weibull equivalent properties)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor*[1]*.*

(i) *The tails of X satisfy*

$$\mathbb{P}(|X| \geq x) \leq \exp\left(-(x/K_1)^{1/\theta}\right) \quad \text{for all } x \geq 0.$$

(ii) *The moments of X satisfy*

$$\|X\|_k \leq K_2 k^\theta \quad \text{for all } k \geq 1.$$

(iii) *The MGF of $|X|^{1/\theta}$ satisfies*

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta}|X|^{1/\theta}\right)\right] \leq \exp(\lambda^{1/\theta} K_3^{1/\theta})$$

for all λ such that $|\lambda| \leq \frac{1}{K_3}$.

(iv) *The MGF of $|X|^{1/\theta}$ is bounded at some point, namely*

$$\mathbb{E}\left[\exp\left(|X|^{1/\theta}/K_4^{1/\theta}\right)\right] \leq 2.$$

Remark 3.1.3. *The constant 2 that appears in some properties in Theorem 3.1.3 does not have any special meaning. It is chosen for simplicity and can be replaced by other absolute constants.*

Distribution	Tails	Moments
Sub-Gaussian	$\mathbb{P}(\|X\| \geq x) \leq \mathrm{e}^{-(x/K_1)^2}$	$\|X\|_k \leq K_2\sqrt{k}$
Sub-Exponential	$\mathbb{P}(\|X\| \geq x) \leq \mathrm{e}^{-x/K_1}$	$\|X\|_k \leq K_2 k$
Sub-Weibull	$\mathbb{P}(\|X\| \geq x) \leq \mathrm{e}^{-(x/K_1)^{1/\theta}}$	$\|X\|_k \leq K_2 k^\theta$

TABLE 3.1: Sub-Gaussian, sub-Exponential and sub-Weibull distributions comparison in terms of tail $P(|X| \geq x)$ and moment condition, with K_1 and K_2 some positive constants. The first two are a special case of the last with $\theta = 1/2$ and $\theta = 1$ respectively.

[1]There exists an absolute constant C such that property i implies property j with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, \ldots, 4$.

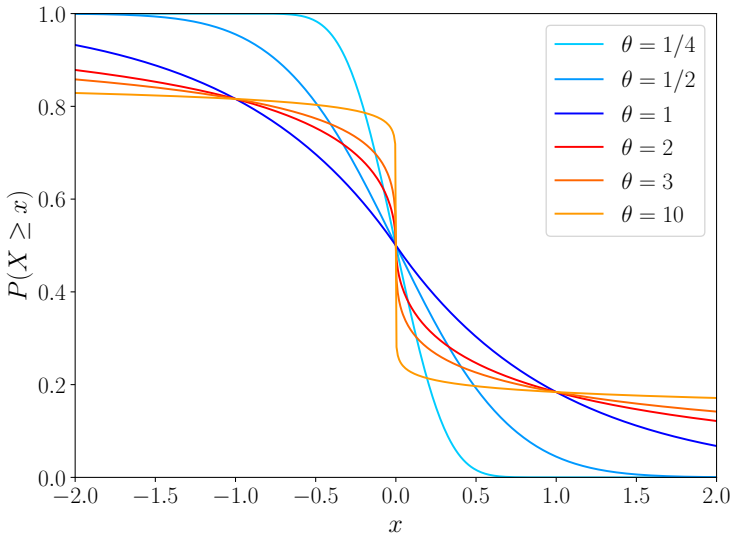


FIGURE 3.3: Illustration of sub-Weibull survival curves on $\mathbb{R}_+$ with varying tail parameters θ.

Informally, the tails of a $\mathsf{subW}(\theta)$ distribution are dominated by (i.e. decay at least as fast as) the tails of a Weibull variable with shape parameter[2] equal to $1/\theta$. Sub-Gaussian and sub-Exponential variables, which are commonly used, are special cases of sub-Weibull random variables with tail parameter $\theta = 1/2$ and $\theta = 1$, respectively, see Table 3.1. Symmetric sub-Weibull distributions (their survival function) are represented in Figure 3.3 for varying tail parameter θ. Since only the tail is relevant for the illustration, the sub-Weibull random variables depicted here consist of Weibull random variables with shape parameter $1/\theta$ and scale parameter 1, truncated at their mode (which is equal to zero if $\theta \geq 1$ and to $(1-\theta)^{1/\theta}$ otherwise), and then symmetrized.

Additional properties

The Sub-Gaussian distribution is known to obey the sub-Exponential distribution definition. It leads to the inclusion of a sub-Gaussian distribution family into a sub-Exponential one. The following proposition generalizes this property for Sub-Weibull distributions with different tail parameters.

Proposition 3.1.8 (Inclusion). *Let θ_1 and θ_2 such that $0 < \theta_1 < \theta_2$ be tail parameters for some sub-Weibull distributed variables. Then the following inclusion holds*

$$\mathsf{subW}(\theta_1) \subset \mathsf{subW}(\theta_2).$$

[2] Weibull distributions are commonly parameterized by a shape parameter κ. Here we use instead $\theta = 1/\kappa$ for the convenience that the larger the tail parameter θ, the heavier the tails of the sub-Weibull distribution.

Proof. For $X \sim \mathsf{subW}(\theta_1)$, there exists some constant $K_2 > 0$ such that for all $k > 0$, $\|X\|_k \leq K_2 k^{\theta_1}$. Since $k^{\theta_1} \leq k^{\theta_2}$ for all $k \geq 1$, this yields $\|X\|_k \leq K_2 k^{\theta_2}$, which by definition implies $X \sim \mathsf{subW}(\theta_2)$. ■

Let a random variable X follow a sub-Weibull distribution with tail parameter θ. Due to the property of inclusion from Proposition 3.1.8, the sub-Weibull definition states an upper bound for the tail. To describe a tail lower bound of X through some sub-Weibull distribution family, i.e. a distribution of X to have the tail heavier than some sub-Weibull, we define an optimal tail parameter for that distribution through an asymptotic equivalence in the moment property 2. of Theorem 3.1.3. Introduce the definition of asymptotic equivalence between numeric sequences:

Definition 3.1.3 (Asymptotic equivalence)**.** *Two positive sequences $(a_k)_k$ and $(b_k)_k$ are called asymptotic equivalent and denoted as $a_k \asymp b_k$ if there exist positive constants d and D such that*

$$d \leq \frac{a_k}{b_k} \leq D, \quad \textit{for all } k \in \mathbb{N}. \tag{3.31}$$

Proposition 3.1.9 (Optimal sub-Weibull tail coefficient and moment condition)**.** *Let $\theta > 0$ and let X be a random variable satisfying the following asymptotic equivalence on moments*

$$\|X\|_k \asymp k^{\theta}.$$

Then X is sub-Weibull distributed with optimal tail parameter θ, in the sense that for any $\theta' < \theta$, X is not $\mathsf{subW}(\theta')$.

It is typically assumed that the random variable X has zero mean. If this is not the case, we can always center X by subtracting the mean. We state in the following lemma that variable centering does not change the optimal tail parameter of a sub-Weibull distribution. The reader is referred to [5] for a proof.

Proposition 3.1.10 (Centering variables)**.** *Centering does not harm tail properties of sub-Weibull distributions. In particular, if a random variable X is sub-Weibull with optimal tail parameter θ, then the same holds for the centered variable $(X - \mathbb{E}[X])$.*

[A4] M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel. Understanding Priors in Bayesian Neural Networks at the Unit Level. *ICML*, 2019

[P5] M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural networks become heavier-tailed with depth. *NeurIPS Bayesian Deep Learning Workshop*, 2018

[P6] M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural network priors at the level of units. *1st Symposium on Advances in Approximate Bayesian Inference*, 2018

3.2 Understanding priors in Bayesian neural networks

3.2.1 Introduction

Neural networks (NNs), and their deep counterparts [139], have largely been used in many research areas such as image analysis [198], signal processing [142], or reinforcement learning [301], just to name a few. The impressive performance provided by such machine learning approaches has greatly motivated research that aims at a better understanding the driving mechanisms behind their effectiveness. In particular, the study of the NNs distributional properties through Bayesian analysis has recently gained much attention.

Bayesian approaches investigate models by assuming a prior distribution on their parameters. Bayesian machine learning refers to extending standard machine learning approaches with posterior inference, a line of research pioneered by works on Bayesian neural networks [241, 221]. There is a large variety of applications, e.g. gene selection [206], and the range of models is now very broad, including e.g. Bayesian generative adversarial networks [293]. See [271] for a review. The interest of the Bayesian approach to NNs is at least twofold. First, it offers a principled approach for modeling uncertainty of the training procedure, which is a limitation of standard NNs which only provide point estimates. A second main asset of Bayesian models is that they represent regularized versions of their classical counterparts. For instance, maximum a posteriori (MAP) estimation of a Bayesian regression model with double exponential (Laplace) prior is equivalent to Lasso regression [321], while a Gaussian prior leads to ridge regression. When it comes to NNs, the regularization mechanism is also well appreciated in the literature, since they traditionally suffer from overparameterization, resulting in overfitting.

Central in the field of regularization techniques is the *weight decay* penalty [199], which is equivalent to MAP estimation of a Bayesian neural network with independent Gaussian priors on the weights. Dropout has recently been suggested as a regularization method in which neurons are randomly turned off [306], and [124] proved that a neural network with arbitrary depth and non-linearities, with dropout applied before every weight layer, is mathematically equivalent to an approximation to the probabilistic deep Gaussian process [91], leading to the consideration of such NNs as Bayesian models.

This section is devoted to the investigation of hidden units prior distributions in Bayesian neural networks under the assumption of independent Gaussian weights.

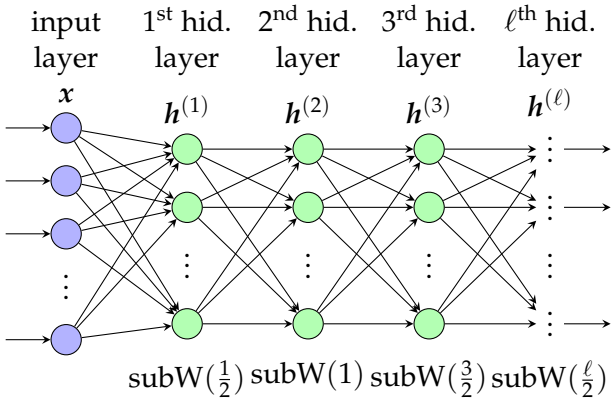


FIGURE 3.4: Neural network architecture and characterization of the ℓ-layer units prior distribution as sub-Weibull distribution with tail parameter $\ell/2$, see Definition 3.2.3.

We first describe a fully connected neural network architecture as illustrated in Figure 3.4. Given an input $\boldsymbol{x} \in \mathbb{R}^N$, the ℓ-th hidden layer unit activations are defined as

$$\boldsymbol{g}^{(\ell)}(\boldsymbol{x}) = \boldsymbol{W}^{(\ell)}\boldsymbol{h}^{(\ell-1)}(\boldsymbol{x}), \quad \boldsymbol{h}^{(\ell)}(\boldsymbol{x}) = \phi(\boldsymbol{g}^{(\ell)}(\boldsymbol{x})), \tag{3.32}$$

where $\boldsymbol{W}^{(\ell)}$ is a weight matrix including the bias vector. A nonlinear activation function $\phi : \mathbb{R} \to \mathbb{R}$ is applied element-wise, which is called nonlinearity, $\boldsymbol{g}^{(\ell)} = \boldsymbol{g}^{(\ell)}(\boldsymbol{x})$ is a vector of pre-nonlinearities, and $\boldsymbol{h}^{(\ell)} = \boldsymbol{h}^{(\ell)}(\boldsymbol{x})$ is a vector of post-nonlinearities. When we refer to either pre- or post-nonlinearities, we will use the notation $\boldsymbol{U}^{(\ell)}$.

We extend the theoretical understanding of feedforward fully connected NNs by studying prior distributions at the units level, under the assumption of independent and normally distributed weights. Our contributions are the following:

(i) As our main contribution, we prove in Theorem 3.2.1 that under some conditions on the activation function ϕ, a Gaussian prior on the weights induces a sub-Weibull distribution on the units (both pre- and post-nonlinearities) with optimal tail parameter $\theta = \ell/2$, see Figure 3.4. The condition on ϕ essentially imposes that ϕ strikes at a linear rate to $+\infty$ or $-\infty$ for large absolute values of the argument, as ReLU does. In the case of bounded support ϕ, like sigmoid or tanh, the units are bounded, making them *de facto* sub-Gaussian[3]

(ii) We offer an interpretation of the main result from a more elaborate regularization scheme at the level of the units in Section 3.2.3.

3.2.2 Bayesian neural networks have heavy-tailed deep units

The deep learning approach uses stochastic gradient descent and error back-propagation in order to fit the network parameters $(\mathbf{W}^{(\ell)})_{1\leq\ell\leq L}$, where ℓ iterates over all network layers. In the Bayesian approach, the parameters are random variables described by probability distributions.

[3] A trivial version of our main result holds, see Remark 3.2.1.

Assumptions on neural network

We assume a prior distribution on the model parameters, that are the weights $\boldsymbol{W}$. In particular, let all weights (including biases) be independent and have zero-mean normal distribution

$$W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2), \tag{3.33}$$

for all $1 \leq \ell \leq L$, $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$, with fixed variance σ_w^2. Given some input $\boldsymbol{x}$, such prior distribution induces by forward propagation (3.32) a prior distribution on the pre-nonlinearities and post-nonlinearities, whose *tail properties* are the focus of this section. To this aim, the nonlinearity ϕ is required to span at least half of the real line as follows. We introduce an extended version of the nonlinearity assumption from [226]:

Definition 3.2.1 (Extended envelope property for nonlinearities)**.** *A nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ is said to obey the extended envelope property if there exist $c_1, c_2 \geq 0$, $d_1, d_2 > 0$ such that the following inequalities hold*

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned} \tag{3.34}$$

The interpretation of this property is that ϕ must shoot to infinity at least in one direction ($\mathbb{R}_+$ or $\mathbb{R}_-$, at least linearly (first line of (3.34)), and also at most linearly (second line of (3.34)). Of course, compactly supported nonlinearities such as sigmoid and tanh do not satisfy the extended envelope property but the majority of other nonlinearities do, including ReLU, ELU, PReLU, and SeLU.

We need to recall the definition of asymptotic equivalence between numeric sequences which we use to describe characterization properties of distributions:

Definition 3.2.2 (Asymptotic equivalence for sequences)**.** *Two sequences a_k and b_k are called asymptotic equivalent and denoted as $a_k \asymp b_k$ if there exist constants $d > 0$ and $D > 0$ such that*

$$d \leq \frac{a_k}{b_k} \leq D, \quad \text{for all } k \in \mathbb{N}. \tag{3.35}$$

The extended envelope property of a function yields the following asymptotic equivalence:

Lemma 3.2.1. *Let a nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ obey the extended envelope property. Then for any symmetric random variable X the following asymptotic equivalence holds*

$$\|\phi(X)\|_k \asymp \|X\|_k, \quad \text{for all } k \geq 1, \tag{3.36}$$

where $\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k}$ is a k-th norm of X.

The proof can be found in the supplementary material.

Main theorem

This section postulates the rigorous result with a proof sketch. In the supplementary material one can find proofs of intermediate lemmas.

Firstly, we define the notion of *sub-Weibull* random variables [200, 328].

Definition 3.2.3 (Sub-Weibull random variable)**.** *A random variable X satisfying for all $x > 0$ and for some $\theta > 0$*

$$\mathbb{P}(|X| \geq x) \leq a \exp\left(-x^{1/\theta}\right), \tag{3.37}$$

is called a sub-Weibull random variable with so-called tail parameter θ, which is denoted by $X \sim \mathsf{subW}(\theta)$.

Sub-Weibull distributions are characterized by tails lighter than (or equally light as) Weibull distributions; in the same way as sub-Gaussian or sub-exponential distributions correspond to distributions with tails lighter than Gaussian and exponential distributions, respectively. Sub-Weibull distributions are parameterized by a positive tail index θ and are equivalent to sub-Gaussian for $\theta = 1/2$ and sub-exponential for $\theta = 1$. To describe a tail lower bound through some sub-Weibull distribution family, i.e. a distribution of X to have the tail heavier than some sub-Weibull, we define the optimal tail parameter for that distribution as the positive parameter θ characterized by:

$$\|X\|_k \asymp k^{\theta}. \tag{3.38}$$

Then X is sub-Weibull distributed with optimal tail parameter θ, in the sense that for any $\theta' < \theta$, X is not sub-Weibull with tail parameter θ' [see 328, for a proof].

The following theorem postulates the main results.

Theorem 3.2.1 (Sub-Weibull units)**.** *Consider a feed-forward Bayesian neural network with Gaussian priors (3.33) and with nonlinearity ϕ satisfying the extended envelope condition of Definition 3.2.1. Then conditional on the input $\boldsymbol{x}$, the marginal prior distribution[4] induced by forward propagation (3.32) on any unit (pre- or post-nonlinearity) of the ℓ-th hidden layer is sub-Weibull with optimal tail parameter $\theta = \ell/2$. That is for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,*

$$U_m^{(\ell)} \sim \mathsf{subW}(\ell/2),$$

where a subW *distribution is defined in Definition 3.2.3, and $U_m^{(\ell)}$ is either a pre-nonlinearity $g_m^{(\ell)}$ or a post-nonlinearity $h_m^{(\ell)}$.*

Proof. The idea is to prove by induction with respect to hidden layer depth ℓ that pre- and post-nonlinearities satisfy the asymptotic moment equivalence

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2} \text{ and } \|h^{(\ell)}\|_k \asymp k^{\ell/2}.$$

The statement of the theorem then follows by the moment characterization of optimal sub-Weibull tail coefficient in Equation (3.38).

According to Lemma 1.1 from from the supplementary material, centering does not harm tail properties, then, for simplicity, we consider zero-mean distributions $W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2)$.

Base step: Consider the distribution of the first hidden layer pre-nonlinearity $g = g^{(1)}$. Since weights $\boldsymbol{W}_m$ follow normal distribution and $\boldsymbol{x}$ is a feature vector, then

[4] We define the *marginal prior distribution* of a unit as its distribution obtained after all other units distributions are integrated out. *Marginal* is to be understood by opposition to *joint*, or *conditional*.

each hidden unit $\boldsymbol{W}_m^\intercal \boldsymbol{x}$ follow also normal distribution

$$g = \boldsymbol{W}_m^\intercal \boldsymbol{x} \sim \mathcal{N}(0, \sigma_w^2 \|\boldsymbol{x}\|^2).$$

Then, for normal zero-mean variable g, having variance $\sigma^2 = \sigma_w^2 \|\boldsymbol{x}\|^2$, holds the equality in sub-Gaussian property with variance proxy equals to normal distribution variance and from Lemma 1.1 in the supplementary material:

$$\|g\|_k \asymp \sqrt{k}.$$

As activation function ϕ obeys the extended envelope property, nonlinearity moments are asymptotically equivalent to symmetric variable moments

$$\|\phi(g)\|_k \asymp \|g\|_k \asymp \sqrt{k}.$$

It implies that first hidden layer post-nonlinearities h have sub-Gaussian distribution or sub-Weibull with tail parameter $\theta = 1/2$ (Definition 3.2.3).

Inductive step: show that if the statement holds for $\ell - 1$, then it also holds for ℓ.

Suppose the post-nonlinearity of $(\ell - 1)$-th hidden layer satisfies the moment condition. Hidden units satisfy the non-negative covariance theorem (Theorem 3.2.2):

$$\text{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0, \text{ for any } s, t \in \mathbb{N}.$$

Let the number of hidden units in $(\ell - 1)$-th layer equals to H. Then according to Lemma 2.2 from the supplementary material, under assumption of zero-mean Gaussian weights, pre-nonlinearities of ℓ-th hidden layer $g^{(\ell)} = \sum_{i=1}^{H} W_{m,i}^{(\ell-1)} h_i^{(\ell-1)}$ also satisfy the moment condition, but with $\theta = \ell/2$

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2}.$$

From the extended envelope property (Definition 3.2.1) post-nonlinearities $h^{(\ell)}$ satisfy the same moment condition as pre-nonlinearities $g^{(\ell)}$. This finishes the proof. ■

Remark 3.2.1. *If the activation function ϕ is bounded, such as the sigmoid or tanh, then the units are bounded. As a result, by Hoeffding's Lemma, they have a sub-Gaussian distribution.*

Remark 3.2.2. *Normalization techniques, such as batch normalization [163] or layer normalization [37], significantly reduce the training time in feed-forward neural networks. Normalization operations can be decomposed into a set of elementary operations. According to Proposition 1.4 from the supplementary material, elementary operations do not harm the distribution tail parameter. Therefore, normalization methods do not have an influence on tail behavior.*

Intermediate theorem

This section states with a proof sketch that the covariance between hidden units in the neural network is non-negative.

