UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo 36

Advancements in Distribution Sampling for Statistical Inference

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Laura Ventura

Co-supervisori: Prof. Pierre E. Jacob, Prof. Robin J. Ryder

Dottoranda: Elena Bortolato

January 10th, 2024

Abstract

The process of inference in a parametric statistical model involves assessing the uncertainty of the surrounding parameters based on an observed sample. Often, the lack of analytical solutions prohibits a direct precise quantification. Monte Carlo (MC) simulations play a central role in understanding this uncertainty by reproducing samples that mimic data-dependent probability distributions or replicating data-generating mechanisms to estimate functionals and specific quantities of interest. This dissertation is dedicated to the advancement of methods for sampling algorithms for statistical inference in different paradigms.

Chapter 2 focuses on general simulation-based strategies for obtaining confidence distributions, confidence curves and confidence densities in non non regular settings, where for instance Bootstrap methods are not directly applicable. Special attention is paid to the treatment of parameter vectors and nuisance parameters, to ensure invariance of the procedures under reparametrizations. The developed techniques are investigated in the context of robust methods and estimating equations. Some extensions are considered for inference with non-parametric tests, with probability semi-metrics. Possible applications can be found in non-inferiority tests and in the realm of likelihood-free inference.

In Chapter 3, we study a method in the context of Approximate Bayesian Computation which is free from the choice of the tuning parameter. The resulting approximation implicitly uses a pseudo-likelihood that exhibits some consistency properties, and is linked to Confidence Distributions and data depth functions.

In Chapter 4, we derive and discuss coupling techniques for Markov Chain Monte Carlo (MCMC) algorithms on submanifolds. They form the basis for the generation of

convergence diagnoses and principal possibilities for the execution of parallel chains. In particular, we describe probabilistic reflection-contract couplings and meeting-inducing couplings, placing the latter in the context of couplings for a broader class of MCMC algorithms with complex proposal mechanisms.

Chapter 5 presents two novel MCMC strategies developed for sampling generic target distributions on R^d. These algorithms utilise ideas derived from MCMC algorithms on manifolds, incorporating geometric information from the target distribution in the problem at hand. In particular, equations specifically relevant to the sampling problem are used to define an artificial submanifold, such as the graph of the target distribution and the contour set.

Finally, Chapter 6 deals with the problem of performing Bayesian inference in the presence of an intractable matching prior distribution. This involves transitioning to a manifold characterized by an estimating equation that encompasses the derivatives of the said intractable matching prior distribution. An application in the context of Bayesian testing with e-values is presented, where the default prior guarantees the invariance properties of the procedure.

Sommario

Il processo di inferenza in un modello statistico parametrico implica valutare l'incertezza dei parametri circostanti basandosi su un campione osservato. Spesso, la mancanza di soluzioni analitiche proibisce una quantificazione precisa diretta.

Le simulazioni Monte Carlo (MC) rivestono un ruolo fondamentale per la comprensione e determinazione di questa incertezza, mediante la generazione empirica di distribuzioni di probabilità o riproducendo i meccanismi di generazione dei dati stessi, al fine di riprodurre e stimare funzionali e quantità specifiche di interesse. Questa tesi è dedicata allo sviluppo di metodi per algoritmi di campionamento per l'inferenza statistica secondo diversi paradigmi.

Il Capitolo 2 si concentra su strategie generali basate su simulazioni per ottenere distribuzioni di confidenza, curve di confidenza e densità di confidenza in contesti non regolari, dove ad esempio i metodi Bootstrap non sono direttamente applicabili. Si presta particolare attenzione al trattamento di vettori di parametri e al caso di parametri di disturbo, per garantire l'invarianza delle procedure sotto riparametrizzazioni. Le tecniche sviluppate sono studiate nel contesto di metodi robusti ed equazioni di stima. Alcune estensioni sono considerate per l' inferenza con test non parametrici con semimetriche di probabilità. Possibili applicazioni si trovano nei test di non inferiorità e nel campo dell'inferenza priva di verosimiglianza.

Nel Capitolo 3 viene introdotto un metodo nel contesto di Approximate Bayesian Computation (ABC) libero dalla scelta del parametri di regolarizzazione o *tuning*. L'approssimazione risultante utilizza implicitamente una funzione di pseudo-verosimiglianza che gode di proprietà di consistenza, grazie anche ad alcuni legami con le distribuzioni di confidenza e le funzioni di profondità dei dati (*data depth*).

Nel Capitolo 4, vengono ottenute e discusse tecniche per l' accoppiamento (*coupling*) per algoritmi MCMC su sottovarietà. Queste tecniche, nel contesto degli algoritmi basati su catene di Markov, costituiscono la base per la derivazione di diagnostiche di convergenza e per la possibilità di ottenere stime non distorte a partire da campioni ottenuti dall'esecuzione di catene parallele. In particolare, descriviamo e proponiamo l' implementazione di schemi di coupling basati su riflessione, che favoriscono l'avvicinamento reciproco delle catene, e meccanismi di coupling che rendono possibile l'incontro delle catene stesse. Questi ultimi schemi vengono infine ricondotti ed estesi ad un contesto e ad algoritmi Markov Chain Monte Carlo (MCMC) più generali, in cui il meccanismo di proposta viene definito complesso, in quanto si articola in più fasi.

Il Capitolo 5 presenta due nuovi algoritmi MCMC sviluppati per campionare distribuzioni di interesse definite in R^d. Questi fanno uso di idee derivate da algoritmi MCMC su sottovarietà, incorporando informazioni geometriche della distribuzione target nel problema di campionamento in considerazione. In particolare, equazioni specificamente rilevanti per il problema del campionamento sono utilizzate per definire artificialmente delle sottovarietà, come il grafico della distribuzione target e le curve di livello della funzione.

Infine, il Capitolo 6 affronta il problema di effettuare inferenza Bayesiana in presenza di una distribuzione a priori di tipo *matching* che risulta intrattabile. L'approccio considerato coinvolge la simulazione su una sottovarietà definita a partire da un'equazione di stima che coinvolge le derivate della suddetta distribuzione a priori intrattabile. Viene presentata un'applicazione nel contesto dei test d'ipotesi Bayesiani con e-values in cui la prior è scelta in modo da garantire le proprietà di invarianza della procedura.

Acknowledgements

First of all, I would like to sincerely thank my supervisor Laura for her support, guidance and trust, for all the growth opportunities and her efforts to value me. I have always felt her sincere care and for that I am extremely thankful. I am infinitely grateful to Pierre for his patience, encouragement and the countless opportunities: stimulating research projects, the teaching experience and, together with Robin, the chance to learn in such an academically and humanly rich environment. Thank you both for organizing discussions and workshops, for your sincere passion for research and your commitment to share knowledge with students and young researchers. A big thank you to all the wonderful people I met over the three years: Anna, Charly, Lorenzo, Ruihua, Mahdi, Giovanni, Antoine, Claudia, Camilla, Luke, Adrien, Theo, for a tea, a coffee, a relaxed chat or a game board, for teaching me something about the world. I would like to thank the CEREMADE lab, especially Christian and Alessandra, for their warm welcome and attention, and also the IDS department of ESSEC for their kindness and hospitality. I thank my friends and fellow "cyclist": Dafne, with whom I shared joys and sorrows on the phone during the first isolation, and then Airbn'bs and airplanes, Francesco, Andrea, Riccardo, Daniele, Parvaneh, Emanuele, Thien Phuc, to my office colleagues, all "Cucconi" students, as well as to Erika, Pietro, Andrea. And then to my 146 office colleagues, Cristian and Nasrin, with whom I have shared the last turbulent months. I would like to thank Tony for his valuable advice, which has always helped me and made me feel encouraged and appreciated. Special thanks to Manuela, my "spiritual godmother", to Marco, my "hablas también con los muros", to Dede especially for the encouragement with the mouse at 3am, to Marco for the Jungian introspections and to everyone for their great patience, understanding and friendship. Special thanks to Sara and Alessandro for their invaluable support, affection and hospitality, and to Enrico for his unconditional love and his attempts to make me stronger in every way. Thanks to my family for bringing a little zest and craziness into my life, especially to my mother for her unwavering patience, and to my neighbors in Paris, especially Jean Jean, my daily alarm clock. I could not have written this thesis without you.

Contents

List of Figures

List of Tables

Introduction

Overview

While crucial for statistical inference, probability distributions and their functionals often defy simple mathematical expressions. This difficulty arises in various scenarios, including sampling distributions in finite sampling regimes, conditional distributions with posterior distributions being a notable example (Barndorff-Nielsen and Cox, 1994; DiCiccio *et al.*, 1993; Martin *et al.*, 2023). Other cases include statistical models with intractable likelihood functions, for instance when latent variables are involved or normalization constants depend on complex integrals (Kent, 1982; Ising, 1924). Similar problems also occur in connection with certain objective prior distributions (Consonni *et al.*, 2018; Leisen *et al.*, 2020). In such cases, determining the distribution of interest requires solving complicated partial differential equations. Further challenges arise when the support of the distribution of interest deviates from the usual Euclidean space. The lack of regularity conditions may not only lead to poor approximations but also render some techniques useless.

When it comes to drawing conclusions on high-dimensional, constrained or complicated domains, computer simulations, with Monte Carlo techniques at their core, play an important role in various statistical paradigms. In this context, a convenient way of representing models is by means of *algorithmic form*, also known as a data generating equation, which can be expressed as follows

$$y = g(u, \theta). \tag{1}$$

This formulation outlines the process by which the data is generated: the function $g(\cdot, \cdot)$ represents a deterministic mapping from the parameter space Θ and the space of random components U to the observation space or output space $\mathcal{Y}$. The random, unobserved variables $u \in U$ are distributed according to a known probability distribution $\rho(u)$, which is independent of θ while the nature of $\theta \in \Theta$ and its adherence to a suitable

probability distribution can vary depending on the paradigm used. Upon defining this structure, simulations can facilitate the recovery of the distribution to be inferred while the chosen inference paradigm guides the process of elaborating and synthesizing the information derived from the observation and comparing it with the data y^{obs}. This is the fundamental premise of simulation-based inference (SBI).

Historically, the representation (1) was first proposed in fiducial inference for group models and structural inference (Fraser, 1961; Bunke, 1975; Fraser, 2004; Dawid and Stone, 1982) and is now widely used in the context of generalized fiducial inference (GFI) (Hannig, 2009; Hannig *et al.*, 2016). The basic concept of GFI is to pair each input u with parameter estimates $\hat{\theta}(y^{\text{obs}}, u)$ such that $g(u, \hat{\theta}(y^{\text{obs}}, u))$ provides the most accurate approximation for y^{obs}. In other words, for each potential realization of the random quantity u, the estimates are defined as $\hat{\theta}(y^{\text{obs}}, u) = \arg\min_\theta \|g(u, \theta)\|$, with $\|\cdot\|$ the L^2 or L^∞ norm. The operation of combining u, originally assumed to be distributed with density $\rho(u)$ to $\hat{\theta}$ allows to define a new distribution on u

$$\psi_\epsilon(u) \propto \rho(u) \cdot \mathbb{I}\{\|g(u, \hat{\theta}(y^{\text{obs}}, u)) - y\| < \epsilon\}, \tag{2}$$

which is supported on a restricted space. The expression (2) then, for $u \sim \psi_\epsilon(u)$ and as $\epsilon \to 0$, induces a distribution on $\hat{\theta}(y^{\text{obs}}, u)$, which is referred to as a generalized fiducial distribution (GFD) Hannig *et al.* (2016).

In Bayesian analysis, the relationship between the target posterior distribution and the form in equation (1) is established by introducing an additional layer, which is given by the prior distribution $\pi(\theta)$ on the parameter space. Considering the observed data, the posterior can then be written as

$$\pi(\theta|y^{\text{obs}}) \propto P\{g(u, \theta) \in \mathcal{G}_\theta(y^{\text{obs}})|\theta\}\pi(\theta), \tag{3}$$

where $\mathcal{G}_\theta(y) = \{u \in U : g(u, \theta) = y\}$ is a section of the data generating manifold, i.e. is a subset of the data generating manifold for fixed θ and $P\{g(u, \theta) \in \mathcal{G}_\theta(y^{\text{obs}})|\theta\}$ can be recognized as the likelihood function, which is

$$\int_{\mathcal{G}_\theta(y^{\text{obs}})} \frac{\rho(u)}{\det\left(\nabla_u g(u, \theta) \nabla_u g(u, \theta)^\top\right)^{1/2}} \lambda_{\mathcal{G}_\theta(y^{\text{obs}})}(du),$$

where $\lambda_{\mathcal{G}_\theta(y^{\text{obs}})}$ is the intrinsic measure of the manifold (for details we refer to Liu *et al.*, 2022).

In the frequentist framework, inference on the parameter θ relies on comparing the sample with all possible outcomes of the model. Thus, the p-value is computed as

$$p - \text{val}(\theta, y^{\text{obs}}) = P\{t(g(u,\theta)) > t(y^{\text{obs}})\},$$

where a function of the data $t : \mathbb{R}^n \mapsto \mathbb{R}$ is introduced to reduce the dimensionality of the problem. A confidence interval of level α can be defined as

$$IC_\alpha(\theta, y^{\text{obs}}) = \{\theta : \exists\ u \mid y^{\text{obs}} = g(\theta, u), t(g(u,\theta)) \in B_\alpha\}, \tag{4}$$

where B_α is a set such that $P\{t(g(u,\theta)) \in B_\alpha\} \geq \alpha$.

There are already several established computational methods that make effective use of representations (3) and (4). A prominent example is the family of Approximate Bayesian Computation (ABC) algorithms, which focus on approximating posterior distributions by selecting parameters that yield simulated pseudo-data that closely resemble the observed data according to certain criteria (Sisson *et al.*, 2018).

Another category of methods for dealing with target distributions, represented in the forms (3) and (2) are Markov Chain Monte Carlo (MCMC) algorithms tailored to sampling on submanifolds (Brubaker *et al.* 2012, Zappa *et al.* 2018, Graham and Storkey 2017, Liu *et al.* 2022). Submanifolds arise, for example, when there is a lack of relaxation in the comparison between simulations and observed data in ABC, but also in conditional tests (see for example Diaconis *et al.* 2013 and Lindqvist *et al.* 2022). In these scenarios, the ability to recover the target distribution largely depends on the convergence of the Markov chains. By combining simulations with other techniques, such as grid search algorithms, and using representations similar to those of (4), a number of methods have recently been proposed to perform frequentist inference in non-regular environments, e.g. in the absence of the likelihood function (Dalmasso *et al.* 2021, Xie and Wang 2022, Wang *et al.* 2022a).

Main contributions of the thesis

The aim of this dissertation is to improve existing methods and introduce new simulation-based procedures tailored to solve some complex inference problems. Chapter 1 provides a general introduction to statistical inference based on different paradigms, accompanied by an overview of Monte Carlo methods and simulation-based inference.

Chapter 2 focuses on general simulation-based strategies for obtaining confidence distributions, confidence curves and confidence densities in non non regular settings,

where for instance Bootstrap methods are not directly applicable. Special attention is paid to the treatment of parameter vectors and nuisance parameters, to ensure invariance of the procedures under reparametrizations. The developed techniques are studied in the context of robust methods and estimating equations. Some extensions are considered for inference with non parametric tests, with probability semi-metrics. Possible applications are in non-inferiority tests and within the realm of likelihood-free inference.

In Chapter 3, we study a method in the context of Approximate Bayesian Computation (ABC) which is free from the choice of the tuning parameter. The resulting approximation implicitly uses a pseudo-likelihood that exhibits some consistency properties, and is linked to link to confidence distributions and data depth functions.

In Chapter 4, we derive and discuss coupling techniques for MCMC algorithms on submanifolds. They form the basis for the generation of convergence diagnoses and principal possibilities for the execution of parallel chains. In particular, we describe probabilistic reflection-contractive couplings and meeting-inducing couplings, placing the latter in the context of couplings for a broader class of MCMC algorithms with complex proposal mechanisms.

Chapter 5 presents two novel MCMC strategies developed for sampling generic target distributions on $\mathbb{R}^d$. These algorithms utilise ideas derived from MCMC algorithms on manifolds, incorporating geometric information from the target distribution in the problem at hand. In particular, equations specifically relevant to the sampling problem are used to define an artificial submanifold, such as the graph of the target distribution and the contour set.

Finally, Chapter 6 deals with the problem of performing Bayesian inference in the presence of an intractable matching prior distribution. This involves transitioning to a manifold characterized by an estimating equation that encompasses the derivatives of the said intractable matching prior distribution. An application in the context of Bayesian testing with e-values is presented, where the default prior guarantees the invariance properties of the procedure.

Chapter 1

Statistical Inference and Monte Carlo Methods

1.1 Distributions in statistics

A parametric statistical model is defined as a collection of probability distributions, represented in a general form as

$$\mathcal{M} := \{p(y|\theta), y \in \mathcal{Y}, \theta \in \Theta \subseteq \mathbb{R}^d\},$$

where y is a random variable in the sample space $\mathcal{Y}$ and θ is the parameter of the model at which the analysis is aimed.

In the context of the model $\mathcal{M}$, the inference process aims to quantify the uncertainty associated with the model's parameters given an observed sample, y^{obs}. Monte Carlo (MC) simulations help to assess and understand this uncertainty by generating an empirical distribution of interest for either y or θ and provide relevant information for inference.

The aim of this Chapter is to give an introduction to the different formalisms used in statistical inference, focusing on posterior distributions and confidence distributions. In addition, we give an overview of MC methods and simulation-based techniques for approximating these inference distributions.

1.2 Bayesian inference

In Bayesian inference, the process encompasses not solely outlining the model but also regarding the parameter space Θ as a probability space, thereby specifying an initial

distribution (prior) $\pi(\theta)$ for θ. Subsequently, after collecting data, this distribution gets updated using Bayes' theorem to form the posterior distribution

$$\pi(\theta|y^{\text{obs}}) = \frac{p(y^{\text{obs}}|\theta) \cdot \pi(\theta)}{p(y^{\text{obs}})}. \tag{1.1}$$

In (1.1) $p(y^{\text{obs}}|\theta) = \mathcal{L}(\theta)$ is the likelihood function, defined as the probability of the observed data y^{obs} conditioned on the parameter point θ. Drawing from the principles of probability theory, this paradigm offers a structured and coherent method to refine beliefs regarding parameters of interest. Another advantage lies in its ability to easily convey information. The normalizing constant in the denominator, termed *marginal likelihood*, requires an integral computation across the parameter space, that is

$$\int_{\theta} p(y^{\text{obs}}|\theta)\pi(\theta)d\theta.$$

The accessibility of the posterior distribution in a closed form is frequently restricted due to lack of analytical solution in either the prior, likelihood, or the integral (see among others Robert *et al.*, 2007) and this aspect prohibits a direct uncertainty quantification in probabilistic terms, through credible intervals or functionals of the form $\mathbb{E}_{\pi}[f(\theta)]$.

Monte Carlo methods and simulation-based techniques offer a workaround for deriving specific posterior distributions, especially when deterministic approximations as quadrature methods are become of difficult application in high dimension.

1.3 Confidence Distributions

Within frequentist inference, the concept analogous to the Bayesian posterior distribution is the Confidence Distribution (CD). The CD does not rely on the choice of a prior distribution, but unlike the Bayesian posterior, its form is not readily derived once the likelihood is computed. Conversely, it arises through a process grounded in repeated sampling principle: the uncertainty inherent the parameters is derived comparing observed data with the epistemic ordering of outcomes from the assumed data generating process $p(y|\theta)$, for given θ.

More precisely, a function $C(\cdot) = C(y, \cdot)$ on $\mathcal{Y} \times \Theta \to [0, 1]$ is called a *Confidence Distribution* for a parameter θ, if

(*i*) for each given $y \in \mathcal{Y}, C(y, \cdot)$ is a cumulative distribution function on Θ;

(*ii*) at the true parameter value $\theta = \theta_0$, $C(\theta_0) = C(y^{\text{obs}}, \theta_0)$, as a function of the sample y^{obs}, follows the Uniform$(0, 1)$ distribution.

A random variable ξ such that

$$\xi | y = y^{\text{obs}} \sim C(\cdot)$$

is called a CD *random variable* and its probability distribution function, called *confidence density*, is given by $\frac{\partial C(\cdot)}{\partial \xi}$. The CD random variable represents the uncertainty in the estimation of the parameter of interest, or can be seen as a random estimator of the parameter of interest. As a consequence of (i) and (ii), if $C(\theta)$ is a CD, $[C^{-1}(\alpha/2), C^{-1}(1-\alpha/2)]$ becomes an equi-tailed $1-\alpha$ confidence interval. In particular, a zero-level equi-tailed confidence interval is called the *confidence median* and denoted by $\tilde{\theta}$. The confidence median is median unbiased, implying that $P_{\theta_0}(\tilde{\theta} > \theta_0) = 0.5$. Furthermore, this estimator, as well as all equi-tailed confidence intervals are naturally equivariant under one-to-one reparametrizations.

A generalization of the CD, in cases where the monotonicity condition (i) does not hold, is given by the *confidence curve*, $cc(\theta) = cc(\theta, y)$ (Xie and Singh, 2013; Schweder and Hjort, 2016). If θ_0 is the true parameter point, then the random variable $cc(\theta_0) = cc(\theta_0, y)$ is designed to have a uniform distribution across the unit interval, i.e.

$$P_{\theta_0}(cc(\theta_0, y) \leq \alpha) = \alpha, \quad \text{for all } \alpha.$$

Thus confidence intervals can be read off, at each desired level. In regular cases, $cc(\theta)$ can be uniquely linked to a full confidence distribution $C(\theta) = C(\theta, y)$, via

$$cc(\theta) = |1 - 2C(\theta, y)| = \begin{cases} 1 - 2C(\theta, y), & \text{if } \theta \leq \tilde{\theta} \\ 2C(\theta, y) - 1, & \text{if } \theta > \tilde{\theta}. \end{cases}$$

Solving $cc(\theta) = 1 - \alpha$ yields two cut-off points for θ, precisely those of a $1 - \alpha$ confidence interval. This property allows to extract confidence intervals at any desired level. Furthermore, the confidence curve allows to identify confidence intervals when they are given by the union of disconnected regions and is well defined for multidimensional parameters.

1.3.1 Likelihood-based CDs

The standard theory for CDs evolves around the use of likelihood methods. This translates into establishing the order within the sample space through the utilization of data reduction summaries that are pivotal likelihood-based quantities. Under regularity conditions, the choice of such pivotal quantities simplifies the process of the construction

of Confidence Distributions and enables making epistemic-probabilistic statements with well-working large-sample approximations. Consider the partition of the d dimensional parameter $\theta = (\psi, \lambda)$, where ψ is a scalar parameter for which inference is required and λ represents the remaining $(d-1)$ nuisance parameters. If $\hat{\psi}$ is the Maximum Likelihood Estimator (MLE) of ψ, then the CD derived from the profile Wald statistic,

$$w_p(\psi) = \frac{\hat{\psi} - \psi}{\sqrt{j_p(\hat{\psi})^{-1}}}, \tag{1.2}$$

with $j_p(\psi)$ profile observed information, coincides with the asymptotic first-order Bayesian posterior distribution for ψ (see e.g. Ruli and Ventura 2021). A pivotal quantity that can reflect asymmetry and likelihood multimodality in the underlying distributions, unlike (1.2) is the log-likelihood ratio. Let $\ell(\theta) = \log \mathcal{L}(\theta)$ be the log-likelihood function for θ, and let $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ be the profile log-likelihood for ψ, where $\hat{\lambda}_\psi$ is the maximum likelihood estimator (MLE) for λ given ψ. The profile log-likelihood ratio $W_p(\psi) = 2(\ell_p(\hat{\psi}) - \ell_p(\psi))$, under mild regularity conditions, has an asymptotic null χ^2_1 distribution. Hence $\Gamma_1(W_p(\psi)) \dot{\sim} \text{Uniform}(0, 1)$, with $\Gamma_1(\cdot)$ denoting the χ^2_1 distribution function, and

$$C(\psi) \doteq \Gamma_1(W_p(\psi))$$

is a first-order asymptotic CD. Similarly, the profile likelihood root

$$r_p(\psi) = \text{sign}(\hat{\psi} - \psi)\sqrt{2(\ell_p(\hat{\psi}) - \ell_p(\psi))} \tag{1.3}$$

can be used to derive a first-order accurate CD, with error $O(n^{-1/2})$. Improved CD inference based on higher-order asymptotics (see, among others, Severini 2000, Reid 2003, Schweder and Hjort (2016, Chap. 7) and Ruli and Ventura 2021). One key pivotal quantity is the modified profile likelihood root, derived as refinement of (1.3)

$$r^*_p(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)} \log \frac{q_p(\psi)}{r_p(\psi)}, \tag{1.4}$$

where the quantity $q_p(\psi)$ is a suitably defined correction term (see, e.g., Severini 2000, Chapter 9). In practice, $r^*_p(\psi)$ allows to obtain asymptotically third-order accurate CDs, i.e. with error of order $O(n^{-3/2})$.