Theorem 3.2.2 (Non-negative covariance between hidden units)**.** *Consider the deep neural network described in, and with the assumptions of, Theorem 3.2.1. The covariance between hidden units of the same layer is non-negative. Moreover, for given ℓ-th hidden layer units $h^{(\ell)}$ and $\tilde{h}^{(\ell)}$, it holds*

$$Cov\left[\left(h^{(\ell)}\right)^{s}, \left(\tilde{h}^{(\ell)}\right)^{t}\right] \geq 0, \text{ where } s, t \in \mathbb{N}.$$

For first hidden layer $\ell = 1$ there is equality for all s and t.

Proof. A more detailed proof can be found in the supplementary material in Section 3.

Recall the covariance definition for random variables X and Y

$$\text{Cov}\left[X, Y\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{3.39}$$

The proof is based on induction with respect to the hidden layer number.

In the proof let us make notation simplifications: $\boldsymbol{w}_m^\ell = \boldsymbol{W}_m^\ell$ and $w_{mi}^\ell = W_{mi}^\ell$ for all $m \in H_\ell$. If the index m is omitted, then $\boldsymbol{w}^\ell$ is some the vectors $\boldsymbol{w}_m^\ell$, w_i^ℓ is i-th element of the vector $\boldsymbol{w}_m^\ell$.

1. First hidden layer. Consider the first hidden layer units $h^{(1)}$ and $\tilde{h}^{(1)}$. The covariance between units is equal to zero and the units are Gaussian, since the weights $\boldsymbol{w}^{(1)}$ and $\tilde{\boldsymbol{w}}^{(1)}$ are from $\mathcal{N}(0, \sigma_w^2)$ and independent. Thus, the first hidden layer units are independent and its covariance (3.39) is equal to 0. Moreover, since $h^{(1)}$ and $\tilde{h}^{(1)}$ are independent, then $\left(h^{(1)}\right)^{s}$ and $\left(\tilde{h}^{(1)}\right)^{t}$ are also independent.

2. Next hidden layers. Assume that the $(\ell-1)$-th hidden layer has $H_{\ell-1}$ hidden units, where $\ell > 1$. Then the ℓ-th hidden layer pre-nonlinearity is equal to

$$g^{(\ell)} = \sum_{i=1}^{H_{\ell-1}} w_i^{(\ell)} h_i^{(\ell-1)}. \tag{3.40}$$

We want to prove that the covariance (3.39) between the ℓ-th hidden layer pre-nonlinearities is non-negative. Let us show firstly the idea of the proof in the case $H_{\ell-1} = 1$ and then briefly describe the proof for any finite $H_{\ell-1} > 1$, $H_{\ell-1} \in \mathbb{N}$.

2.1 One hidden unit. In the case $H_{\ell-1} = 1$, the covariance (3.39) sign is the same as of the expression

$$\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2(s_1+t_1)}\right] - \mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2s_1}\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2t_1}\right],$$

since the weighs are zero-mean distributed, its moments are equal to zero with an odd order. According to Jensen's inequality for convex function f, we have $\mathbb{E}[f(x_1, x_2)] \geq f(\mathbb{E}[x_1], \mathbb{E}[x_2])$. Since a function $f(x_1, x_2) = x_1 x_2$ is convex for $x_1 \geq 0$ and $x_2 \geq 0$, then, taking $x_1 = \left(h^{(\ell-1)}\right)^{2s_1}$ and $x_2 = \left(h^{(\ell-1)}\right)^{2t_1}$, we have the condition we need (3.41) being satisfied.

2.1. H hidden units. Now let us consider the covariance between pre-nonlinearities (3.40) for $H_{\ell-1} = H > 1$. Raise the sum in the brackets to the power

$$\Big(\sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)}\Big)^s = \\ = \sum_{s_H=0}^{s} C_s^{s_H} \left(w_H^{(\ell)} h_H^{(\ell-1)}\right)^{s_H} \Big(\sum_{i=1}^{H-1} w_i^{(\ell)} h_i^{(\ell-1)}\Big)^{s-s_H}.$$

And the same way for the second bracket $\left(\sum_{i=1}^{H} \tilde{w}_i^{(\ell)} h_i^{(\ell-1)}\right)^t$. Notice that binomial terms will be the same in the minuend and the subtrahend terms of (3.39). So the covariance in our notations can be written in the form of

$$\mathrm{Cov}\left[\Big(\sum_{i=1}^{H_{\ell-1}} w_i^{(\ell)} h_i^{(\ell-1)}\Big)^s, \Big(\sum_{i=1}^{H_{\ell-1}} \tilde{w}_i^{(\ell)} h_i^{(\ell-1)}\Big)^t\right] = \\ = \sum\sum C\left(\mathbb{E}\left[AB\right] - \mathbb{E}\left[A\right]\mathbb{E}\left[B\right]\right),$$

where C-terms contain binomial coefficients, A-terms — all possible products of hidden units in $\left(g^{(\ell)}\right)^s$ and B-terms — all possible products of hidden units in $\left(\tilde{g}^{(\ell)}\right)^t$. In order for the covariance to be non-negative, it is sufficient to show that the difference $\mathbb{E}\left[AB\right] - \mathbb{E}\left[A\right]\mathbb{E}\left[B\right]$ is non-negative. Since the weights are Gaussian and independent, we have the following equation, omitting the superscript for simplicity,

$$\mathbb{E}\left[AB\right] = W\tilde{W} \cdot \mathbb{E}\left[\prod_{i=1}^{H} h_i^{s_i+t_i}\right],$$

$$\mathbb{E}\left[A\right]\mathbb{E}\left[B\right] = W\tilde{W} \cdot \mathbb{E}\left[\prod_{i=1}^{H} h_i^{s_i}\right] \mathbb{E}\left[\prod_{i=1}^{H} h_i^{t_i}\right],$$

where $W\tilde{W}$ is the product of weights moments

$$W\tilde{W} = \prod_{i=1}^{H} \mathbb{E}\left[w_i^{s_i}\right] \mathbb{E}\left[\tilde{w}_i^{t_i}\right].$$

For $W\tilde{W}$ not equal to zero, all the powers must be even. Now we need to prove

$$\mathbb{E}\left[\prod_{i=1}^{H/2} h_i^{2(s_i+t_i)}\right] \geq \mathbb{E}\left[\prod_{i=1}^{H/2} h_i^{2s_i}\right] \mathbb{E}\left[\prod_{i=1}^{H/2} h_i^{2t_i}\right]. \tag{3.41}$$

According to Jensen's inequality for convex functions, since a function $f(x_1, x_2) = x_1 x_2$ is convex for $x_1 \geq 0$ and $x_2 \geq 0$, then, taking $x_1 = \prod_{i=1}^{H/2} h_i^{2s_i}$ and $x_2 = \prod_{i=1}^{H/2} h_i^{2t_i}$, the condition from (3.41) is satisfied.

3. Post-nonlinearities.

Let show the proof for the ReLU nonlinearity.

The distribution of the ℓ-th hidden layer pre-nonlinearity $g^{(\ell)}$ is the sum of symmetric distributions, which are products of Gaussian variables $w^{(\ell)}$ and the non-negative ReLU output, i.e. the $(\ell-1)$-th hidden layer post-nonlinearity $h^{(\ell-1)}$. Therefore, $g^{(\ell)}$

follows a symmetric distribution and the following inequality

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} gg'\,p(g,g')\,dg\,dg' \geq \\ \geq \int_{-\infty}^{+\infty} g\,p(g)\,dg \cdot \int_{-\infty}^{+\infty} g'\,p(g')\,dg'$$

implies the same inequality for a positive part

$$\int_{0}^{+\infty}\int_{0}^{+\infty} gg'\,p(g,g')\,dg\,dg' \geq \\ \geq \int_{0}^{+\infty} g\,p(g)\,dg \cdot \int_{0}^{+\infty} g'\,p(g')\,dg'.$$

Notice that the equality above is the ReLU function output and for a symmetric distribution we have

$$\int_{0}^{+\infty} x\,p(x)\,dx = \frac{1}{2}\mathbb{E}\left[|X|\right]. \tag{3.42}$$

That means if the non-negative covariance is proven for pre-nonlinearities, for post-nonlinearities it is also non-negative. We omit the proof for the other nonlinearities with the extended envelope property, since instead of precise equation (3.42), the asymptotic equivalence for moments will be used for a positive part and for a negative part — precise expectation expressions which depend on certain nonlinearity. ■

Convolutional neural networks

Convolutional neural networks [123, 203] are a particular kind of neural network for processing data that has a known grid-like topology, which allows to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the neural network. Neurons in such networks are arranged in three dimensions: width, height and depth. There are three main types of layers that can be concatenated in these architectures: convolutional, pooling, and fully-connected layers (exactly as seen in standard NNs). The convolutional layer computes dot products between a region in the inputs and its weights. Therefore, each region can be considered as a particular case of a fully-connected layer. Pooling layers control overfitting and computations in deep architectures. They operate independently on every slice of the input and reduces it spatially. The most commonly functions used in pooling layers are *max pooling* and *average pooling*.

Proposition 3.2.1. *The operations: 1.* max pooling *and 2.* averaging *do not modify the optimal tail parameter θ of sub-Weibull family. Consequently, the result of Theorem 3.2.1 carries over to convolutional neural networks.*

The proof can be found in the supplementary material.

Corollary 3.2.1. *Consider a convolutional neural network containing convolutional, pooling and fully-connected layers under assumptions from Section 3.2.2. Then a unit of ℓ-th hidden layer has sub-Weibull distribution with optimal tail parameter $\theta = \ell/2$, where ℓ is the number of convolutional and fully-connected layers.*

Proof. Proposition 3.2.1 implies that the pooling layer keeps the tail parameter. From discussion at the beginning of the section, the result of Theorem 3.2.1 is also applied to convolutional neural networks where the depth is considered as the number of convolutional and fully-connected layers. ■

3.2.3 Regularization scheme on the units

Our main theoretical contribution, Theorem 3.2.1, characterizes the marginal prior distribution of the network units as follows: when the depth increases, the distribution becomes more heavy-tailed. In this section, we provide an interpretation of the result in terms of regularization at the level of the units. To this end, we first briefly recall shrinkage and penalized estimation methods.

Short digest on penalized estimation

The notion of penalized estimation is probably best illustrated on the simple linear regression model, where the aim is to improve prediction accuracy by shrinking, or even putting exactly to zero, some coefficients in the regression. Under these circumstances, inference is also more *interpretable* since, by reducing the number of coefficients effectively used in the model, it is possible to grasp its salient features. Shrinking is performed by imposing a penalty on the size of the coefficients, which is equivalent to allowing for a given budget on their size. Denote the regression parameter by $\beta \in \mathbb{R}^p$, the regression sum-of-squares by $R(\beta)$, and the penalty by $\lambda L(\beta)$, where L is some norm on $\mathbb{R}^p$ and λ some positive tuning parameter. Then, the two formulations of the regularized problem

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda L(\beta), \text{ and}$$
$$\min_{\beta \in \mathbb{R}^p} R(\beta) \text{ subject to } L(\beta) \leq t,$$

are equivalent, with some one-to-one correspondence between λ and t, and are respectively termed the *penalty* and the *constraint* formulation. This latter formulation provides an interesting geometrical intuition of the shrinkage mechanism: the constraint $L(\beta) \leq t$ reads as imposing a total budget of t for the parameter size in terms of the norm L. If the ordinary least squares estimator $\hat{\beta}^{\text{ols}}$ lives in the L-ball with surface $L(\beta) = t$, then there is no effect on the estimation. In contrast, when $\hat{\beta}^{\text{ols}}$ is outside the ball, then the intersection of the lowest level curve of the sum-of-squares $R(\beta)$ with the L-ball defines the penalized estimator.

The choice of the L norm has considerable effects on the problem, as can be sensed geometrically. Consider for instance $\mathcal{L}^q$ norms, with $q \geq 0$. For any $q > 1$, the associated $\mathcal{L}^q$ norm is differentiable and contours have a round shape without sharp angles. In that case, the penalty effect is to shrink the β coefficients towards 0. The most well-known estimator falling in this class is the *ridge* regression obtained with $q = 2$, see Figure 3.5 top-left panel. In contrast, for any $q \in (0, 1]$, the $\mathcal{L}^q$ norm has some non differentiable points along the axis coordinates, see Figure 3.5 top-right and bottom panels. Such critical points are more likely to be hit by the level curves of the sum-of-squares $R(\beta)$, thus setting exactly to zero some of the parameters. A

very successful approach in this class is the Lasso obtained with $q = 1$. Note that the problem is computationally much easier in the convex situation which occurs only for $q \geq 1$.

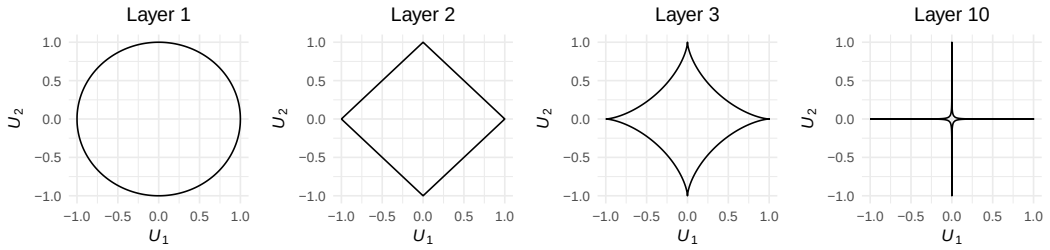


FIGURE 3.5: $\mathcal{L}^{2/\ell}$-norm unit balls (in dimension 2) for layers $\ell = 1, 2, 3$ and 10.

MAP on weights W is weight decay

These penalized methods have a simple Bayesian counterpart in the form of the maximum a posteriori (MAP) estimator. In this context, the objective function R is the negative log-likelihood, while the penalty L is the negative log-prior. The objective function takes on the form of sum-of-squared errors for regression under Gaussian errors, and of cross-entropy for classification.

For neural networks, it is well-known that an independent Gaussian prior on the weights

$$\pi(\boldsymbol{W}) \propto \prod_{\ell=1}^{L} \prod_{i,j} \mathrm{e}^{-\frac{1}{2}(W_{i,j}^{(\ell)})^2}, \tag{3.43}$$

is equivalent to the weight decay penalty, also known as ridge regression:

$$L(\boldsymbol{W}) = \sum_{\ell=1}^{L} \sum_{i,j} (W_{i,j}^{(\ell)})^2 = \|\boldsymbol{W}\|_2^2, \tag{3.44}$$

where products in (3.43) and sums in (3.44) involving i and j above are over $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$, H_0 and H_L representing respectively the input and output dimensions.

MAP on units U

Now moving the point of view from *weights* to *units* leads to a radically different shrinkage effect. Let $U_m^{(\ell)}$ denote the m-th unit of the ℓ-th layer (either pre- or post-nonlinearity). We prove in Theorem 3.2.1 that conditional on the input $\boldsymbol{x}$, a Gaussian prior on the weights translates into some prior on the units $U_m^{(\ell)}$ that is marginally sub-Weibull with optimal tail index $\theta = \ell/2$. This means that the tails of $U_m^{(\ell)}$ satisfy

$$\mathbb{P}(|U_m^{(\ell)}| \geq u) \leq \exp\left(-u^{2/\ell}/K_1\right) \quad \text{for all } u \geq 0, \tag{3.45}$$

for some positive constant K_1. The exponent of u in the exponential term above is optimal in the sense that Equation (3.45) is not satisfied with some parameter θ' smaller than $\ell/2$. Thus, the marginal density of $U_m^{(\ell)}$ on $\mathbb{R}$ is approximately proportional to

$$\pi_m^{(\ell)}(u) \approx \mathrm{e}^{-|u|^{2/\ell}/K_1}. \tag{3.46}$$

The joint prior distribution for all the units $\boldsymbol{U} = (U_m^{(\ell)})_{1\leq\ell\leq L, 1\leq m\leq H_\ell}$ can be expressed from all the marginal distributions by Sklar's representation theorem [303] as

$$\pi(\boldsymbol{U}) = \prod_{\ell=1}^{L}\prod_{m=1}^{H_\ell} \pi_m^{(\ell)}(U_m^{(\ell)})\, C(F(\boldsymbol{U})), \tag{3.47}$$

where C represents the copula of $\boldsymbol{U}$ (which characterizes all the dependence between the units) while F denotes its cumulative distribution function. The penalty incurred by such a prior distribution is obtained as the negative log-prior,

$$\begin{aligned} L(\boldsymbol{U}) &= -\sum_{\ell=1}^{L}\sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(\boldsymbol{U})), \\ &\overset{(a)}{\approx} \sum_{\ell=1}^{L}\sum_{m=1}^{H_\ell} |U_m^{(\ell)}|^{2/\ell} - \log C(F(\boldsymbol{U})), \\ &\approx \|\boldsymbol{U}^{(1)}\|_2^2 + \|\boldsymbol{U}_1^{(2)}\|_1 + \cdots + \|\boldsymbol{U}^{(L)}\|_{2/L}^{2/L} \\ &\quad - \log C(F(\boldsymbol{U})), \end{aligned} \tag{3.48}$$

where (a) comes from (3.46). The first L terms in (3.48) indicate that some shrinkage operates at every layer of the network, with a penalty term that approximately takes the form of the $\mathcal{L}^{2/\ell}$ norm at layer ℓ. Thus, the deeper the layer, the stronger the regularization induced at the level of the units, as summarized in Table 3.2.

Layer	Penalty on $\boldsymbol{W}$	Approximate penalty on $\boldsymbol{U}$	
1	$\|\boldsymbol{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\boldsymbol{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\boldsymbol{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$

TABLE 3.2: Comparison of Bayesian neural network penalties on weights $\boldsymbol{W}$ and units $\boldsymbol{U}$.

3.2.4 Experiments

We illustrate the result of Theorem 3.2.1 on a 100 layers MLP. The hidden layers of neural network have $H_1 = 1000$, $H_2 = 990$, $H_3 = 980$, ..., $H_\ell = 1000 - 10(\ell - 1)$, ..., $H_{100} = 10$ hidden units, respectively. The input $\boldsymbol{x}$ is a vector of features from $\mathbb{R}^{10^4}$. Figure 3.6 represents the tails of first three, 10th and 100th hidden layers pre-nonlinearity marginal distributions in logarithmic scale. Units of one layer have the same sub-Weibull distribution since they share the same input and prior on the

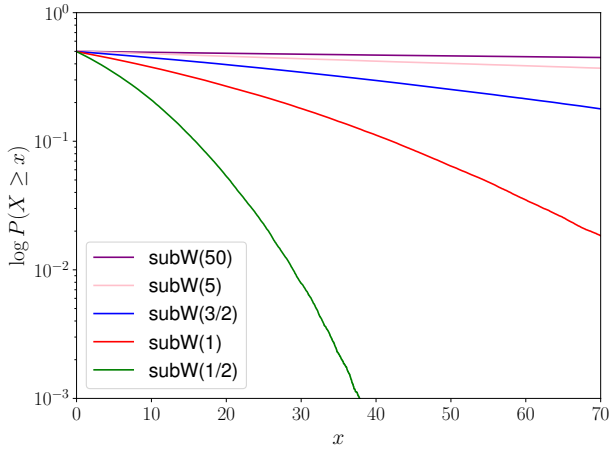


FIGURE 3.6: Illustration of layers $\ell = 1, 2, 3, 10$ and 100 hidden units (pre-nonlinearities) marginal prior distributions. They correspond respectively to subW$(1/2)$, subW(1), subW$(3/2)$, subW(5) and subW(50).

corresponding weights. The curves are obtained as histograms from a sample of size 10^5 from the prior on the pre-nonlinearities, which is itself obtained by sampling 10^5 sets of weights $\boldsymbol{W}$ from the Gaussian prior (3.33) and forward propagation via (3.32). The input vector $\boldsymbol{x}$ is sampled with independent features from a standard normal distribution once for all at the start. The nonlinearity ϕ is the ReLU function. Being a linear combination involving symmetric weights $\boldsymbol{W}$, pre-nonlinearities $\boldsymbol{g}$ also have a symmetric distribution, thus we visualize only their distribution on $\mathbb{R}_+$.

Figure 3.6 corroborates our main result. On the one hand, the prior distribution of the first hidden units is Gaussian (green curve), which corresponds to a subW$(1/2)$ distribution. On the other hand, deeper layers are characterized by heavier-tailed distributions. The deepest considered layer (100th, violet curve) has an extremely flat distribution, which corresponds to a subW(50) distribution.

3.2.5 Discussion

Despite the ubiquity of deep learning throughout science, medicine and engineering, the underlying theory has not kept pace with applications for deep learning. In this section, we have extended the state of knowledge on Bayesian neural networks by providing a characterization of the marginal prior distribution of the units. [226] and [204] proved that unit distributions have a Gaussian process limit in the wide regime, i.e. when the number of hidden units tends to infinity. We showed that they are heavier-tailed as depth increases, and discussed this result in terms of a regularizing mechanism at the level of the units. We anticipate that the Gaussian process limit of sub-Weibull distributions in a given layer for increasing width could be also recovered through a modification of the Central Limit Theorem for heavy-tailed distributions, see [200].

Since initialization and learning dynamics are key in modern machine learning in order to properly tune deep learning algorithms, a good implementation practice

requires a proper understanding of the prior distribution at play and of the regularization it incurs.