In some contexts, the absence of regularity conditions, the presence of intractable

likelihood functions, or small sample sizes related to the complexity of the model, prohibit the use of asymptotic pivotal quantities, which are not available or may not hold true. A way to obtain instead approximations of pivotal distributions and related confidence distribution is by relying on computer simulations, as Monte Carlo methods and resampling techniques.

1.4 Monte Carlo and simulation-based inference

Monte Carlo methods refer to procedures that aims at estimating integrals using finite samples from stochastic simulations (Devroye, 1985). For a generic quantity of interest, expressed as an expectation

$$I = \int f(x)\pi(x)dx = \mathbb{E}_\pi[f(x)] < \infty, \tag{1.5}$$

if $x_1, \ldots, x_R$ are independent realizations sampled from a distribution with density π, then the estimator $\hat{I} = R^{-1}\sum_{j=1}^{R} f(x_j)$ converges in probability to I as the number of stochastic realizations (R) increases. Furthermore, if the variance of I is finite, the estimator converges almost surely, at rate $R^{-1/2}$, from the Central Limit Theorem (see e.g. Robert and Casella 1999).

Using independent realizations is not the sole approach for estimating (1.5). Dependent realizations, particularly from Markov chains, can also be employed for this purpose, and under appropriate conditions the result is validated by the Ergodic Theorem. This extension gives rise to Markov Chain Monte Carlo (MCMC) methods. Despite their asymptotic validity, the determination of the convergence rate of dependent sampling methods is more complicated. Another important difference between the independent and dependent samplers lies in their parallelization potential. For independent samplers, the process of generating a sample of R replicates can be easily sped up by executing operations in parallel, while for dependent samplers this is less natural and requires *ad hoc* strategies (see e.g. Wang and Dunson 2013, Heng and Jacob 2019,Scott *et al.* 2022).

1.4.1 Independent samplers

Inversion generation

The Inverse generation method leverages the definition of the quantile function, $\Pi^{-1}(u) := \inf\{\Pi(x) \geq u\}$, i.e. the inverse of the cumulative distribution function

(CDF), to transform uniform random variables into variables following the desired distribution Π. If $u \sim \text{Uniform}(0,1)$, then $x = \Pi^{-1}(u) \sim \pi$. This method represents a basic form of an algorithmic model of form $y = g(u, \theta)$, in which the function g is replaced by the quantile function. Thanks to its generality, it allows easily sampling from univariate distributions, while extending this method to the multivariate case is possible in a limited number of problems, since it requires the complete knowledge of the CDF.

Rejection sampling

The Accept-Reject method, or rejection sampling, relies on a instrumental distribution with density $\tilde{\pi}$ to obtain samples from π. In its simplest form, Accept-Reject is rooted in the observation that the density π can be obtained marginalizing a uniform random variable in the the interval $[0, \pi(x)]$:

$$\pi(x) = \int_0^{\pi(x)} du.$$

Thus, sampling uniformly (u^*, x^*) and choosing x^* s.t. $u^* \sim \text{Uniform}(0, \pi(x))$ produces marginally samples with density π. If instead of sampling uniformly in the space, an instrumental distribution $\tilde{\pi}$ is chosen, the target can be written as

$$\pi(x) \propto \int_0^{M\tilde{\pi}(x)} \tilde{\pi}(v)dv,$$

provided that $M\tilde{\pi}(x) \geq \tilde{\pi}(x)$ for all x. The pseudo-code for this method is presented in Algorithm 1. Despite this method can be used to sample in d-dimensional space, it becomes inefficient in high dimensions due to the complexity of covering the distribution's support.

Algorithm 1 Accept-Reject Algorithm

Input: target density π, instrumental density $\tilde{\pi}$, M s.t. $M\tilde{\pi}(x) \geq \tilde{\pi}(x)$

for j in $1, \ldots, R$ **do**

Sample $x \sim \tilde{\pi}$ and $u \sim \text{Uniform}[0,1]$

  if $u \leq \pi(x)/M\tilde{\pi}(x)$ **then** accept x

  else reject

  end if

end for

Importance sampling and resampling

Importance sampling (Kahn and Marshall, 1953) involves generating independent random variables from an instrumental distribution $\tilde{\pi}$ and construct estimates for the test function (1.5) based on a weighted sample. The validity of the technique can be shown by rewriting the integral (1.5) as

$$I = \int f(x) \frac{\pi(x)}{\tilde{\pi}(x)} \tilde{\pi}(x) dx,$$

where $\tilde{\pi}(x)$ is any density function dominating $\pi(x)$. Note that importance sampling does not yield a sample directly distributed from $\pi(x)$. Nevertheless, it is possible to resample values from the instrumental distribution with multinomial weights proportional to the ratio $\frac{\pi(x)}{\tilde{\pi}(x)}$ the original sample from the instrumental distribution. This further step is called Importance resampling and is the basis of advanced techniques as Sequential Monte Carlo (Chopin, 2002; Del Moral *et al.*, 2007; Chopin and Papaspiliopoulos, 2020; Dai *et al.*, 2022).

1.4.2 Dependent samplers

Dependent samplers relate to Markov Chain Monte Carlo (MCMC) algorithms. The fundamental idea behind this class of algorithms is to design a Markov chain, $(x_t)_{t\in\mathbb{N}}$, i.e. a sequential processes where the future state depends solely on the present state, that converges in the time limit of the sequence, $t \to \infty$ to the target distribution $\pi(x)$. We recall some basic properties and conditions to ensure that the dependent sample drawn from a Markov chain can be used to estimate functions of type (1.5); for more details we refer to Robert and Casella (1999).

Markov chains

A *Markov kernel* on a topological space $\mathcal{X}$ is defined as a function $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$, where $K(x, \cdot) \in \mathcal{B}(\mathcal{X})$ for all $x \in \mathcal{X}$ and $K(\cdot, A)$ is a measurable function for every $A \in \mathcal{B}(\mathcal{X})$. A sequence of random variables $(x_t)_{t\in\mathbb{N}}$, represents a Markov chain with transition kernel K on X if $P(x_{t+1} \in A|x_1, \ldots, x_t) = P(x_{t+1} \in A|x_t) = K(x_t, A)$ for all $t \in \mathbb{N}$ and $A \in \mathcal{B}(\mathcal{X})$. If the transition kernel is the same for all t, the chain is said *homogeneous*. The evolution of homogeneous Markov chains can be completely described by the initial distribution π_0, and their Markov kernel K.

A necessary condition for a transition kernel K to have π as a limiting distribution is that π serves as an *invariant* or *stationary* distribution of K, that writes as $\pi(A) =$

$\pi(K(A)) = \int_X K(x, A)\pi(dx)$ for all $A \in \mathcal{B}(\mathcal{X})$. A condition which is easier to verify is the *detailed balance condition*, which implies that the Markov chain is *reversible*, $K(x, dy)\pi(dx) = K(y, dx)\pi(dy)$, where also $y \in \mathcal{X}$. The reversibility of a transition kernel K with respect to π implies that K has π as an invariant distribution, although the converse statement is not generally true. Finally, for the invariant distribution being unique, a chain must be irreducible, aperiodic and positive recurrent.

A chain is *irreducible* if starting at any point in X, there is a non-zero probability of moving to any set with positive measure, after a finite number of steps, i.e. there exists t such that $K_t(x, A) > 0$ for all $x \in X$ and for every $A \in \mathcal{B}(X)$ such that the set (A) is of positive probability under a dominating measure, where $K_t(x, A) = \int K_{t-1}(y, A)K(x, dy)$. A chain is *aperiodic* if state transitions do not occur in a strictly periodic manner, i.e. there do not exist cyclically ordered $s > 1$ subsets of $\mathcal{X}$, such that $K(A_i, A_j) = 1$, $j = i+1 \leq s$ and $K(A_s, A_1) = 1$. An irreducible Markov chain is said to be *recurrent* if for every set $A \in \mathcal{B}(X)$, of positive measure $E[\mathbb{1}_{X_t \in A}] = \infty$. Moreover, if the expected first return time to A is finite, the chain is said *positive recurrent*.

For a Markov chain (x_t) which is aperiodic, positive recurrent and has π as an invariant distribution, a CLT, called Ergodic Theorem, applies:

$$\frac{1}{N}\sum_{t=1}^{N} f(x_t) \xrightarrow[N\to\infty]{} I,$$

with I defined in (1.5), with asymptotic variance given bu

$$v(K, f) = \mathrm{Var}_\pi(f(X_0)) + 2\sum_{t=1}^{\infty} \mathrm{Cov}_\pi(f(X_0), f(X_t)) \quad (1.5),$$

where the subscript π stands for $X_0 \sim \pi$. This well-known expression results from simple calculations of $\lim_{t\to\infty} \mathrm{var}_\pi(T^{-1/2}\sum_{t=0}^{T-1} f(X_t))$.

1.4.3 Markov Chain Monte Carlo

Metropolis-Rosembluth-Teller-Hastings (MRTH)

The Metropolis-Hastings algorithm, as introduced by Metropolis *et al.* (1953), and further elaborated by Hastings (1970), operates by iteratively proposing new states for the Markov chain via a proposal distribution $q(x, dy)$ in form of a Normal centered at the current state and accepting these proposals with a carefully computed acceptance probability, α that involves the ratio of the target density at proposed and current states and the ratio of the proposal distributions from and to the states as summarized in the

pseudo-code of Algorithm 2. The transition kernel related to the procedure can be written as

$$K(x, dy) = \alpha(x, y)q(x, dy) + (1 - \alpha'(x))\delta_x,$$

where $\alpha'(x) = \int_X \alpha(x, y)q(x, dy)$ and δ_x is a Dirac at x.

Algorithm 2 Metropolis-Rosembluth-Teller-Hastings algorithm

Input: starting value $x^{(0)}$, proposal $q(, dx)$,target distribution $\pi(x)$
for j in $1, \ldots, R$ **do**
  Sample $x^* \sim q(x^{(j-1)}, dx^*)$, $u \sim \text{Uniform}(0, 1)$
  Compute the acceptance probability $\alpha(x, x^*) = \min\left(1, \frac{\pi(x^*)q(x^*, dx^{(j-1)})}{\pi(x^{(j-1)})q(x^{(j-1)}, dx^*)}\right)$
  if $u \leq \alpha$ **then**
    $x^{(j)} = x^*$
  else $x^{(j)} = x^{(j-1)}$
  end if
end for

Algorithm 2 can be extended to use other acceptance probabilities (see e.g. Barker 1965, Tierney 1998, Andrieu *et al.* 2020).

Gibbs sampler

The Gibbs sampler is a widely used technique for sampling from high-dimensional distributions (Verdinelli and Wasserman, 1991; Carlin and Gelfand, 1991). The idea is to separate blocks of components of the state space and update each of them, fixing the current values of the other components by exploiting the conditional distributions. It is therefore necessary to define a partition of variables and the corresponding tractable conditionals.

Depending on how the parameter space is partitioned, several Gibbs samplers can be defined for the same target distribution. A popular variant of the Gibbs sampler is the Metropolis-in-Gibbs method, in which some (or all) of the conditional updates are replaced by MRTH proposals and acceptance steps with a single block of components. If the partition of the blocks consists of singlets, as in algorithm 3, these are often referred to as *full conditionals*.

Algorithm 3 Gibbs sampling algorithm (systematic scan)

Input: starting value $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$, target distribution $\pi(x)$

for j in $1, \ldots, R$ **do**
  Sample $x_1^{(j)}$ from $\pi(x_1|x_{-1}^{(j-1)})$
  for $k = 2$ to d **do**
    Sample $x_k^{(j)}$ from $\pi(x_k|x_{-k}^{(j)})$
  end for
end for

Slice sampler

The slice sampler was introduced by Neal (2003) and serves for sampling from a target distribution without rejecting values, similarly to the Gibbs sampler. It operates in two steps: first, given a current state x_t, a random height u_t is chosen uniformly from the interval $[0, \pi^*(x_t)]$, where $\pi^*(x_t)$ represents the value of the (un-normalized) target distribution at x_t. This height creates a horizontal plane on the graph of the target distribution. Subsequently, within the region under the hyperplane at height $\pi^*(x_t)$, a new point x_{t+1} is sampled uniformly from the region where $\pi(x_{t+1}) \geq \pi(x_t)$, see Algorithm 4.

Algorithm 4 Slice sampler

Input: current state x_t, target distribution $\pi(x)$
for j in $1, \ldots, R$ **do**
  Sample $u_j \sim \text{Uniform}[0, \pi(x_j)]$.
  Define $S_j = \{x : \pi(x) \geq u_j\}$.
  Sample x_{j+1} uniformly in S_j.
  Update x_j to x_{j+1}
end for

Metropolis adjusted Langevin Algorithm

The Metropolis-adjusted Langevin Algorithm (MALA), pioneered by Ermak (1975) and Doll and Dion (1976) integrates the Langevin dynamics proposal into the Metropolis-Hastings framework with the goal of proposing new states that match the local geometry of the target distribution and aid in the efficient exploration of high-dimensional spaces (see also Roberts and Tweedie, 1996). This involves leveraging gradient information from the log-target distribution, to obtain a proposal kernel of the form $q(x, dy) = \text{Normal}(x + \frac{1}{2}\sigma\nabla \log \pi(x), \sigma^2)$, to be used in Algorithm 2.

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) (Duane *et al.*, 1987; Neal, 1999; Betancourt *et al.*, 2017) represents an advanced MCMC technique that leverages Hamiltonian dynamics to move and explore the space of interest. HMC introduces an auxiliary variable for each parameter, called the velocity, v, creating a joint system with potential and kinetic energy functions. This results in the Hamiltonian function:

$$H(x, v) = -\log \pi(x) + \frac{v^T M^{-1} v}{2},$$

where M is the Hamiltonian mass matrix associated with the target distribution π.

The system evolves according to Hamilton's equations:

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H}{\partial v} = \frac{\partial H}{\partial v} = M^{-1} v, \\ \frac{dv}{dt} &= -\frac{\partial H}{\partial x} = \frac{\partial \log \pi(x)}{\partial x}. \end{aligned}$$

Throughout this evolution, the total energy of the system, as governed by the Hamiltonian equations, remains constant, thus adhering to the principle of energy conservation, while the determinant of the transformation is 1, thus the distribution $\exp\{-H(x, v)\}$ is preserved, and consequently $\pi(x)$. In practice, the integration is performed numerically, for a discretization step, η, called *timestep*. The resulting position x_K after K integration steps is used as a candidate proposal for the next state of the chain, before being accepted or rejected according to a Metropolis-Hastings type of mechanism. The mechanism for generating a proposal, using the leapfrog integrator, is summarized in Algorithm 5.

Algorithm 5 Hamiltonian Monte Carlo with leapfrog integrator

Input: Current state x, potential $H(x)$, timestep η, leapfrog steps K.

for j in $1, \ldots, R$ **do**

set $x_1 = x_j$

Sample $v \sim \text{Normal}(0, M)$

Simulate the Hamiltonian dynamics for $k \leq K$ using the leapfrog integrator:

$$x_{k+1} = x_k + \eta v_k - \frac{\eta^2}{2}\nabla H(x_k)$$
$$v_{k+1} = v_k - \frac{\eta}{2}\nabla H(x_k) - \frac{\eta}{2}\nabla H(x_{k+1}),$$

Use x_K as the proposal for the next state of the Markov chain.

end for

1.4.4 Couplings of Markov chains

A coupling of two distributions π and $\tilde{\pi}$ refers to a joint distribution $\Gamma(\pi, \tilde{\pi})$ that preserves π and $\tilde{\pi}$ as its marginals. Coupling methods in the MCMC context have historically played a fundamental role in the study of theoretical convergence properties. A key concept to describe convergence is the distance of total variation (TV). The TV distance between distributions π and $\tilde{\pi}$ is defined as

$$\|\pi - \tilde{\pi}\|_{\text{TV}} = \sup_A |\pi(A) - \tilde{\pi}(A)|,$$

and corresponds to the maximum difference in probabilities assigned to any event by the distributions π and $\tilde{\pi}$. It can also be expressed as follows:

$$\|\pi - \tilde{\pi}\|_{\text{TV}} = \inf_{(x,y)\in\Gamma(\pi,\tilde{\pi})} \{P(x \neq y)\}, \tag{1.6}$$

where the infimum is reached in accordance with the maximum coupling, i.e. a coupling $\Gamma^{\max}(\pi, \tilde{\pi})$ where the probability that $P\{x = y\}$ when $x \sim \pi$ and $y \sim \tilde{\pi}$ is maximum. Couplings for Markov chains involves creating a joint (Markovian) process (x_t, y_t), where x_t and y_t represent states of the chains at time t, and where the process evolves such that the marginal chains target the assigned distributions $(\pi, \tilde{\pi})$. A coupling is successful when there exists a random variable $\tau \geq 1$ such that $x_t = y_t$ for $t \geq \tau$, indicating that the chains meet and remain together or *faithful* after meeting. From a practical point of view, couplings can also be implemented and provide useful and versatile methodological tools for monitoring convergence and for post-processing MCMC outputs. Among

others, Jacob *et al.* (2020) have shown how to successfully construct coupled kernels for various MCMC algorithms for which meeting time τ has a finite expected value.

Unbiased estimator

One potential application of constructing couplings for MCMC algorithms is to compute an unbiased estimate of a functional, as described in Equation 1.5. This possibility was demonstrated by Glynn and Rhee (2014). Writing the expectation as a telescopic sum, for all $k \geq 0$,

$$\mathbb{E}_\pi[f(x)] = \lim_{t\to\infty} \mathbb{E}[f(x_t)] = \mathbb{E}[f(x_k)] + \sum_{j=1}^{\infty} \mathbb{E}[f(x_{k+jL}) - f(x_{k+(j-1)})],$$

since for all $t \geq 0, x_t$ and y_t have the same distribution, the equivalence can be expressed as

$$\mathbb{E}[f(x_k)] + \sum_{j=1}^{\infty} \mathbb{E}[f(x_{k+j}) - f(y_{k+(j-1)})].$$

Finally, by exchanging of the expectation and the limit,

$$\mathbb{E}\left[f(x_k) + \sum_{j=1}^{\infty}(f(x_{k+j}) - f(y_{k+(j-1)}))\right], \tag{1.7}$$

for which it is natural to derive an unbiased estimator from two coupled chains, x_t and y_t, run for a finite time. Indeed, for $t > \tau$ the contributions in the sum (1.7) will be null.

Convergence

A second important result on couplings of Markov chains is related to convergence diagnostics. Under suitable assumptions, exploiting triangular inequalities from 1.6 Biswas *et al.* (2019) show

$$\|\pi_t - \pi\|_{\mathrm{TV}} \leq \mathbb{E}\left[\max\left(0, \left\lceil \frac{\tau - \ell - t}{\ell} \right\rceil\right)\right],$$

where ℓ denotes a delay between coupled chains and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. These upper bounds yield estimates for the total variation distance and can be extended to also bound the 1-Wasserstein distance.

Asymptotic variance estimation

Consider a function of interest f and a function g that satisfies the equation $(I - K)g = f - \pi(f)$, where K is the Markov operator and I is the identity. If f and g belong to $L^2(\pi)$, the following relations hold with the asymptotic variance $v(K, f)$ of a Markov chain,

$$v(K, f) = \mathbb{E}_\pi[(g(x) - Kg(x_0))^2] = 2\pi((f - \pi(f)) \cdot g) - \pi((f - \pi(f))^2) = \tag{1.8}$$

$$= -v(\pi, f) + 2\pi((f\pi(f)) \cdot g). \tag{1.9}$$

where $v(\pi, f)$ is the variance of the function f for $x \sim \pi$. In particular, considering as the g function $g = \sum_{t=0}^{\infty} K^t f(x_t) - K^t f(y_t)$, which can be estimated by $\hat{g} = \sum_{t=0}^{\tau-1} f(x_t) - f(y_t)$, with x_t and y_t coupled chains, and combining it in (1.8) with an estimator of $\pi(f)$, one can derive the estimator of the asymptotic variance, as shown by Douc *et al.* (2022).

1.5 Resampling methods

Bootstrap procedures were first introduced in 1979 by Efron (Efron, 1979). This resampling-based methodology offers a potent way to quantify uncertainty under the frequentist framework, even when traditional assumptions about data distribution are uncertain or violated. Bootstrap methods are particularly valuable for scenarios involving skewed data, distributions with heavy tails, or cases where the sampling distribution is unknown. The core idea is the following: instead of relying solely on the observed data, bootstrapping involves repeatedly resampling the original data with replacement. This creates numerous simulated datasets, each representing a plausible alternative to the original sample. By examining how a statistic of interest varies across these Bootstrap samples, it is possible to gain insights into its sampling distribution and construct confidence intervals, perform hypothesis tests, and more. The procedure is characterized by two key phases, the first of which is the actual *bootstrapping phase*.

1. **Resampling - Bootstrapping**: A significant number of Bootstrap samples are drawn from the observed data set using a resampling mechanism.

2. **Statistic Computation**: For each Bootstrap sample generated, the statistic of interest is calculated. This statistic can include measures of central tendency, dispersion or parameter estimates for the model under consideration.

When iteratively computing statistics of interest for these synthetic datasets, the goal is to approximate the true sampling distribution associated with these specific statistics. The final approximation is called the Bootstrap distribution, which in turn functions as an approximate confidence distribution. Bootstrap procedures can be divided into two main families depending on how the resampling phase is performed: non-parametric and parametric methods. The choice between these two families of Bootstrap methods depends on the specific characteristics of the data and the analytical goals, with parametric Bootstrap offering the advantages of complete model-based inference and non-parametric Bootstrap excelling for more flexible and distribution-free settings.

1.5.1 Non-parametric Bootstrap

The non-parametric Bootstrap makes no *a priori* assumptions about the underlying data distribution. It is a model-free approach that relies solely on the observed data. and is basically quite simple and at the same time profound and versatile. It involves the crucial step of generating numerous surrogate datasets through resampling with replacement from the observed data, see Algorithm 6. This process mimics the stochastic nature of sampling and captures the inherent variability of the data. The elegance of the non-parametric Bootstrap lies in its ability to transform empirical data into a self-contained universe of possible realizations.

Algorithm 6 Non-Parametric Bootstrap

1: **Input:** Observed data $y^{\text{obs}} = \{y_1, y_2, \ldots, y_n\}$, number of resamples B.
2: **for** $b = 1, 2, \ldots, B$ **do**
3: Create a Bootstrap dataset y_b^* by sampling from y^{obs} with replacement n observations.
4: **end for**
5: **return** Bootstrap samples $y_1^*, y_2^*, \ldots, y_B^*$.

1.5.2 Parametric Bootstrap

The Parametric Bootstrap method operates under the foundational assumption that a specific parametric model characterizes the data under observation. In the parametric Bootstrap, the first step involves the estimation of model parameters based on the observed data. In the context of the assumed model $p(y|\theta)$, the parametric Bootstrap replaces the empirical distribution function with the plug-in estimator $p(y|\hat{\theta})$, where $\hat{\theta} = \theta(y_{1:n})$ generally represents the maximum likelihood estimator (MLE). Apart from this modification, the technique is very similar to its non-parametric counterpart in all

other aspects. Resampling is then performed from the fitted model using the estimated parameters, as summarised in Algorithm 7. This approach is advantageous when prior knowledge or theoretical considerations justify the assumption of a particular parametric distribution. The choice of the specific statistic of interest for deriving the Bootstrap

Algorithm 7 Parametric Bootstrap

1: **Input:** Observed data y^{obs}, parametric model $p(y|\theta)$, number of resamples B
2: Estimate the parameters of the parametric model using the original data: $\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta|y^{\text{obs}})$, where $\mathcal{L}(\theta|y^{\text{obs}})$ is the likelihood function.
3: **for** $b = 1, 2, \ldots, B$ **do**
4: Create a Bootstrap resample $y_b^* \sim p(y|\hat{\theta})$, where $p(y|\hat{\theta})$ is the fitted distribution
5: **end for**
6: **return** Bootstrap resamples $y_1^*, y_2^*, \ldots, y_B^*$

distribution has an impact in determining the precision and accuracy of the inference results, nonetheless these are influenced by the model and the underlying population distribution.

1.5.2.1 Improved Bootstrap

The selection of specific statistics, such as pre-pivoted quantities or bias-corrected versions of the Bootstrap, can yield confidence intervals with superior properties compared to standard percentile-based intervals, obtained by Bootstrap resamples. The common idea behind the construction of improved intervals is based on the observation that the percentile Bootstrap intervals have higher precision when the estimate has symmetric distribution properties.

t-Bootstrap. Let ψ denote the scalar parameter of interest, $\hat{\psi}$ the estimate derived from the original sample, and $\hat{\psi}^*$ an estimate obtained after Bootstrap resampling. Consider a monotone transformation of the parameter $\psi \mapsto m(\psi)$ and let $q(\psi, y) = (m(\psi) - m(\hat{\psi}))/\hat{\tau}$ an approximate studentized-pivot, where and $\hat{\tau}$ is a suitable estimate of the pivot standard deviation. Let $Q(\cdot)$ be the distribution function of $q(\psi, x)$. Then, a confidence distribution for the parameter of interest is $C(h(\psi)) = Q\left((m(\psi) - m(\hat{\psi}))/\hat{\tau}\right)$. When $Q(\cdot)$ is unknown, it can be estimated via bootstrapping. Let $m(\psi^*)$ and $\hat{\tau}^*$ be the result of bootstrapping, then the $Q(\cdot)$ distribution can be estimated as $\hat{Q}$, via bootstrapped values of $q^* = q(\psi^*, x^*) = (m(\psi) - m(\hat{\psi}^*))/\hat{\tau}^*$. The approximate CD is then

$$C_{t-\text{boot}}(\psi) = \hat{Q}\left(\frac{m(\psi) - m(\hat{\psi})}{\hat{\tau}}\right).$$

This Bootstrap method applies even when $q(\psi, y)$ is not a perfect pivot, but is especially successful when it is, because q^* then has exactly the same distribution $Q(\cdot)$ as $q(\psi, x)$.

Bias correction. Define $K_B[x] := P\{\hat{\psi}^* \leq x\}$, where P represents probability conditioned on the observed sample. A uncorrected $1-\alpha$ lower confidence interval derived from the Bootstrap can be obtained as $\psi_L = K_B^{-1}[\alpha]$. Indeed, suppose there exists a monotone increasing transformation m such that $\phi = m(\psi)$ follows a normal distribution centered around $m(\hat{\psi} + z_0)$. This normal distribution can be used to construct an unbiased confidence interval, subsequently adjusted through a back-transformation to achieve an almost-unbiased confidence interval. Given such an m, the $1-\alpha$ lower confidence bound for ψ becomes $\psi_{mL} = m^{-1}(m(\hat{\psi}) + z_0 + z_\alpha)$, with z_α denoting the α-th percentile of a standard $N(0,1)$ distribution, where z_0 is termed the "bias" associated with m. For K_B defined as above,

$$K_B[\hat{\psi}] = P\{(m(\hat{\psi}^*) - m(\hat{\psi}) + z_0) \leq z_0\} = \Phi(z_0),$$

with Φ denoting the standard Normal distribution. Consequently, $z_0 = \Phi^{-1}[K_B[\hat{\psi}]]$. Furthermore, for any $0 < \alpha < 1$,

$$1-\alpha = \Phi(-z_\alpha) = P\{(m(\hat{\psi}^*) - m(\hat{\psi}) + z_0) \leq z_\alpha\} = P\{\hat{\psi}^* \leq m^{-1}(m(\hat{\psi}) - z_0 - z_\alpha)\}.$$

Similarly, $K_B^{-1}[\alpha] = m^{-1}(m(\hat{\psi}) - z_0 - z_\alpha)$. This imply that a Bootstrap bias corrected (BC) confidence distributions is obtained back-transforming a standard normal z,

$$CD_{\text{BC-boot}} = K_B^{-1}[\Phi[2z_0 + z]].$$

While the error in Bootstrap confidence intervals is in most cases $O_p(n^{-1})$, it becomes of order $O_p(n^{-3/2})$ with bias correction procedures and for $t-$Bootstrap (see Diciccio and Romano 1988, DiCiccio and Efron 1996 and Chapter 7 of Schweder and Hjort 2016). .

1.5.3 Permutations

The Non-parametric Bootstrap method exhibits a close alignment with the concept of permutations, introduced by Fisher (Fisher, 1936) and further developed since then (Pitman, 1937, see e.g.). In permutation tests, data points are systematically rearranged in all enumerable ways to assess the variability in the statistic of interest. Non-parametric bootstrapping and permutations, are grounded in the same fundamental principle of empirical data resampling, offering a means to explore and quantify uncertainty without necessitating stringent parametric assumptions. These methods share a model-free

nature, meaning that their resampling procedures don't rely on specific parametric assumptions. This shared core concept underscores their versatility and wide applicability in a diverse array of statistical analyses.

A crucial distinction among the two methods is that Bootstrap tests rest upon the foundational assumption of independence among observations. Conversely, permutation tests adopt the less strict assumption of exchangeability. Thus, dependencies among observations are permitted, provided that the order of observations can be freely rearranged without affecting the essential statistical characteristics of the dataset. In asymptotic regime, permutation tests exhibit higher power than the non parametric Bootstrap, often equivalent to that of the most powerful parametric test Albers *et al.* (1978). For a broader discussion, which goes beyond the scope of the introductory Chapter, see Good (2004), Pesarin and Salmaso (2010) and references therein.

1.5.4 Monte Carlo beyond tractable models

Monte Carlo methods and Markov Chain Monte Carlo (MCMC) techniques are not exclusively confined to Bayesian inference when sampling posterior distributions. They also prove invaluable in scenarios where directly evaluating the likelihood of the model is intractable. This situation arises, for example, when the distribution assumed for the data is known only up to a normalizing constant, a value that is parameter-dependent. In such scenarios, Monte Carlo methods can be used to obtain point estimators and p-values.

In this context, Simulation-based inference and Likelihood-Free Inference approaches have arisen as a fundamental suite of techniques for models where it is possible to generate simulations $y \in Y$ at various parameter values $\theta \in \Theta$. This field has experienced significant expansion in recent times.These methods emphasize leveraging the same process responsible for producing observed data to generate pseudo-observations across various parameter and aim to approximate posterior distributions when computing the probability density function of the model is either computationally infeasible or intractable. In particualr, two primary families of approaches have been delineated within this domain: Synthetic Likelihood, which directly formulates a likelihood function, and Approximate Bayesian Computation (ABC), that relies mainly on measuring the difference between simulated and observed data. In the following sections, we will provide a more detailed description of these distinct approaches.

MCMC for p-values

Besag and Clifford (1989) describes simulation-based methods for generating p-values within a MCMC procedure. The *"parallel runs method"* or *"hub-and-spoke"* (Barber and Janson, 2022) consists of running a reversible Markov chain backward from a state $x^{(1)}$ for r steps using the transition kernel Q, leading to $x^{(0)}$. Then, starting from $x^{(0)}$, the chain is run forward for r steps, and this process is repeated $m-1$ times independently to obtain states $x^{(2)}, \ldots, x^{(m)}$ that are referred to as contemporaneous to $x^{(1)}$. These states $(x^{(1)}, \ldots, x^{(m)})$ have an exchangeable joint distribution, π. Therefore, the p-values are calculated under the assumption that $x^{(1)}$ comes from π. The rank of $u^{(1)}$ among $u^{(1)}, \ldots, u^{(m)}$ (where $u = u(x)$ is a test statistic function of x) is uniformly distributed, which allows the calculation of the standard p value. The parameter r must be chosen sufficiently large to effectively explore the state space.

A second method, called *"serial run method"* was also introduced: let $x^{(1)}$ be a random draw from π. In this case, a chain with a stationary distribution π is created and observations $y^{(1)}, \ldots, y^{(m)}$ are made at intervals of r steps, running the chain backward and forward. These observations are converted into values $u(y^{(1)}), \ldots, u(y^{(m)})$, which represent the test statistic. To create a legitimate p-value, the goal is to place $u(x^{(1)})$ at the dth position, where d is randomly drawn from a uniform distribution between 1 and m. If $u(y^{(d)}) = u(x^{(1)})$, then, marginally over d (but not conditionally), the rank of the observed test statistic m values would be uniformly distributed. This observed rank can be used as a valid p-value. In practice, this involves sampling d first and then running the chain forward from $y^{(d)} = x^{(1)}$ to obtain $y^{(d+1)}, \ldots, y^{(m)}$, and running it backward to obtain $y^{(d-1)}, \ldots, y^{(1)}$.

Monte Carlo MLE

The Monte Carlo maximum likelihood estimation (Geyer, 1991), aims to find the maximum likelihood estimate $(\hat{\theta})$ when the likelihood function, $p(y^{\text{obs}}|\theta) = \frac{h(y^{\text{obs}}|\theta)}{c(\theta)}$. is intractable for the presence of a normalizing constant, $c(\theta)$, defined as $c(\theta) = \int h(y;\theta)dy$, which, together with its derivatives, cannot be calculated. The estimation is performed by maximizing the ratio:

$$\hat{\theta} = \arg\max_{\theta \in \theta} \left(\ln \left(\frac{h(y^{(0)}|\theta)}{h(y^{(0)}|\bar{\theta})} \right) - \ln \left(\frac{c(\theta)}{c(\bar{\theta})} \right) \right),$$

where the following empirical mean can be used to approximate the ratio of the normalizing constant

$$\frac{c(\theta)}{c(\bar{\theta})} \approx \frac{1}{m} \sum_{t=1}^{m} \frac{h(y^{(t)}; \theta)}{h(y^{(t)}; \bar{\theta})}.$$

The latter estimator is recognizably based on importance sampling.

Indirect inference

One of the initial attempts to address inferential problems in absence of the likelihood function can be traced back to Gourieroux *et al.* (1993) and it revolves around the utilization of an auxiliary model denoted as $\mathcal{M}^*(\phi)$, where ϕ represents an auxiliary parameter. The procedure involves fitting this auxiliary model to the data to obtain a point estimator, denoted as $\hat{\phi}$. Subsequently, a binding function $\phi(\theta)$ connecting the original and auxiliary models is considered, and this mapping is estimated through simulation. This entails performing the following iterative process: for a given grid of parameter values ($\theta_j \in \Theta$), R datasets are simulated and auxiliary models are subsequently estimated. The outcomes of these model fittings for the same θ_j are then averaged, resulting in $\phi(\theta_j) = R^{-1} \sum_{r=1}^{R} \phi_r(\theta_j)$. To draw inferences, the aim is finding the parameter value within the grid that is most likely to have generated our estimated model. In other words, the parameter, $\hat{\theta}$ that minimizes a loss function, typically the Mahalanobis distance. The complete procedure is reported in Algorithm 8

Algorithm 8 Indirect Inference

Input: grid of values θ_j, $j = 1, \ldots, J$
for j in $1, \ldots, J$ **do**
  for $r = 1$ to R **do**
    Simulate dataset r
    Compute $\hat{\phi}_r(\theta_j)$
  end for
  Compute $\phi(\theta_j) = \frac{1}{R} \sum_{r=1}^{R} \hat{\phi}_r(\theta_j)$
  Compute $\hat{\theta} = \underset{j}{\operatorname{argmin}} D(\phi(\theta_j), \hat{\phi})$
end for

Variants, based on different estimators for $\hat{\theta}$ were suggested by Smith Jr (1993) and Gallant and Tauchen (1996), see (Sisson *et al.*, 2018, Chapter 7) for a review.

Approximate Bayesian Computation

ABC algorithms (Rubin 1984, Tavaré *et al.* 1997, Marin *et al.* 2012) operate by sampling values from the parameter spaces and focusing on values that are able to

generate simulated data close enough to the observed data. Commonly, a distance function which involves a set of summary statistics to reduce the dimensionality of observed and simulated data is employed for this comparison and the summary statistics are assumed to be informative for the model. Parameter values porposed are accepted if the distance is such that $\delta(t(y^*), t(y^{obs})) < \varepsilon$, for small ε. Parameter values otherwise are rejected. The resulting sample of θ^* obtained is drawn from an approximation of the posterior distribution $\pi(\theta|y)$, given by

$$\pi^{\text{ABC}}_{\text{rej}}(\theta|y) = \frac{\pi(\theta)p(t(y)|\theta)\mathbb{I}_{\delta(t(y),t(y^{obs}))<\varepsilon}}{\int_\Theta \pi(\theta)p(t(y)|\theta)\mathbb{I}_{\delta(t(y),t(y^{obs}))<\varepsilon}d\theta},$$

Variations and extension of the rejection sampler, based on techniques like Markov Chain Monte Carlo (Marjoram *et al.*, 2003), Sequential Monte Carlo (Toni *et al.*, 2009; Del Moral *et al.*, 2012), Gibbs-type Clarté *et al.* (2021) allow to achieve lower ε values with the same computational resources.

The resulting posterior inference depends on three types of errors or approximations: the finite number of simulations, the amount of tolerance ε together with the associated distance δ and finally the non-sufficiency of the summary statistics. To minimize the loss of information, one can gradually include additional elements in the set of summary statistics, such as higher order sampling moments (Fearnhead and Prangle, 2012). However, the increased number of summary statistics leads to the curse of dimensionality in the evaluation of distances (Blum *et al.*, 2013).

Recent alternatives to the evaluation of summary statistics are based on the direct comparison of distributions by distance metrics (Bernton *et al.*, 2019; Legramanti *et al.*, 2022) or divergence estimators derived by adversarial learning (Wang *et al.*, 2022b). .

Bayesian synthetic likelihood

The use of synthetic or surrogate likelihood (SL) functions for summary statistics represents an alternative to ABC, as introduced in Wood (2010) and further study in Price *et al.* (2018). This approach aims at approximating the likelihood function of the summary statistics by repeatedly simulating R independent samples from the same parameter θ^*, and fitting a multivariate Gaussian density, obtaining $\hat{\mu}(\theta), \hat{\Sigma}(\theta)$ as the sample moments of $t(y^*)$. The approximated posterior has form

$$\pi^{\text{SL}}(\theta|y) = \frac{\pi(\theta)N(t(y)|\hat{\mu}(\theta), \hat{\Sigma}(\theta))}{\int_\Theta \pi(\theta)N(t(y)|\hat{\mu}(\theta), \hat{\Sigma}(\theta))d\theta}.$$

The challenge posed by the curse of dimensionality in ABC, when dealing with distances computed in high dimensions, is somewhat alleviated by introducing a parametric form for the likelihood function. Non parametric alternatives approaches, that use kernel density estimation instead of the Gaussian fit can be more accurate than the SL when the summary statistics are low-dimensional, but tend to get worse as the dimension increases, see also Grazian and Fan (2020) and Drovandi and Frazier (2022).

Chapter 2

Confidence distributions computing

Simulation-based inference (SBI) plays a central role in modern computational statistics and across various scientific disciplines. This heightened interest can be attributed to the use of generative models, which aim to simulate data from complex mechanisms, as stochastic differential equations in epidemiological models or in ecology studies. In such cases, the data generating process defines the probabilistic model and in turns the likelihood function only implicitly. This restricts the application of statistical inference methods reliant on direct likelihood evaluation, hence why this setup is also referred to as likelihood-free inference (LFI).

The cornerstone of SBI is Monte Carlo simulation, which declinates, depending on the specific problem at hand into various techniques such as Bootstrapping (Efron 1979, DiCiccio and Efron 1996, Efron 2003) and permutation methods (Anderson and Robinson, 2001), Sequential Monte Carlo and Particle Filters (Chopin and Papaspiliopoulos, 2020, e.g.), Approximate Bayesian Computation (ABC) (Beaumont *et al.* 2002, Marin *et al.* 2012), Indirect Inference (II) (Gourieroux *et al.* 1993, Genton and Ronchetti 2003), Synthetic likelihood (Wood 2010, Price *et al.* 2018, An *et al.* 2020), Certain techniques, such as non-parametric Bootstrap and permutations, are essentially model-free. In contrast, others are fundamentally model-based.

This Chapter discusses the construction of confidence distributions, confidence curves and confidence densities with finite coverage properties through the use of simulation methods. The use of simulations avoids the need to rely on approximations of pivotal quantities and thus on asymptotic assumptions about the size of the data. The method is simple and can be applied to both regular models and less regular situations as it is likelihood-free.

We discuss some issues such as the choice of summary statistics, invariance under reparametrizations, covering cases with a scalar parameters of interest as well as scenarios with nuisance parameters and parameter vectors.

2.1 Confidence distributions, curves and densities computing

Confidence distributions (Xie and Singh 2013, Schweder and Hjort 2016, Hjort and Schweder 2018) are a complete tool for performing frequentist inference, as they can summarize all inference results for a parameter of interest based on an assumed parametric model. They provide point estimates and allow the assessment of their accuracy for testing hypotheses. Similar to posterior distributions of the Bayesian framework, confidence distributions convey the set of confidence intervals at an arbitrary level, and include all intervals that hold the specified confidence level, along with measures of evidence for fixed intervals in parameter space, and finally allow comparison of inference results for the parameter of interest with results from multiple analyses.

Consider a sample $y = (y_1, \ldots, y_n)$ of size n from a random variable Y with assumed parametric model $f(y;\theta)$, indexed by a $d-$dimensional parameter θ. Let $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of primary interest and λ represents the remaining $(d-1)$ nuisance parameters. A recent definition of a confidence curve $cc(\psi) = cc(\psi, y)$ for ψ can be found, among others, in Xie and Singh (2013). Let $\theta_0 = (\psi_0, \lambda_0)$ the true parameter point. Then, the random variable $cc(\psi_0) = cc(\psi_0, Y)$ should have a uniform distribution on the unit interval and

$$P_{\theta_0}(cc(\psi_0, Y) \leq \alpha) = \alpha, \quad \text{for all } \alpha.$$

Thus confidence intervals can be read off, at each desired level. When α tends to zero the confidence interval tends to a single point, say $\tilde{\psi}$, the zero-confidence level estimator of ψ or confidence median. In regular cases, $cc(\psi)$ is decreasing to the left of $\tilde{\psi}$ and increasing to the right, in which case the confidence curve $cc(\psi)$ can be uniquely linked to a full confidence distribution $C(\psi) = C(\psi, y)$, via

$$cc(\psi) = |1 - 2C(\psi, y)| = \begin{cases} 1 - 2C(\psi, y), & \text{if } \psi \leq \tilde{\psi} \\ 2C(\psi, y) - 1, & \text{if } \psi \geq \tilde{\psi}. \end{cases}$$

With $C(\psi)$ a CD, $[C^{-1}(\alpha/2), C^{-1}(1-\alpha/2)]$ becomes an equi-tailed confidence interval

of level $1-\alpha$. Also, solving $cc(\psi) = 1-\alpha$ yields cut-off points for ψ, identifying the extremes of a $1-\alpha$ confidence interval. Finally, by differentiating the CD, the confidence density, $cd(\psi)$ is obtained, where also point estimators can be easily read off, see Figure 2.1 for an illustration.

A general recipe to derive a CD is based on the inversion of a pivotal quantity. Suppose $q(\psi; y)$ is a monotone increasing function in ψ, with a distribution not depending on the underlying parameter, i.e. $q(\psi; y)$ is a pivot (Barndorff-Nielsen and Cox, 1994). Thus $Q(x) = P_\theta(q(\psi; Y) \leq x)$ does not depend on ψ, which implies that

$$C(\psi) = Q(q(\psi; y)) \tag{2.1}$$

is a CD. The corresponding confidence density for ψ is

$$cd(\psi) = \frac{\partial Q(q(\psi; y))}{\partial q(\psi; y)} \frac{\partial q(\psi; y)}{\partial \psi}.$$

If the natural pivot is decreasing in ψ, then $C(\psi) = 1 - Q(q(\psi; y))$.

In the preface of their book, "Confidence, probability and likelihood", (Schweder and Hjort, 2016) state "The price to be paid for an epistemic distribution not based on a prior is that in most models only approximate confidence distributions are available, and they might be more computationally demanding than the Bayesian posterior". Indeed, exact confidence intervals are analytically available just for some specific models for which a closed form for the pivotal distribution is available. In most cases, while higher order approximations of pivotal quantities may be derived (see e.g. Ruli and Ventura 2021) an exact pivot does not exist. Thus, for understanding the behaviour of the designed statistic, simulations are generally appealed.

In practice, retrieving a continuous function on the parameter space that serves as CD in a simulation-based setting is perceived as a difficult task, analogous to that of computing confidence intervals. Indeed, while obtaining a Monte Carlo p-value is straightforward, by simulating synthetic realizations y_r^* $(r = 1, \ldots, R)$ from the model under the null hypothesis, $\mathcal{H} : \psi = \psi_0$, with the auxilium of a statistic $t(\cdot)$ through

$$p - \text{val}(\psi_0) = \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}_{\{t(y_r^*) > t(y^{\text{obs}})\}},$$

constructing confidence requires thorough computation. The same issue is encountered in constrained Bootstrap (Diciccio *et al.* 2001, Lee and Young 2005), where a distinct fitted distribution at each potential parameter value for the parameter of interest ψ is

obtained, while for nuisance parameters constrained estimators $\hat{\lambda}_\psi$ are used. For such procedure there is no immediate way to obtain confidence intervals, but some computational solutions have been proposed, including the use of stochastic search techniques, such as the Robbins-Monro method (see Lee and Young 2005 and reference therein). Another possibility stands in using the Neyman construction of confidence intervals: for a given significance level, tests across the parameter space are conducted, and intervals encompass the values for which the null hypothesis is not rejected. This method was recenlty adopted in Likelihood-free setup, in particular Dalmasso *et al.* (2021) and Masserano *et al.* (2022) propose to train a classification machine learning algorithm to calibrate a testing machinery for a fixed significance α, and then obtain confidence intervals with this procedure. One drawback there is that the entire training phase must be repeated to obtain confidence intervals for all levels, which is computationally extremely demanding. In the context of simulation-based inference, a recently proposed method, called Repro sampling (Xie and Wang, 2022), diverges from directly using hypothesis tests with external calibration. The method employs a general mapping functions based on an algorithmic model representation and a grid search algorithm is run along with the computation of the empirical distribution of the mapping functions, for each significance level.

Similarly, when the aim is deriving a CD, an alternative procedure (see e.g. Garcia-Angulo and Claeskens 2022) involves computing p-values for multiple values within a parameter space range and afterwards interpolating the results. This implies that for a fixed computational budget it is required to allocate a number of simulations for each parameter value, (N_θ) and a number of parameters ($|\Theta^*|$) to be considered. Furthermore, the interpolation step might not be easy in more than one dimension.