We hope that our results will open avenues for further research. Firstly, Theorem 3.2.1 regards the *marginal* prior distribution of the units, while a full characterization of the joint distribution of all units $\boldsymbol{U}$ remains an open question. More specifically, a precise description of the copula defined in Equation (3.47) would provide valuable information about the dependence between the units, and also about the precise geometrical structure of the balls induced by that penalty. Secondly, the interpretation of our result (Section 3.2.3) is concerned with the maximum a posteriori of the units, which is a point estimator. One of the benefits of the Bayesian approach to neural networks lies in its ability to provide a principled approach to uncertainty quantification, so that an interpretation of our result in terms of the full posterior distribution would be very appealing. Lastly, the practical potentialities of our results are many: to better comprehend the regularizing mechanisms in deep neural networks will contribute to design and understand strategies to avoid overfitting and improve generalization.

[A3] J. Arbel, M. Crispino, and S. Girard. Dependence properties and Bayesian inference for asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 174, 2019

3.3 Dependence properties of asymmetric copulas

3.3.1 Introduction

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a continuous random vector with d-variate cumulative distribution function (cdf) F, and let F_j, $j \in \{1, \ldots, d\}$, be the marginal cdf of X_j. According to Sklar's theorem [303], there exists a unique d-variate function $C : [0,1]^d \to [0,1]$ such that

$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \quad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

The function C is referred to as the copula associated with F. It is the d-dimensional cdf of the random vector $(F_1(X_1), \ldots, F_d(X_d))$ with uniform margins on $[0,1]$.

A copula is said to be symmetric (or exchangeable) if for any $\boldsymbol{u} \in [0,1]^d$, and for any permutation $(\sigma_1, \ldots, \sigma_d)$ of the first d integers $\{1, \ldots, d\}$, it holds that $C(u_1, \ldots, u_d) = C(u_{\sigma_1}, \ldots, u_{\sigma_d})$. The assumption of exchangeability may be unrealistic in many domains, including quantitative risk management [99], reliability modeling [339], and oceanography [344]. The urge for asymmetric copula models in order to better account for complex dependence structures has recently stimulated research in several directions, including [288, 6, 103, 339, 104]. We focus here on a simple yet general method for building asymmetric copulas introduced by [207, Theorem 2.1, Property (i)]:

Theorem 3.3.1. *[Liebscher, 207] Let $C_1, \ldots, C_K : [0,1]^d \to [0,1]$ be copulas, $g_j^{(k)} : [0,1] \to [0,1]$ be increasing functions such that $g_j^{(k)}(0) = 0$ and $g_j^{(k)}(1) = 1$ for all $k \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, d\}$. Then,*

$$\boldsymbol{u} \in [0,1]^d \mapsto \tilde{C}(\boldsymbol{u}) = \prod_{k=1}^{K} C_k(g_1^{(k)}(u_1), \ldots, g_d^{(k)}(u_d)) \tag{3.49}$$

is also a copula under the constraint that

$$\prod_{k=1}^{K} g_j^{(k)}(u) = u \quad \text{for all } u \in [0,1], j \in \{1, \ldots, d\}. \tag{3.50}$$

Theorem 3.3.1 provides a generic way to construct an asymmetric copula $\tilde{C}$, henceforth referred to as *Liebscher copula*, starting from a sequence of symmetric copulas $C_1, \ldots, C_K$. This mechanism was first introduced by [187] in the particular case where $K = 2$ and with the functions $g_j^{(k)}$ assumed to be power functions, for each $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, K\}$, that satisfy condition (3.50). The class of Liebscher copulas covers a broad range of dependencies and benefits from tractable bounds on dependence coefficients of the bivariate marginals [207, 208, 227]. However,

there are two main reasons why the practical implementation of this approach is not straightforward: (i) it is not immediate to construct functions that satisfy condition (3.50); and (ii) the product form complicates the density computation even numerically, which makes it difficult to perform likelihood inference on the model parameters [227].

The aim of this section is to deepen the understanding of Liebscher's construction in order to overcome drawbacks (i) and (ii). Our contributions in this regard are three-fold. First, we provide theoretical properties of the asymmetric copulas in (3.49), including exact expressions of tail dependence indices, thus complementing the partial results of [207, 208]. Second, we give an iterative representation of (3.49) which has the advantage to relax assumption (3.50) by automatically satisfying it. Third, we develop an inferential procedure and a sampling scheme that rely on the newly developed iterative representation.

The Bayesian paradigm proves very useful for inference in our context as it overcomes the problematic computation of the maximum likelihood estimate, which requires the maximization of a very complicated likelihood function (see recent contributions [326, 253]). General Bayesian sampling solutions in the form of Markov chain Monte Carlo are not particularly well-suited neither since they require the evaluation of that complex likelihood. Instead, we resort to Approximate Bayesian computation (ABC), a technique dedicated to models with complicated, or intractable, likelihoods (see [224, 285, 184] for recent reviews). ABC requires the ability to sample from the model, which is straightforward with our iterative representation of Liebscher copula. The adequacy of ABC for inference in copula models was leveraged by [143], although in the different context of empirical likelihood estimation. A reversed approach to ours is followed by [205], who make use of copulas in order to adapt ABC to high-dimensional settings.

Since its introduction, the construction by Liebscher has received much attention in the copula literature (e.g., [295, 105, 201]). However, most studies have been limited to simple cases where the product in (3.49) has only two terms. We hope that this paper will contribute to the further spreading of Liebscher's copulas, because it allows to exploit their full potential by: (i) better understanding their properties; (ii) providing a novel construction, which facilitates their use with an arbitrary number K of terms in (3.49); and (iii) giving a strategy to make inference on them.

On top of what has been presented above, an additional contribution of this paper is to derive specific results for the subclass of Liebscher's copula when two or more comonotonic copulas are combined, which we call comonotonic-based Liebscher copula. This subclass is characterized by an arbitrary number of singular components. To the best of our knowledge, this is the first paper to investigate this copula's properties and to provide an inference procedure.

The section is organized as follows. Section 3.3.2 provides some theoretical results concerning the properties of asymmetric Liebscher copulas, also presenting the novel

iterative construction. In Section 3.3.3, we introduce and analyze the comonotonic-based Liebscher copula. Section 3.3.4 is dedicated to the inference strategy. It demonstrates our approach on simulated data and provides a comparison with a likelihood-based approach for a class of Liebscher copulas where maximum likelihood estimation is feasible. We conclude with a short discussion in Section 3.3.5. The reader is referred to [3] for proofs.

3.3.2 Properties of Liebscher copula

In this section, some new properties of the copula $\tilde{C}$ are established, complementing the ones in [207, 208]. Sections 3.3.2 and 3.3.2 are dedicated to (tail) dependence properties. For the sake of simplicity, we focus on the case $d = 2$ of bivariate copulas. Some stability properties of Liebscher's construction are highlighted in Section 3.3.2. Finally, an alternative construction to Liebscher copula (3.49) is introduced in Section 3.3.2.

Tail dependence

The *lower and upper tail dependence functions*, denoted by $\Lambda_L(C;\cdot)$ and $\Lambda_U(C,\cdot)$ respectively, are defined for all $(x,y) \in \mathbb{R}_+^2$ by

$$\Lambda_L(C;x,y) = \lim_{\varepsilon\to 0} \frac{C(\varepsilon x, \varepsilon y)}{\varepsilon}, \quad \text{and} \quad \Lambda_U(C;x,y) = x + y + \lim_{\varepsilon\to 0} \frac{C(1-\varepsilon x, 1-\varepsilon y) - 1}{\varepsilon},$$

where C is a given bivariate copula, see for instance [178]. Note that these limits exist under a bivariate regular variation assumption, see [283], Section 5.4.2 for details. When they exist, these functions are homogeneous ([178], Proposition 2.2), *i.e.*, for all $t \in (0,1]$ and $(x,y) \in \mathbb{R}_+^2$, $\Lambda(C;tx,ty) = t\Lambda(C;x,y)$, where Λ is equal to Λ_L or Λ_U. The *lower and upper tail dependence coefficients*, denoted by $\lambda_L(C)$ and $\lambda_U(C)$ respectively, are defined as the conditional probabilities that a random vector associated with a copula C belongs to lower or upper tail orthants given that a univariate margin takes extreme values:

$$\lambda_L(C) = \lim_{u\to 0} \frac{C(u,u)}{u}, \quad \lambda_U(C) = 2 - \lim_{u\to 1} \frac{C(u,u)-1}{u-1}.$$

These coefficients can also be interpreted in terms of the tail dependence functions: $\lambda_L(C) = \Lambda_L(C;1,1)$ and $\lambda_U(C) = \Lambda_U(C;1,1)$. Conversely, in view of the homogeneity property, the behavior of the tail dependence functions on the diagonal is determined by the tail dependence coefficients: $\Lambda_L(C;t,t) = \lambda_L(C)t$ and $\Lambda_U(C;t,t) = \lambda_U(C)t$ for all $t \in (0,1]$. The tail dependence functions for Liebscher copula are provided by Proposition 3.3.1 which, in view of the previous remarks, allows us to derive the tail dependence coefficients in Corollary 3.3.1. Some of these results rely on the notion of (univariate) regular variation. Recall that a positive function g is said to be regularly varying with index γ if $g(xt)/g(x) \to t^\gamma$ as $x \to \infty$ for all $t > 0$, see [57].

Proposition 3.3.1. *Let $(x,y) \in \mathbb{R}_+^2$ and consider $\tilde{C}$ the bivariate copula defined by (3.49) with $d = 2$.*

(i) *Lower tail, symmetric case. Assume that* $g_1^{(k)} = g_2^{(k)}$ *is a regularly varying function with index* $\gamma^{(k)} > 0$ *for all* $k \in \{1, \dots, K\}$. *Then,*

$$\Lambda_L(\tilde{C}; x, y) = \prod_{k=1}^{K} \Lambda_L(C_k; x^{\gamma^{(k)}}, y^{\gamma^{(k)}})$$

and, necessarily, $\sum_{k=1}^{K} \gamma^{(k)} = 1$.

(ii) *Lower tail, asymmetric case. Suppose there exists* $k_0 \in \{1, \dots, K\}$ *such that* $g_1^{(k_0)}(\varepsilon)/g_2^{(k_0)}(\varepsilon) \to 0$ *as* $\varepsilon \to 0$. *Then,*

$$\Lambda_L(\tilde{C}; x, y) = 0.$$

(iii) *Upper tail, general case. Assume that, for all* $k \in \{1, \dots, K\}$, $g_1^{(k)}$ *and* $g_2^{(k)}$ *are differentiable at 1, with derivative at 1 denoted by* $\mathrm{d}_1^{(k)}$ *and* $\mathrm{d}_2^{(k)}$ *respectively. Then*

$$\Lambda_U(\tilde{C}; x, y) = \sum_{k=1}^{K} \Lambda_U(C_k; \mathrm{d}_1^{(k)} x, \mathrm{d}_2^{(k)} y)$$

and, necessarily, $\sum_{k=1}^{K} \mathrm{d}_j^{(k)} = 1$, *for* $j \in \{1, 2\}$.

(iv) *Upper tail, particular case. If, in addition to* (iii), $d_1^{(k)} = d_2^{(k)} =: d^{(k)}$ for all $k \in \{1, \dots, K\}$, then

$$\Lambda_U(\tilde{C}; x, y) = \sum_{k=1}^{K} d^{(k)} \Lambda_U(C_k; x, y)$$

and, necessarily, $\sum_{k=1}^{K} d^{(k)} = 1$.

Let us note that the functions $g_j^{(k)}$ considered by [207] and indexed by (I-III) in his Section 2.1 all satisfy the assumptions of Proposition 3.3.1. The following result complements [207, Proposition 2.3] and [208, Proposition 0.1] which provide bounds on the tail dependence coefficients. Here instead, explicit calculations are provided.

Corollary 3.3.1. *Let* $\tilde{C}$ *be the bivariate copula defined by (3.49) with* $d = 2$.

(i) *Lower tail, symmetric case. Under the assumptions of Proposition 3.3.1*(i), $\lambda_L(\tilde{C}) = \prod_{k=1}^{K} \lambda_L(C_k)$.

(ii) Lower tail, asymmetric case. Under the assumptions of Proposition 3.3.1(ii), $\lambda_L(\tilde{C}) = 0$.

(iii) Upper tail, general case. Under the assumptions of Proposition 3.3.1(iii), $\lambda_U(\tilde{C}) = \sum_{k=1}^{K} \Lambda_U(C_k; d_1^{(k)}, d_2^{(k)})$ and, necessarily, $\sum_{k=1}^{K} d_j^{(k)} = 1$, for $j \in \{1, 2\}$.

(iv) Upper tail, particular case. Under the assumptions of Proposition 3.3.1(iv), $\lambda_U(\tilde{C}) = \sum_{k=1}^{K} d^{(k)} \lambda_U(C_k)$ and, necessarily, $\sum_{k=1}^{K} d^{(k)} = 1$.

It appears that the lower and upper tail dependence coefficients have very different behaviors. In the case (i) of a symmetric copulas, the lower tail dependence coefficient λ_L is the product of the lower tail dependence coefficients associated with the components. Besides, $\lambda_L = 0$ as soon as a component k_0 has functions $g_1^{(k_0)}$ and $g_2^{(k_0)}$ with different behaviors at the origin (case (ii)). At the opposite, the upper tail dependence coefficient does not vanish even though a component k_0 has functions $g_1^{(k_0)}$

and $g_2^{(k_0)}$ with different behaviors at 1 (case (iii)). In the particular situation where all components $k \in \{1, \ldots, K\}$ have functions $g_1^{(k)}$ and $g_2^{(k)}$ with the same behavior at 1 (case (iv)), λ_U is a convex combination of the upper tail dependence coefficients associated with the components.

Dependence

Let (X, Y) be a pair of random variables with continuous margins and associated copula C.

- X and Y are said to be *totally positive of order 2*, TP_2 (see [177]), if for all $x_1 < y_1$, $x_2 < y_2$,

  $$\Pr(X \leq x_1, Y \leq x_2)\Pr(X \leq y_1, Y \leq y_2) \geq \Pr(X \leq x_1, Y \leq y_2)\Pr(X \leq y_1, Y \leq x_2).$$

  Since this can be equivalently written in terms of C, we will write in short that C is TP_2.

- X and Y are said to be *positively quadrant dependent* (PQD) if

  $$\Pr(X \leq x, Y \leq y) \geq \Pr(X \leq x)\Pr(Y \leq y) \text{ for all } (x, y).$$

  Since this property can be characterized by the copula property $C \geq \Pi$ where Π denotes the independence copula, see for instance [244, Paragraph 5.2.1], we shall write for short that C is PQD. The *negatively quadrant dependence* (NQD) property is similarly defined by $C \leq \Pi$.

- X and Y are said to be *left-tail decreasing* (LTD) if

  $$\begin{aligned} &\Pr(X \leq x | Y \leq y) \text{ is a decreasing function of } y \text{ for all } x, \\ &\Pr(Y \leq y | X \leq x) \text{ is a decreasing function of } x \text{ for all } y. \end{aligned} \tag{3.51}$$

  From [244, Theorem 5.2.5], this property can be characterized by the copula properties

  $$\begin{aligned} &C(u, v)/u \text{ is decreasing in } u \text{ for all } v \in [0, 1], \\ &C(u, v)/v \text{ is decreasing in } v \text{ for all } u \in [0, 1], \end{aligned} \tag{3.52}$$

  and we shall thus write that C is LTD. The *left-tail increasing* property (LTI) is similarly defined by reversing the directions of the monotonicity in (3.51) and (3.52).

- X and Y are said to be *stochastically increasing* (SI) if

  $$\begin{aligned} &\Pr(X > x | Y = y) \text{ is an increasing function of } y \text{ for all } x, \\ &\Pr(Y > y | X = x) \text{ is an increasing function of } x \text{ for all } y. \end{aligned} \tag{3.53}$$

  From [244, Corollary 5.2.11], this property can be characterized by the copula properties

  $$\begin{aligned} &C(u, v) \text{ is a concave function of } u \text{ for all } v \in [0, 1], \\ &C(u, v) \text{ is a concave function of } v \text{ for all } u \in [0, 1], \end{aligned} \tag{3.54}$$

and we shall thus write that C is SI. The *stochastically decreasing* (SD) property is similarly defined by replacing increasing by decreasing in (3.53) and concave by convex in (3.54).

In the next proposition, we show that under mild conditions, the above dependence properties are preserved under Liebscher's construction, thus complementing LTD and TP_2 properties established in [207, Proposition 2.2].

Proposition 3.3.2. *If copulas $C_1, \ldots, C_K$ all satisfy any of the properties defined above,* TP_2, PQD, NQD, LTD, LTI, SI *or* SD, *then the same is satisfied for the Liebscher copula $\tilde{C}$ defined in* (3.49)*—for* SI *(respectively* SD*), the $g_j^{(k)}$ functions in Theorem 3.3.1 are additionally required to be concave functions (respectively convex functions) and the copulas C_k to be twice differentiable, $k \in \{1, \ldots, K\}$, $j \in \{1, \ldots, d\}$.*

Stability properties

Let us focus on the situation where the functions $g_j^{(k)}$ of Theorem 3.3.1 are power functions: for all $j \in \{1, \ldots, d\}$, $k \in \{1, \ldots, K\}$ and $t \in [0, 1]$, let

$$g_j^{(k)}(t) = t^{p_j^{(k)}}, \quad p_j^{(k)} \in (0, 1), \sum_{\ell=1}^{K} p_j^{(\ell)} = 1. \tag{3.55}$$

Recall that a copula $C_\#$ is said to be max-stable if for all integer $n \geq 1$ and $(u_1, \ldots, u_d) \in [0, 1]^d$:

$$C_\#^n(u_1^{1/n}, \ldots, u_d^{1/n}) = C_\#(u_1, \ldots, u_d).$$

From [127, Proposition 3], it is clear that associating max-stable copulas C_k with power functions (3.55) in Liebscher construction (3.49) still yields a max-stable copula. The goal of this paragraph is to investigate to what extent this result can be generalized. Our first result establishes the stability of a family of Liebscher copulas built from homogeneous functions. More specifically, each copula $C_k(\cdot)$ in (3.49) is rewritten as $C(\cdot \mid \theta_k)$ where

$$C(\cdot \mid \theta_k) := \prod_{i=1}^{m} \varphi_i^{\theta_{ik}}(\cdot), \tag{3.56}$$

with $\theta_k = (\theta_{1k}, \ldots, \theta_{mk})^T$ and $\varphi_i : [0, 1]^d \to [0, 1]$, $i \in \{1, \ldots, m\}$.

Proposition 3.3.3. *For all $j \in \{1, \ldots, d\}$ and $t \in [0, 1]$, let $g_j^{(k)}$ be given by (3.55) where $p_j^{(k)} = p^{(k)}$ for all $k \in \{1, \ldots, K\}$. Let $m > 0$ and for all $i \in \{1, \ldots, m\}$ introduce $\varphi_i : [0, 1]^d \to [0, 1]$ such that $\ln \circ \varphi_i \circ \exp$ is homogeneous of degree λ_i. For all $k \in \{1, \ldots, K\}$, assume that $C(\cdot \mid \theta_k)$ in (3.56) is a copula for some $\theta_k \in \mathbb{R}^m$. Then, copula (3.49) is given for all $K \geq 1$ by*

$$\tilde{C}^{(K)}(\cdot) = C(\cdot \mid \tilde{\theta}_K),$$

with $\tilde{\theta}_K = (\tilde{\theta}_{1K}, \ldots, \tilde{\theta}_{mK})^T$, and for all $i \in \{1, \ldots, m\}$,

$$\tilde{\theta}_{iK} = \sum_{k=1}^{K} \theta_{ik} (p^{(k)})^{\lambda_i}.$$

Example 3.3.1 (Gumbel-Barnett copula C_k)**.** *Let C_k be the Gumbel-Barnett copula [244, Table 4.1]. It can be written as*

$$C_k(\boldsymbol{u}) = C(\boldsymbol{u} \mid \theta_k) = \prod_{j=1}^{d} u_j \exp\left(\theta_k \prod_{j=1}^{d} \ln(1/u_j)\right) = \varphi_1^{\theta_{1k}}(\boldsymbol{u})\varphi_2^{\theta_{2k}}(\boldsymbol{u})$$

with $\theta_{1k} = 1$, $\theta_{2k} = \theta_k \geq 1$,

$$\varphi_1(\boldsymbol{u}) = \prod_{j=1}^{d} u_j \;\; \textit{and} \;\; \varphi_2(\boldsymbol{u}) = \exp\left(\prod_{j=1}^{d} \ln(1/u_j)\right).$$

It thus fulfills the assumptions of Proposition 3.3.3 with $m = 2$, $\lambda_1 = 1$ and $\lambda_2 = d$.

Example 3.3.2 (Extreme-value copula C_k)**.** *Extreme-value copulas exactly correspond to max-stable copulas and are characterized by their tail-dependence function L as:*

$$C_{\#}(u_1, \ldots, u_d) = \exp(-L(\ln(1/u_1), \ldots, \ln(1/u_d))),$$

where $L : \mathbb{R}_+^d \to \mathbb{R}_+$ is homogeneous of degree 1, see for instance [150]. It is thus clear that every max-stable copula C_k fulfills the assumptions of Proposition 3.3.3 with $m = 1$, $\theta_{1k} = 1$, $\lambda_1 = 1$ and $\ln \circ \varphi_1 \circ \exp(\boldsymbol{t}) = -L(-\boldsymbol{t})$, $t \in \mathbb{R}_-^d$. Moreover,

$$\tilde{\theta}_{1K} = \sum_{k=1}^{K} p^{(k)} = 1$$

and thus $\tilde{C}^{(K)} = C$ for all $K \geq 1$.

It appears that max-stable copulas can be considered as fixed-points of Liebscher's construction (3.49). The next result shows that, under mild assumptions, they are the only copulas verifying this property.

Proposition 3.3.4. *For all $j \in \{1, \ldots, d\}$, let $g_j^{(k)}$ be given by (3.55) where $p_j^{(k)} = p^{(k)}$ for all $k \in \{1, \ldots, K\}$. Assume $C_k = C$ for all $k \in \{1, \ldots, K\}$ and let $\tilde{C}^{(K)}$ be the copula defined by (3.49). Then, $\tilde{C}^{(K)} = C$ for all $K \geq 1$ and for all sequences $p^{(1)}, \ldots, p^{(K)} \in (0, 1)$ such that $\sum_{k=1}^{K} p^{(k)} = 1$ if and only if C is max-stable.*

To complete the links with max-stable copulas, let us consider the situation where $C_k = C$ and $p_j^{(k)} = 1/K$ for all $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, K\}$. Liebscher's construction thus yields

$$\tilde{C}^{(K)}(\boldsymbol{u}) := \tilde{C}(\boldsymbol{u}) = C^K(u_1^{1/K}, \ldots, u_d^{1/K}),$$

which is the normalized cdf associated with the maximum of K independent uniform random vectors distributed according to the cdf C. Therefore, as $K \to \infty$, $\tilde{C}^{(K)}$ converges to a max-stable copula under standard extreme-value assumptions on C.

An iterative construction

Let $\mathcal{F}$ be the class of increasing functions $f : [0, 1] \to [0, 1]$ such that $f(0) = 0$, $f(1) = 1$ and Id/f is increasing, where Id denotes the identity function. For all $k \geq 1$ let C_k be a d-variate copula and $f_j^{(k)} \in \mathcal{F}$ for all $j \in \{1, \ldots, d\}$, with the assumption

$f_j^{(1)}(t) = 1$ for all $t \in [0,1]$. We propose the following iterative construction of copulas. For all $\boldsymbol{u} \in [0,1]^d$, consider the sequence defined by

$$\tilde{C}^{(1)}(\boldsymbol{u}) = C_1(\boldsymbol{u}), \tag{3.57}$$

$$\tilde{C}^{(k)}(\boldsymbol{u}) = C_k\left(\frac{u_1}{f_1^{(k)}(u_1)}, \ldots, \frac{u_d}{f_d^{(k)}(u_d)}\right) \tilde{C}^{(k-1)}\left(f_1^{(k)}(u_1), \ldots, f_d^{(k)}(u_d)\right), \ k \geq 2. \tag{3.58}$$

We prove in [14] that $\tilde{C}^{(k)}$ is a d-variate copula, for all $k \geq 1$.

Let $K \geq 1$. For all functions $f^{(1)}, \ldots, f^{(K)} : [0,1] \to [0,1]$ and $i, j \in \{1, \ldots, K\}$, let us introduce the notation

$$\bigodot_{k=i}^{j} f^{(k)} := f^{(i)} \circ \ldots \circ f^{(j)} \text{ if } i \leq j \text{ and } \bigodot_{k=i}^{j} f^{(k)} := \mathrm{Id} \text{ otherwise.} \tag{3.59}$$

The next result shows that there is a one-to-one correspondence between copulas built by the iterative procedure (3.57), (3.58) and Liebscher copulas, reported in Theorem 3.3.1.

Proposition 3.3.5. *The copula $\tilde{C}^{(K)}$, $K \geq 1$ defined iteratively by (3.57), (3.58) is a Liebscher copula. It can be rewritten as*

$$\tilde{C}^{(K)}(\boldsymbol{u}) = \prod_{k=1}^{K} C_k\left(g_1^{(K-k+1,K)}(u_1), \ldots, g_d^{(K-k+1,K)}(u_d)\right) \tag{3.60}$$

for all $\boldsymbol{u} \in [0,1]^d$ where, for all $j \in \{1, \ldots, d\}$ and $K \geq 1$,

$$g_j^{(1,K)} = \mathrm{Id} / f_j^{(K)}, \tag{3.61}$$

$$g_j^{(k,K)} = \bigodot_{i=K-k+2}^{K} f_j^{(i)} \Bigg/ \bigodot_{i=K-k+1}^{K} f_j^{(i)}, \ k \in \{2, \ldots, K\}. \tag{3.62}$$

Conversely, each Liebscher copula (defined in Theorem 3.3.1) can be built iteratively from (3.57), (3.58).

Let us note that the iterative construction (3.57), (3.58) thus provides a way to build functions (3.61), (3.62) that automatically fulfill Liebscher's constraints (3.50) of Theorem 3.3.1. As a consequence, the construction (3.57), (3.58) also gives an iterative way to sample from a Liebscher copula (3.49), described in detailed in Algorithm 3.3.1 (see the proof of Lemma **??** in **??** for a theoretical justification).

Algorithm 3.3.1 Iterative sampling scheme for Liebscher copula (3.49)

Input $\left[f_j^{(k)}\right]_{k,j}, (C_k)_k$ ▷ functions in $\mathcal{F}$ appearing in (3.58), and copulas

$(X_1^{(1)}, \ldots, X_d^{(1)}) \sim C_1$

For $k = 2, \ldots, K$

$(Y_1, \ldots, Y_d) \sim C_k$ independently of $(X_1^{(k-1)}, \ldots, X_d^{(k-1)})$

For $j \in \{1, \ldots, d\}$

$X_j^{(k)} = \max\left(\left(f_j^{(k)}\right)^{-1}(X_j^{(k-1)}), \left(\mathrm{Id}/f_j^{(k)}\right)^{-1}(Y_j)\right)$

Output $\boldsymbol{X} = (X_1^{(K)}, \ldots, X_d^{(K)}) \sim \tilde{C}$

Example 3.3.3 (Power functions $f_j^{(k)}$)**.** *Let functions $f_j^{(k)}$ be power functions in the form of*

$$f_j^{(k)}(t) = t^{1-a_j^{(k)}}, \quad a_j^{(1)} = 1,\ a_j^{(k)} \in (0,1),\ \text{for all } k \geq 2, \tag{3.63}$$

for all $j \in \{1, \ldots, d\}$ and $t \in [0,1]$. From Proposition 3.3.5, $g_j^{(k)}(t) = g_j^{(k,K)}(t) = t^{p_j^{(k,K)}}$ with

$$\begin{cases} p_j^{(1,K)} = a_j^{(K)}, \\ p_j^{(k,K)} = a_j^{(K-k+1)} \displaystyle\prod_{i=K-k+2}^{K} (1 - a_j^{(i)}), \ \text{if } 2 \leq k \leq K, \end{cases} \tag{3.64}$$

for all $K \geq 1$ and $j \in \{1, \ldots, d\}$. Note that, by construction,

$$\sum_{k=1}^{K} p_j^{(k,K)} = 1,$$

for all $j \in \{1, \ldots, d\}$ and thus $\left(p_j^{(1,K)}, \ldots, p_j^{(K,K)}\right)$ can be interpreted as a discrete probability distribution on $\{1, \ldots, K\}$. Besides, let $\tilde{a}_j^{(k,K)} := a_j^{(K+1-k)}$ for all $k \in \{1, \ldots, K\}$. Equations (3.64) can be rewritten as

$$\begin{cases} p_j^{(1,K)} = \tilde{a}_j^{(1,K)}, \\ p_j^{(k,K)} = \tilde{a}_j^{(k,K)} \displaystyle\prod_{i=1}^{k-1} (1 - \tilde{a}_j^{(i,K)}), \ \text{if } 2 \leq k \leq K, \end{cases}$$

which corresponds to the so-called stick-breaking construction [300].

3.3.3 The comonotonic-based Liebscher copula

We analyze here in more details the Liebscher copula obtained by combining $K \geq 2$ comonotonic (also called Fréchet) copulas defined by $C(u,v) = \min(u,v)$. We here focus on the bivariate case $d = 2$, although some of the derivations carry over to the general d-dimensional case. We consider the specific case of Example 3.3.3, where the functions in Liebscher's construction are power functions, $g_j^{(k)}(t) = t^{p_j^{(k,K)}}$ with $p_j^{(k,K)} \in [0,1]$, $j \in \{1,2\}$ as in (3.64). Assuming that K is fixed and limiting ourselves

to $d = 2$, we denote for notation simplicity $p_k := p_1^{(k,K)}$ and $q_k := p_2^{(k,K)}$ for $k \in \{1, \ldots, K\}$. Recall that, in view of (3.50), $\sum_{k=1}^K p_k = \sum_{k=1}^K q_k = 1$. Under the above assumptions, the comonotonic-based Liebscher copula denoted by $\tilde{C}_{\mathrm{CL}}$ has the form

$$\tilde{C}_{\mathrm{CL}}(u,v) = \prod_{k=1}^{K} \min(u^{p_k}, v^{q_k}), \qquad (u,v) \in [0,1]^2. \tag{3.65}$$

In the particular case where $K = 2$, it is referred to as the BC2 copula by [222] and it is proved that any bivariate extreme-value copula with arbitrary discrete dependence measure can be represented as the geometric mean of BC2 copulas, which corresponds to the situation where K is even in (3.65). We also refer to [323] for further links with extreme-value theory.

Geometric description of $\tilde{C}_{\mathrm{CL}}$

For all $k \in \{1, \ldots, K\}$, introduce $r_k = p_k/q_k \in [0, \infty]$. For notation simplicity, we shall let $r_0 = 0$ and $r_{K+1} = \infty$. Since the above product (3.65) is commutative, one can assume without loss of generality that the sequence $(r_k)_{0 \le k \le K+1}$ is nondecreasing. The copula $\tilde{C}_{\mathrm{CL}}$ can be easily expressed on the partition of the unit square $[0,1]^2$ defined by the following moon shaped subsets (see the illustration in Fig. 3.7 with $K = 2$)

$$\mathcal{A}_k = \left\{(u,v) \in [0,1]^2 : u^{r_{k+1}} < v \le u^{r_k}\right\}, \quad k \in \{0, \ldots, K\}. \tag{3.66}$$

Proposition 3.3.6. *Let $\tilde{C}_{\mathrm{CL}}$ be the comonotonic-based Liebscher copula defined in (3.65). Then, for all $(u,v) \in [0,1]^2$,*

$$\tilde{C}_{\mathrm{CL}}(u,v) = \sum_{k=0}^{K} u^{1-\bar{p}_k} v^{\bar{q}_k} \mathbb{1}[(u,v) \in \mathcal{A}_k], \tag{3.67}$$

where $\bar{x}_k = x_1 + \ldots + x_k$, with the convention that $\bar{x}_0 = 0$. Moreover, the singular component of $\tilde{C}_{\mathrm{CL}}$ is

$$\tilde{S}_{\mathrm{CL}}(u,v) = \sum_{k=1}^{K} \min(p_k, q_k) \min(u, v^{1/r_k})^{\max(1, r_k)}. \tag{3.68}$$

The singular component $\tilde{S}_{\mathrm{CL}}$ and the absolute continuous component $\tilde{A}_{\mathrm{CL}} = \tilde{C}_{\mathrm{CL}} - \tilde{S}_{\mathrm{CL}}$ weights are $\sum_{k=1}^K \min(p_k, q_k)$ and $1 - \sum_{k=1}^K \min(p_k, q_k)$, respectively.

A key property of the comonotonic-based Liebscher copula (3.65) is the presence of multiple singular components lying on the curves $v = u^{r_k}$ with associated weights $\min(p_k, q_k)$, $k \in \{1, \ldots, K\}$. As an illustrative example, let us consider the bivariate comonotonic-based Liebscher copula defined with $p_1 = 1 - p_2 = 1/3$, $q_1 = 1 - q_2 = 3/4$,

$$\tilde{C}_{\mathrm{CL}}(u,v) = \min(u^{1/3}, v^{3/4}) \min(u^{2/3}, v^{1/4}), \qquad (u,v) \in [0,1]^2, \tag{3.69}$$

which entails $r_0 = 0, r_1 = 4/9, r_2 = 8/3$ and $r_3 = \infty$. The moon shaped subsets of the partition of the unit square $[0,1]^2$ are represented on Fig. 3.7, and the expressions

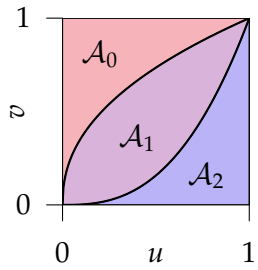


FIGURE 3.7: Partition of the unit square defined in (3.66), for the copula of Equation (3.69).

of $\tilde{C}_{\mathrm{CL}}$ and the singular component $\tilde{S}_{\mathrm{CL}}$ are as follows:

$$\tilde{C}_{\mathrm{CL}}(u,v) = \begin{cases} u & \text{on } \mathcal{A}_0, \\ u^{2/3}v^{3/4} & \text{on } \mathcal{A}_1, \\ v & \text{on } \mathcal{A}_2, \end{cases} \qquad \text{and} \qquad \tilde{S}_{\mathrm{CL}}(u,v) = \frac{1}{3}\min\left(u, v^{9/4}\right) + \frac{1}{4}\min\left(u^{8/3}, v\right).$$

Appropriate choices of pairs $(p_k, q_k)_{k=1,\dots,K}$ may lead to a number of singular components ranging from 0 to K. The independence copula (no singular component) is obtained for instance with $p_i = b_j = 1$ for a given pair (i,j), $i \neq j$, the comonotonic copula (one singular component) is obtained by choosing $p_k = q_k$, $\forall k \in \{1,\dots,K\}$, and a copula with exactly K singular components can be obtained provided that $0 < r_1 < r_2 < \cdots < r_K < \infty$. See Fig. 3.8 and Fig. 3.9 for illustrations. As a comparison, [88] copula given by $C(u,v) = (uv)^{1-\theta}\min(u,v)^{\theta}$, $\theta \in [0,1]$ is limited to a single singular component, necessarily on the diagonal $v = u$. Similarly, [225] copula is defined by $C(u,v) = \min(u^{1-\alpha}v, uv^{1-\beta})$, $(\alpha,\beta) \in [0,1]^2$ and has only one singular component located on the curve $v = u^{\alpha/\beta}$. The proposal by [201] based on singular mixture copulas includes comonotonic-based Liebscher copula (3.65) in the particular case when $K = 2$ but is limited to two singular components. Finally, Sibuya copulas [161] is a very general family of copulas: Let us point out that, in the bivariate case, a non-homogeneous Poisson Sibuya copula allows for only one singular component, this singular component being supported by a curve with very flexible shape (see Remark 4.2 in the previously referenced work for further details).

Dependence and association properties of $\tilde{C}_{\mathrm{CL}}$

We consider here several measures of dependence and measures of association between the components of the bivariate comonotonic-based Liebscher copula (3.65). Some of these measures are already dealt with in great generality in Section 3.3.2, while some others seem to be tractable only in the comonotonic-based Liebscher copula case: Blomqvist's medial correlation coefficient, Kendall's τ and Spearman's ρ.

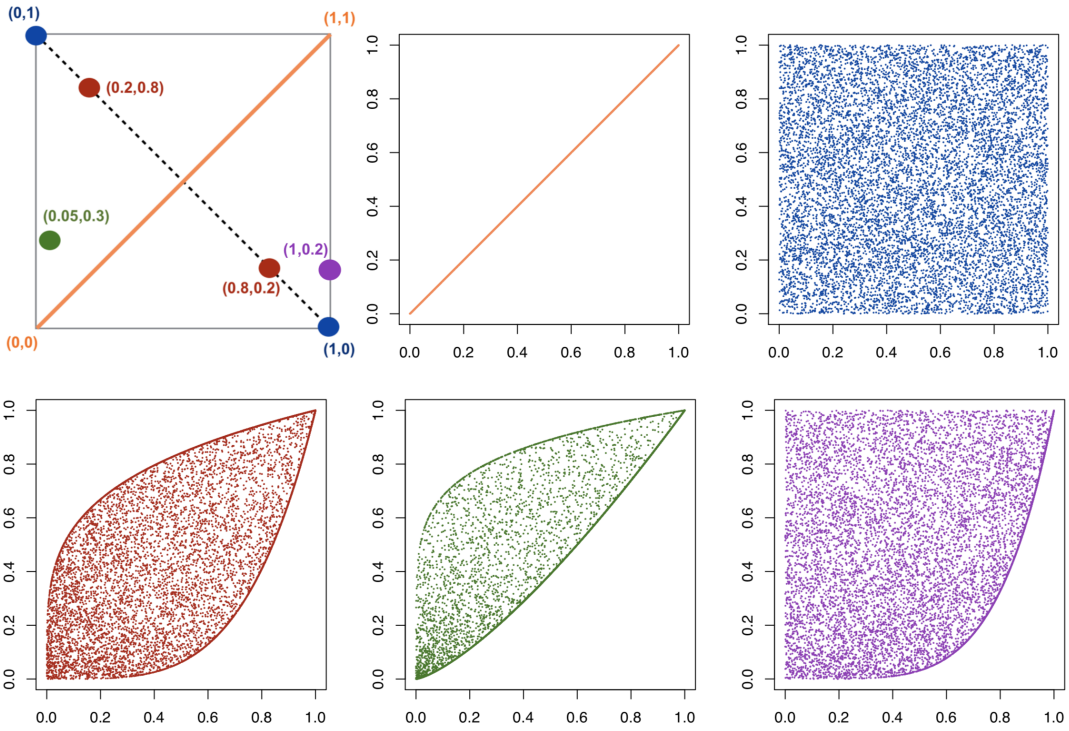


FIGURE 3.8: Top-left: representation of the $p \times q$ square unit space. Other five panels: scatter plots of $n = 10^4$ data points sampled from comonotonic-based Liebscher copula with $K = 2$. Choices for parameters (p, q) (such that $p_1 = p$, $p_2 = 1 - p$, $q_1 = q$, $q_2 = 1 - q$) are summarized on the top-left panel. Complete dependence (top-middle), complete independence (top-right), symmetric (bottom-left), asymmetric (bottom-middle), degenerate asymmetric (bottom-right).

Tail dependence Recall that for the comonotonic copula C_C, it holds $\Lambda_U(C_C; \cdot, \cdot) = \min(\cdot, \cdot)$. Then, Corollary 3.3.1 yields

$$\lambda_L(\tilde{C}_{CL}) = \prod_{k=1}^{K} \mathbb{1}(p_k = q_k), \qquad \lambda_U(\tilde{C}_{CL}) = \sum_{k=1}^{K} \min(p_k, q_k).$$

In other words, the lower tail dependence coefficient is non zero only in the case when $p_k = q_k$ for all $k \in \{1, \ldots, K\}$, where $\tilde{C}_{CL}$ boils down to the comonotonic copula, while the upper tail dependence coefficient coincides with the weight of the singular component $\tilde{S}_{CL}$ (see Proposition 3.3.6). These results were also established in [222, Lemma 3], in the particular case where $K = 2$.

Dependence It is well-known that the comonotonic copula C_F fulfills the following positive dependence properties defined in Section 3.3.2, namely it is TP_2, PQD, LTD and SI [244]. According to Proposition 3.3.2, we can thus conclude that the comonotonic-based Liebscher copula $\tilde{C}_{CL}$ also satisfies these positive dependence properties.