2.2 Monte Carlo based CDs

In contrast to the approaches mentioned in this introduction, we propose to use a rejection sampler to build Monte Carlo-based confidence distributions. The algorithm we consider is strongly inspired by Approximate Bayesian Computation (ABC) methods, (Marin *et al.*, 2012, among others). The proposed scheme, allows to align with the underlying probabilistic framework while constructing confidence distributions, while interpolation methods and stochastic search techniques used for constructing intervals in this setting don't completely leverage the nature of the problem. Also, the difference among ABC and the methodology studied here mirrors the difference between Bayesian and frequentist inference. The Bayesian inference process aims at ordering the parameter

space by level of agreement with the observed data, directly using the likelihood function. Consequently, the ABC posterior relies on the probability or likelihood of of obtaining simulated datasets that match (or closely resemble) the observed data,

$$\pi^{\text{ABC}}(\theta|y^{\text{obs}}) \propto Pr(y^* = y^{\text{obs}}|\theta), \text{ if } y^* \sim p(y|\theta), \theta \sim \pi(\theta). \tag{2.2}$$

In the frequentist approach, inference reflects the stochastic ordering of the sample space with respect to the assumed null hypothesis, reducing the data through a statistic, $t(y)$. The CD is therefore naturally linked to a p-value function, that writes

$$CD(\theta) = p - \text{val}(\theta) = Pr(t(y) > t(y^{\text{obs}})|\theta). \tag{2.3}$$

Note that (2.3), the event of interest has a positive probability mass, thanks to the inequality, while in (2.2) the probability of such occurrence is zero, except when the sample space is discrete, thus the equality is typically only approximately satisfied.

As rejection sampling can be used to approximate the posterior in ABC(2.2), the same technique can be successfully used to estimate a function proportional to the CD (2.3), see Algorithm 12. The other difference between the two settings lies in the use of the proposal distribution, which in the second case can be a uniform function in a sufficiently large parameter range without the need for transformations after changing the parametrization. With the confidence distribution, it should be noted that the density estimated by sampling must be monotonic, as the result is otherwise proportional to a confidence curve. Since it is also expected that $\sup C(\theta) = \sup cc(\theta) = 1$, the normalized function, i.e. the function in the right co-domain $(0, 1)$, can be recovered from the density estimate by calculating the empirical supremum of the density estimate and normalizing the function by this amount. If the estimated functional satisfies the monotonicity condition (ii), a confidence density can also be obtained. We propose two computational strategies for differentiating the CD, which are reported in Algorithms 10 and 11, respectively. The first procedure (10) is symmetric to the classical inverse generation method and relies on the inversion of the CD: starting from the estimated monotone CD, specifically form equi-spaced grid of values $G = \{\frac{j}{R}, j = 0, \ldots, R\}$, the corresponding parameter values, namely the quantiles of the distribution are matched to obtain CD-random variables. The second (11) directly operates differentiating the function at the density estimates level based on a Gaussian kernel. This additional differentiation step isn't strictly essential for conducting inference. However, it enables the straightforward identification of the shortest possible interval among all feasible intervals. Similar to the Highest Posterior Density (HPD) intervals in the Bayesian

framework, this might be referred to as Higher Confidence Density (HCD) intervals.

Algorithm 9 Accept-reject confidence curve/distribution computing for a scalar parameter

Input: proposal $p(\psi)$, summary statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$.

for $j \in 1, \ldots, R$ **do**

Sample $\psi_j^* \sim p(\psi)$ and $y_j^* \sim f(y; \psi_j^*)$

Compute $t_j^* = t(y_j^*)$

Accept ψ_j^* if $t_j^* \geq t^{\text{obs}}$ else reject

end for

return ψ^* with density $\propto cc_R(\psi)$ a confidence curve/distribution

Algorithm 10 Confidence density via inversion

Input: ψ with density $C_R(\psi)$, grid of values $G = \{\frac{j}{R}, j = 0, \ldots, R\}$

Compute the density estimation of $\hat{C}(\psi)$,

Normalize $\hat{C}(\psi) = \hat{C}(\psi)/\max\{\hat{C}(\psi)\}$

for $j \in 1, \ldots, R$ **do**

Obtain the CD-random varaibles: $\psi_j^* = \hat{C}^{-1}(G_j)$

end for

return ψ^* distributed as $\widehat{cd}_R(\psi)$.

Algorithm 11 Confidence density via differentiation

Input: ψ with density $C_R(\psi)$, bandwith h, $\phi(\cdot)$ Gaussian density function

for $j \in 1, \ldots, R$ **do**

Compute $\widehat{cd}(\psi_j) = \frac{1}{h \sum 1\{\psi \neq \psi_j\}} \sum_{\psi \neq \psi_j} \frac{2(\psi - \psi_j)}{h} \phi(\frac{\psi - \psi_j}{h})$

end for

resample ψ^* with weights $\widehat{cd}(\psi_j)$

return ψ^* distributed as $\widehat{cd}_R(\psi)$.

2.2.1 Computational details

One observation on the choice of the proposal distribution for unbounded parameter space pertains the individuation of the proposal region. If the region chosen for the proposal is symmetric with respect to the median of the CD, the expected number of accepted values will be $R/2$. In practice, if the empirical proportion of rejections is far from $1/2$, the proposal mechanism can be improved focusing on a different region. This

also allows to easily identify the confidence median with precision of order $O(R^{-1/2})$. This precision, as in other Monte Carlo methods can be improved using quasi Monte Carlo methods to $O(R^{-1})$ in one dimensional settings (Robert, 1994).

Instead, for the two step procedure using linear interpolation, let us denote with Θ^* the set containing the values of the parameter of interest used for the simulations and its cardinality by $|\Theta^*|$. Then, for a fixed Simulation budget R it is needed to allocate a number of points as distinct values on the parameter space, $|\Theta^*|$ and for each value a number of Monte Carlo draws to obtain, N_θ, such that $R = |\Theta^*| \times N_\theta$. Pointwise, on the selected values $\theta^* \in \Theta^*$ the precision will be of order $O(N_\theta^{-1/2})$, while after a linear interpolation, the error is summed to that depending on the distance between the data points, which is $O((1/|\Theta^*|)^2) = O((N_\theta/R)^{-2})$.

Another observation concerns bounded parameter spaces. To maintain a consistent precision using the Kernel Density Estimation (KDE), we propose expanding the bounded space $\mathcal{B}$ by considering an interval $\mathcal{B} \pm \epsilon$ as the range from which proposal values are drawn. If these values are incompatible with the model, replacement should only occur when they are utilized in simulating the data. This approach ensures that the density estimation result at the boundaries of the parameter space reaches its asymptot and prevents the values from decreasing due to numerical limitations.

2.2.2 Choice of summary statistics

The validity of the CD, in terms of I type errors and empirical frequentist coverages does not depend on a specific choice of a summary statistic. However, the spread of the CD, is related to the variability of the statistic, and to the power of a related hypothesis test. Thus, sufficient summary statistics, when available, are to be preferred. In some parametric models, when these are not easily obtained, a more general strategy is using estimating equations,as proposed by Ruli *et al.* (2020). These can be written as $g(y^*, \psi^*)$, and can be compared to $g(y^{\text{obs}}, \hat{\psi})$, where ψ^*, y^* are simulated. This choice is particularly convenient with respect to the use of MLE since it allows to avoid running an optimization algorithm at each simulation. In absence of a parametric model available in closed form, either a relevant statistic elicited by expert, or an estimating equations of an auxiliary model can be used, similarly to the ABC context Beaumont *et al.* (2002), with the difference that only one statistic can be handled at a time for obtaining a CD.

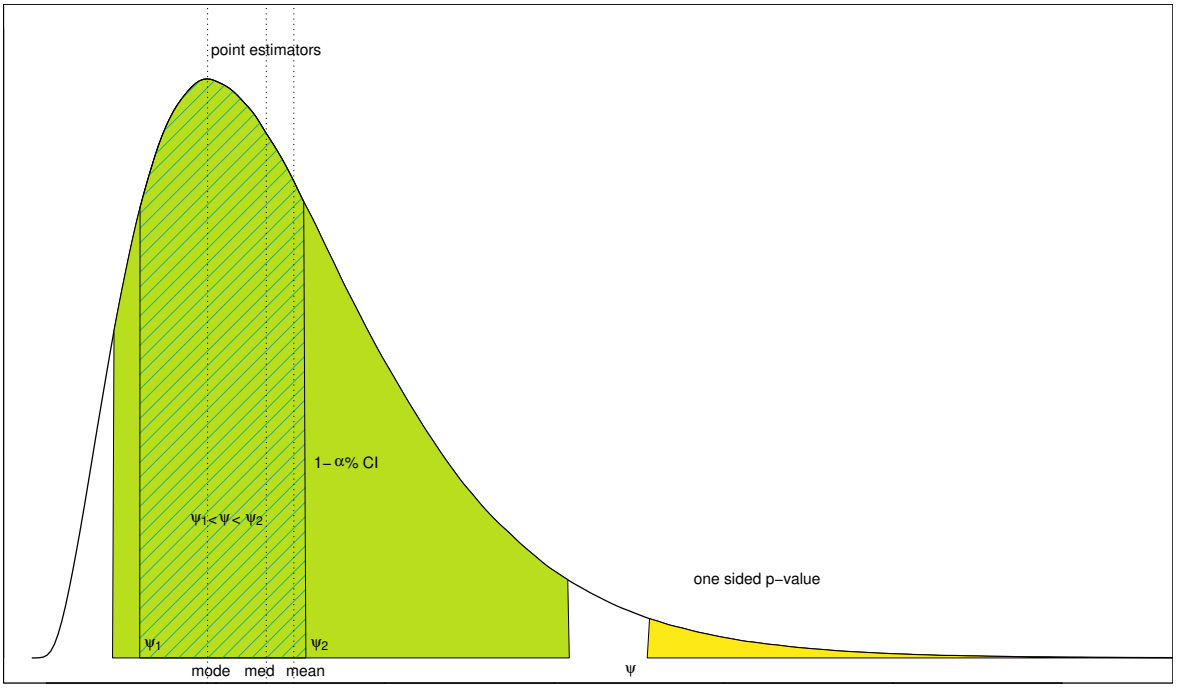

FIGURE 2.1: Illustration of inference summaries for a scalar parameter of interest ψ using a confidence density: point estimators (mode, median, mean), $(1-\alpha)\%$ quantile-type confidence intervals, one-sided p-value and measure of evidence for "$\psi_1 < \psi < \psi_2$".

2.3 Examples: scalar parameter

2.3.1 Bernoulli

Consider a simple example with a Bernoulli model, where $y \sim \text{Bernoulli}(\theta)$, $\theta \in [0, 1]$. The sufficient statistic for this model is the sum of the observations: $\sum_{i=1}^{n} y_i$. Inference with a small sample size in this context poses challenges. For instance, the empirical coverage of Wald-type intervals is known to be lower than the nominal coverage, while for other methods such as Clopper-Pearson intervals, it tends to be higher than the nominal (Gonçalves *et al.*, 2012). Additionally, the extremes of the intervals can fall outside the parameter space. Parametric Bootstrap is also unsuitable if the summary statistic assumes zero as a possible value. Despite the simplicity of the model, situations where the binomial distribution encounters zero observations have strong connection to to the issue of zero-total-event studies in odds ratio models in meta-analysis, which is difficult to handle.

In this simulation study 3×10^3 Monte Carlo replications were performed. In each experiment 10^4 proposals are drawn from an equi-spaced sequence in $[0, 1]$. In each experiment the true value of the parameter θ was set to 0.1.

We compared two types of confidence intervals: Highest Confidence Density (HCD) and quantile-based (Q) intervals. In each case, we computed the empirical length and coverage of 90% confidence intervals and evaluated the II type error rate of corresponding tests against specific alternative hypotheses: $\mathcal{H}_0^* : \theta_0 = 0.15$, $\mathcal{H}_0^{**} : \theta_0 = 0.2$.

Both methods controlled for the I-type error (false positive rate) as expected. However, HCD intervals were generally shorter, incorporating lower values and neglecting values on the right tail of the confidence density. This characteristic makes them more likely to reject false null hypotheses with true values greater than the assumed value under the null hypothesis, resulting in higher power compared to Q-type intervals.

It is important to note that HCD intervals are not invariant to transformations of the statistical model.

$\theta_0 = 0.1$	$n = 10$				$n = 20$			
	length	I type	II type		length	I type	II type	
			$\theta = 0.15$	$\theta = 0.2$			$\theta = 0.15$	$\theta = 0.2$
HCD	0.292	0.091	0.959	0.707	0.205	0.084	0.868	0.612
Q	0.326	0.089	0.961	0.987	0.224	0.102	0.891	0.868

TABLE 2.1: Length of 90% level confidence intervals, I and II type errors for the Bernoulli model based on 3000 replications.

To assess the impact of non-invariance of Highest Confidence Density (HCD)-type intervals compared to quantile-based (Q-type) intervals under changes in parametrization, we employed the transformation $\phi = \log(\frac{\theta}{1-\theta})$.

We computed the sum of absolute errors (SAE) of the 90% confidence interval limits (lower: L, upper: U) calculated for the true value $\theta = 0.2$ and sample size $n = 20$. The SAE is obtained as $|\log(\frac{\theta_L}{1-\theta_L}) - \phi_L| + |\log(\frac{\theta_U}{1-\theta_U})$ -$\phi_U|$. This calculation was repeated over 100 simulations and results are reported in Table 2.2 (median and inter-quantile range). While both methods exhibit errors due to numerical limitations, the results indicate greater stability for quantile-based intervals compared to HCD intervals in this scenario.

sae	1st Qu.	Median	3rd Qu.
HCD	0.61	0.65	0.68
Q	0.07	0.10	0.17

TABLE 2.2: Comparison of higher confidence density-type (HCD) and quantile-type (Q) intervals: impact on SAE for 90% confidence limits over 100 replications.

2.3.2 Uniform

We consider as a second example the Uniform model, $y \sim \text{Uniform}(\theta_1, \theta_2)$, and assume θ_2 known. For a non parametric bootstrap it is trivial to understand that the confidence intervals won't contain the true value, since the MLE for the minimum in the bootstrap replicates is at least the sample minimum, hence constantly biased. In the parametric case, the initial guess bootstrap is also biased for the same reason. In this case, the bias corrected intervals are not computable since the involved parameters are infinite. The percentiles bootstrap intervals are constantly 0. In table 2.3 we report a short simulation study that shows the poor performance of bootstrap within the situation described when computing 95% confidence intervals compared to the method proposed.

type	$n = 10$	$n = 50$
basic	90.91	94.61
norm	89.95	94.11
CD-HCD	95.56	95.53
CD-Q	95.29	95.19

TABLE 2.3: Coverage of bootstrap and CD-based intervals for the Uniform's minimum problem.

2.3.3 Sum of lognormals

Consider the sum of lognormals model, given by:

$$\bar{y} = n^{-1} \sum_{i=1}^{n} e^{z_i}, \quad z_i \sim Z \sim N(\mu, 1), \quad i = 1, \ldots, n,$$

and a draw from the model for $\mu = 0$, with $\bar{y}^{\text{obs}} = 1.439$, obtained by $n = 5$ realizations. When $n > 2$, an analytical form for the distribution of the random variable $\bar{y}$ is not available; thus, furnishing a confidence interval for μ, despite the formal simplicity of the model, is not straightforward in finite sample regime. We consider 10^5 proposal values for μ uniformly in the interval $(-4, 4)$ and obtain the CD based on $\bar{y}$ together with an approximate CD based on the Vanilla ABC algorithm suggested in Thornton *et al.* (2022) with tolerance $\epsilon = 0.01$ (Figure 2.2). Then, we run a simulation study in which we compute the 95% confidence intervals obtained again with CDs and with approximate likelihoods via ABC with a small tolerance ($\epsilon = 0.01$). The resulting empirical coverage based on 10^4 simulations and 10^4 proposal values for each of them is 94.9% (se = 0.2%), against ABC's 77.1% (se = 0.3%).

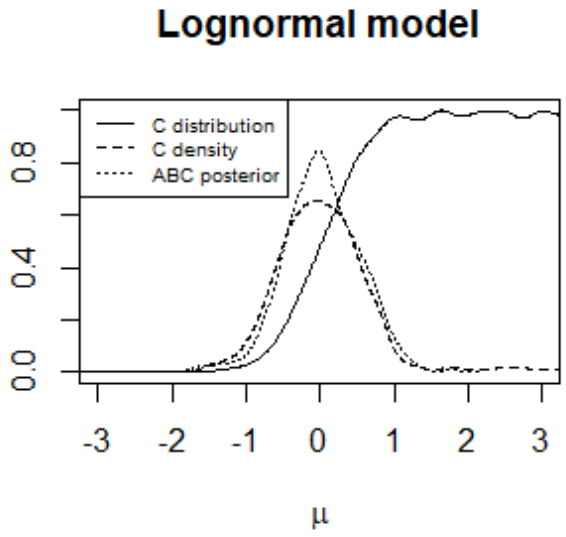

FIGURE 2.2: Confidence distribution, confidence density and ABC posterior based on a sample from the sum of lognormals model, with $\bar{y}^{\text{obs}} = 1.439$ and $n = 5$.

2.3.4 Application to fusion inference

One potential application of the proposed methodology is the so called *fusion inference* framework (Cunen and Hjort, 2022, and references therein), or *multiverse-analysis* (Steegen *et al.*, 2016) which aims at combining results obtained from different sources of information or competing methods to enhance the robustness of the analysis. Indeed, when inference is based on different data-reducing statistics, and confidence distributions are derived via asymptotic pivots, these extra levels of approximation introduce arbitrariness and results will depend also on the choice of the pivot. In contrast, when confidence distributions are obtained through simulations, the observed differences would primarily be attributed to the choice of the statistic utilized rather than being influenced by the quality of approximation for each pivot. We briefly illustrate how to combine confidence through Implied Likelihoods, a method suggested by Efron (1993). This involves the construction of a fictitious second dataset, doubling the original $y^{II} = (y_n, y_n)$, whose likelihood function is $\mathcal{L}(\theta; y^{II}) \propto \mathcal{L}(\theta; y_n)^2$. The implied likelihood is retrieved by

$$\mathcal{L}(\theta) = \frac{c(\theta; y^{II})}{c(\theta; y_n)}, \tag{2.4}$$

where $c(\theta; y)$ is the confidence density. In practice, it is possible to perform this step resampling the CD-random variables obtained from the analysis of the doubled dataset with importance weights given by the distribution of the CD-random variables obtained

on the original data. Importantly, the Implied Likelihood does not depend on the parametrization used or on the model. Other recombination strategies can be considered; see for instance Singh *et al.* (2005) and Liu *et al.* (2014).

Suppose three independent laboratories have obtained the same measurement for a quantity of interest from a sample of size $n = 18$. The objective is to combine these results to make inferences about a parameter μ, representing the mean of the phenomenon. The first lab reports as a summary of the data the mean ($S_1 = 31.04$). The second lab synthesises the results with the median ($S_2 = 30.69$), accounting for the possible influence of few outliers, while the third lab considered the logarithm of the measures, trying to correct for some observed asymmetric behaviour, and furnishes the mean of the logarithms ($S_3 = 3.43$). From the diverse summaries/sources, one can obtain the CDs and the implied likelihoods (IL) and combine them as components of a mixture, as exemplified in Figure 2.3. The figure shows a strong alignment between the full likelihood, available for the first lab's analysis, the implied likelihood of first and third labs.

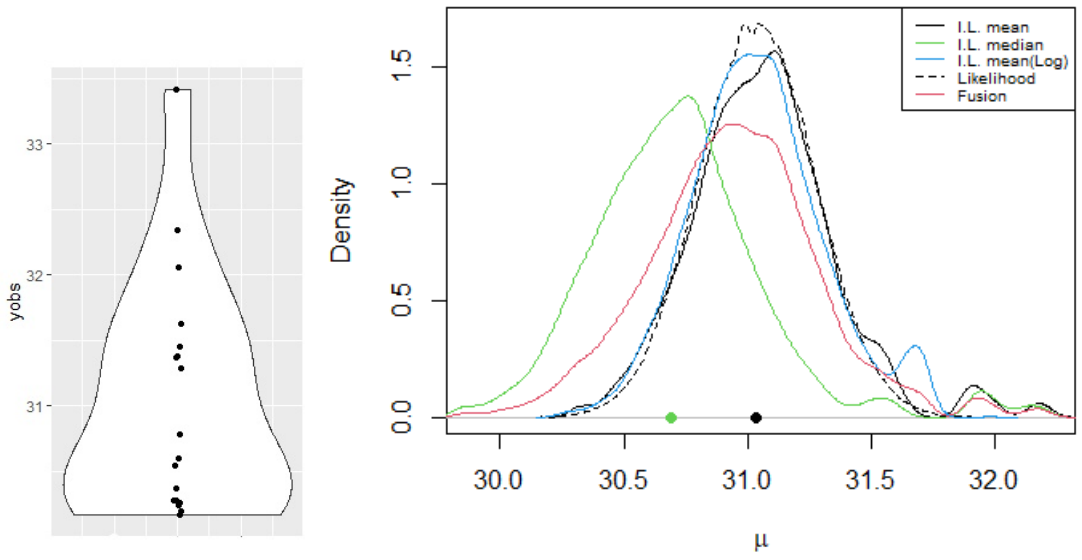

FIGURE 2.3: Fusion data example: original sample (*left*) and combined likelihood (red line), obtained by fusion of three ILs, related to different CDs (*right*). Black and green dots represent the statistics given by the first and the second lab (S_1, S_2), the third is not in the same scale of the x-axis ($S_3 = 3.43$).

2.4 Treatment of nuisance parameters

Algorithm 12 Accept-reject confidence curve/distribution computing with nuisance parameters

Input: proposal $p(\psi, \lambda)$, profile estimating equation $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$, where y^{obs} is the observed sample.

for $j \in 1, \ldots, R$ **do**
  Sample $(\psi_j^*, \lambda_j^*) \sim p(\psi, \lambda)$ and $y_j^* \sim f(y; \psi_j^*, \lambda_j^*)$
  Compute $t_j^* = t(y_j^*)$
  Accept ψ_j^* if $t_j^* \geq t^{\text{obs}}$ else reject
end for
return ψ^* with density $\propto cc_R(\psi)$ a confidence curve/distribution

In frequentist inference, there are several ways to handle nuisance parameters (Basu, 1975). Perhaps the two most common approaches are: 1) to plug-in nuisance quantities with constrained estimates, a technique known as *profiling*; 2) to integrate over the parameter space. Importantly, profiled quantities usually exhibit invariance properties under model reparametrization, making them a preferred choice. With the partition $\theta = (\psi, \lambda)$, the estimating equation is similarly partitioned as $g(y; \theta) = (g_\psi(y; \theta), g_\lambda(y; \theta))$, and a profile estimating equation for ψ has general form

$$g_\psi(y; \psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ represents the constrained estimate of λ for each value ψ considered. To obtain a confidence distribution in the presence of nuisance parameters, one solution is to consider a profile estimating equation as above and a corresponding acceptance rule

$$g_\psi(y; \psi^*, \lambda^*) > g_\psi(y^{\text{obs}}; \psi^*, \hat{\lambda}_\psi^*), \tag{2.5}$$

where we denote as $\hat{\lambda}_\psi^* = \hat{\lambda}_\psi(y^{\text{obs}})$ the constrained estimated nuisance parameter in the observed sample. This form can be referred to as the *profiled-plug-in method* and the CD obtained by (2.5) corresponds to

$$CD(\psi) \propto \int \int 1_{g_\psi(y;\psi,\lambda) > g_\psi(y^{\text{obs}};\psi,\hat{\lambda}_\psi)} p(y|\psi, \lambda) d\lambda dy.$$

The estimator $\hat{\lambda}_\psi$ does not depend on y, as it is a deterministic function of y^{obs} and ψ. Thus, if κ is a one-to-one transformation of the nuisance parameter, which is by a slight

abuse of notation $\lambda = \lambda(\kappa)$, the CD transforms as follows when the parameterization is changed

$$CD(\psi) \propto \int\int 1_{g_\psi(y;\psi,\kappa) > g_\psi(y^{\text{obs}};\psi,\hat{\kappa}_p si)} p(y|\psi,\kappa) \frac{\partial \lambda(\kappa)}{\partial \kappa} d\kappa dy.$$

In the latter expression, integrating out in y, one obtains a quantity, that is equivariant to κ or λ, since it is related to the distribution of the full score of the model where only ψ is unknown. Thus the CD is invariant.