Stability properties It is easily seen that comonotonic-based Liebscher copula (3.65) is max-stable. Proposition 3.3.4 thus entails that this copula is stable with respect

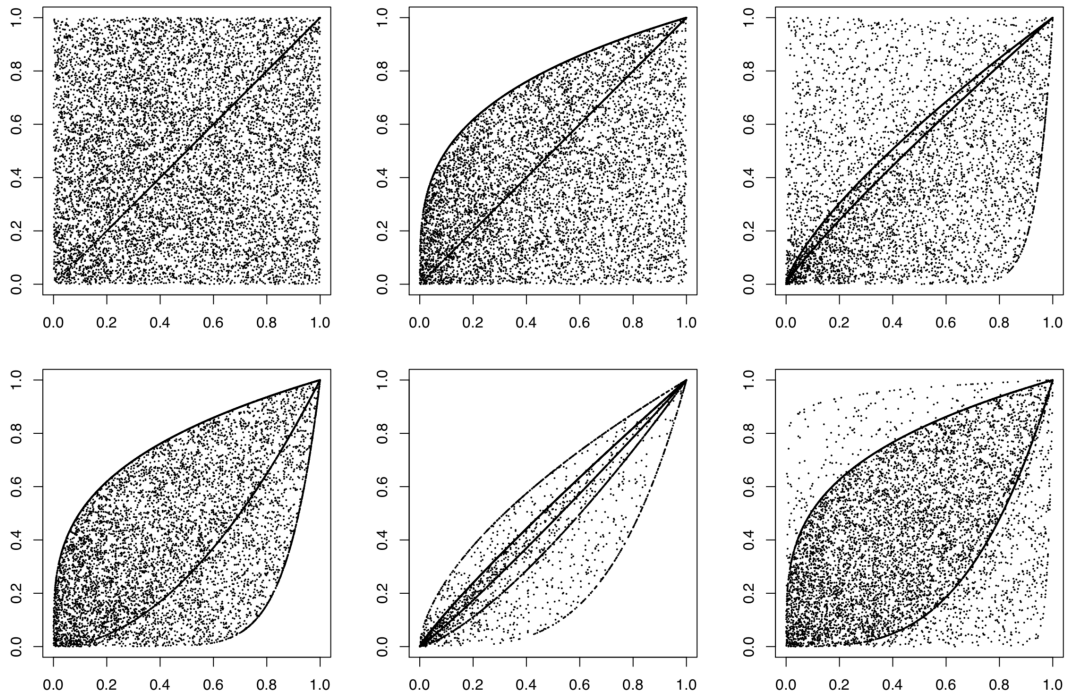


FIGURE 3.9: Scatter plots of $n = 10^4$ data points sampled from comonotonic-based Liebscher copula with $K > 2$. Top-left: $K = 3$, $r_k \in \{0, 1, \infty\}$; Top-middle: $K = 3$, $r_k \in \{0.3, 1, \infty\}$; Top-right: $K = 4$, $r_k \in \{0, 0.7, 0.9, 17\}$; Bottom-left: $K = 4$, $r_k \in \{0.3, 1.9, 8.3, 8.3\}$; Bottom-middle: $K = 5$, $r_k \in \{0.6, 0.9, 1.1, 1.4, 3.6\}$; Bottom-right: $K = 6$, $r_k \in \{0.04, 0.3, 2.8, 3.3, 4.2, 58\}$.

to Liebscher's construction. Another consequence is that comonotonic-based Liebscher construction (3.65) can be interpreted as a possible cdf for modelling bivariate maxima.

Dependence coefficients The β-Blomqvist's medial correlation coefficient ([244], Paragraph 5.1.4) defined by $\beta(C) = 4C(\frac{1}{2}, \frac{1}{2}) - 1$, Kendall's τ ([244], Paragraph 5.1.1) and Spearman's ρ ([244], Paragraph 5.1.2) defined by

$$\tau(C) = 4 \int_{[0,1]^2} C(u, v)\, \mathrm{d}C(u, v) - 1 \quad \text{and} \quad \rho(C) = 12 \int_{[0,1]^2} C(u, v)\, \mathrm{d}u\mathrm{d}v - 3$$

are provided in next proposition.

Proposition 3.3.7. *Blomqvist's medial correlation coefficient, Kendall's τ and Spearman's ρ for the comonotonic-based Liebscher copula (3.65) are respectively given by*

$$\beta(\tilde{C}_{\mathrm{CL}}) = 2^{\sum_{k=1}^{K} \min(p_k, q_k)} - 1,$$

$$\tau(\tilde{C}_{\mathrm{CL}}) = 1 - \sum_{k=1}^{K-1} \frac{(1 - \bar{p}_k)\bar{q}_k(r_{k+1} - r_k)}{(\bar{q}_k r_k + (1 - \bar{p}_k))(\bar{q}_k r_{k+1} + (1 - \bar{p}_k))},$$

$$\rho(\tilde{C}_{\mathrm{CL}}) = \frac{12(1 + r_1 + r_1 r_K)}{(2 + r_1)(1 + 2r_K)} - 3 + \sum_{k=1}^{K-1} \frac{r_{k+1} - r_k}{((1 + \bar{q}_k)r_k + (2 - \bar{p}_k))((1 + \bar{q}_k)r_{k+1} + (2 - \bar{p}_k))},$$

where $\bar{x}_k = x_1 + \ldots + x_k$.

It appears that β-Blomqvist medial correlation coefficient is closely related to the upper tail dependence coefficient: $\beta(\tilde{C}_{\mathrm{CL}}) = 2^{\lambda_U(\tilde{C}_{\mathrm{CL}})} - 1$. Besides, in the particular case where $K = 2$, these results coincide the ones of Lemma 2 of [222]: the Kendall's τ can be simplified as $\tau(\tilde{C}_{\mathrm{CL}}) = p_1 + q_2 = \lambda_U(\tilde{C}_{\mathrm{CL}})$. No similar simplification seems to be possible for Spearman's ρ.

For the special case of copula (3.69), we have $\lambda_L(\tilde{C}_{\mathrm{CL}}) = 0$, $\lambda_U(\tilde{C}_{\mathrm{CL}}) = \tau(\tilde{C}_{\mathrm{CL}}) = p_1 + q_2 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} \approx 0.583$, $\beta(\tilde{C}_{\mathrm{CL}}) = 2^{7/12} - 1 \approx 0.498$ and $\rho(\tilde{C}_{\mathrm{CL}}) \approx 0.298$.

Iterative construction for $\tilde{C}_{\mathrm{CL}}$

Algorithm 3.3.1 can be simplified when specified to comonotonic-based Liebscher setting since: (i) sampling from the comonotonic copula is straightforward and only requires sampling from the uniform distribution $\mathcal{U}(0,1)$, and (ii) power functions benefit from an explicit inverse. The specific sampling procedure for this construction is described in detail as Algorithm 3.3.2.

Algorithm 3.3.2 Iterative sampling scheme for comonotonic-based Liebscher copula (3.65)

Input $\left[a_j^{(k)}\right]_{(k,j)}$ ▷ exponents of power functions $f_j^{(k)}$ $X_j^{(1)} \sim \mathcal{U}(0,1)$ for each $j = 1, \dots, d$
For $k = 2, \dots, K$
Sample $Y \sim \mathcal{U}(0,1)$, independently of $X_1^{(k-1)}, \dots, X_d^{(k-1)}$
For $j \in \{1, \dots, d\}$
Compute $X_j^{(k)} = \max\left(\left(X_j^{(k-1)}\right)^{\frac{1}{1-a_j^{(k)}}}, Y^{\frac{1}{a_j^{(k)}}}\right)$
Output $\boldsymbol{X} = (X_1^{(K)}, \dots, X_d^{(K)}) \sim \tilde{C}_{\mathrm{CL}}$

3.3.4 Bayesian inference

In this section, we provide a simple strategy to make Bayesian inference on any Liebscher copula based on an Approximate Bayesian computation algorithm (ABC, see for instance [224, 285, 184] for reviews). ABC is a "likelihood-free" method usually employed for inference of models with intractable likelihood: it enables to perform approximate Bayesian analysis on any statistical model from which it is possible to sample new data, without the need to explicitly evaluate the likelihood function.

Let $\boldsymbol{X}_{\mathrm{obs}} = \{\boldsymbol{X}_{\mathrm{obs},1}, \dots, \boldsymbol{X}_{\mathrm{obs},n}\}$ be the observed data, where $\boldsymbol{X}_{\mathrm{obs},i} = (X_{\mathrm{obs},i,1}, \dots, X_{\mathrm{obs},i,d})$, $i \in \{1, \dots, n\}$, and assume that the statistical model for $\boldsymbol{X}_{\mathrm{obs}}$ is described by a likelihood function $\mathcal{L}_\theta$ with parameter θ which is to be inferred. The basic scheme of one step of ABC is the following:

(i) Sample θ from the prior distribution $\pi(\theta)$;

(ii) Given θ, sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ from $\mathcal{L}_\theta$, and set $\boldsymbol{X} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$;

(iii) If $\boldsymbol{X}$ is too different from $\boldsymbol{X}_{\mathrm{obs}}$, discard θ, otherwise, keep θ.

The outcome of the ABC algorithm is a sample of values of the parameter θ approximately distributed according to its posterior distribution. The basic (rejection)

ABC approach in point 3. amounts to *a priori* specifying a tolerance level $\epsilon > 0$, and then keeping θ if $d(\boldsymbol{X}, \boldsymbol{X}_{\text{obs}}) < \epsilon$ for some distance $d(\cdot, \cdot)$ between samples. Another common approach employed in this section consists in selecting the tolerance level ϵ as a fixed quantile of the distances $d(\boldsymbol{X}, \boldsymbol{X}_{\text{obs}})$. More specifically, Steps 1. to 3. are repeated M' times, out of which M are retained, yielding a quantile of order M/M' [285]. In other words, the M retained parameters are those associated with the smallest values of the distance $d(\boldsymbol{X}, \boldsymbol{X}_{\text{obs}})$. In this section, we choose as distance between samples the Hilbert distance introduced by [52], henceforth denoted by $d_H(\cdot, \cdot)$. The Hilbert distance is an approximation of the Wasserstein distance between empirical probability distributions which preserves the desirable properties of the latter in the context of ABC, while being considerably faster to compute in multivariate data settings. More precisely, given two samples, $y_{1:n}$ and $z_{1:n}$, the Hilbert distance of order 1 associated with the Euclidean distance (henceforth simply referred to as the Hilbert distance) is defined as $d_H(y_{1:n}, z_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} ||y_i - z_{\sigma(i)}||$, where $\sigma(i) = \sigma_z \circ \sigma_y^{-1}(i)$ for all $i = 1, \ldots, n$, and σ_y and σ_z are the permutations obtained by mapping the vectors $y_{1:n}$ and $z_{1:n}$ through their projection via the Hilbert space-filling curve [294] and sorting the resulting vectors in increasing order (see [52] for details).

The choice for ABC is motivated by two main reasons. First, it is nontrivial in general to derive the likelihood of copulas, especially for Liebscher copulas which involve differentiating a product of K terms. All the more, the specific case of comonotonic-based Liebscher copulas induces up to K singular components which precludes a general evaluation of the likelihood. Second, sampling new data (step 2. above) from a Liebscher copula is straightforward and fast thanks to the iterative procedure of Algorithm 3.3.1 (Section 3.3.2).

Section 3.3.4 introduces the ABC procedure in the case of the Liebscher copula (3.49) and describes the prior distributions on the model parameters. The methodology is then illustrated on two data generating distributions: Section 3.3.4 focuses on the well-specified setting where the data are sampled from the comonotonic-based Liebscher copula; we show that the estimation procedure performs well in this case. Then, Section 3.3.4 investigates the misspecified setting where the data are sampled from a noisy version of the model; we show that the estimation procedure still performs well, but the estimation accuracy may deteriorate for too large values of the noise. Finally, Section 3.3.4 compares our proposed ABC approach to a likelihood-based technique.

ABC inference for Liebscher copulas

The description of the ABC procedure is first completed by specifying the prior distributions on the model parameters. For simplicity, we here focus on the case of the d-dimensional Liebscher copula with the functions $f_j^{(k)}(\cdot)$, $k \in \{1, \ldots, K\}$, $j \in \{1, \ldots, d\}$ of Algorithm 3.3.1 chosen as the power functions (3.63) introduced in Example 3.3.3. The parameters of the $f_j^{(k)}(\cdot)$ functions are collected in a $K \times d$ matrix $A = [a_j^{(k)}]_{k,j}$, where $a^{(1)} = (a_1^{(1)}, \ldots, a_d^{(1)}) = (1, \ldots, 1)$. Since the $(K-1)d$ free parameters are constrained to $a_j^{(k)} \in (0, 1)$ for $2 \leq k \leq K$, we simply choose, by symmetry,

independent and uniform distributions $a_j^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$. More flexible distributions like the Beta distribution could be thought of in order to reflect some prior knowledge on these parameters. Additionally, note that different functions $f_j^{(k)}(\cdot)$ would simply require setting prior distributions adapted to the parameters used.

The number of iterative steps K is also considered as a parameter of the model. Independently of parameters A, K is assigned a Zipf distribution, $K \sim \text{Zipf}(\xi) + 1$, for $\xi > 1$. Such a distribution is supported on integers $k \geq 2$ and has probability mass function $\mathbb{P}(K = k)$ proportional to $(k-1)^{-\xi}$. We further choose the parameter ξ to be equal to 2, which insures that 90% of the prior mass for K is supported on most realistic values $2 \leq K \leq 6$. This can be changed depending on applications at hand. Another option, useful in case where some prior information is available on K, is to adopt a Binomial distribution (translated, such that $K \geq 2$). The choice of the two hyper-parameters of the Binomial density could then be set as a function of prior knowledge, such as prior mode and confidence, that one may be able to elicit thanks to expert knowledge or previous studies.

We are now ready to state the main ABC inference procedure as Algorithm 3.3.3.

Algorithm 3.3.3 ABC inference for Liebscher copulas

Input $\boldsymbol{X}_{\text{obs}}, M', M, (C_k)_k$.
For $s = 1, \ldots, M'$
$K^{(s)} \sim \text{Zipf}(\xi) + 1$ ▷ sample number of iterations in construction (3.49)
For $j \in \{1, \ldots, d\}$ *and* $k \in \{2, \ldots, K^{(s)}\}$
$a_j^{(1)} = 1$
$a_j^{(k)} \sim \mathcal{U}(0,1)$ ▷ sample copula parameters
$A^{(s)} = [a_j^{(k)}]_{j,k}$ ▷ set parameters for Liebscher's construction
$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \tilde{C}_{\text{L}}^{(K^{(s)})}$ ▷ sample data using Algo. 3.3.1 with power functions and $A = A^{(s)}$
$\boldsymbol{X}^{(s)} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ ▷ set synthetic data
$d_H^{(s)} = d_H(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}^{(s)})$ ▷ compute Hilbert distances
Compute d^*: the quantile of order M/M' of the distances $[d_H^{(s)}]_{s=1}^{M'}$
Output $\{(\boldsymbol{X}^{(m)}, A^{(m)}, K^{(m)}) : \ d_H^{(m)} < d^*\}_{m=1}^{M}$ ▷ return M parameters with smallest d_H from $\boldsymbol{X}_{\text{obs}}$

In general, the sequence of copulas $(C_k)_k$ depends on some sequence of parameters $(\gamma_k)_k$. In such a case, Algorithm 3.3.3 can be easily amended by adding a step consisting in sampling γ_k parameters from some prior distribution to be set based on available prior information or expert knowledge. In the following section, we focus on the case of comonotonic-based Liebscher copulas, which do not depend on any additional parameter. Thus, the sampling step for new data $\boldsymbol{X}^{(s)}$ in Algorithm 3.3.3 is performed with the iterative construction of Algorithm 3.3.2 tailored to comonotonic-based Liebscher copulas.

Numerical illustrations with comonotonic-based Liebscher copulas

This section provides two illustrations of the inferential procedures described so far. The first investigates a setting where data are sampled from the comonotonic-based Liebscher model, while the second is concerned with observations from a noisy version of it. The code is implemented in R using the `copula` package [342] and `winference` package [51] for the Hilbert distance implementation [52].

Well-specified setting: data from comonotonic-based Liebscher copula We generate $n = 500$ data points from a 2-dimensional comonotonic-based Liebscher copula (3.65) with $d = 2$, varying values of $K \geq 2$ and of the parameters in the matrix A, using Algorithm 3.3.2. The estimation is then performed with the ABC procedure summarized in Algorithm 3.3.3.

Our method provides the full (approximate) posterior distribution of the parameters of interest, making possible to select any strategy to summarize them, possibly driven by the application at hand. One can for instance compute the posterior distribution of the Spearman's ρ, and, thanks to the retained samples, any other quantity of interest. Here, the performance of the estimation procedure is assessed basing on the following three summary statistics: (i) Kendall's distribution function $\mathcal{K}(t) = \Pr_C(C(U,V) < t)$, (ii) Spearman's ρ index of association, introduced in Section 3.3.3, and for which an explicit closed form is obtained for comonotonic-based Liebscher copula in Proposition 3.3.7, and (iii) an asymmetry measure, since it is a central motivation of the present work. More specifically, the Cramér-von Mises test statistics $\mathrm{E}_C[(C(U,V) - C(V,U))^2]$ defined in [128] has been selected since it emerged as a powerful statistic to test the symmetry of a copula. Following the strategy of [128], the approximate p-values associated with the symmetry test, performed both on the observed sample $\boldsymbol{X}_{\text{obs}}$ and on the retained samples $\boldsymbol{X}_m, m \in \{1, \ldots, M\}$, are computed on the basis of 250 bootstrap replicates.

The results obtained on a single simulation experiment are displayed on Fig. 3.10, where $n = 500$ data points were simulated from the comonotonic-based Liebscher copula (3.65) with $d = 2$ and $K = 3$ (top-left panel). The number of ABC iterations was set to $M' = 10^4$, of which $M = 300$ were retained (resulting in a quantile of order 3%). The empirical Kendall's distribution functions of the observed and retained samples are compared on the top-right panel; The posterior distribution of ρ is compared to the empirical Spearman's ρ of the observed sample on the bottom-left panel; Finally, the posterior distribution of the approximate p-values is displayed on the bottom-right panel. Let us highlight that the estimating procedure provides distributions around the true values in the three considered cases.

We then vary the generating number of iterative steps K and compute the average relative errors $\eta_{\mathcal{K}}$ and η_ρ for $\mathcal{K}$ and ρ between the values computed on the observed sample and on the M samples retained by ABC:

$$\eta_{\mathcal{K}} = \frac{1}{M}\sum_{m=1}^{M} \frac{\|\hat{\mathcal{K}}_{\text{obs}} - \hat{\mathcal{K}}_m\|_1}{\|\hat{\mathcal{K}}_{\text{obs}}\|_1}, \quad \text{and} \quad \eta_\rho = \frac{1}{M}\sum_{m=1}^{M} \frac{|\hat{\rho}_{\text{obs}} - \hat{\rho}_m|}{|\hat{\rho}_{\text{obs}}|}, \tag{3.70}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm. In order to take care of the randomness involved in sampling the parameters in the matrix A, the previous procedure has been replicated 20 times based on 20 independent data samples repetitions. The average relative errors $\eta_{\mathcal{K}}$ and η_ρ in (3.70) were therefore averaged over the 20 independent samples, and reported as $\bar{\eta}_{\mathcal{K}}$ and $\bar{\eta}_\rho$ in Table 3.3 (first two rows), along with standard deviations in parentheses. As for the asymmetry test, we computed for each of the 20 data replications the fraction of times (out of M) that the same decision is taken ('reject' vs 'do not reject') at the 5% level, based on the approximate p-values computed on $\boldsymbol{X}_m$ and $\boldsymbol{X}_{\text{obs}}$. The obtained values were averaged over the 20 independent data samples repetitions and reported in the third row of Table 3.3 as $\bar{\text{f}}_{\text{test}}$.

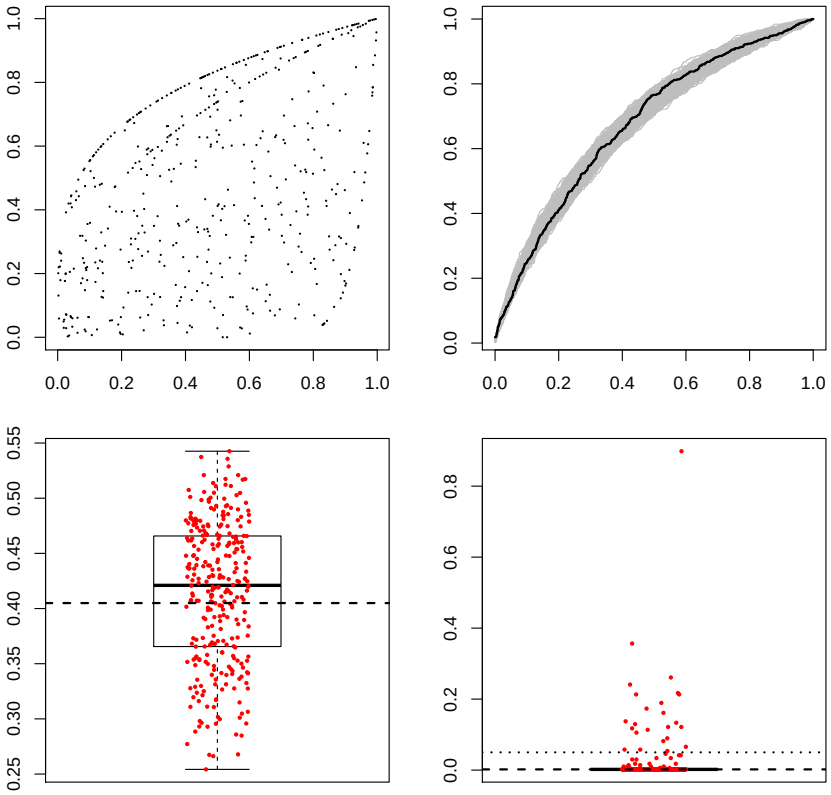


FIGURE 3.10: Results from a simulation experiment with $K = 3$. Top-left: $n = 500$ data points simulated from a comonotonic-based Liebscher copula with $d = 2$, $K = 3$ and $r_k \in \{0.26, 0.44, 17.32\}$. Top-right: empirical Kendall's distribution function $\hat{\mathcal{K}}$ of the observed (black) and retained (gray) samples; Bottom-left: boxplot of the posterior distribution of ρ (the dashed line corresponds to the empirical Spearman's ρ of the observed sample); Bottom-right: boxplot of the posterior distribution of the approximate p-values (the dashed line corresponds to $\boldsymbol{X}_{\text{obs}}$, and the dotted line is the 5% threshold).

TABLE 3.3: First two rows: average relative errors (3.70) for Kendall's distribution function and Spearman's ρ between the observed sample and the samples retained by the ABC procedure, for varying K (columns). Third row: fraction of times that the same decision is taken ('reject' vs 'do not reject', at the 5% level) based on $\boldsymbol{X}_m$ and $\boldsymbol{X}_{\text{obs}}$. The results are averaged over 20 independent repetitions. Standard deviations in parentheses. All values are in %.

K	**2**	**3**	**4**	**5**
$\bar{\eta}_{\mathcal{K}}$	1.99 (0.16)	2.32 (0.49)	2.38 (0.40)	2.37 (0.56)
$\bar{\eta}_{\rho}$	5.44 (3.98)	8.16 (5.62)	6.34 (3.74)	9.66 (4.12)
$\bar{\text{f}}_{\text{test}}$	16.8 (18.1)	19.4 (22.0)	9.2 (13.2)	13.3 (18.1)

Table 3.3 suggests a general trend: the larger K is, the more difficult the estimation

is. However, the estimation procedure yields satisfactory results for all cases considered.

Misspecified setting: data from a noisy comonotonic-based Liebscher copula In this section, we generate data from a noisy version of the comonotonic-based Liebscher copula, and demonstrate that our inference procedure works well even if the data are not sampled from the exact model (so-called misspecified setting). In order to sample data from such a noisy model, a slightly changed version of Algorithm 3.3.2 is used in which the parameters A are not fixed. Instead, they are sampled from a beta distribution with given variance σ_a^2 (interpreted as the error variance) around some fixed value corresponding to the zero noise version. The latter is illustrated on Fig. 3.11: a sample of $n = 10^3$ data points from a 2-dimensional comonotonic-based Liebscher copula is depicted on the top-left panel with $K = 2$ iterative steps, and with the two parameters of the power functions set to $a_1^{(2)} = 0.4$ and $a_2^{(2)} = 0.8$ (recall that $a_1^{(1)} = a_2^{(1)} = 1$). The remaining five panels correspond to samples from comonotonic-based Liebscher copula with increasing noise variance $\sigma_a^2 = 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$.

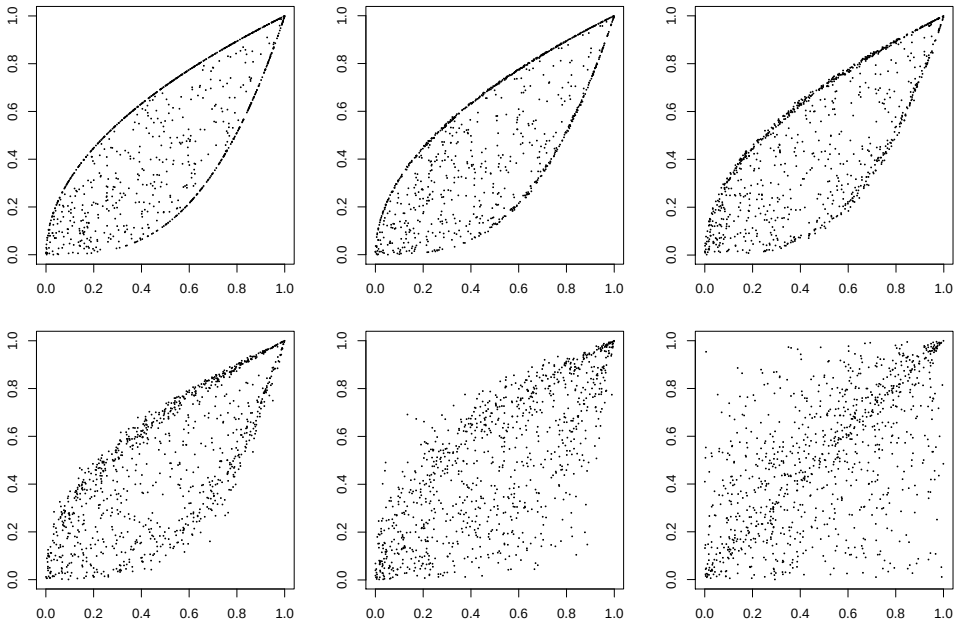


FIGURE 3.11: Samples from the exact comonotonic-based Liebscher copula with $d = K = 2$, $a_1^{(2)} = 0.4$, $a_2^{(2)} = 0.8$ (top-left panel) and from five noisy versions of it, where, for each data point $i \in \{1, \ldots, n\}$, $a_{i1}^{(2)} \sim \mathcal{B}(\alpha_1, \beta_1)$ and $a_{i2}^{(2)} \sim \mathcal{B}(\alpha_2, \beta_2)$ such that $\mathrm{E}(a_{i1}^{(2)}) = 0.4$, $\mathrm{E}(a_{i2}^{(2)}) = 0.8$, and increasing noise variance $\sigma_a^2 = \mathrm{var}(a_{i1}^{(2)}) = \mathrm{var}(a_{i2}^{(2)}) = 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ (other panels). Sample size $n = 10^3$.

The inference results obtained from Algorithm 3.3.3 (ran with $M' = 10^4$ ABC iterations of which $M = 300$ were retained) are reported in Table 3.4. Unsurprisingly, it appears that the larger the noise variance is, the more difficult is the estimation. Again, the estimation procedure yields good results for all cases considered. Let

us note that the results reported in Table 3.4 are not averaged over 20 independent replications like in the previous section. The reason is that, here, the interest is in illustrating how the procedure deteriorates with the increasing noise in the observed data. Therefore, it is sufficient to illustrate the results of the analysis on a single dataset.

TABLE 3.4: First two rows: average relative errors (3.70) for Kendall's distribution function and Spearman's ρ between the observed sample and the samples retained by the ABC procedure, for growing noise (columns). Third row: fraction of times that the same decision is taken at the 5% level based on $\boldsymbol{X}_m$ and $\boldsymbol{X}_{\text{obs}}$. All values are in %.

σ_a^2	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
$\eta_{\mathcal{K}}$	1.7	1.68	1.77	1.68	2.03	2.81
η_ρ	4.07	3.77	4.66	4.30	4.88	13.00
f_{test}	9.00	1.33	15.7	3.33	2.33	7.00

Comparison of ABC with likelihood-based estimation

In this section, the inference based on our ABC procedure is compared with the likelihood-based method provided in the `copula` R package [342].

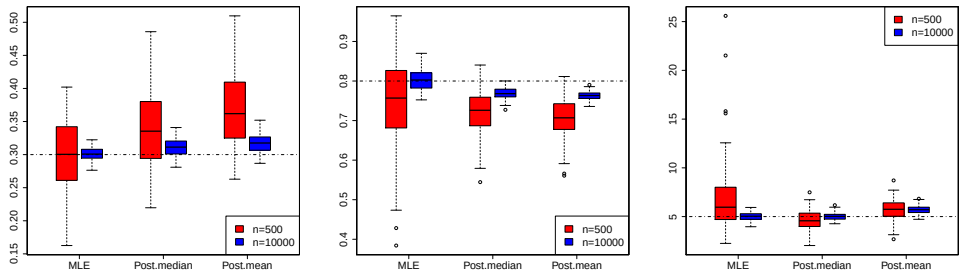
As a matter of fact, in some specific cases it is possible to derive the density of Liebscher copulas, and hence, to perform likelihood-based inference on them. We work here with the following bivariate Liebscher copula obtained by combining Clayton copula $C_1(u,v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}$ with the independence copula $C_2(u,v) = uv$, and by using power functions (Example 3.3.3),

$$C(u,v) = C_1(u^p, v^q) C_2(u^{1-p}, v^{1-q}). \tag{3.71}$$

In this illustration we set $\theta = 5$, $p = 0.3$, and $q = 0.8$, and estimate these parameters on 100 data sets independently sampled from copula (3.71), with both our ABC procedure and the optimization procedure implemented in the `fitCopula` function of the `copula` package. The above procedure is repeated for samples of size $n = 500$ and $n = 10\,000$. In order to compare the results of our Bayesian procedure with the likelihood-based one, which provides only point estimates (maximum likelihood estimator, MLE), the posterior distribution of the model parameters is summarized into two point estimates: the posterior mean and the posterior median.

The results are plotted in Fig. 3.12, where each boxplot is made out of the 100 estimated values. In each panel, three quantities are plotted: `MLE`, which refers to the likelihood-based estimation, `Post.median`, which refers to the posterior median ABC estimation, and `Post.mean`, which refers to the posterior mean ABC estimation. It appears from the three plots that, as expected, the MLE is asymptotically unbiased. This can be seen by noticing the convergence to the true value as n increases. On the other hand, both the mean and the median of the posterior distribution obtained with our ABC procedure show some bias. Unexpectedly, the results from MLE show a larger variance than ABC in some cases, particularly when $n = 500$. This could be

due to difficulties in the optimisation procedure of the fitCopula R function which, for small sample sizes, could have problems converging.



(A) Parameter p (first coordinate)

(B) Parameter q (second coordinate)

(C) Parameter θ (Clayton copula)

FIGURE 3.12: Results from the comparison. (A): boxplot of the parameter p; (B): boxplot of the parameter q; (C): boxplot of the parameter θ. Red corresponds to samples of size $n = 500$, blue to samples of size $n = 10\,000$. The horizontal dotted line represents the true value.

3.3.5 Discussion

In this section, we have studied the class of asymmetric copulas first introduced by [187], and developed in its general form by [207]. Some new theoretical properties of these copulas were provived, including novel closed form expressions for its tail dependence coefficients, thus complementing the partial results of [207] and [208]. An iterative procedure is also introduced to flexibly sample from these copulas, which makes it easy to apply an Approximate Bayesian computation procedure to make inference on them.

New Perspectives and Future Research

I conclude with some future research directions.

Perspectives with applications in Ecology

I plan to work on methodological developments in Ecology with Wilfried Thuiller and students Daria Bystrova and Giovanni Poggiato. The modelling of species distribution plays an important role in both theoretical and applied ecology: given a set of species occurrence, the aim is to infer its spatial distribution over a given territory. Joint Species Distribution Models (JSDM) aim to model and predict the distribution of species jointly at each location, to capture both abiotic and biotic dependencies. However, JSDMs have several limitations. First, they provide estimates for residual correlations between the species, and these residual correlations are interpreted as species interactions. However, this assumption is not true in general. An objective is to evaluate the performance of JSDMs for detecting the signal of biotic interactions using abundance data. A possible direction is the study of conditional prediction and its performance regarding different types of interactions, including missing data imputations. A second problem of JSDMs is the curse of dimensionality as models incorporate a large number of parameters [317]. There is a need for new algorithms for dimension reduction and variable selection. The data related imperfections such as sparse information in space and time measurements, and different observational scales would be addressed with machine learning approaches adapted to an ecology framework. For instance, Bayesian nonparametric priors, such as Dirichlet process, Dependent Dirichlet process and Hierarchical Dirichlet process, could be used to address different data-related questions and scenarios. The high-dimension problems would be addressed with the use of shrinkage priors [55].

Perspectives with deep learning

Although deep learning provides state-of-the-art algorithm for prediction in almost any application field, a major essentially unsolved problem is how to handle uncertainty quantification — all the more when applications are consequential as is often the case today. If this could in principle be solved by the posterior distribution of Bayesian neural networks, there remain many open questions and research directions. One is to design scalable sampling algorithms. This is made challenging due to the very high dimensionality of the parameter space, up to several millions in modern deep neural networks. Recent proposals in this direction [167] suggest to

focus the inferential procedure on low-dimensional subspaces of parameter space, for instance on the first principal components of the stochastic gradient descent trajectory. A second open question is the assessment of credibility sets derived from the posterior distribution of Bayesian neural networks. A theoretical framework for such evaluation is known as the Bernstein–von Misses (BvM) theorem which implies, for regular parametric models, that the posterior distribution of the parameter centered at the maximum likelihood estimator (MLE) converges to a normal distribution given as the limit of the renormalized MLE. The main thrust of this theorem is to show that asymptotically, a Bayesian posterior distribution and (frequentist) sampling probabilities agree, hence the posterior can be used to build approximate confidence regions. Refer to [310] for details and discussions in the Bayesian nonparametric setting. Although Bayesian neural networks are parametric objects, the dimension of their parameter space is often huge. Also, recent results on approximation theory [296, 46] suggest how their size (in terms of depth) should be scaled to the data size for optimal approximation properties. A major complication for deriving a BvM theorem for deep neural networks is the complexity of their likelihood function.

Index of Collaborators

Remark: **Bold** denotes co-authors.

Publication List

All the papers are available on my website : https://www.julyanarbel.com/publications

Articles in peer-reviewed journals

[A1] **Julyan Arbel** and Stefano Favaro. Approximating predictive probabilities of Gibbs-type priors. **Sankhyā, forthcoming**, 2019.

[A2] **Julyan Arbel**, Olivier Marchal, and Hien D Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. **ESAIM: Probability & Statistics, forthcoming**, 2019.

[A3] **Julyan Arbel**, Marta Crispino, and Stéphane Girard. Dependence properties and Bayesian inference for asymmetric multivariate copulas. **Journal of Multivariate Analysis**, 174, 2019.

[A4] Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and **Julyan Arbel**. Understanding Priors in Bayesian Neural Networks at the Unit Level. **ICML**, 2019.

[A5] Caroline Lawless and **Julyan Arbel**. A simple proof of Pitman–Yor's Chinese restaurant process from its stick-breaking representation. **Dependence Modeling**, 7, 2019.

[A6] **Julyan Arbel**, Pierpaolo De Blasi, and Igor Prünster. Stochastic approximations to the Pitman–Yor process. **Bayesian Analysis**, 14(3):753–771, 2019.

[A7] Olivier Marchal and **Julyan Arbel**. On the sub-Gaussianity of the Beta and Dirichlet distributions. **Electronic Communications in Probability**, 22:1–14, 2017.

[A8] **Julyan Arbel** and Igor Prünster. A moment-matching Ferguson & Klass algorithm. **Statistics and Computing**, 27(1):3–17, 2017.

[A9] **Julyan Arbel**, Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. **Statistica Sinica**, 27:839–858, 2017.

[A10] **Julyan Arbel**, Kerrie Mengersen, and Judith Rousseau. Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. **Annals of Applied Statistics**, 10(3):1496–1516, 2016.

[A11] **Julyan Arbel** and Vianney Costemalle. Estimation of immigration flows : reconciling two sources by a Bayesian approach (in French). **Économie et Statistique**, 483-484-485:121–149, 2016.

[A12] **Julyan Arbel**, Antonio Lijoi, and Bernardo Nipoti. Full Bayesian inference with hazard mixture models. **Computational Statistics & Data Analysis**, 93:359–372, 2016.

[A13] **Julyan Arbel**, Kerrie Mengersen, Ben Raymond, Tristrom Winsley, and Catherine King. Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of Antarctic microbial communities to fuel contaminated soil. **Ecology and Evolution**, 5(13):2633–2645, 2015.

[A14] **Julyan Arbel**, Ghislaine Gayraud, and Judith Rousseau. Bayesian optimal adaptive estimation using a sieve prior. **Scandinavian Journal of Statistics**, 40(3):549–570, 2013.

Book (as an editor)

[B1] Didier Fraix-Burnet, Stéphane Girard, **Julyan Arbel**, and Jean-Baptiste Marquette, editors. *Statistics for Astrophysics: Bayesian Methodology*. EDP Sciences, 2018.

Book Chapters

[C1] **Julyan Arbel**. ***Clustering Milky Way's Globulars: a Bayesian Nonparametric Approach***, chapter in Statistics for Astrophysics: Bayesian Methodology. EDP Sciences. Editors: Didier Fraix-Burnet, Stéphane Girard, Julyan Arbel and Jean-Baptiste Marquette, 2018.

[C2] Kerrie Mengersen, Clair Alston, **Julyan Arbel**, and Earl Duncan. ***Applications in Industry***, chapter in Handbook of mixture analysis. CRC Press, Editors: Gilles Celeux, Sylvia Früwirth-Schnatter, and Christian P. Robert, 2018.

[C3] Guillaume Kon Kam King, **Julyan Arbel**, and Igor Prünster. ***Bayesian Statistics in Action***, chapter A Bayesian nonparametric approach to ecological risk assessment, pages 151–159. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017.

[C4] **Julyan Arbel** and Igor Prünster. ***Bayesian Statistics in Action***, chapter On the truncation error of a superposed gamma process, pages 11–19. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017.

[C5] **Julyan Arbel**, Antonio Lijoi, and Bernardo Nipoti. ***Bayesian Statistics from Methods to Models and Applications***, chapter Bayesian Survival Model based on Moment Characterization, pages 3–14. Springer Proceedings in Mathematics & Statistics, Volume 126. Springer International Publishing, Editors: Sylvia Frühwirth-Schnatter et al., 2015.

Discussions

[D1] **Julyan Arbel**, Riccardo Corradin, and Michal Łewandowski. Discussion of "Bayesian Cluster Analysis: Point Estimation and Credible Balls", by Wade and Ghahramani. **Bayesian Analysis**, 13:559–626, 2018.

[D2] **Julyan Arbel**. Discussion of "Sparse graphs using exchangeable random measures" by Caron and Fox. **Journal of the Royal Statistical Society. Series B**, 79, 2017.

[D3] **Julyan Arbel** and Christian P. Robert. Discussion of "Statistical modelling of citation exchange between statistics journals" by Varin, Cattelan and Firth. **Journal of the Royal Statistical Society. Series A**, 179:41–42, 2016.

[D4] **Julyan Arbel** and Igor Prünster. Discussion of "Sequential Quasi-Monte Carlo" by Gerber and Chopin. **Journal of the Royal Statistical Society. Series B**, 77: 559–560, 2015.

[D5] **Julyan Arbel** and Bernardo Nipoti. Discussion of "Bayesian Nonparametric Inference – Why and How" by Müller and Mitra. **Bayesian Analysis**, 8(02): 326–328, 2013.

[D6] Christian P. Robert and **Julyan Arbel**. Discussion of "Sparse Bayesian regularization and prediction" by Polson and Scott. **Bayesian Statistics 9**, 2009.

Proceedings

[P1] Veronica Munoz Ramirez, Florence Forbes, **Julyan Arbel**, Alexis Arnaud, and Michel Dojat. Quantitative MRI Characterization of Brain Abnormalities in 'de novo' Parkinsonian Patients. **IEEE International Symposium on Biomedical Imaging (ISBI)**, 2019.

[P2] Fabien Boux, Florence Forbes, **Julyan Arbel**, and Emmanuel Barbier. Estimation de paramètres IRM en grande dimension via une reégression inverse. **Congrès de la Société Française de Résonance Magnétique en Biologie et Médecine (SFRMBM)**, 2019.

[P3] Fabien Boux, Florence Forbes, **Julyan Arbel**, and Emmanuel Barbier. Apprentissage de dictionnaires par régression: application à l'IRM vasculaire. **Congrès National de l'Imagerie du Vivant (CNIV)**, 2019.

[P4] Fabien Boux, Florence Forbes, **Julyan Arbel**, and Emmanuel Barbier. Dictionary-free MR fingerprinting parameter estimation via inverse regression. **International Society for Magnetic Resonance in Medicine (ISMRM)**, 2018.

[P5] Mariia Vladimirova, **Julyan Arbel**, and Pablo Mesejo. Bayesian neural networks become heavier-tailed with depth. **NeurIPS Bayesian Deep Learning Workshop**, 2018.

[P6] Mariia Vladimirova, **Julyan Arbel**, and Pablo Mesejo. Bayesian neural network priors at the level of units. **1st Symposium on Advances in Approximate Bayesian Inference**, 2018.

[P7] Hongliang Lü, **Julyan Arbel**, and Florence Forbes. Bayesian Nonparametric Priors for Hidden Markov Random Fields. **50^e Journées de la Statistique de la SFdS**, 2018.

[P8] **Julyan Arbel** and Jean-Bernard Salomond. Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models. **NeurIPS Practical Bayesian Nonparametrics workshop**, 2016.

[P9] **Julyan Arbel** and Igor Prünster. Truncation error of a superposed gamma process in a decreasing order representation. **NeurIPS Advances in Approximate Bayesian Inference workshop**, 2016.

[P10] Guillaume Kon Kam King, **Julyan Arbel**, and Igor Prünster. Bayesian Nonparametric Density Estimation in Ecotoxicology. **48^e Journées de la Statistique de la SFdS**, 2016.

[P11] **Julyan Arbel**, Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. Discovery Probabilities when Uncertainty Matters. **48^e Journées de la Statistique de la SFdS**, 2016.