Figure 2.4 shows the contours of the likelihood function and the CDs for the shape parameter of a Weibull model - i.e. $y \sim \text{Weibull}(\gamma, \beta)$ - for a sample of $n = 5$ realizations and for the shape parameter α of a generalized exponential distribution - i.e. $y \sim Gexp(\alpha, \lambda)$- with $n = 20$. For the parameters β and λ we considered the transformation $\mu = 1/\beta$ and $\nu = 1/\lambda$ for which the CDs overlap. We also show the comparison with a CD obtained for the Constrained Bootstrap, which is also invariant.

When the likelihood is intractable, the constrained maximization involved in the profiling operation is not feasible. In these cases it is possible to resort a parametric auxiliary model with a corresponding surrogate likelihood and estimating equation. In some cases, finally, the maximization is computationally expensive. For instance, for retrieving a solution with M-estimating functions, the numerical computation of an integral is required, thus instead of the constrained estimator, $\hat{\lambda}^*_\psi$, one can consider a generalized version with $\hat{\lambda}$ equal to the global maximizer.

2.5 Models with nuisance parameters: examples

2.5.1 Adjusted score function

We consider the problem of obtaining confidence intervals from modified score approaches, introduced by Firth (1993). One of the main disadvantages is related to the fact that obtaining Wald type intervals is that they can be outside of the parameter space. Also, the empirical coverages can be far from the nominal for precision parameters. The simulation-based approach does not rely on large sample approximation and allows to recover the exact distribution of the pivot. The adjusted score functions in a scalar parameter case are generally of form $U(\theta, y) + m(\theta)$, (Kosmidis *et al.*, 2020) where U is precisely the score function and $m(\theta)$ depends on the model but is data independent, involving expected values of second and third derivatives of the score. Thus, the sampling distribution of the adjusted score is identical to the sampling distribution of the modified score. When the parameter is multidimensional instead $\theta = (\psi, \lambda)$, the profiled version of the modified score function depends on data dependent estimates $\hat{\lambda}$

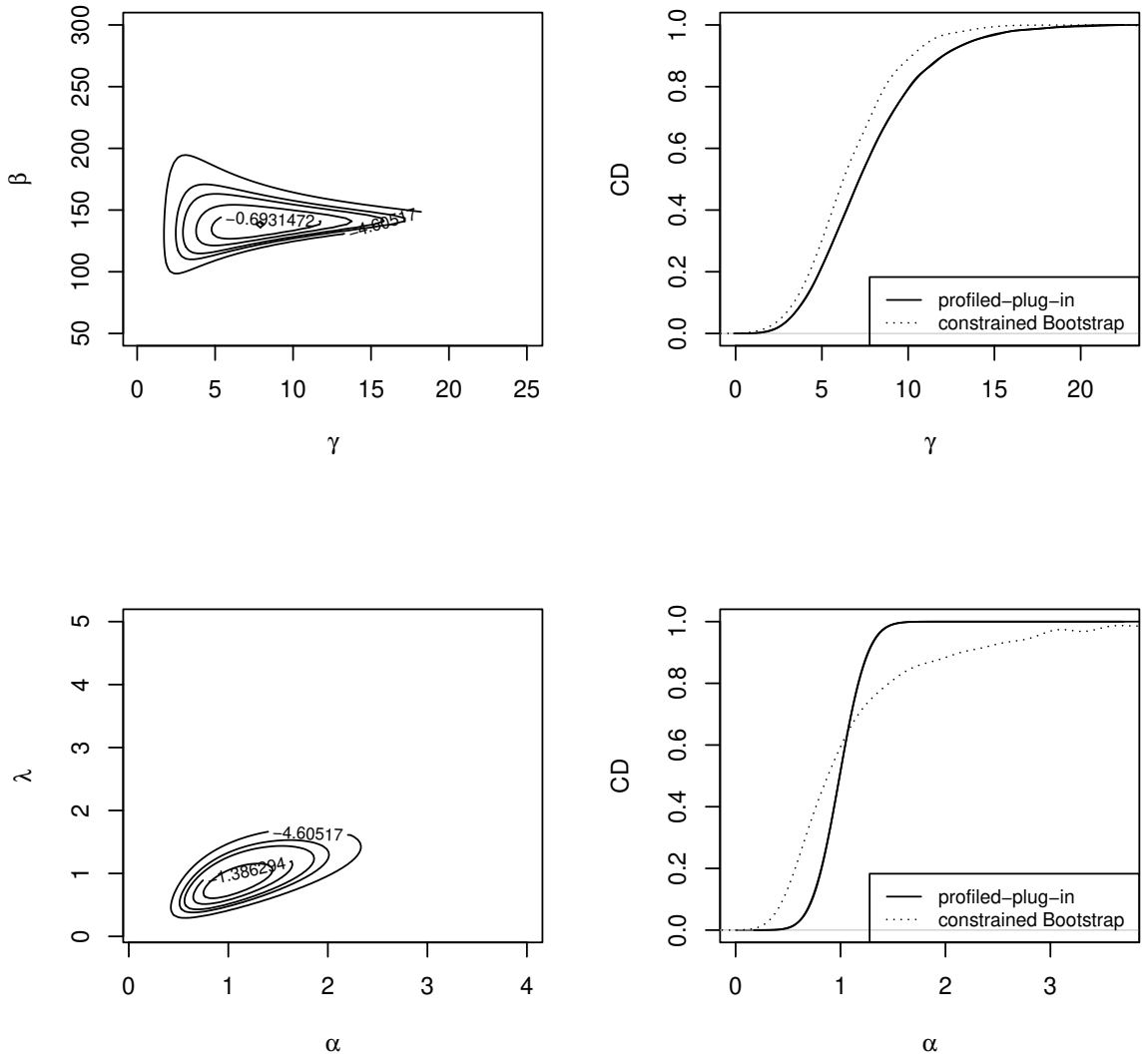

FIGURE 2.4: Contour plots for the likelihood function and CDs for the shape parameter of a Weibull (*above*) and Generalized exponential model (*below*) under transformation of nuisance parameters. CDs obtained after reparametrizations overlap with original.

thus the sampling distribution is different from that derived from the standard score function.

As an example, we consider a bivariate Gaussian regression model $y \sim N_2(\mu, \Sigma)$ where $\mu = X\beta$, $\beta \in \mathbb{R}^p$ and

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

We focus on the parameter ρ. The sample is of size $n = 40$, $p = 5$ regression coefficients are used, and the true parameter point is $\beta = (0.500.530.300.101.16)$, $\sigma^2 = 1$, $\rho = 0.9$.

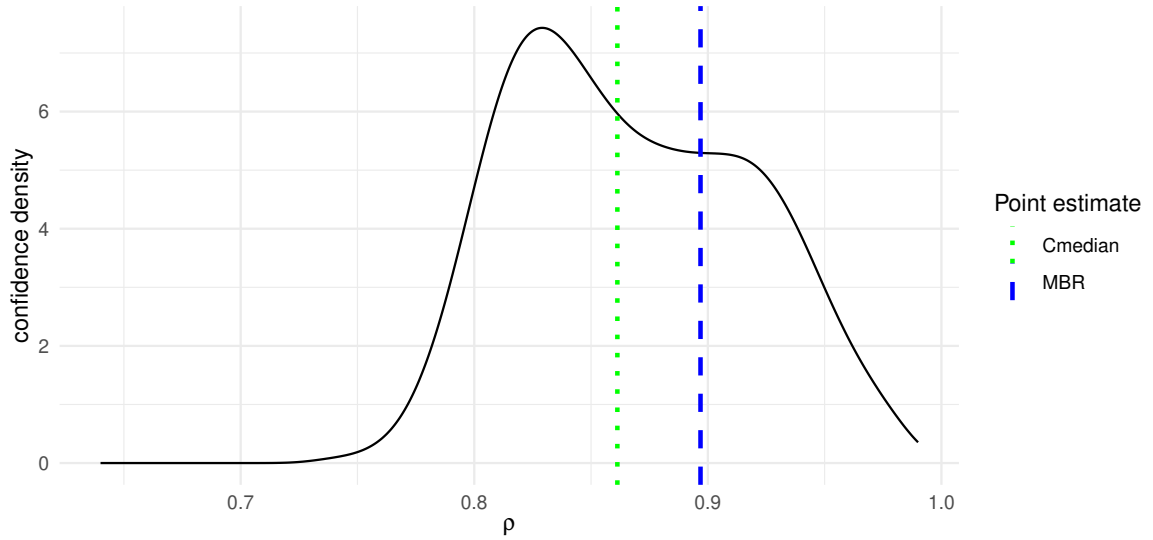

FIGURE 2.5: Adjusted score example: confidence density obtained from the profile modified score function.

The empirical means are $\bar{y} = (1.431.45)$ and the empirical covariance matrix is $S = \begin{bmatrix} 12.91 & 12.40 \\ 12.40 & 12.53 \end{bmatrix}$.

For obtaining the CD, we draw 20000 proposals uniformly in a hypercube centered on the point estimates and at least 4 standard deviations apart. In Figure 2.5 the confidence density with the confidence median and the point estimate obtained from the solution of the modified score is reported. The 95% equi-tailed confidence interval based on the CD is [0.78 0.96], while with the Wald approximation is [0.83 0.96]. The simulation-based method in this case is able to capture the asymmetry of the CD.

2.5.2 CD from M-estimating functions

The standard theory for CDs evolves around the use of likelihood methods for a scalar parameter of interest ψ of a parametric model. However, it is well-known that likelihood-based methods are not robust when the assumed distribution is just an approximate parametric model or in the presence of deviant values in the observed data. In this case, it may be preferable to base inference on procedures that are more resistant, that is which specifically take into account the fact that the assumed models used by the analysts are only approximate. In order to produce statistical procedures that are stable with respect to small changes in the data or to small model departures, robust statistical methods can be considered (see, e.g., Ronchetti and Huber 2009, Heritier

and Ronchetti 1994,Farcomeni and Ventura 2012 and references therein). In particular, we consider the class of M-estimators, that includes among others the maximum composite likelihood estimator (see e.g., Varin *et al.* 2011), estimators based on proper scoring rules (see, e.g., Dawid *et al.* 2016, and references therein), and classical robust estimators (see e.g. Ronchetti and Huber 2009). Under broad regularity conditions, an M-estimator $\tilde{\theta}$ is the solution of the unbiased estimating equation $g(\theta)$ and it is asymptotically normal, with mean θ and covariance matrix $V(\theta) = K(\theta)^{-1}J(\theta)(K(\theta)^{-1})^\top$, where $K(\theta) = E_\theta(\partial g(\theta)/\partial\theta^\top)$ and $J(\theta) = E_\theta(g(\theta)g(\theta)^\top)$ are the sensitivity and the variability matrices, respectively. The matrix $V_g(\theta) = V(\theta)^{-1}$ is known as the Godambe information and its form is due to the failure of the information identity since, in general, $K(\theta) \neq J(\theta)$. Let us denote with $G(\theta) = \sum_{i=1}^n G(y_i;\theta)$ the function such that $g(\theta)$ is the gradient vector, i.e. $g(y;\theta) = \partial G(y;\theta)/\partial\theta$.

From the general theory of M-estimators, the influence function (IF) of the estimator $\tilde{\theta}$ is given by

$$IF(y;\tilde{\theta}) = K(\theta)^{-1}g(y;\theta), \tag{2.6}$$

and it measures the effect on the estimator $\tilde{\theta}$ of an infinitesimal contamination at the point y, standardised by the mass of the contamination. The estimator $\tilde{\theta}$ is B-robust if and only if $g(y;\theta)$ is bounded in y. Note that the IF of the MLE is proportional to the score function; therefore, in general, MLE has unbounded IF, i.e. it is not B-robust.

Paralleling likelihood-based results, asymptotic robust inference on the scalar parameter of interest ψ can be based on first-order pivots, extending the theory of robust scoring rules discussed in Hjort and Schweder (2018) and Ruli *et al.* (2022). With the partition $\theta = (\psi, \lambda)$, consider the further partitions

$$K = \begin{bmatrix} K_{\psi\psi} & K_{\psi\lambda} \\ K_{\lambda\psi} & K_{\lambda\lambda} \end{bmatrix}, \quad K^{-1} = \begin{bmatrix} K^{\psi\psi} & K^{\psi\lambda} \\ K^{\lambda\psi} & K^{\lambda\lambda} \end{bmatrix},$$

and similarly for V_g and V_g^{-1}. Finally, let $\tilde{\lambda}_\psi$ be the constrained M-estimate of λ, let $\tilde{\theta}_\psi = (\psi, \tilde{\lambda}_\psi)$, and let $\tilde{\psi}$ be the ψ component of $\tilde{\theta}$. Then, a profile Wald-type statistic for the ψ may be defined as

$$w_R(\psi) = (\tilde{\psi} - \psi)(\tilde{V}_g^{\psi\psi})^{-1/2},$$

and it has an asymptotic $N(0,1)$ null distribution. Similarly, the profile score-type statistic

$$w_{sR}(\psi) = g_\psi(\tilde{\theta}_\psi)^\top K^{\psi\psi}(V_g^{\psi\psi})^{-1}K^{\psi\psi}g_\psi(\tilde{\theta}_\psi)$$

has an asymptotic χ^2_1 null distribution, while the asymptotic distribution of the profile ratio-type statistic for ψ, given by $W_R(\psi) = 2\left(G(\tilde{\theta}_\psi) - G(\tilde{\theta})\right)$, is $\nu\chi^2_1$, where $\nu = (\tilde{K}^{\psi\psi})^{-1}\tilde{V}_g^{\psi\psi}$. In view of this, for the adjusted profile ratio-type statistic to first-order it holds

$$W_R^{adj}(\psi) = \frac{W_R(\psi)}{\nu} \;\dot{\sim}\; \chi^2_1.$$

Finally, the adjusted profile root, analogous to (1.3), can be defined as

$$r_R(\psi) = \text{sign}(\tilde{\psi} - \psi)\sqrt{W_R^{adj}(\psi)},$$

which has an asymptotic standard normal distribution. For the general theory of robust tests see Heritier and Ronchetti (1994).

Similarly to the likelihood based CDs, a recipe to derive an asymptotic CD from robust M-estimating functions is the following. Let us denote with $q_R(\psi; y)$ a robust pivotal quantity, such as the profile Wald-type statistic $w_R(\psi)$ or the adjusted profile scoring rule root $r_R(\psi)$. Then,

$$C_R^w(\psi) \doteq \Phi\left((\psi - \tilde{\psi})(\tilde{V}_g^{\psi\psi})^{-1/2}\right) \tag{2.7}$$

and

$$C_R^r(\psi) \doteq \Phi\left(\text{sign}(\psi - \tilde{\psi})\sqrt{W_R^{adj}(\psi)}\right) \tag{2.8}$$

are first-order asymptotic CDs, and the corresponding confidence densities are, respectively,

$$cd_R^w(\psi) \doteq \frac{\phi\left((\psi - \tilde{\psi})(\tilde{V}_g^{\psi\psi})^{-1/2}\right)}{\sqrt{\tilde{V}_g^{\psi\psi}}}$$

and

$$cd_R^r(\psi) \doteq \phi\left(\text{sign}(\psi - \tilde{\psi})\sqrt{W_R^{adj}(\psi)}\right)\left|\frac{\partial W_R^{adj}(\psi)^{1/2}}{\partial\psi}\right|.$$

Note that the Wald-type based confidence density $cd_R^w(\psi)$ coincides with the asymptotic first-order robust Bayesian posterior distribution for ψ (see, e.g. Greco *et al.* 2008 and

Ventura and Racugno 2016).

In practice, using for instance (2.8), the confidence median is $\tilde{\psi}$ and an $(1-\alpha)$ equi-tailed confidence interval can be obtained as $\{\psi||r_R(\psi)| \leq z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}\}$ is the $(1-\alpha/2)$-quantile of the standard normal density. When testing, for instance, $H_0 : \psi = \psi_0$ against $H_1 : \psi < \psi_0$, the p-value is $p = C_R^r(\psi_0)$, while when testing $H_0 : \psi = \psi_0$ against $H_1 : \psi \neq \psi_0$ the p-value is $p = 2(1 - \Phi(|r_R(\psi_0)|))$. Furthermore, a measure of evidence for a statement of the form "$\psi_1 < \psi < \psi_2$" can be computed as $C_R^r(\psi_2) - C_R^r(\psi_1)$.

To study the stability of robust CDs, let us write the robust pivotal quantity more generally as $q_R(\psi; T(\hat{F}_n))$, where $\hat{F}_n$ is the empirical distribution function and $T(F)$ is the functional defined by the unbiased M-estimating equation $\int g(y; T(F))\, dF(y) = 0$, where $F = F(y; \theta)$ is the assumed parametric model. In CD inference the tail area, given by $C_R(\psi) = \Phi(q_R(\psi; T(\hat{F}_n))$, plays a central role and thus we can consider the tail area influence function (see, e.g., Field and Ronchetti 1990, and Ronchetti and Ventura 2001), given by

$$TAIF(y; T) = \frac{\partial}{\partial \varepsilon} \Phi(q_R(\psi; T(F_\varepsilon)))\Big|_{\varepsilon=0}, \tag{2.9}$$

where $F_\varepsilon = (1-\varepsilon)F + \epsilon\Delta_y$ and Δ_y is the probability measure which puts mass 1 at the point y. The $TAIF(y; T)$ thus describes the normalized influence on the CD tail area of an infinitesimal observation at y and, by considering its supremum, it can be used to evaluate the maximum bias of the tail area on the ε-neighborhood of F. It can be shown that

$$TAIF(y; T) \;=\; \phi(q_R(\psi; T(F)))\, \frac{\partial q_R(\psi; T(F))}{\partial T(F)}\; \frac{\partial T(F_\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=0}, \tag{2.10}$$

where the last term in (2.10) is the IF (2.6) of the M-estimator. Thus, the tail area influence function for the CD tail area at the statistical model F is proportional to the M-estimating function and this gives an immediate handle on robustness. Furthermore, it is bounded with respect to y when the M-estimating function is bounded.

The application of (2.7) and (2.8) in the particular context of a robust scoring rule has been discussed in Ruli *et al.* (2022). In particular, the Tsallis score (Tsallis, 1988) is considered, which is given by

$$G(y; \theta) = (\gamma - 1) \int f(y; \theta)^\gamma\, dy - \gamma f(y; \theta)^{\gamma-1}, \quad \gamma > 1,$$

with corresponding unbiased M-estimating function $g(\theta) = \partial G(y; \theta)/\partial\theta$ (Ghosh and

Basu 2013,Dawid *et al.* 2016), and with the parameter γ which gives a trade-off between efficiency and robustness.

In this section, we study and compare simulation-based approach for computing CDs based on robust M-estimating functions, to Bootstrap and a method based on a frequentist reinterpretation of the ABC machinery (see, e.g., Bee *et al.* 2017, Ruli *et al.* 2020, Thornton *et al.* 2022), whose properties have been derived by Rubio and Johansen (2013) in a general setup. The idea consists in generating candidate parameter values from an uniform distribution, computing a robust suitable summary statistic using the simulated data and then accepting only the parameter values such that the corresponding summary statistic is "close" to its observed counterpart (see Algorithm 13).

Algorithm 13 Accept-reject robust ABC

Input: proposal $p(\psi, \lambda)$, number of iterations R, robust summary statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$, where y^{obs} is the observed sample, tolerance ε, distance $\rho(\cdot;\cdot)$

for $j \in 1, \ldots, R$ **do**

Sample $(\psi_j^*, \lambda_j^*) \sim p(\psi, \lambda)$ and $y_j^* \sim f(y; \psi_j^*, \lambda_j^*)$

Compute $t_j^* = t(y_j^*)$

Accept ψ_j^* if $\rho(t_j^*; t^{\text{obs}}) \leq \varepsilon$ else reject

end for

resample the accepted (ψ^*, λ^*) with probability $\propto 1/p(\psi^*, \lambda^*)$

return robust approximate normalized pseudo-likelihood $\propto$ confidence density $\hat{cd}_R^{abc}(\psi)$

In Algorithm 13, the summary statistics of Soubeyrand and Haon-Lasportes (2015) or of Ruli *et al.* (2016, 2020) can be used. In particular, the first one is based directly on the M-estimator $\tilde{\psi}$ as the summary statistic $t(y)$ and a, possibly rescaled, distance among the observed and the simulated value of the statistic. In the second one, a rescaled version of the M-estimating function $g(\theta)$, evaluated at a fixed value of the parameter, is used as a summary statistic $t(y)$; this avoids repeated evaluations of the consistency correction involved in the M-estimating function, which is instead necessary for the Bootstrap. For a single parameter of interest, we propose to use instead the profile estimating equation, and plugging in the value of proposals λ^* for nuisance parameters used to generate pseudo-data as equation 2.5. Note also that when using the M-estimator as a summary statistic, the algorithm for solving the estimating equation might not converge after a prefixed number of iterations, thus causing additional noise

in the results. The treatment of the nuisance parameters resembles that of a generalized profile likelihood (Severini and Wong, 1992). Note that, assuming the regularity assumptions of Soubeyrand and Haon-Lasportes (2015) and the usual regularity conditions on M-estimators (Ronchetti and Huber, 2009, Chapter 4), then for $n \to \infty$ the robust confidence densities derived via simulation are asymptotically equivalent to the Wald-type confidence density $cd_R^w(\psi)$. Moreover, following Ruli *et al.* (2020), if $g(y;\theta)$ is bounded in y, i.e. if the M-estimator is B-robust, then asymptotically the posterior mode, as well as other posterior summaries of the robust confidence density $\hat{cd}_R^{abc}(\psi)$ have bounded IF.

2.5.3 Applications to non-inferiority tests

The aim of this section is to introduce and apply CDs inference in the context of non-inferiority testing, in which interest is in establishing if a new product is not unacceptably worse than a product already in use. Applications of non-inferiority testing has revealed an attractive problem in medical statistics, biostatistics, statistical quality control and engineering statistics, among others. Here we focus in non-inferiority clinical trials where the aim is to show that an experimental treatment is not (much) worse than a standard treatment. Clinical practice, however, is not the only field of application of these tests: in comparing the performance of sensors in industrial environment, for instance, the margin may be linked to some difference in costs due to sensor functioning. Other applications can be found in machine learning literature, where instead the meaningful margin is related to the accuracy or to the speed in classification tasks.

In the process of evaluating the efficacy of an experimental treatment, it is common to develop studies in which the two arms are the new and the standard therapy, respectively, rather than the new and the placebo. This is because it is considered unethical to deprive patients from a therapy that has already been proven to be beneficial. The underlying research hypothesis to be verified is that new therapies have equivalent or non-inferior efficacies to the ones currently in use. Both non-inferiority and superiority tests are examples of directional (one-sided) tests (see, e.g., D'Agostino Sr *et al.* 2003, Rothmann *et al.* 2011 and references therein). In particular, the *non-inferiority test* wants to test that the treatment mean μ_N is not worse than the reference mean μ_S by more than a given equivalence margin δ. The actual direction of the hypothesis depends on the response variable being studied. This question can be formulated into a test procedure for which the null hypothesis is

$$H_0 : \mu_S - \mu_N \geq \delta,$$

where $\delta > 0$ is the equivalence margin, when higher values of the response variable mean better results, versus

$$H_1 : \mu_S - \mu_N < \delta.$$

The scalar parameter of interest in this context is thus $\psi = \mu_S - \mu_N$, and non-inferiority is claimed when the null hypothesis is rejected.