[P12] **Julyan Arbel**, Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. On Bayesian nonparametric inference for discovery probabilities. **Proceedings of the 48th Meeting of the Italian Statistical Society**, 2016.

[P13] **Julyan Arbel**, Kerrie Mengersen, and Judith Rousseau. On diversity under a Bayesian nonparametric dependent model. **Proceedings of the 47th Meeting of the Italian Statistical Society**, 2014.

Submitted preprints

[S1] **Julyan Arbel**, Olivier Marchal, and Bernardo Nipoti. On the Hurwitz zeta function with an application to the exponential-beta distribution. **Submitted**, 2019.

[S2] Hongliang Lü, **Julyan Arbel**, and Florence Forbes. Bayesian Nonparametric Priors for Hidden Markov Random Fields. **Under major revision, Statistics and Computing**, 2019.

[S3] **Julyan Arbel**, Guillaume Kon Kam King, Antonio Lijoi, Luis E. Nieto-Barajas, and Igor Prünster. BNPdensity: Bayesian nonparametric mixture modeling in R. **Submitted**, 2019.

[S4] Hien D Nguyen, **Julyan Arbel**, Hongliang Lü, and Florence Forbes. Approximate Bayesian computation via the energy statistic. **Submitted**, 2019.

[S5] Mariia Vladimirova and **Julyan Arbel**. Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. **Submitted**, 2019.

[S6] **Julyan Arbel**, Riccardo Corradin, and Bernardo Nipoti. Dirichlet process mixtures under affine transformations of the data. **Submitted**, 2018.

Ph.D. Thesis

[T1] **Julyan Arbel**. Contributions to Bayesian nonparametric statistics. **Ph.D. Thesis, Université Paris-Dauphine**, 2013.

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012.

[2] L. Al Labadi and M. Zarepour. On simulations from the two-parameter Poisson-Dirichlet process and the normalized Inverse-Gaussian process. *Sankhya*, 76-A:158–176, 2014.

[3] M. Albughdadi, L. Chaâri, J. Tourneret, F. Forbes, and P. Ciuciu. A Bayesian nonparametric hidden Markov random model for hemodynamic brain parcellation. *Signal Processing*, 135:132–146, 2017.

[4] T. Aldenberg and J. S. Jaworska. Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and environmental safety*, 46(1):1–18, may 2000.

[5] G. Alexits and I. Földes. *Convergence problems of orthogonal series*. Pergamon Press New York, 1961.

[6] A. Alfonsi and D. Brigo. New families of copulas based on periodic functions. *Communications in Statistics-Theory and Methods*, 34(7):1437–1447, 2005.

[7] J. Arbel. Contributions to Bayesian nonparametric statistics. *Ph.D. Thesis, Université Paris-Dauphine*, 2013.

[8] J. Arbel. Faà di Bruno's note on eponymous formula, trilingual version. *arXiv*, 2016.

[9] J. Arbel. Discussion of "Sparse graphs using exchangeable random measures" by Caron and Fox. *Journal of the Royal Statistical Society. Series B*, 79, 2017.

[10] J. Arbel. *Clustering Milky Way's Globulars: a Bayesian Nonparametric Approach*, chapter in Statistics for Astrophysics: Bayesian Methodology. EDP Sciences. Editors: Didier Fraix-Burnet, Stéphane Girard, Julyan Arbel and Jean-Baptiste Marquette, 2018.

[11] J. Arbel, R. Corradin, and M. Łewandowski. Discussion of "Bayesian Cluster Analysis: Point Estimation and Credible Balls", by Wade and Ghahramani. *Bayesian Analysis*, 13:559–626, 2018.

[12] J. Arbel, R. Corradin, and B. Nipoti. Dirichlet process mixtures under affine transformations of the data. *Submitted*, 2018.

[13] J. Arbel and V. Costemalle. Estimation of immigration flows : reconciling two sources by a Bayesian approach (in French). *Économie et Statistique*, 483-484-485:121–149, 2016.

[14] J. Arbel, M. Crispino, and S. Girard. Dependence properties and Bayesian inference for asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 174, 2019.

[15] J. Arbel, P. De Blasi, and I. Prünster. Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*, 14(3):753–771, 2019.

[16] J. Arbel and S. Favaro. Approximating predictive probabilities of Gibbs-type priors. *Sankhyā, forthcoming*, 2019.

[17] J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858, 2017.

[18] J. Arbel, G. Gayraud, and J. Rousseau. Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*, 40(3):549–570, 2013.

[19] J. Arbel, G. Kon Kam King, A. Lijoi, L. E. Nieto-Barajas, and I. Prünster. BNPdensity: Bayesian nonparametric mixture modeling in R. *Submitted*, 2019.

[20] J. Arbel, A. Lijoi, and B. Nipoti. *Bayesian Statistics from Methods to Models and Applications*, chapter Bayesian Survival Model based on Moment Characterization, pages 3–14. Springer Proceedings in Mathematics & Statistics, Volume 126. Springer International Publishing, Editors: Sylvia Frühwirth-Schnatter et al., 2015.

[21] J. Arbel, A. Lijoi, and B. Nipoti. Full Bayesian inference with hazard mixture models. *Computational Statistics & Data Analysis*, 93:359–372, 2016.

[22] J. Arbel, O. Marchal, and H. D. Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability & Statistics, forthcoming*, 2019.

[23] J. Arbel, O. Marchal, and B. Nipoti. On the Hurwitz zeta function with an application to the exponential-beta distribution. *Submitted*, 2019.

[24] J. Arbel, K. Mengersen, B. Raymond, T. Winsley, and C. King. Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of Antarctic microbial communities to fuel contaminated soil. *Ecology and Evolution*, 5(13):2633–2645, 2015.

[25] J. Arbel, K. Mengersen, and J. Rousseau. Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Annals of Applied Statistics*, 10(3):1496–1516, 2016.

[26] J. Arbel and B. Nipoti. Discussion of "Bayesian Nonparametric Inference – Why and How" by Müller and Mitra. *Bayesian Analysis*, 8(02):326–328, 2013.

[27] J. Arbel and I. Prünster. Discussion of "Sequential Quasi-Monte Carlo" by Gerber and Chopin. *Journal of the Royal Statistical Society. Series B*, 77:559–560, 2015.

[28] J. Arbel and I. Prünster. Truncation error of a superposed gamma process in a decreasing order representation. *NeurIPS Advances in Approximate Bayesian Inference workshop*, 2016.

[29] J. Arbel and I. Prünster. A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17, 2017.

[30] J. Arbel and I. Prünster. *Bayesian Statistics in Action*, chapter On the truncation error of a superposed gamma process, pages 11–19. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017.

[31] J. Arbel and C. P. Robert. Discussion of "Statistical modelling of citation exchange between statistics journals" by Varin, Cattelan and Firth. *Journal of the Royal Statistical Society. Series A*, 179:41–42, 2016.

[32] J. Arbel and J.-B. Salomond. Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models. *NeurIPS Practical Bayesian Nonparametrics workshop*, 2016.

[33] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, May 2011.

[34] R. Argiento, I. Bianchini, and A. Guglielmi. A blocked Gibbs sampler for NGG-mixture models via a priori truncation. *Statistics and Computing*, 26(3):641–661, 2016.

[35] R. Argiento, I. Bianchini, and A. Guglielmi. Posterior sampling from ε-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547, 2016.

[36] J. A. Awkerman, S. Raimondo, and M. G. Barron. Development of species sensitivity distributions for wildlife using interspecies toxicity correlation models. *Environmental Science and Technology*, 42(9):3447–3452, may 2008.

[37] J. Ba, J. Kiros, and G. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[38] S. Bacallado, M. Battiston, S. Favaro, and L. Trippa. Sufficientness postulates for Gibbs-type priors and hierarchical generalizations. *Statistical Science*, 32(4):487–500, 2017.

[39] S. Bacallado, S. Favaro, and L. Trippa. Bayesian nonparametric analysis of reversible Markov chains. *The Annals of Statistics*, 41(2):870–896, 2013.

[40] S. Bacallado, S. Favaro, and L. Trippa. Looking-backward probabilities for Gibbs-type exchangeable random partitions. *Bernoulli*, 21(1):1–37, 2015.

[41] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, 2012.

[42] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

[43] E. Barrios, A. Lijoi, L. E. Nieto-Barajas, and I. Prünster. Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334, 2013.

[44] F. Bassetti, R. Casarin, and F. Leisen. Beta-product dependent Pitman–Yor processes for Bayesian inference. *Journal of Econometrics*, 180(1):49 – 72, 2014.

[45] M. Battiston, S. Favaro, D. M. Roy, and Y. W. Teh. A characterization of product-form exchangeable feature probability functions. *The Annals of Applied Probability*, 28(3):1423–1448, 2018.

[46] B. Bauer and M. Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

[47] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683-1775)*, pages 370–418, 1763.

[48] M. Beal and Z. Ghahramani. The variational Bayesian EM Algorithm for incomplete data: with application to scoring graphical model structures. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics*, pages 453–464. Oxford University Press, 2003.

[49] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287, 2017.

[50] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013.

[51] E. Bernton and P. E. Jacob. *winference: Approximate Bayesian Computation with the Wasserstein distance*, 2018. R package version 0.1.2.

[52] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:235–269, 2019.

[53] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 192–236, 1974.

[54] A. Beygelzimer, S. Kakadet, J. Langford, S. Arya, D. Mount, and S. Li. *FNN: Fast Nearest Neighbor Search miller and Applications*, 2013. R package version 1.1.3.

[55] A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

[56] A. Bibi, M. Alfadly, and B. Ghanem. Analytic expressions for probabilistic moments of PL-DNN with Gaussian input. In *CVPR*, 2018.

[57] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27. Cambridge University Press, 1989.

[58] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 03 2006.

[59] S. Boltz, É. Debreuve, and M. Barlaud. High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18:1266–1283, 2009.

[60] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[61] F. Boux, F. Forbes, J. Arbel, and E. Barbier. Dictionary-free MR fingerprinting parameter estimation via inverse regression. *International Society for Magnetic Resonance in Medicine (ISMRM)*, 2018.

[62] F. Boux, F. Forbes, J. Arbel, and E. Barbier. Estimation de paramètres IRM en grande dimension via une reégression inverse. *Congrès de la Société Française de Résonance Magnétique en Biologie et Médecine (SFRMBM)*, 2019.

[63] A. Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, pages 929–953, 1999.

[64] T. Broderick, J. Pitman, and M. I. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

[65] V. V. Buldygin and I. V. Kozachenko. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Society, Providence, Rhode Island, 2000.

[66] V. V. Buldygin and Y. V. Kozachenko. Sub-Gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489, 1980.

[67] V. V. Buldygin and K. Moskvichova. The sub-Gaussian norm of a binary random variable. *Theory of probability and mathematical statistics*, 86:33–49, 2013.

[68] R. Burden and J. Faires. Numerical Analysis (5th edn), 1993.

[69] T. Campbell, J. H. Huggins, J. P. How, and T. Broderick. Truncated random measures. *Bernoulli*, 25(2):1256–1288, 2019.

[70] A. Canale, R. Corradin, and B. Nipoti. BNPmix: an R package for Bayesian nonparametric modelling via Pitman–Yor mixtures. *Submitted*, 2019.

[71] A. Canale, A. Lijoi, B. Nipoti, and I. Prünster. On the Pitman–Yor process with spike and slab base measure. *Biometrika*, 104(3):681–697, 2017.

[72] O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[73] F. Caron. Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, pages 2051–2059, 2012.

[74] F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.

[75] F. Caron, Y. W. Teh, and T. B. Murphy. Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2):1145–1181, 2014.

[76] I. Castillo. Pólya tree posterior distributions on densities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(4):2074–2102, 2017.

[77] L. Chaari, T. Vincent, F. Forbes, M. Dojat, and P. Ciuciu. Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach. *IEEE Transactions on Medical Imaging*, 32(5):821–837, 2013.

[78] D. Chandler. *Introduction to modern statistical mechanics*. Oxford University Press, New York, Oxford, 1987.

[79] S. P. Chatzis. A Markov random field-regulated Pitman-Yor process prior for spatially constrained data clustering. *Pattern Recognition*, 46(6):1595–1603, 2013.

[80] S. P. Chatzis and G. Tsechpenakis. The infinite hidden Markov random field model. *IEEE Transactions on Neural Networks*, 21(6):1004–1014, 2010.

[81] C. Chen, N. Ding, and W. Buntine. Dependent hierarchical normalized random measures for dynamic topic modeling. *International Conference on Machine Learning*, 2012.

[82] Y. Cho and L. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, 2009.

[83] R. Cont and P. Tankov. *Financial modelling with jump processes*. Chapman & Hall / CRC Press, London, 2008.

[84] A. Corduneanu and C. M. Bishop. Variational Bayesian Model Selection for Mixture Distributions. In *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pages 27–34. Morgan Kaufmann, 2001.

[85] D. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–202, 1972.

[86] R. Crackel and J. Flegal. Bayesian inference for a flexible class of bivariate beta distributions. *Journal of Statistical Computation and Simulation*, 87:295–312, 2017.

[87] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[88] C. M. Cuadras and J. Augé. A continuous general multivariate distribution and its properties. *Communications in Statistics-Theory and Methods*, 10(4):339–353, 1981.

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

[90] D. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I*. Springer-Verlag, New York, 2003.

[91] A. Damianou and N. Lawrence. Deep Gaussian processes. In *AISTATS*, 2013.

[92] A. DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advance*. Springer, New York, 2011.

[93] P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2015.

[94] P. De Blasi, S. Favaro, and P. Muliere. A class of neutral to the right priors induced by superposition of beta processes. *Journal of Statistical Planning and Inference*, 140(6):1563–1575, June 2010.

[95] P. De Blasi, A. Lijoi, and I. Prünster. An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica*, 23:1299–1322, 2013.

[96] P. De Blasi, G. Peccati, and I. Prünster. Asymptotics for posterior hazards. *The Annals of Statistics*, 37(4):1906–1945, 2009.

[97] C. Dellacherie and P. Meyer. *Probability and Potential B: Theory of Martingales*. North-Holland, Amsterdam, 1980.

[98] L. Devroye. Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 19(4):18, 2009.

[99] E. Di Bernardino and D. Rullière. On an asymmetric extension of multivariate Archimedean copulas based on quadratic form. *Dependence Modeling*, 4(1):328–347, 2016.

[100] K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2(2):183–201, 1974.

[101] F. Doshi, K. Miller, J. V. Gael, and Y. W. Teh. Variational inference for the indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 137–144, 2009.

[102] C. C. Drovandi and A. N. Pettitt. Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55:2541–2556, 2011.

[103] F. Durante. Construction of non-exchangeable bivariate distribution functions. *Statistical Papers*, 50(2):383–391, 2009.

[104] F. Durante, S. Girard, and G. Mazo. Copulas based on Marshall–Olkin machinery. In U. Cherubini, F. Durante, and S. Mulinacci, editors, *Marshall-Olkin Distributions. Advances in Theory and Applications*, volume 141 of *Springer Proceedings in Mathematics and Statistics*, pages 15–31. Springer, 2015.

[105] F. Durante and G. Salvadori. On the construction of multivariate extreme value models via copulas. *Environmetrics: The official journal of the International Environmetrics Society*, 21(2):143–161, 2010.

[106] D. Duvenaud, O. Rippel, R. Adams, and Z. Ghahramani. Avoiding pathologies in very deep networks. In *AISTATS*, 2014.

[107] R. Dykstra and P. Laud. A Bayesian nonparametric approach to reliability. *The Annals of Statistics*, 9(2):356–367, 1981.

[108] ECHA. Characterisation of dose concentration-response for environment. In *Guidance on information requirements and chemical safety assessment*, number May, chapter R.10. European Chemicals Agency, Helsinki, 2008.

[109] S. Elder. Bayesian adaptive data analysis guarantees from subgaussianity. *arXiv preprint: arXiv:1611.00065*, 2016.

[110] I. Epifani, A. Guglielmi, and E. Melilli. Moment-based approximations for the law of functionals of Dirichlet processes. *Applied Mathematical Sciences*, 3(20):979–1004, 2009.

[111] I. Epifani, A. Lijoi, and I. Prünster. Exponential functionals and means of neutral-to-the-right priors. *Biometrika*, 90(4):791–808, 2003.

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

[113] S. Favaro, A. Lijoi, R. H. Mena, and I. Prünster. Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):993–1008, 2009.

[114] S. Favaro, A. Lijoi, C. Nava, B. Nipoti, I. Prünster, and Y. W. Teh. On the Stick-Breaking Representation for Homogeneous NRMIs. *Bayesian Analysis*, 11(3):697–724, 09 2016.

[115] S. Favaro, A. Lijoi, and I. Prünster. A new estimator of the discovery probability. *Biometrics*, 68(4):1188–1196, 2012.

[116] S. Favaro and S. G. Walker. Slice sampling σ-stable poisson-kingman mixture models. *Journal of Computational and Graphical Statistics*, 22(4):830–847, 2013.

[117] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[118] T. S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, pages 615–629, 1974.

[119] T. S. Ferguson and M. J. Klass. A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643, 1972.

[120] F. Forbes and N. Peyrard. Hidden Markov Random Field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101, 2003.

[121] V. E. Forbes and P. Calow. Species Sensitivity Distributions Revisited: A Critical Appraisal. *Human and Ecological Risk Assessment*, 8(3):473–492, 2002.

[122] D. Fraix-Burnet, S. Girard, J. Arbel, and J.-B. Marquette, editors. *Statistics for Astrophysics: Bayesian Methodology*. EDP Sciences, 2018.

[123] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*. Springer, 1982.

[124] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

[125] A. E. Gelfand. Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, pages 145–161, 1996.

[126] A. E. Gelfand and A. Kottas. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11(2):289–305, 2002.

[127] C. Genest, K. Ghoudi, and L.-P. Rivest. Comment on "Understanding relationships using copulas" by E.W. Frees and E.A. Valdez. *North American Actuarial Journal*, 2(1):143–149, 1998.

[128] C. Genest, J. Nešlehová, and J.-F. Quessy. Tests of symmetry for bivariate copulas. *The Annals of the Institute of Statistical Mathematics*, 64(4):811–834, 2012.

[129] M. Gerber and N. Chopin. Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579, 2015.

[130] Z. Ghahramani and T. L. Griffiths. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482, 2005.

[131] S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[132] J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis: Theory and Methods*. Springer, New York, 2006.

[133] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer, New York, 2003.

[134] A. Gnedin. A species sampling model with finitely many types. *Electronic Communications in Probability*, 15:79–88, 2010.

[135] A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

[136] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.

[137] A. V. Gnedin et al. The bernoulli sieve. *Bernoulli*, 10(1):79–96, 2004.

[138] A. V. Gnedin, A. M. Iksanov, P. Negadajlov, and U. Rösler. The bernoulli sieve revisited. *The Annals of Applied Probability*, 19(4):1634–1655, 2009.

[139] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[140] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, Reading, 1994.

[141] J. Grandell. *Mixed Poisson Processes*. Monographs on Statistics and Applied Probability. Springer US, 1997.

[142] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[143] C. Grazian and B. Liseo. Approximate Bayesian inference in semiparametric copula models. *Bayesian Analysis*, 12(4):991–1016, 2017.

[144] P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[145] A. Gretton, K. M. Bogwardt, M. Rasch, B. Scholkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, 2007.

[146] A. Gretton, K. M. Bogwardt, M. J. Rasch, B. Scholkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[147] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, 2009.

[148] J. E. Griffin. An adaptive truncation method for inference in Bayesian nonparametric models. *Statistics and Computing*, 26(1-2):423–441, 2016.

[149] J. E. Griffin and S. G. Walker. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011.

[150] G. Gudendorf and J. Segers. Extreme-value copulas. In P. Jaworski, F. Durante, W. Härdle, and T. Rychlik, editors, *Copula Theory and Its Applications*, volume 198 of *Lecture Notes in Statistics*, pages 127–145. Springer, 2010.

[151] A. Gut. *Probability : a graduate course*. Springer texts in statistics. Springer, 2nd ed edition, 2013.

[152] S. Hayou, A. Doucet, and J. Rousseau. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning*, 2019.

[153] W. He, N. Qin, X. Kong, W. Liu, W. Wu, Q. He, C. Yang, Y. Jiang, Q. Wang, B. Yang, and F. Xu. Ecological risk assessment and priority setting for typical toxic pollutants in the water from Beijing-Tianjin-Bohai area using Bayesian matbugs calculator (BMC). *Ecological Indicators*, 45:209–218, 2014.

[154] C. Heaukulani and D. M. Roy. Gibbs-type Indian buffet processes. *Bayesian Analysis*, 2018.

[155] T. Herlau, M. N. Schmidt, and M. Mørup. Completely random measures for modelling block-structured sparse networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 4260–4268. Curran Associates, Inc., 2016.

[156] N. L. Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

[157] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, April 2010.

[158] N. L. Hjort, C. Holmes, P. Muller, and S. G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, 2010.

[159] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[160] M. Hofert. Efficiently sampling nested Archimedean copulas. *Computational Statistics & Data Analysis*, 55(1):57–70, 2011.