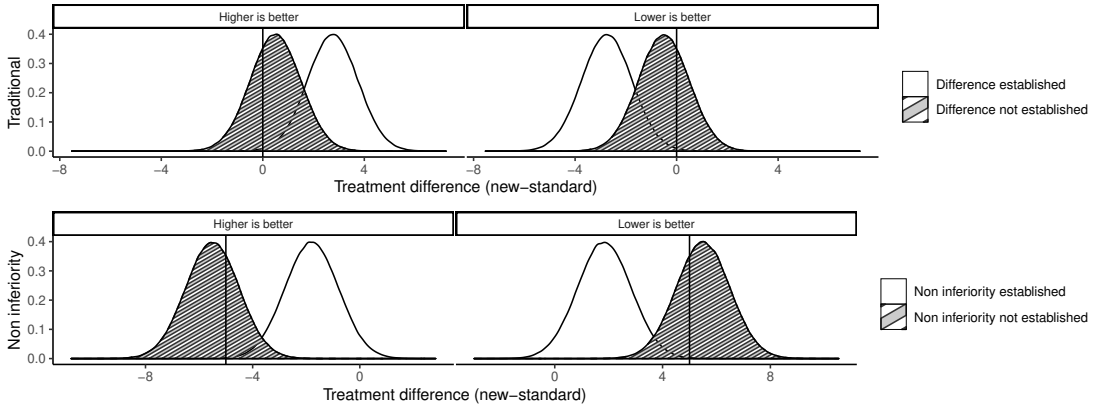

FIGURE 2.6: Testing procedure with traditional comparative studies and non-inferiority studies using confidence densities: vertical lines represent the equivalence/non-inferiority margin ($\delta = -5$).

The equivalence margin δ corresponds to the practical acceptable difference and should be pre-specified before the data is recorded (see e.g. Garrett 2003). An overly conservative margin might result in a high risk of not being able to claim non-inferiority when it actually is non-inferior. Conversely, overly liberal margins could result in a high risk of claiming non-inferiority when it actually is not non-inferior. A reasonable margin would be best derived from a combination of factors: the expected event rate, the duration of follow-up, and the number and nature of the events. However, arbitrary clinical judgment and the sponsor budget are of a great influence, resulting in a somewhat subjective non-inferiority margin. It is not clear in some situations how to perform the choice, and multiple thresholds could be plausible; in this respect, CDs are particularly useful to perform sensitivity analyses. Indeed, in this situation a confidence distribution on the difference $\psi = \mu_S - \mu_N$ will simultaneously show the evidence of the p-value against the null for a series of values δ, and decide for a reasonable δ with the nominal control of the rejection level and possible alternatives.

Here we consider an example of trial where higher levels of the response variable mean that the new treatment is effective. The aim is verifying that the new treatment (N) is not unacceptably worse to the standard (S). Let us assume that $n = 80$ patients

are randomized into two groups, and the model for the data is assumed to be

$$Y_S = \mu_N + \psi + u, \quad Y_N = \mu_N + u, \quad u \sim N(0, \sigma^2). \tag{2.11}$$

The normal distribution on the error term is often the basis of statistical analyses in medicine, genetics and in related sciences. Under this assumption, parametric inferential procedures based on the sample means, standard deviations, two-samples t-test, and so on, are the most efficient. However, it is well known that they are not robust when the normal distribution is just an approximate parametric model or in the presence of deviant values in the observed data (see, e.g., Farcomeni and Ventura 2012). In the framework described by (2.11), we inspect the effect of adding some contamination in the data of the new treatment group. In particular, in the contaminated scenario, 10% of the error terms in the new treatment group are half-Cauchy distributed (see Figure 2.7).

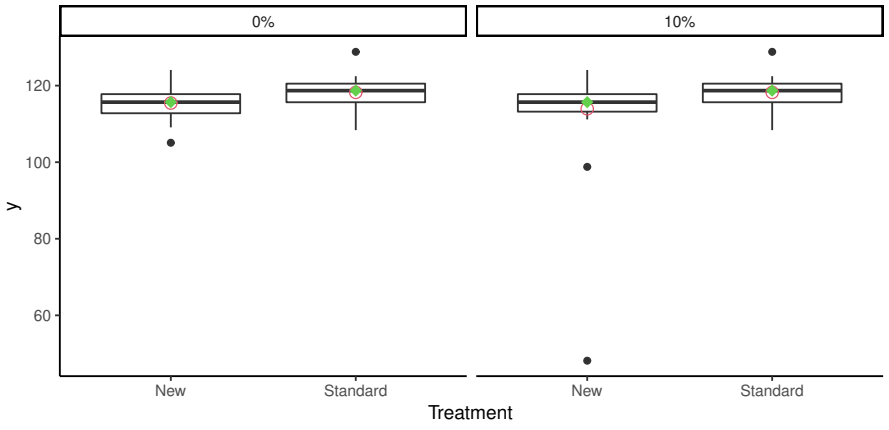

FIGURE 2.7: Non inferiority testing example: boxplots of recorded values for the new treatment group and the standard group under two scenarios. Red dots represent group means, while green dots represent group medians.

It is of interest to compare CDs inference for ψ based on the following approaches (abbreviations are also used in Figure 2.8 and in the following) used to derive confidence densities:

1. exact classical Wald-type confidence density based on $w_p(\psi)$, which is related to the classical two sample t-test (Wald/Mean)

2. robust asymptotic Wald-type confidence density $cd_R^w(\psi)$ based on the Huber's estimator (Wald/M-test)

3. approximate confidence density based on ABC (Algorithm 13) with robust Huber's estimator as summary statistics (ABC/M-est)

4. approximate confidence density based on ABC (Algorithm 13) with the robust Huber's estimating equation as summary statistic (ABC/M-EE)

5. simulated confidence density (Algorithm 12) based on the robust Huber's estimator (CDensity/M-est)

6. simulated confidence density (Algorithm 12) based on the robust Huber's estimating equation (CDensity/M-EE)

7. approximate confidence density based on ABC (Algorithm 13) with the difference of medians as summary statistics (ABC/Median)

8. simulated confidence density (Algorithm 12) based on the difference of medians (CDensity/Median)

9. parametric bootstrap confidence density (Boot/Basic)

10. parametric bootstrap with normal intervals confidence density (Boot/Norm)

11. parametric bootstrap with percentiles confidence density (Boot/Perc)

The nominal value of the mean difference between the treatment effects is ψ_0 is fixed to 2.6, and for simulation-based confidence distribution as well as for those obtained by the ABC-type algorithm we used 10^5 proposals and a tolerance level of 0.1. In the Huber's estimator we fix the tuning constant which controls the desired degree of robustness to 1.345, which imply that the estimator is 5% less efficient than the corresponding MLE under the assumed model.

From the resulting confidence densities illustrated in Figure 2.8 we note that, when the data come from the central model (left column) all the confidence densities are in reasonable agreement, even if the confidence densities based on the median behave slightly worse, with a greater variability. When the data are contaminated (right column), the non-robust confidence density (Wald/Mean) is less trustworthy as it drifts away from the true parameter value (green dotted line). This is not the case however for the robust confidence densities, which remain centred around the true parameter value. We further note that in the contaminated case, the robust confidence densities based on the M-estimating equation (ABC/M-EE and CDensity/M-EE) display the smallest variability. For all these confidence densities, Table 2.4 gives the measures of evidence for the statement "$\psi > \delta$", with the equivalence margin δ taken equal to 4 (black dotted line in Figure 2.8). As a benchmark, one can consider the measure derived by the exact $t-$distribution of the classical Wald-type confidence density in the non contaminated

case, which is 0.08. The results, without and with the contamination mirror the behaviour of the confidence densities in Figure 2.8, in particular the non-robustness of the likelihood-based confidence density (Wald/Mean). The most stable results under contamination seem to be those obtained with M-EE approaches (0.09 with ABC/M-EE and 0.05 with CD/M-EE). The same analysis could be done in principle for any margin δ.

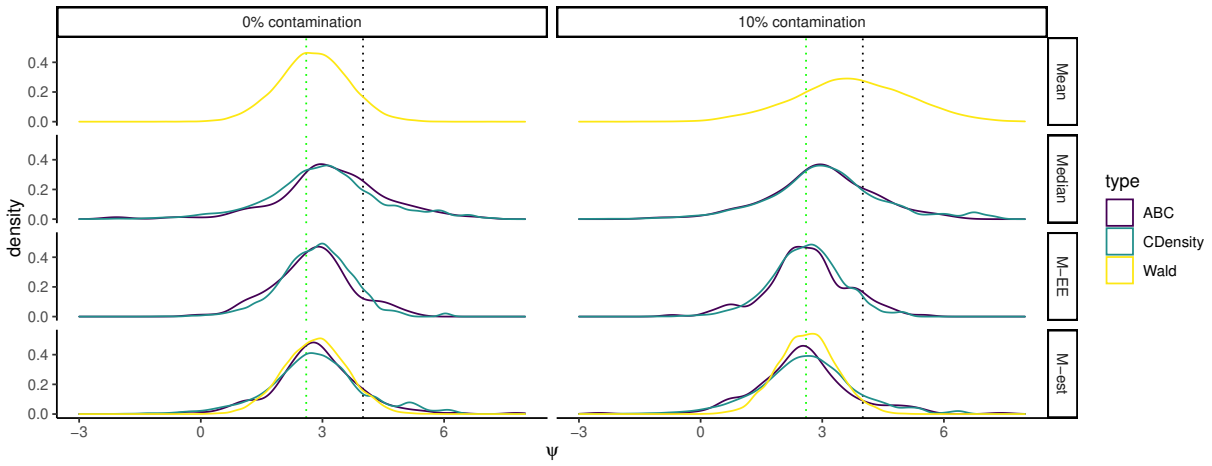

FIGURE 2.8: Confidence densities for ψ based on 10^5 proposals, without (*left column*) and with contamination (*right column*). Results for different choices of statistics are reported for each row, with inferential techniques represented by different colors. The black vertical dotted line represents the margin δ, the green one indicates ψ_0.

Method	0% cont.	10% cont.
Wald/Mean	0.08	0.42
Wald/M-test	0.07	0.03
ABC/Median	0.24	0.20
ABC/M-EE	0.11	0.09
ABC/M-est	0.12	0.09
CDensity/Median	0.20	0.21
CDensity/M-EE	0.08	0.05
CDensity/M-est	0.14	0.11

TABLE 2.4: Confidence measures of evidence associated to Figure 2.8, for the null hypothesis $H_0 : \psi > \delta$, with $\delta = 4$ without and with contamination.

Simulation study

For investigating the behaviour of the several types of confidence densities, we perform a simulation study under two sample sizes settings, $n = 40, 80$ (20, 40 per group) and for each of them we investigate two scenarios: one in which the assumptions of the

model in (2.11) are met by the true data generating mechanism, and the second one where 10% of the error terms in the new treatment group are half-Cauchy distributed (as in Figure 2.7). The families of methods to derive the confidence densities considered are the same as in the example above, and confidence distributions construction is based again on exact and asymptotic pivotal quantities or simulation-based. For Rejection-ABC-type confidence distributions the tolerance for the discrepancy was set to 0.1, the true value of parameter of interest was set to $\psi_0 = 2.6$, and the Huber's tuning constant to 1.345. The proposals for ψ were drawn from a Uniform random variable in $[-3, 9]$, for the parameter μ_N we sample from a Uniform in $[110, 130]$, while for σ we generated values from a Uniform $[1, 8]$. For the simulations, 4000 values were generated from the proposals and a total of 2000 simulations were performed.

We compare the empirical coverages of 90% and 95% equi-tailed confidence intervals. Results are synthetized in Tables 2.5 and 2.6. We also report in Tables 2.7 and 2.8 the error associated to confidence median point estimators, in terms of bias ($b = \sum_{r=1}^{R} \tilde{\theta}_r - \theta_0$), probability of underestimation ($PU = \sum_{r=1}^{R} 1_{\{\tilde{\theta}_r < \theta_0\}}$) and I type error with $\alpha = 0.05$. We note that, under the central model, the Wald/Mean CD shows a good performance, as well as some robust CDs (Wald/M-test, CDensity/M-EE and ABC/M-EE). With contaminated data, the Wald/Mean CD tends to be affected by contamination, whereas the robust CDs perform substantially better, with the CDs based on M-estimating equations being preferred over those based on M-estimators. Asymptotic symmetric confidence densities based on Wald-type robust CDs and ABC-type confidence densities seem to be affected more by bias than the simulated CDs (see Tables 2.7 and 2.8). Note finally that ABC-type results, even if behaving well, depend on a tolerance choice, hence the results may degradate when the latter is not well calibrated.

As a final remark, note that an interesting aspect of this simulation study was the difference among the approach of using robust M-estimating functions instead of robust M-estimates, especially in the treatment of nuisance parameters.

Real data application

A class of problems requiring similar considerations to those of non-inferiority tests, i.e. sensitivity analysis with respect to the reference margin δ, is that of superiority studies (see Figure 2.9).

Here we analyze the data collected in a randomized controlled trial (see Carhart-Harris *et al.* 2021 and Nayak *et al.* 2023) with the aim of assessing the superiority of a new therapy with psilocybin (P) versus that with escitalopram (E), in treating major depressive disorder. The dataset contains the scores obtained by $n = 57$ patients on a

Contamination	0%		10%	
$n = 40$	95% CI	90% CI	95% CI	90% CI
Wald/Mean	93.9	89.1	97.1	94.0
Wald/M-test	93.7	88.4	94.1	88.3
ABC/Median	97.1	93.4	97.7	93.7
ABC/M-EE	92.7	87.2	93.7	88.9
ABC/M-est	97.0	93.1	97.6	93.9
CDensity/Median	99.5	97.6	99.2	98.0
CDensity/M-EE	95.8	90.5	96.7	92.1
CDensity/M-est	99.4	97.3	99.2	98.0
Boot/basic	93.4	88.0	92.3	86.1
Boot/Norm	93.5	88.2	92.4	86.0
Boot/Perc	93.4	87.9	92.3	86.1

TABLE 2.5: Empirical coverages in a simulation study without and with 10% contamination and $n = 40$.

Contamination	0%		10%	
$n = 80$	95% CI	90% CI	95% CI	90% CI
Wald/Mean	95.5	90.0	95.9	92.2
Wald/M-test	95.1	89.6	93.9	87.9
ABC/Median	97.2	93.5	97.3	93.5
ABC/M-EE	93.3	86.9	92.7	87.3
ABC/M-est	97.5	93.6	97.2	93.9
CDensity/Median	99.1	97.6	99.2	97.5
CDensity/M-EE	95.9	89.4	96.4	92.5
CDensity/M-est	99.1	97.3	99.2	97.7
Boot/Basic	94.1	89.5	92.3	87.5
Boot/Norm	94.2	89.5	92.4	87.4
Boot/Perc	94.3	89.6	92.3	87.5

TABLE 2.6: Empirical coverages in a simulation study without and with 10% contamination and $n = 80$.

questionnaire, before and after a 6-week period of therapy. The model considered for the scores at the time of follow-up (FU) is the following

$$y_{\text{FU}} = \beta_0 + \beta_1 y_{\text{BL}} + \beta_2 \text{P} + u, \quad u \sim N(0, \sigma^2),$$

where y_{BL} represents the value at the baseline and P is a dummy variable that equals 1 if the subject belongs to the group treated with the new therapy (psilocybin), and thus the coefficient relates to the additional change with respect to the control group (escitalopram) after the therapy. A reduction of the score indicates a clinical improvement; thus superiority is claimed if the estimate of the coefficient β_2 is sufficiently lower than 0. In particular, in order to conclude in favour of meaningful superiority, the clinicians

Contamination	0%			10%		
$n = 40$	$\|b\|$	PU	I type err.	$\|b\|$	PU	I type err.
Wald/Mean	0.01	0.51	0.06	5.57	0.65	0.03
Wald/M-test	0.00	0.51	0.06	0.23	0.42	0.08
ABC/Median	0.03	0.52	0.01	0.09	0.46	0.01
ABC /M-EE	0.00	0.51	0.03	0.23	0.42	0.03
ABC/M-est	0.01	0.51	0.01	0.23	0.42	0.01
CDensity/Median	0.15	0.55	0.01	0.03	0.51	0.01
CDensity/M-EE	0.11	0.55	0.03	0.11	0.46	0.03
CDensity/M-est	0.13	0.56	0.01	0.09	0.46	0.01
Boot/Basic	0.84	0.75	0.06	1.07	0.79	0.10
Boot/Norm	0.84	0.75	0.06	1.07	0.79	0.10
Boot/Perc	0.84	0.75	0.06	1.07	0.79	0.09

TABLE 2.7: Measures of stability of CDs: absolute bias ($|b|$), probability of underestimation (PU) and I type error ($\alpha = 0.05$) of confidence estimators (medians) in the simulation study with $n = 40$.

Contamination	0%			10%		
$n = 80$	$\|b\|$	PU	I type err.	$\|b\|$	PU	I type err.
Wald/Mean	0.02	0.49	0.05	1.76	0.58	0.05
Wald/M-test	0.01	0.49	0.05	0.19	0.42	0.08
ABC/Median	0.02	0.50	0.02	0.05	0.49	0.02
ABC/M-EE	0.02	0.48	0.03	0.19	0.42	0.03
ABC/M-est	0.01	0.49	0.01	0.19	0.42	0.02
CDensity/Median	0.07	0.53	0.01	0.02	0.51	0.02
CDensity/M-EE	0.08	0.53	0.03	0.11	0.45	0.03
CDensity/M-est	0.08	0.54	0.01	0.11	0.46	0.02
Boot/Basic	0.03	51.1	0.05	0.39	0.33	0.09
Boot/Norm	0.03	51.1	0.05	0.39	0.33	0.09
Boot/Perc	0.03	51.1	0.05	0.39	0.33	0.09

TABLE 2.8: Measures of stability of CDs: absolute bias ($|b|$), probability of underestimation (PU) and I type error ($\alpha = 0.05$) of confidence estimators (medians) in the simulation study with $n = 80$.

considered as reference a margin $\delta = -5.3$. It is of interest to provide stable measures of evidence for the statement "$\beta_2 > \delta$", with $\delta = -5.3$ (H_0).

The MLE for the parameter β_2 and its standard error are, respectively, -5.32 and 1.44, while the robust counterparts are -6.18 and 1.33. Note that after removing two outliers the MLE become -6.23, with standard error 1.35. We resume the whole confidence densities based on Wald-type methods together with simulated confidence densities based on Huber's estimators and Huber's estimating equations in Figure 2.10. As it can be noted the classical confidence density (Wald/Mean) is shifted to the right, because of the presence of outliers. Evidence measures for different margins are reported

in Table 2.9. Using the margin chosen by the clinicians (-5.3) there is no evidence of superiority at level $\alpha = 0.1$; however note that the measure of evidence computed with the Wald-type confidence density (Wald/Mean) is the double of the ones computed with the robust confidence densities. With a margin of $\delta = -3.5$ all the robust procedure would agree in claiming superiority with $\alpha = 0.1$, while according to classical Wald-type confidence density (Wald/Mean) there would not be enough evidence to conclude

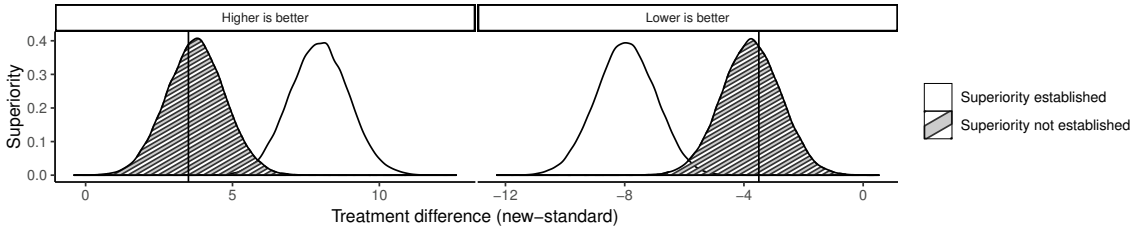

FIGURE 2.9: Example of making inference with confidence densities in superiority test, with margin $\delta = -3.5$.

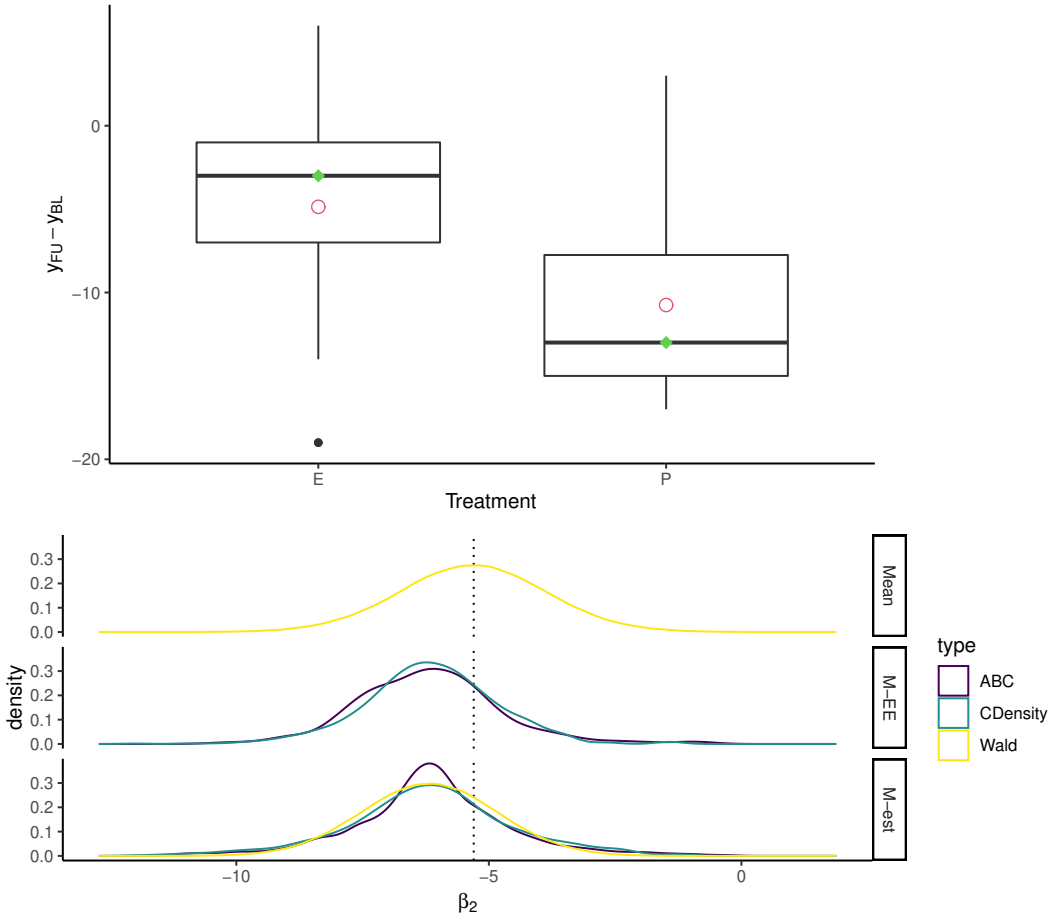

FIGURE 2.10: Real data example: boxplots representing pre-post differences of scores in a group of subjects treated with psilocybin (P) versus escitalopram (E) (*left*), accompanied by confidence densities for the parameter β_2, indicating the difference in efficacy of the two therapies (*right*).

superiority.

$-\delta$	**-3.5**	**-4**	**-4.5**	**-5**	**-5.3**
Wald/Mean	0.11	0.18	0.29	0.41	0.49
Wald/M-test	0.02	0.05	0.10	0.19	0.26
ABC/M-est	0.00	0.00	0.14	0.14	0.29
ABC/M-EE	0.08	0.12	0.18	0.24	0.30
CDensity/M-est	0.03	0.09	0.14	0.21	0.28
CDensity/M-EE	0.07	0.14	0.17	0.22	0.28

TABLE 2.9: Measures of evidence for the hypothesis "$\beta_2 > \delta$" for several margins.

2.6 Vector parameter

For a parameter vector, it is possible to obtain simulation-based confidence curves, resorting to a global statistic for the model, as the likelihood ratio of the model, or that of an a auxiliary model. Denote log-likelihood ratio test as $W(y, \theta, \hat{\theta}) = 2\ell(\hat{\theta}, y) - (\ell(\theta, y)$. Denoting with $\theta^*, y^*, \hat{\theta}^*$ simulated parameters, simulated data and estimated parameters from simulated data, respectively, the proposed θ^* are accepted if

$$W(y^*, \theta^*, \hat{\theta}^*) > W(y^{\text{obs}}, \theta, \hat{\theta}). \tag{2.12}$$

We could generalize the procedure, if the inference focuses simultaneously on multiple parameters in presence of nuisance parameters as well. Letting ψ be the vector parameter of interest and λ the nuisance components, the proposed parameters will be accepted if

$$W(y^*, \psi^*, \hat{\psi}^* \lambda_{\psi^*}) > W(y^{\text{obs}}, \psi, \hat{\psi}, \hat{\lambda}_{\psi}).$$

The general algorithm is given in 14.

Algorithm 14 Accept-reject confidence curve computing with parameter vector.

Input: Uniform proposal $p(\theta)$, number of iterations R, global statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$, where y^{obs} is the observed sample.

for $j \in 1, \ldots, R$ **do**
  Sample $(\theta_j^*) \sim p(\theta)$ and $y_j^* \sim f(y; \theta_i^*)$
  Compute $t_j^* = t(y_j^*)$
  Accept ψ_j^* if $t_j^* \geq t^{\text{obs}}$ else reject
end forreturn θ^* with density $\propto cc_R(\theta)$ a confidence curve.

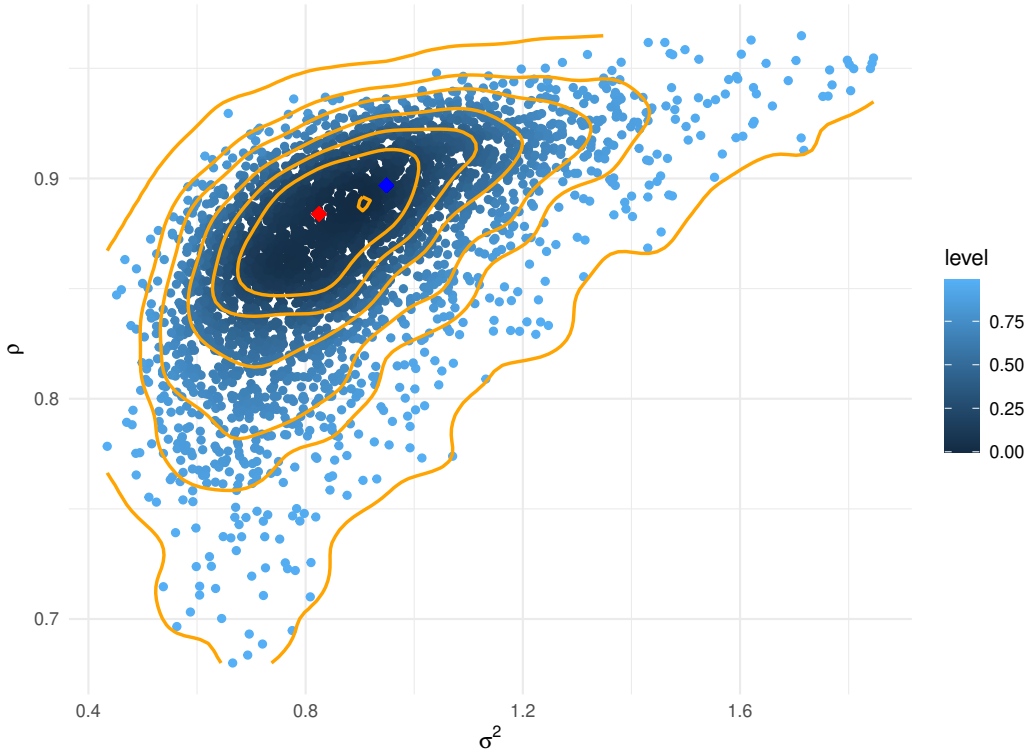

FIGURE 2.11: Bivariate confidence curve and regions for (σ^2, ρ) in the bivariate normal model.

2.7 Examples with parameter vector

2.7.1 Multivariate normal with modified score functions

In the same setting of Example 2.5.1, we consider the construction of a confidence curve based on the log-likelihood ratio test. Such confidence curve is illustrated in Figure 2.12. We also display the maximum likelihood estimator (red mark), and the median bias reduced estimator (blue mark) and confidence regions of levels $[0, 0.2, 0.4, 0.6, 0.8, 0.9, 99]$ for (σ^2, ρ), obtained by successive density estimation. Here regression parameters $(\beta_1, \ldots, \beta_p)$ are treated as nuisance components.

2.7.2 SIR Epidemic Model

Consider a SIR (Susceptible-Infectious-Recovered) model, widely utilized in epidemiology. The SIR model is a compartimental model defined on three distinct populations:

- "S" represents individuals who are susceptible to the disease,
- "I" represents individuals who are infectious,
- "R" represents individuals who have recovered.

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13
y_t	3	8	28	75	221	281	255	235	190	125	70	28	12	5

TABLE 2.10: Data of influenza outbreak in England (Anonymous, 1978).

The transmission of the disease occurs through interactions at times defined by the occurrences of a Poisson process. Susceptibles become infected at a rate determined by the product of the infectious contact rate β and the number of infectious individuals, denoted as "I". Infectious individuals, on the other hand, can recover from the disease at a rate denoted as γ.

$$\begin{aligned} N &= S + I + R \\ dS &= -\frac{\beta \cdot I}{N} \cdot S \\ dI &= \frac{\beta \cdot I}{N} \cdot S - \gamma \cdot I \\ dR &= \gamma \cdot I. \end{aligned}$$

Moreover N is the total population and is considered fixed.

We consider a dataset documenting the incidence of influenza cases within a boarding school located in England. These data were originally reported as a graphical representation in a study by Anonymous (1978). The specific numerical values presented in Table 2.10 have been extracted by G. de Vries (2006). The aim of this example is to show how to perform inference with a confidence curve in a multi-parameter model where the likelihood function is intractable. Let $g(\theta) = g(\theta, u_1)$ be the generator of the deterministic version of the SIR model, i.e. with constant number of events and an initial number of infected u_1 equal to one. As a surrogate pivotal quantity, we consider the normalized Residual Sum of Squares of the number of predicted infected, given by

$$\mathrm{RSS}(\theta, y) - \mathrm{RSS}(\hat{\theta}, y) = [y - g(\theta)]^\top [y - g(\theta)] - [y - g(\hat{\theta})]^\top [y - g(\hat{\theta})], \tag{2.13}$$

where $\theta = (\beta, \gamma)$ and $\hat{\theta}$ the minimizer of the RSS. Note that (2.13) is used in analogous way to (2.12), as a global statistic based on a surrogate model. It can be seen as a test based on a quasi-likelihood where the loss function is the residual sum of squares.

Then, the confidence curve is obtained retaining parameter values θ^* for which the following inequality holds:

$$\mathrm{RSS}(\theta^*, y^*) - \mathrm{RSS}(\hat{\theta}^*, y^*) > \mathrm{RSS}(\theta, y) - \mathrm{RSS}(\hat{\theta}, y),$$

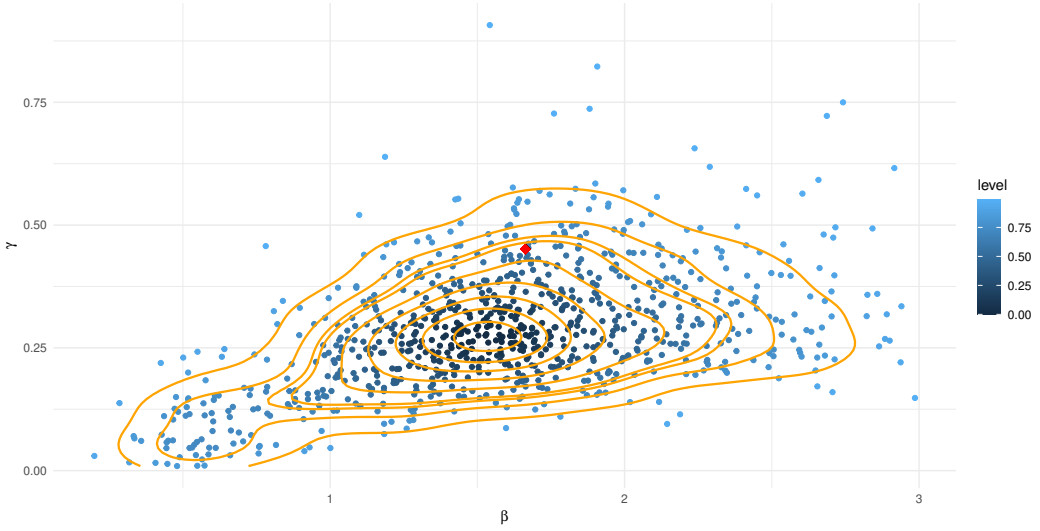

FIGURE 2.12: Bidimensional confidence curve and equi-spaced confidence regions (with levels 0.1-0.9) associated to the epidemic data. The point estimate is marked in red.

where pseudo-datasets y^* are obtained by θ^* according to Algorithm 12.

2.8 Confidence distributions based on integral probability semimetrics

In conclusion, we mention the possibility to resort to a non-parametric derivation of CDs based on integral probability semimetrics (Müller, 1997) or pseudo-metrics (Ronchetti and Huber, 2009, Chapter 2). These divergences are classically associated to the concept of stability, used as global tests and studied in the context of misspecified models where the meaningfulness of model features is uncertain, wherealse directly comparing the distributions happens to be more natural (see for instance Bernton *et al.* 2019 and Legramanti *et al.* 2022).

For mitigating the effects of small departures from model assumptions, and possible dramatic changes of inferential conclusions due to unconvenient choices of pivotal quantities and summaries, the use of non parametric procedures may be an alternative. In particular, we focus on the Kolmogorov-Smirnov distance (d_{KS}) and the Wasserstein distance (d_W) for one-dimensional distributions, defined respectively as

$$d_{KS}(P,Q) = \sup_y |P(y) - Q(y)|,$$

$$d_W(P,Q) = \int_Y |P(y) - Q(y)| dy.$$

The discrepancy furnishes global indication of potential agreement between the two distributions P and Q, analogously to a likelihood ratio test as in the likelihood-based inference for correctly specified models. For estimating the distances, empirical cumulative distribution functions ($\hat{P}_n(y)$ and $\hat{Q}_n(y)$, respectively) are used.

In general models, the non-asymptotic distributions of the statistics $d_{KS}(\cdot,\cdot)$ and $d_W(\cdot,\cdot)$ are complex, and numerical methods are employed to compute exact p-values. Here we suggest to rely on Algorithm 13 to derive CDs, once identifyed as summary statistics suitable discrepancies with distribution stochastically monotone in a scalar parameter of interest θ. In particular, let us consider the observed sample y^{obs} and a fixed reference sample y^{ref}, drawn from a completely known model $f(y;\theta^{\text{ref}})$. A sequence of unilateral tests can be built by using as observed summary statistic in Algorithm 12 the quantity $d(y^{\text{obs}}, y^{\text{ref}}) = d(\hat{P}(y^{\text{obs}}), \hat{Q}(y^{\text{ref}}))$, where $d(\cdot,\cdot)$ may be the Kolmogorov-Smirnov distance (d_{KS}) or the Wasserstein distance (d_W). Then, the CD is obtained with the Accept-Reject scheme of Algorithm 12, evaluating

$$Pr_{\theta^*}(d(y^*, y^{\text{ref}}) > d(y^{\text{obs}}, y^{\text{ref}})),$$

where y^* is simulated from the central model $y^* \sim f(y;\theta^*)$. Also, by Algorithm 10 a confidence density can be retrieved. To obtain a proper confidence distribution, the distribution of the summary statistic should be stochastically ordered in the parameter of interest. Hence it is convenient to draw y^{ref} from the model $f(y;\theta')$, with θ' being the supremum of the proposal distribution support in Algorithm 12.

Otherwise, a serie of bilateral tests, directly comparing $d(y^*, y^{\text{obs}})$ to zero, without a reference sample, can also be performed, for obtaining a confidence curve instead of a proper confidence distribution.

2.8.1 Example: "Non parametric" CDs

As in Legramanti *et al.* (2022), we consider a contamination study. The data $y^{\text{obs}} = (y_1, \ldots, y_n)$, with sample size $n = 100$, are realizations of a Gaussian random variable $N(\theta, 1)$, with nominal value $\theta_0 = 1$. Within this setting, some scenarios of contamination are investigated: for each one a percentage of observations is substituded with the most extreme positive realization of a Cauchy of the same size. The amount of contamination here is $(5\%, 10\%, 15\%)$, respectively. In particular, using Algorithm 12 we simulated uniformly θ in $[-3, 3]$ and used as a pivot the distance $d(y^{\text{obs}}, y^{\text{ref}})$, where y^{ref} is drawn from $N(3, 1)$.

As shown in Figure 2.13, although the sample mean is dragged, the confidence distributions remain close to and concentrated around the nominal value of 1. In particular, the test based on the Wasserstein distance is higly stable up to the 15% of contamination. Compared to the approximate posteriors, the CD based on Wasserstein distance seems even more stable.

Let us denote with $\tilde{\theta}^m$ the confidence median and let us focus on the Wasserstein distance. Under the non contaminated sample (y_{θ_0}) the confidence median satisfies

$$Pr(d_W(y_{\tilde{\theta}^m}, y_{\theta^{ref}}) > d_W(y_{\theta_0}, y_{\theta^{ref}})) = 0.5.$$

When considering a ϵ-contaminated sample ($y_{\theta_0^{c_\epsilon}}$, with $\epsilon < 1\%$ of the data are not generated from the assumed model), we look for θ^* that satisfies

$$Pr(d_W(y_{\theta^*}, y_{\theta^{ref}}) > d_W(y_{\theta_0^{c_\epsilon}}, y_{\theta^{ref}})) = 0.5. \tag{2.14}$$

The difference $\theta^* - \tilde{\theta}^m$ is the shift due to the contamination. Writing $d_W(y_{\theta_0^{c_\epsilon}}, y_{\theta^{ref}})$ as

$$d_W(y_{\theta_0^{c_\epsilon}}, y_{\theta^{ref}}) = (1-\epsilon) d_W(y_{\theta_0}, y_{\theta^{ref}}) + \epsilon \cdot d_W(c, y_{\theta^{ref}}),$$

we can rewrite (2.14) as

$$Pr\left(d_W(y_{\theta^*}, y_{\theta^{ref}}) > d_W(y_{\theta_0}, y_{\theta^{ref}}) + \underbrace{\epsilon[d_W(c, y_{\theta^{ref}}) - d_W(y_{\theta_0}, y_{\theta^{ref}})]}_{\Delta} \right) = 0.5.$$

As the term $\Delta \to 0$, the confidence median is recovered. In particular this happens in the trivial case, when $\epsilon \to 0$ or if θ^{ref} minimizes $d_W(c, y_{\theta^{ref}}) - d_W(y_{\theta_0}, y_{\theta^{ref}})$, that means it parametrizes the model which corresponds to the barycenter between the central one and the model that generates the contamination. The optimal value cannot be known in advance, but as an initial guess a nonrobust estimate could be considered.

For analysing the behaviour of resulting confidence densities under the extreme case in which the contamination amount is $\epsilon = 0.2$, the data y^{obs} are still realizations of a Gaussian random variable $N(1,1)$ and for the contamination a percentage of observations is substituded with realizations from a Cauchy. For the derivation of the CDs, we consider different choices for the reference parameter $\theta^{ref} = (3, 4, 5, 6, 10, 20, 40, 100)$ (see boxplot of the data and confidence densities in Figure 2.14).

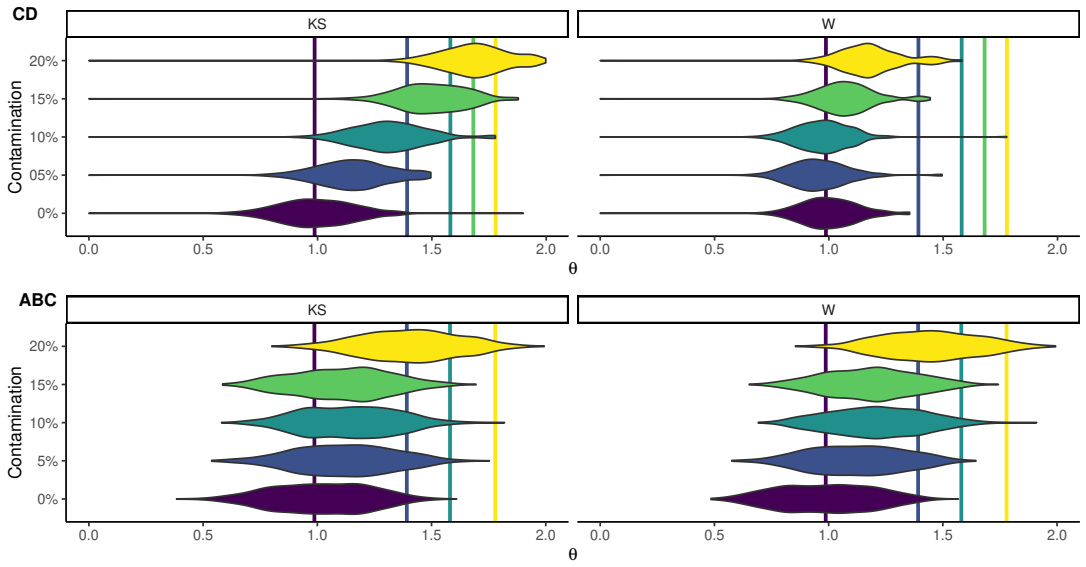

FIGURE 2.13: Confidence densities and ABC posterior based on Kolmogorv-Smirnov (KS) and Wassserstein (W) distances; vertical lines represent the sample means for increasing level of contamination.

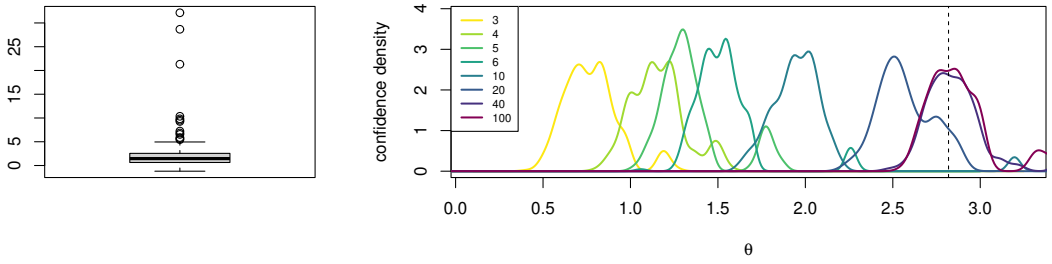

FIGURE 2.14: Boxplot representing the data (*left*) and confidence densities based on the Wasserstein distance with 20% of the data ($\epsilon = 0.2$) from a contamination model and different choices for the reference parameter. The empirical mean (dotted line) is 2.81, while the true value is 1, and the empirical mean of the uncontaminated sample (1-ϵ $\times$ 100 % of the data) is 1.16.

2.9 Discussion

Summarily, the accuracy of confidence distributions, curves, and densities presented in this Chapter is dependent solely on Monte Carlo errors, regardless of sample size constraints or deviations from Normality. While comparing the procedure to a parametric bootstrap is natural, certain differences exist. A key point of contrast is the absence of a preliminary estimate (e.g., maximum likelihood estimate) for simulation, allowing the method to be suitable for complex models, although estimates might be necessary

for nuisance parameters. In such cases, the confidence distributions, along with derived confidence intervals or p-values, remain invariant under reparametrizations. In scenarios lacking direct estimates for nuisance components, alternatives for eliminating them based on marginalization are easily implementable. Additionally, in contrast to bootstrap, the method doesn't fail in absence of regularity conditions.

One further advantage to methods for obtaining reliable confidence intervals in simulation-based inference recently proposed by Dalmasso *et al.* (2021) and Xie and Wang (2022), lies in computational efficiency. The presented procedure allows acquiring confidence intervals for all levels simultaneously, and eliminates the need to derive an empirical distribution for every potential parameter space value. The method's validity, in terms of controlling Type I errors, remains unaffected by the statistic chosen, unlike ABC procedures (Li and Fearnhead, 2018). Nonetheless, a precise selection would yield shorter confidence intervals and more meaningful outcomes in any scenario. Related to the choice of the summary statistic, one interesting aspect is the possibility of combining CDs obtained from different sources or partial information.

As a final remark, we mention the possibility to adopt non parametric criteria and statistics, other than just centrality measures, for deriving confidence distributions for a scalar parameter of interest in presence of contamination. A central parametric model is assumed but the observed data are evaluated in terms of non parametric pseudo-distances from a reference model, directly based on the empirical cdfs. At present our preliminary study is limited to models with a scalar parameter of interest, since adapting these kind of test procedures to more complex models to deal with nuisance parameters may require non parametric point estimation instead of the use of a reference model.

Chapter 3

Box ABC

3.1 Introduction

In this chapter, we introduce an algorithm designed to approximate the posterior distribution. Unlike traditional ABC schemes, our approach incorporates a probabilistic acceptance rule, eliminating the need for selecting and tuning the threshold parameter ε. Although the method is not able to recover the true posterior, it implicitly makes use of a pseudo-likelihood that enjoys some consistency guarantees and has some connections to data depth functions (see e.g. Liu, 1990). The method is currently developed to address problems where multiple summary statistics are involved, but only one parameter is unknown. Possible extensions are suggested in the final discussion.

3.1.1 ABC and the role of ε

In many applications, the evaluation of the likelihood function is either very computationally expensive or not feasible. Likelihood-free methods, such as Approximate Bayesian Computation (ABC) and Bayesian Synthetic Likelihood (BSL), have emerged as invaluable tools in Bayesian inference. These methodologies leverage simulation-based approaches to approximate the likelihood function. BSL, pioneered by Wood (2010) and extended by Price *et al.* (2018), facilitates inference by constructing a synthetic likelihood for the summary statistics from simulated data, thus a parametric approximation of the likelihood function. ABC, introduced by Rubin (1984) and Tavaré *et al.* (1997), aims to generate simulated data

sets (*pseudo-data*) that mimic the observed data without the need for explicit likelihood computation. For this reason, the approximation is considered non-parametric, as a particular form of the function is not given. Intuitively, if the synthetic data match the observed data within a certain tolerance limit ε, it is likely that the model parameters used in the simulations are plausible for the model under consideration and in turns, their likelihood function. The distance between pseudo and actual data is generally assessed using on a set of summary statistics that are intended to be informative for the model. More in detail, let us assume that it is possible to generate data from the model $p(y|\theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^p$ and let θ_0 be the true value of θ, so that $p(y|\theta_0)$ is the true data generating process. We denote with y^{obs} the observed data, of size n, with $t : \mathbb{R}^n \to \mathbb{R}^d$ a collection of summary statistics of $d < n$ components, with $t^{\text{obs}} = t(y^{\text{obs}})$ the observed summary statistics and with $\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ a distance function. The idea of the simplest (rejection) ABC algorithm (Algorithm 15) is to compare simulated data y^*, from $p(y|\theta^*)$, with θ^* generated from the prior proposal $\pi(\theta)$, to the observed data and accept the proposed values θ^* such that the latter discrepancy is relatively small, i.e. $\delta(t(y^*), t(y^{\text{obs}})) < \varepsilon$, with $\varepsilon > 0$ controlling the amount of the approximation. The resulting obtained sample of θ^* is drawn from an approximation of the posterior distribution $\pi(\theta|y^{\text{obs}})$, given by

$$\pi_\varepsilon^{\text{ABC}}(\theta|t(y^{\text{obs}})) = \frac{\pi(\theta)p(t(y)|\theta)\mathbb{I}_{\delta(t(y),t(y^{\text{obs}}))<\varepsilon}}{\int_\Theta \pi(\theta)p(t(y)|\theta)\mathbb{I}_{\delta(t(y),t(y^{\text{obs}}))<\varepsilon}d\theta},$$

and lies in a neighbourhood of the posterior distribution. If $t(y)$ is sufficient, as $\varepsilon \to 0$ the distribution $\pi_\varepsilon^{\text{ABC}}(\theta|t(y^{\text{obs}}))$ converges to the true posterior (Biau *et al.*, 2015).