[161] M. Hofert and F. Vrins. Sibuya copulas. *Journal of Multivariate Analysis*, 114:318–337, 2013.

[162] J. Hron, A. Matthews, and Z. Ghahramani. Variational Bayesian dropout: pitfalls and fixes. In *International Conference on Machine Learning*, 2018.

[163] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

[164] H. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[165] H. Ishwaran and L. James. Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. *Journal of the American Statistical Association*, 99(465):175–190, 2004.

[166] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–283, 2002.

[167] Izmailov, Pavel and Maddox, Wesley J and Kirichenko, Polina and Garipov, Timur and Vetrov, Dmitry and Wilson, Andrew Gordon. Subspace inference for bayesian deep learning. *arXiv preprint arXiv:1907.07504*, 2019.

[168] L. James. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics*, 33(4):1771–1799, 2005.

[169] L. F. James. Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*, 2002.

[170] L. F. James. Stick-breaking PG(α ,ζ)-Generalized Gamma Processes. *arXiv preprint arXiv:1308.6570*, 2013.

[171] L. F. James, A. Lijoi, and I. Prünster. Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120, 2006.

[172] L. F. James, A. Lijoi, and I. Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

[173] A. Jara. Applied Bayesian non-and semi-parametric inference using DPpackage. *SpherWave: An R Package for Analyzing Scattered Spherical Data by Spherical Wavelets*, page 17, 2007.

[174] A. Jara, T. Hanson, F. Quintana, P. Müller, and G. Rosner. DPpackage: Bayesian semi and nonparametric modelling in R. *Journal of Statistical Software*, 40(5):1, 2011.

[175] A. Jara, E. Lesaffre, M. De Iorio, and F. Quintana. Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics*, 4(4):2126–2149, 2010.

[176] B. Jiang, T.-Y. Wu, and W. H. Wong. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[177] H. Joe. *Multivariate models and dependence concepts*. Chapman and Hall/CRC, 1997.

[178] H. Joe, H. Li, and A. K. Nikoloulopoulos. Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, 101(1):252–270, 2010.

[179] T. D. Johnson, Z. Liu, A. J. Bartsch, and T. E. Nichols. A Bayesian non-parametric Potts model with application to pre-surgical fMRI data. *Statistical Methods in Medical Research*, 22(4):364–381, 2013.

[180] D. S. Jones, L. W. Barnthouse, G. W. Suter II, R. A. Efroymson, J. M. Field, and J. J. Beauchamp. Ecological risk assessment in a large river-reservoir: 3. Benthic invertebrates. *Environmental Toxicology and Chemistry*, 18(4):599–609, 1999.

[181] M. C. Jones. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6:70–81, 2009.

[182] M. I. Jordan. Hierarchical models, nested models and completely random measures. *Frontiers of Statistical Decision Making and Bayesian Analysis: in Honor of James O. Berger. New York: Springer*, 2010.

[183] G. Karabatsos. A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, 2016.

[184] G. Karabatsos and F. Leisen. An approximate likelihood perspective on ABC methods. *Statistics Surveys*, 12:66–104, 2018.

[185] M. J. Kearns and L. K. Saul. Large deviation methods for approximate probabilistic inference. In *UAI*, 1998.

[186] B. J. Kefford, G. L. Hickey, A. Gasith, E. Ben-David, J. E. Dunlop, C. G. Palmer, K. Allan, S. C. Choy, and C. Piscart. Global scale variation in the salinity sensitivity of riverine macroinvertebrates: Eastern Australia, France, Israel and South Africa. *PLoS ONE*, 7(5):e35224, jan 2012.

[187] A. Khoudraji. *Contributions à l'étude des copules et à la modélisation des valeurs extrêmes bivariées*. PhD thesis, Université Laval Québec, Canada, 1995.

[188] D. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, 2015.

[189] J. F. C. Kingman. Completely random measures. *Pacific J. Math.*, 21(1):59–78, 1967.

[190] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 1–22, 1975.

[191] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978.

[192] J. F. C. Kingman. *Poisson processes*, volume 3. Oxford university press, 1993.

[193] K.-R. Koch. *Introduction to Bayesian Statistics*. Springer, Heidelberg, 2007.

[194] G. Kon Kam King, J. Arbel, and I. Prünster. *Bayesian Statistics in Action*, chapter A Bayesian nonparametric approach to ecological risk assessment, pages 151–159. Springer Proceedings in Mathematics & Statistics, Volume 194. Springer International Publishing, Editors: Raffaele Argiento et al., 2017.

[195] G. Koop, D. J. Poirier, and J. L. Tobias. *Bayesian Econometric Methods*. Cambridge University Press, Cambridge, 2007.

[196] V. S. Koroljuk and Y. V. Borovskich. *Theory of U-Statistics*. Springer, Dordrecht, 1994.

[197] S. Kotz and J. R. Van Dorp. *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific, Singapore, 2004.

[198] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[199] A. Krogh and J. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, 1991.

[200] A. K. Kuchibhotla and A. Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

[201] D. Lauterbach and D. Pfeifer. Some extensions of singular mixture copulas. In M. Hallin, D. Mason, D. Pfeifer, and J. Steinebach, editors, *Mathematical Statistics and Limit Theorems*, pages 271–286. Springer, 2015.

[202] C. Lawless and J. Arbel. A simple proof of Pitman–Yor's Chinese restaurant process from its stick-breaking representation. *Dependence Modeling*, 7, 2019.

[203] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[204] J. Lee, J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri. Deep neural networks as Gaussian processes. In *International Conference on Machine Learning*, 2018.

[205] J. Li, D. J. Nott, Y. Fan, and S. A. Sisson. Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. *Computational Statistics & Data Analysis*, 106:77–89, 2017.

[206] F. Liang, Q. Li, and L. Zhou. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972, 2018.

[207] E. Liebscher. Construction of asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 99(10):2234–2250, 2008.

[208] E. Liebscher. Erratum to "Construction of asymmetric multivariate copulas" [J. Multivariate Anal. 99 (2008) 2234–2250]. *Journal of Multivariate Analysis*, 102(4):869–870, 2011.

[209] A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291, 2005.

[210] A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

[211] A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

[212] A. Lijoi and B. Nipoti. A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association*, 109(506):802–814, 2014.

[213] A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian nonparametrics*, pages 80–136. Cambridge University Press, Cambridge, 2010.

[214] A. Lijoi, I. Prünster, and S. G. Walker. Investigating nonparametric priors with Gibbs structure. *Statistica Sinica*, 18(4):1653, 2008.

[215] A. Lijoi, I. Prünster, and S. G. Walker. Posterior analysis for some classes of nonparametric models. *Journal of Nonparametric Statistics*, 20(5):447–457, 2008.

[216] J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate Bayesian computation. *Systems Biology*, 60:e60–e82, 2017.

[217] A. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

[218] A. Lo and C. Weng. On a class of Bayesian nonparametric estimates. II. Hazard rate estimates. *The Annals of the Institute of Statistical Mathematics*, 41(2):227–245, 1989.

[219] H. Lü, J. Arbel, and F. Forbes. Bayesian Nonparametric Priors for Hidden Markov Random Fields. *Under major revision, Statistics and Computing*, 2019.

[220] S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

[221] D. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[222] J.-F. Mai and M. Scherer. Bivariate extreme-value copulas with discrete Pickands dependence measure. *Extremes*, 14(3):311–324, 2011.

[223] O. Marchal and J. Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017.

[224] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computation methods. *Statistics and Computing*, 22:1167–1180, 2012.

[225] A. W. Marshall and I. Olkin. A generalized bivariate exponential distribution. *Journal of Applied Probability*, 4(2):291–302, 1967.

[226] A. Matthews, M. Rowland, J. Hron, R. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, volume 1804.11271, 2018.

[227] G. Mazo, S. Girard, and F. Forbes. A class of multivariate copulas based on products of bivariate copulas. *Journal of Multivariate Analysis*, 140:363–376, 2015.

[228] D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911, 2003.

[229] D. A. McAllester and R. E. Schapire. On the Convergence Rate of Good-Turing Estimators. In *COLT*, pages 1–6, 2000.

[230] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1996.

[231] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984.

[232] K. Mengersen, C. Alston, J. Arbel, and E. Duncan. *Applications in Industry*, chapter in Handbook of mixture analysis. CRC Press, Editors: Gilles Celeux, Sylvia Früwirth-Schnatter, and Christian P. Robert, 2018.

[233] J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 113:340–356, 2018.

[234] J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113:340–356, 2018.

[235] D. Molchanov, A. Ashukha, and D. Vetrov. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, 2017.

[236] P. Muliere and L. Tardella. Approximating distributions of random functionals of ferguson-dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297, 1998.

[237] V. Munoz Ramirez, F. Forbes, J. Arbel, A. Arnaud, and M. Dojat. Quantitative MRI Characterization of Brain Abnormalities in 'de novo' Parkinsonian Patients. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019.

[238] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1:16, 2007.

[239] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, 2012.

[240] C. Navarrete, F. A. Quintana, and P. Mueller. Some issues in nonparametric Bayesian modeling using species sampling models. *Statistical Modelling*, 8(1):3–21, 2008.

[241] R. Neal. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report CRG-TR-92-1, University of Toronto, 1992.

[242] R. Neal. *Bayesian learning for neural networks*. Springer, 1996.

[243] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan, editor, *Lear. in Graph. Mod.*, pages 355–368. 1998.

[244] R. B. Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[245] H. D. Nguyen, J. Arbel, H. Lü, and F. Forbes. Approximate Bayesian computation via the energy statistic. *Submitted*, 2019.

[246] Y. Ni, P. Müller, Y. Zhu, and Y. Ji. Heterogeneous reciprocal graphical models. *Biometrics*, 74(2):606–615, 2018.

[247] L. E. Nieto-Barajas. Bayesian semiparametric analysis of short– and long–term hazard ratios with covariates. *Comput. Stat. Data. An.*, 71:477–490, 2014.

[248] L. E. Nieto-Barajas and I. Prünster. A sensitivity analysis for Bayesian nonparametric density estimators. *Statistica Sinica*, 19:685–705, 2009.

[249] L. E. Nieto-Barajas, I. Prünster, and S. G. Walker. Normalized random measures driven by increasing additive processes. *The Annals of Statistics*, 32(6):2343–2360, 2004.

[250] L. E. Nieto-Barajas, I. Prünster, and S. G. Walker. Normalized random measures driven by Increasing Additive Processes. *The Annals of Statistics*, 32(6):2343–2360, 2004.

[251] L. E. Nieto-Barajas and S. G. Walker. Markov Beta and Gamma Processes for Modelling Hazard Rates. *Scandinavian Journal of Statistics*, 29(3):413–424, 2002.

[252] L. E. Nieto-Barajas and S. G. Walker. Bayesian nonparametric survival analysis driven by Lévy driven Markov processes. *Statistica Sinica*, 14:1127–1146, 2004.

[253] S. Ning and N. Shephard. A nonparametric Bayesian approach to copula estimation. *Journal of Statistical Computation and Simulation*, 88(6):1081–1105, 2018.

[254] J. Nolan. *Stable distributions: models for heavy-tailed data*. Birkhauser Boston, 2003.

[255] R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, J. Hron, D. Abolafia, J. Pennington, and J. Sohl-Dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019.

[256] P. Orbanz and J. M. Buhmann. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45, 2008.

[257] P. Orbanz and S. Williamson. Unit–rate poisson representations of completely random measures. *Electronic Journal of Statistics*, 5:1354–1373, 2011.

[258] J. W. Paisley, D. M. Blei, and M. I. Jordan. Stick-breaking beta processes and the poisson process. In *International Conference on Artificial Intelligence and Statistics*, pages 850–858, 2012.

[259] O. Papaspiliopoulos and G. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169, 2008.

[260] M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[261] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

[262] A. Perry, A. S. Wein, and A. S. Bandeira. Statistical limits of spiked tensor models. *arXiv preprint arXiv:1612.07728*, 2016.

[263] G. Pisier. Subgaussian sequences in probability and Fourier analysis. *arXiv preprint: arXiv:1607.01053*, 2016.

[264] J. Pitman. The two-parameter generalization of Ewens random partition structure. Technical report, Technical Report 345, Dept. Statistics, UC Berkeley, 1992.

[265] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

[266] J. Pitman. Poisson-Kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003.

[267] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002.

[268] J. Pitman. *Combinatorial stochastic processes*, volume 1875. Springer-Verlag, 2006.

[269] J. Pitman and Y. Yakubovich. Extremes and gaps in sampling from a GEM random discrete distribution. *Electronic Journal of Probability*, 22, 2017.

[270] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

[271] N. G. Polson and V. Sokolov. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.

[272] L. Posthuma, G. W. Suter II, and P. T. Trass. *Species sensitivity distributions in ecotoxicology*. CRC press, 2002.

[273] S. J. Press. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Wiley, Hoboken, 2003.

[274] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798, 1999.

[275] S. B. Provost. Moment-based density approximants. *Mathematica J.*, 9(4):727–756, 2005.

[276] G. Puccetti. An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, 451(1):132 – 145, 2017.

[277] A. Quetelet. Sur quelques propriétés curieuses que présentent les résultats d'une série d'observations, faites dans la vue de déterminer une constante, lorsque les chances de rencontrer des écarts en plus et en moins sont égales et indépendantes les unes des autres. *Bulletins de l'Académie royale des sciences, des lettres et des beaux-arts de Belgique*, 19:303–317, 1852.

[278] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–246, 2013.

[279] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[280] C. R. Rao. *Linear statistical inference and its applications*. Wiley, New York, 1965.

[281] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[282] E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

[283] S. I. Resnick. *Extreme values, regular variation and point processes*. Springer, 2013.

[284] C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag, 2007.

[285] C. P. Robert. Approximate Bayesian computation, an introduction. In D. Fraix-Burnet, S. Girard, J. Arbel, and J.-B. Marquette, editors, *Statistics for Astrophysics: Bayesian Methodology*, pages 77–112. EDP Sciences, 2018.

[286] C. P. Robert and J. Arbel. Discussion of "Sparse Bayesian regularization and prediction" by Polson and Scott. *Bayesian Statistics 9*, 2009.

[287] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, New York, 2004.

[288] J. A. Rodrıguez-Lallena and M. Úbeda-Flores. A new class of bivariate copulas. *Statistics & Probability Letters*, 66(3):315–325, 2004.

[289] J. Rosiński. Series representations of Lévy processes from the perspective of point processes. In *Lévy processes–Theory and Applications, Barndorff-Nielsen, Mikosch and Resnick*, pages 401–415. Boston, 2001.

[290] D. M. Roy. The continuum-of-urns scheme, generalized beta and Indian buffet processes, and hierarchies thereof. *arXiv preprint arXiv:1501.00208*, 2014.

[291] D. B. Rubin. Bayesian justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12:1151–1172, 1984.

[292] M. Ruggiero, S. G. Walker, and S. Favaro. Alpha-diversity processes and normalized inverse-Gaussian diffusions. *The Annals of Applied Probability*, 23(1):386–425, 2013.

[293] Y. Saatci and A. Wilson. Bayesian GAN. In *Advances in Neural Information Processing Systems*, 2017.

[294] H. Sagan. *Space-filling curves*. Springer-Verlag New York, 1994.

[295] G. Salvadori and C. De Michele. Multivariate multiparameter extreme value models and return periods: A copula approach. *Water Resources Research*, 46(10):1–11, 2010.

[296] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

[297] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41:2263–2291, 2013.

[298] P. K. Sen. Almost sure convergence of generalized U-statistics. *The Annals of Probability*, 5:287–290, 1977.

[299] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[300] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[301] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[302] S. A. Sisson, Y. Fan, and M. A. Beaumont, editors. *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton, 2019.

[303] M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

[304] C. G. Small. *Expansions and asymptotics for statistics*. CRC Press, 2010.

[305] J. Sodjo, A. Giremus, N. Dobigeon, and J.-F. Giovannelli. A generalized Swendsen-Wang algorithm for Bayesian nonparametric joint segmentation of multiple images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1882–1886, La Nouvelle Orléans, LA, United States, Mar. 2017. IEEE.

[306] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[307] J. Stoehr. A review on statistical inference methods for discrete Markov random fields. *arXiv e-prints*, page arXiv:1704.03331, Apr 2017.

[308] E. B. Sudderth and M. I. Jordan. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*, pages 1585–1592, 2008.

[309] G. W. Suter II, L. W. Barnthouse, R. A. Efroymson, and H. Jager. Ecological Risk Assessment in a Large River–Reservoir: 2. Fish Community. *Environmental Toxicology and Chemistry*, 18(4):589–598, 1999.

[310] B. Szabó, A. W. van der Vaart, J. van Zanten, et al. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

[311] G. Szegő. *Orthogonal polynomials*. American Mathematical Society Colloquium Publications, 1967.

[312] G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5:1–16, 2004.

[313] G. J. Szekely and M. L. Rizzo. Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.

[314] G. J. Szekely and M. L. Rizzo. The energy of data. *Annual Review of Statistics and Its Application*, 4(447-479), 2017.

[315] S. Tavaré. On the history of ABC. In S. A. Sisson, Y. Fan, and M. A. Beaumont, editors, *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton, 2019.

[316] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518, 1997.

[317] D. Taylor-Rodriguez, K. Kaufeld, E. M. Schliep, J. S. Clark, and A. E. Gelfand. Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Analysis*, 12(4):939–967, 2017.

[318] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[319] Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, pages 1838–1846. Curran Associates, Inc., 2009.

[320] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *International conference on artificial intelligence and statistics*, pages 564–571, 2007.

[321] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 267–288, 1996.

[322] F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142, 1951.

[323] W. Trutschnig, M. Schreyer, and J. Fernández-Sánchez. Mass distributions of two-dimensional extreme-value copulas and related results. *Extremes*, 19(3):405–427, 2016.

[324] A. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009.

[325] R. Unnikrishnan, C. Pantofaru, and M. Hebert. A measure for objective evaluation of image segmentation algorithms. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 34–34, Sep. 2005.

[326] L. D. Valle, F. Leisen, and L. Rossini. Bayesian non-parametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):523–548, 2018.

[327] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1):61–81, Apr 2005.

[328] M. Vladimirova and J. Arbel. Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Submitted*, 2019.

[329] M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural network priors at the level of units. *1st Symposium on Advances in Approximate Bayesian Inference*, 2018.

[330] M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural networks become heavier-tailed with depth. *NeurIPS Bayesian Deep Learning Workshop*, 2018.

[331] M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel. Understanding Priors in Bayesian Neural Networks at the Unit Level. *ICML*, 2019.

[332] J. Voss. *An Introduction to Statistical Computing: A Simulation-based Approach*. Wiley, Chichester, 2014.

[333] S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54, 2007.

[334] S. G. Walker and P. Damien. Miscellanea. Representations of Lévy processes without Gaussian components. *Biometrika*, 87(2):477–483, 2000.

[335] S. G. Walker and P. Damien. Representations of Lévy processes without Gaussian components. *Biometrika*, 87(2):477–483, 2000.

[336] B. Wang, G. Yu, J. Huang, and H. Hu. Development of species sensitivity distributions and estimation of HC(5) of organochlorine pesticides with five statistical approaches. *Ecotoxicology*, 17(8):716–724, nov 2008.

[337] C. Wang and D. M. Blei. Truncation-free stochastic variational inference for Bayesian nonparametric models. In *Advances in Neural Information Processing Systems*, NIPS'12, pages 413–421, 2012.

[338] Y. Wang, F. Wu, J. P. Giesy, C. Feng, Y. Liu, N. Qin, and Y. Zhao. Non-parametric kernel density estimation of species sensitivity distributions in developing water quality criteria of metals. *Environmental Science and Pollution Research*, 22(18):13980–13989, 2015.

[339] S. Wu. Construction of asymmetric copulas and its application in two-dimensional reliability modelling. *European Journal of Operational Research*, 238(2):476–485, 2014.

[340] D. Xu, F. Caron, and A. Doucet. Bayesian nonparametric image segmentation using a generalized Swendsen-Wang algorithm. *ArXiv e-prints*, Feb. 2016.

[341] F.-L. Xu, Y.-L. Li, Y. Wang, W. He, X.-Z. Kong, N. Qin, W.-X. Liu, W.-J. Wu, and S. E. Jorgensen. Key issues for the development and application of the species sensitivity distribution (SSD) model for ecological risk assessment. *Ecological Indicators*, 54:227–237, 2015.

[342] J. Yan. Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software*, 21(4):1–21, 2007.

[343] B. A. Zajdlik, D. G. Dixon, and G. Stephenson. Estimating Water Quality Guidelines for Environmental Contaminants Using Multimodal Species Sensitivity Distributions: A Case Study with Atrazine. *Human and Ecological Risk Assessment*, 15(3):554–564, 2009.

[344] Y. Zhang, C.-W. Kim, M. Beer, H. Dai, and C. G. Soares. Modeling multivariate ocean data using asymmetric copulas. *Coastal Engineering*, 135:91–111, 2018.

[345] A. Zygmund. An individual ergodic theorem for non-comutative transformations. *Acta Scientiarum Mathematicarum (Szeged)*, 14:103–110, 1951.