Algorithm 15 Accept-reject ABC

Input: prior $\pi(\theta)$, number of iterations R, summary statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$, ε, distance $\delta(\cdot;\cdot)$

for j in $1, \dots, R$ **do**
  Sample $\theta_j^* \sim \pi(\theta)$ and $y_j^* \sim p(y; \theta_j^*)$
  Compute $t_j^* = t(y_j^*)$
  Accept θ_j^* if $\delta(t_j^*; t^{\text{obs}}) \leq \varepsilon$ else reject
end for
return accepted θ^*

The posterior inference involves multiple layers of approximation. First, the verification of sufficiency is not straightforward, since it is not possible to isolate terms that are sufficient without the likelihood function. For this reason, the resulting

ABC posterior is often referred to as *reduced* or *partial posterior*. Alternatives, such as non-parametric density estimation or approaches that bypass the need for summary statistics, offer greater flexibility. However, these methods can be inefficient, especially in high-dimensional spaces and when using non-informative priors (Grazian and Fan, 2020).Secondly, there is the interplay between the tolerance ε and the distance measure δ. Setting the tolerance ε exactly to zero is not feasible, given that the event corresponding to a null distance holds a measure of zero, unless the data are discrete. There is also the Monte Carlo error, which arises from computational constraints. Of particular interest is the trade-off due to the number of accepted values and the relaxation of ε. In particular, there exists an inverse relationship between the tolerance level and the required number of simulations needed to generate and retain sufficiently large Monte Carlo samples from the posterior. Consequently, the choice of ε significantly determines the accuracy of the derived reduced posterior when working with a fixed computational budget. There are generally two strategies for choosing such a tolerance level: the first is to set $\varepsilon = Q_\alpha(\delta_1, \ldots, \delta_R)$, i.e. the empirical α quantile of the measured distances, in advance. This allows to bound the Monte Carlo error and the resulting approximation depends mainly on the prior-posterior discrepancy. A second possibility is to let ε decrease sequentially, which turns out to be more computationally intensive as the final acceptance rate is unknown and eventually tends to zero. Therefore, the first strategy is preferable. Finally, because of the curse of dimensionality, in practice, ε must also increase with the dimension of the summary statistic.

In practice, without infinite computational resources, it is sometimes difficult to quantify the discrepancy between the estimated and exact posterior distributions, and the practitioner's choice remains problem dependent, mainly related to available computational resources. As a result, re-analyzing data sets of different or larger sizes may lead to inconsistencies. The process of tuning the value of ε in applications has also led to a body of research aimed at automating this choice by improving the mechanism for proposing parameter values without compromising computational efficiency. Recent work employs Metropolis-Hastings, Sequential Monte Carlo or Gibbs-like strategies within the ABC algorithm, to achieve increasingly precise target distributions or to evaluate smaller dimensional summary statistics (see for example Del Moral *et al.*, 2012; Simola *et al.*, 2021; Clarté *et al.*, 2021; Karabatsos, 2021).

3.2 Box-ABC: scalar case

Here, we aim to approximate the likelihood function without assuming a closed model for the summary statistics, similarly to ABC. However, we consider an approximated posterior where in the acceptance rule is not used a distance function and the choice and tuning of a threshold parameter. The goal remains to maintain consistency with high acceptance rates. To avoid the need for defining a distance metric and subsequently choosing a tolerance level, we adopt an alternative approach. For every sampled value θ^* drawn from the prior distribution, we generate two pseudo-samples, denoted as y_1^* and y_2^*. The acceptance of θ^* is contingent upon the condition

$$t(y^{\text{obs}}) \in B,$$

where $B := B(t(y_1^*), t(y_2^*))$ represents a "box", or more formally a hyper rectangle with edges defined by the coordinates of the summary statistics of the two pseudo-samples. In cases where the summary statistic is scalar, this "box" corresponds to an interval defined by $t(y_1^*)$ and $t(y_2^*)$ as its endpoints. In this case, the proposed θ^* is accepted if and only if

$$t(y^{\text{obs}}) \in [t_{(1)}, t_{(2)}], \tag{3.1}$$

where $t_{(1)} = t(y^*)_{(1)}$, $t_{(2)} = t(y^*)_{(2)}$ correspond to the (scalar) ordered summary statistics. By accepting according to rule (3.1), the algorithm returns a series of values distributed according to a a pseudo-posterior, that will be denoted as $\pi^{\text{box}}(\theta|y)$.

More precisely, the target distribution of the procedure is of form

$$\pi^{\text{box}}(\theta|y) = \frac{\mathcal{L}^{\text{box}}(\theta)\pi(\theta)}{\int_\theta \mathcal{L}^{\text{box}}(\theta)\pi(\theta)d\theta}, \tag{3.2}$$

with

$$\mathcal{L}^{\text{box}}(\theta) \propto F_t(t^{\text{obs}}|\theta)[1 - F_t(t^{\text{obs}}|\theta)],$$

where $\mathcal{L}^{\text{box}} : \Theta \to \mathbb{R}^+$ is a pseudo-likelihood and the function $F_t(t^{\text{obs}}|\theta)$ is the cumulative density function of $t(y)$, once fixed the value of the summary statistic.

Assumption 1. The statistic $t : \mathcal{Y} \to \mathcal{T} \in$ is one-dimensional, and $0 < Var(t(y)|\theta) < \infty$ for π-almost all θ.

The assumption that $Var(t(y)|\theta) > 0$ for π-almost all θ ensures that the intervals of the form $[t_{(1)}, t_{(2)})$ have positive probability of being non-empty.

Lemma 3.1. *Under Assumption 1, the procedure outputs samples from* $\pi^{box}(\theta|t^{obs})$*, defined in* (3.2)

Proof. Let $\theta \in \Theta$, and consider a pair of statistics following the pushed-forward distribution induced by the summary statistic t applied to $y \sim p(y|\theta)$. i.e. $(t_1,\ t_2) \overset{iid}{\sim} t_{\#}p(y|\theta)$ We compute the probability of acceptance of θ as follows:

$$\begin{aligned}
&\Pr(t_{(1)} \leq t^{\text{obs}} < t_{(2)}|\theta) = \Pr(t_1 \leq t^{\text{obs}} < t_2|\theta) + \Pr(t_1 > t^{\text{obs}} \geq t_2|\theta),\\
&= \Pr(t_1 \leq t^{\text{obs}}, t^{\text{obs}} < t_2|\theta) + \Pr(t_1 > t^{\text{obs}}, t^{\text{obs}} \geq t_2|\theta)\\
&= F_t(t^{\text{obs}}|\theta)[1 - F_t(t^{\text{obs}}|\theta)] + F_t(t^{\text{obs}}|\theta)[1 - F_t(t^{\text{obs}}|\theta)]\\
&= 2F_t(t^{\text{obs}}|\theta)[1 - F_t(t^{\text{obs}}|\theta)]\\
&\propto F_t(t^{\text{obs}}|\theta)[1 - F_t(t^{\text{obs}}|\theta)].
\end{aligned}$$

We obtain the target distribution by a usual rejection sampling argument. □

As in ABC, and in opposition to BSL, the approximation to the parametric model can be considered non parametric, because no assumptions are made on the shape of the model for the summary statistics, which only depends on the data generating process. Similarly to ABC and BSL, the approximate posterior will reflect the possible insufficiency of the summary statistics used.

Algorithm 16 Accept-reject Box-ABC

Input: prior $\pi(\theta)$, number of iterations R, summary statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$

for $j \in 1, \ldots, R$ **do**
  Sample $\theta_j^* \sim \pi(\theta)$ and $y_j^{*1}, y_j^{*2} \sim f(y; \theta_j^*)$
  Compute $t_j^{*1} = t(y_j^{*1}), t_j^{*2} = t(y_j^{*2})$
  Accept θ_j^* if $t^{\text{obs}} \in B(t(y^{*1}), t(y^{*2}))$,
end for
return accepted θ^*

$B(t(y^{*1}), t(y^{*2}))$ is the box/hypercube having as edges the values of two samples' summary statistics.

3.2.1 Variants: R samples with external interval

The procedure can be modified by sampling $R > 2$ datasets and subsequently accept the corresponding generating parameter value if

$$t_{(1)}^* \leq t_n \leq t_{(R)}^*.$$

Hereafter, we denote as $F_{t_1 t_R}$ the joint distribution of the ordered statistics and $f_{t_1 t_R}$ the joint density. Hence,

$$Pr(t_{(1)} \leq t^{\text{obs}}) \leq t(y^{(R)*}) = \int_{t^{\text{obs}}}^{\infty} \int_{-\infty}^{t^{\text{obs}}} f_{t_1 t_R} dt_1 dt_R.$$

Following basics reasoning (see also Ahsanullah *et al.*, 2013)

$$\begin{aligned}
Pr(t_{(1)} \leq t^{\text{obs}}) \leq t(y^{(R)*}) &= \int_{t^{\text{obs}}}^{\infty} \int_{-\infty}^{t^{\text{obs}}} f_{t_1} f_{t_R} [F_{t_R} - F_{t_1}]^{R-2} dt_1 dt_R \\
&= \int_{t^{\text{obs}}}^{\infty} \int_{-\infty}^{t^{\text{obs}}} f_{t_1} f_{t_R} \sum_{k=0}^{R-2} \binom{R-2}{k} F_{t_R}^k (-F_{t_1})^{R-k-2} dt_1 dt_R \\
&= \sum_{k=0}^{R-2} [\binom{R-2}{k} \int_{t^{\text{obs}}}^{\infty} \int_{-\infty}^{t^{\text{obs}}} f_{t_1} f_{t_R} F_{t_R}^k (-F_{t_1})^{R-k-2} dt_1 dt_R] \\
&= \sum_{k=0}^{R-2} [\binom{R-2}{k} \int_{-\infty}^{t^{\text{obs}}} f_{t_1} (-F_{t_1})^{R-k-2} dt_1 \int_{t^{\text{obs}}}^{\infty} f_{t_R} F_{t_R}^k dt_R].
\end{aligned}$$

With integration by parts of the undefined integral, $\int f_{t_R} F_{t_R}^k dt_R$ we obtain

$$\int f_{t_R} F_{t_R}^k dt_R = F_{t_R}^k F_{t_R} - \int F_{t_R} k F_{t_R}^{k-1} f_{t_R} dt_R = F_R^k F_{t_R} - k \int f_{t_R} F_{t_R}^k dt_R$$

where one can recognize inside the last term the original function. Thus, $(1 + k) \int F_{t_R}^k dt_R = F_{t_R}^{k+1}$, and finally $\int F_{t_R}^k dt_R = F_{t_R}^{k+1}/(1 + k)$. Using these results in the defined integral,

$$= \sum_{k=0}^{R-2} [\binom{R-2}{k} \frac{[1 - F_{t_R}(t_n)^{k+1}]}{(1+k)} \int_{-\infty}^{t_n} f_{t_1} (-F_{t_1})^{R-k-2} dt_1].$$

Similarly, consider the remaining-to-be-integrated part

$$\begin{aligned}
\int f_{t_1} (-F_{t_1})^{R-k-2}) dt_1 &= F_1 (-F_{t_1})^{R-k-2} - \int F_1 (R-k-2)(-1)^{R-K-3} F_1^{R-k-3} f_{t_1} dt_1 \\
&= F_1 (-F_{t_1})^{R-k-2} - (-1)^{R-K-3} (R-k-2) \int F_1^{R-k-2} f_{t_1} dt_1,
\end{aligned}$$

with analogous reasoning it follows

$$\int f_{t_1} (-F_{t_1})^{R-k-2}) dt_1 (1 + (-1)^{R-K-3} (R-k-2)) = F_1 (-F_{t_1})^{R-k-2}$$

and

$$\int f_{t_1}(-F_{t_1})^{R-k-2})dt_1 = \frac{F_1(-F_{t_1})^{R-k-2}}{(1+(-1)^{R-K-3}(R-k-2))}.$$

Hence,

$$\begin{aligned}
&\Pr(t_{(1)} \leq t(y_n) \leq t(y^{(R)*})) \\
&= \sum_{k=0}^{R-2} \binom{R-2}{k} \frac{[1-F_R(t_n)^{k+1}]}{(1+k)} \frac{F_1(-F_{t_1})^{R-k-2}}{(1+(-1)^{R-K-3}(R-k-2))} \\
t_1, t_R \text{iid} &= \sum_{k=0}^{R-2} \binom{R-2}{k} \frac{[1-F(t_n)^{k+1}]}{(1+k)} \frac{F(t_n)(-1)^{R-k-2}F(t_n)^{R-k-2}}{(1+(-1)^{R-K-3}(R-k-2))} \\
&= \sum_{k=0}^{R-2} \binom{R-2}{k} \frac{[1-F(t_n)^{k+1}]}{(1+k)} \frac{(-1)^{R-k-2}F(t_n)^{R-k-1}}{(1+(-1)^{R-K-3}(R-k-2))} \\
&= \sum_{k=0}^{R-2} (-1)^{R-k-2} \binom{R-2}{k} \frac{1}{(1+k)} \frac{[F(t_n)^{R-k-1}-F(t_n)^R]}{(1+(-1)^{R-K-3}(R-k-2))}
\end{aligned}$$

which is a polynomial function, not corresponding again to the likelihood amount, and can be recognized to allow for acceptance easier.

3.2.2 Variants: R samples with internal interval

We can also study the behaviour of the algorithm when $R > 2$ samples are the proposed value θ^* is accepted if the summary statistic falls in the central interval, i.e. $[t_{(R/2)}, t_{(R/2+1)}]$. In this case, the pseudo-likelihood is proportional to $[F_t(t_n|\theta)(1-F_t(t_n|\theta))]^{R/2}$. Indeed, let us recall that the joint probability density function of two ordered statistics (r, s) among R is

$$\begin{aligned}
f_{t_r,t_s}(t_r,t_s) = \frac{R!}{(r-1)!(s-r-1)!(R-s)!} f_{t_r}(t_r) f_{t_s}(t_s) F_{t_r}(t_r)^{r-1} \\
[F_{t_s}(t_s) - F_{t_r}(t_r)]^{s-r-1} [1-F_{t_s}(t_s)]^{R-s}.
\end{aligned}$$

To us, $s = r+1$, $R-s-1 = r-1$ and $r = R/2$ since we consider the two median values among R. Hence, by integrating over the right support, we have

$$\int_{-\infty}^{t_n} \int_{t_n}^{\infty} \frac{R!}{(R/2-1)!(R/2-1)!} f_{t_r}(t_r) f_{t_s}(t_s) F_{t_r}(t_r)^{R/2-1} [1-F_{t_r}(t_r)]^{R/2-1} dt_r dt_s. \tag{3.3}$$

For instance, with $R = 4$, we have that this equals

$$24 \int_{-\infty}^{t_n} \int_{t_n}^{\infty} f_{t_r}(t_r) f_{t_s}(t_s) F_{t_r}(t_r)[1 - F_{t_r}(t_r)] dt_r dt_s$$

with similar reasoning as before, we obtain

$$\begin{aligned}
&= 24 \frac{F^2(t_n)}{2} [\int_{t_n}^{\infty} f_{t_s} dt_s - \int_{t_n}^{\infty} f_{t_s} F_{t_s} dt_s] \\
&= 24 \frac{F^2(t_n)}{2} [(1 - F(t_n)) - \frac{(1 - F(t_n)^2)}{2}] \\
&= 24 \frac{F^2(t_n)}{2} [\frac{1}{2} - F(t_n)(1 - \frac{F(t_n)}{2})] \\
&= 24 \frac{F^2(t_n)}{4} - \frac{F^3(t_n)}{2} + \frac{F^4(t_n)}{4} \\
&= 6F^2(t_n)(1 - F(t_n))^2 \\
&\propto (F(t_n)[1 - F(t_n)])^2,
\end{aligned} \tag{3.4}$$

where we neglect the multiplicative constants. Now we can generalize the result to the case of R samples, by induction. For $R = 2$ and 4 it is true that

$$[F(t_n)(1 - F(t_n))]^{R/2}.$$

Assume it is true for R^*, we then show that is even valid for $R^* + 2$. This means we have samples $R/2$ pairs of samples and we have observed t^{obs} greater than $R/2$ and lower than $R/2$ of the summaries. Then, we sample independently another pair of summaries and accept if and only if t^{obs} one of the summaries is grater than the observed and ne is lower, with probability $F(t^{\text{obs}})(1 - F(t^{\text{obs}}))$. Since the drawn are independent from the previous we will get

$$F(t_n)(1 - F(t_n))^{R^*/2} F(t_n)(1 - F(t_n)) = [F(t_n)(1 - F(t_n))]^{(R^*+2)/2}.$$

3.2.3 Link to confidence distributions

In order to study the properties of the approximate posterior, $\pi^{\text{box}}(\theta|y)$ it is convenient to note that the function $F_t(t^{\text{obs}}|\theta)$ is a Confidence Distribution. Indeed, a function $H_n(\cdot) = H_n(y, \cdot)$ on $\mathcal{Y} \times \Theta \to [0, 1]$ is a CD for θ if (see e.g. Xie and Singh, 2013):

C1 for each given $y \in \mathcal{Y}$, $H_n(\cdot)$ is a cumulative distribution function on Θ;

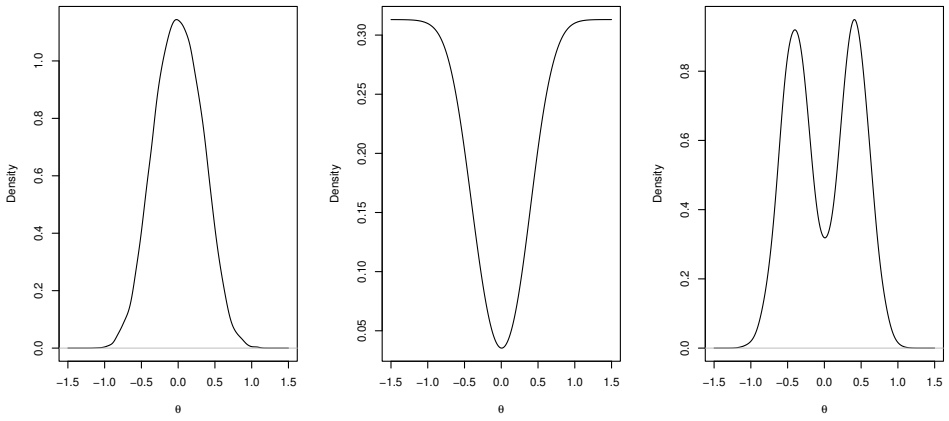

FIGURE 3.1: Instance of a confidence curve $cc(t^{\text{obs}}|\theta)$, its complementary $1-cc(t^{\text{obs}}|\theta)$ and their product.

C2 at $\theta = \theta_0$, $H_n(\theta_0) = H_n(y^{\text{obs}}, \theta_0)$, as a function of the sample y^{obs}, follows a Uniform$[0, 1]$ distribution.

Remark 3.2. If condition **C1** requires that $F_t(t^{\text{obs}}|\theta)$ is stochastically increasing in θ, for obtaining a CD, in Box-ABC the direction of the stochastic ordering does not matter since the expression handles symmetrically both cases. Thus, we can extend **C1** to accomodate cases in which $1 - H_n(\cdot)$ is a cumulative distribution function on Θ.

Remark 3.3. Note that the assumption can be checked *a posteriori* by running the accept-reject procedure described in Algorithm 16. If this does not hold, the pseudo-likelihood corresponds to an approximation taking form $cc(t^{\text{obs}}|\theta)(1 - cc(t^{\text{obs}}|\theta))$, with $cc(\cdot)$ a confidence curve (Schweder and Hjort, 2016). In this case, there is no guarantee of a miximizer.Figure 3.1 illustrates an instance of $cc(t^{\text{obs}}|\theta)$ (first panel), $1-cc(t^{\text{obs}}|\theta)$, and their product (third panel). The confidence median in this case is not the maximizer of the resulting pseudo-likelihood.

Remark 3.4. Since any statistic gives valid $F_t(t^{\text{obs}}|\theta)$, we can derive some frequentist properties, for $\mathcal{L}^{\text{box}}(\theta|t)$ that will be valid despite the choice of the statistic.

Remark 3.5. The concentration of $\mathcal{L}^{\text{box}}(\theta|t^{\text{obs}})$ is related to the spread of $F_t(t^{\text{obs}}|\theta)$ and thus to the power of a test based on t (see e.g. Schweder and Hjort, 2016, and refernces therein).

From this standpoint, it may be noted that in the scalar approach, the adoption of such a rule, along with the generation of twice the data, does not offer substantial information compared to deriving a Confidence distribution or confidence density,

as presented in Chapter 2 of this thesis. However, the potential advantage will become more apparent in the subsequent discussion, wherein the extension of the box-ABC to incorporate multiple summary statistics unveils a more streamlined and easily adaptable framework to that of CDs.

3.2.4 Properties of maximum pseudo-likelihood estimators

Median Unbiasedness. The maximizer of the pseudo-likelihood is exactly median unbiased. Define,

$$\tilde{\theta} = \text{argmax } \mathcal{L}^{box}(\theta), \mathcal{L}^{box}(\theta) = F_t(t^{\text{obs}}|\theta)[1 - F_t(t^{\text{obs}}|\theta)],$$

where the $F_t(t^{\text{obs}}|\theta)$ is normalized, i.e. $0 < F_t(t^{\text{obs}}|\theta) < 1$. It is easy to recognize that $\tilde{\theta}$ corresponds to the confidence median, i.e. $\tilde{\theta}$ is such that $F_t(t^{\text{obs}}|\tilde{\theta}) = \frac{1}{2}$. For the latter it holds that if θ_0 is the true value, $Pr(\tilde{\theta} > \theta_0) = 1/2$, for any n.

Asymptotic equivalence to MLE. Under standard regularity assumptions, the MLE $\hat{\theta}$ satisfies $Pr(\hat{\theta} > \theta_0) \underset{n\to\infty}{\to} 1/2$,. From this we get an asymptotic equivalence between $\tilde{\theta}$ and $\hat{\theta}$.

Asymptotic distribution. To facilitate the study of the asymptotic distribution of $\sqrt{n}(\tilde{\theta}-\theta_0)$, we employ a Central Limit Theorem argument under the assumption 1 that allows us to consider the convergence of the statistic $t(y)$ to a standard normal distribution. Denoting the variance of $t(y)$ as V the Confidence Distribution and the standard asymptotic posterior distribution have the form

$$C(\theta) \to \Phi(1/V^{1/2}(\theta - \theta_0)),$$

$$\pi(\theta|y) \to 1/V^{1/2}\phi(1/V^{1/2}(\theta - \theta_0)),$$

where Φ and ϕ are the cdf and the pdf, respectively, of a standard Normal distribution.

Lemma 3.6. *Under the asymptotic regime, the the pseudo-likelihood can be approximated by*

$$\frac{\Phi(V^{-1/2}(\theta - \theta_0))(1 - \Phi(V^{-1/2}(\theta - \theta_0)))}{\int_\Theta \Phi(V^{-1/2}(\theta - \theta_0))(1 - \Phi(V^{-1/2}(\theta - \theta_0)))d\theta},$$

where the normalizing constant is given by

$$\int_{\Theta} \Phi(V^{-1/2}(\theta-\theta_0))(1-\Phi(V^{-1/2}(\theta-\theta_0)))d\theta = (V/\pi)^{1/2}.$$

In particular, if the variance of the summary statistic is of order $O(n^{-1})$, the normalizing constant will be $O(n^{1/2})$ and the concentration of the box-pseudo-likelihood will scale at rate $O(n^{-1})$, since the numerator is $O(n^{-1/2})$. Hence, for large n, the concentration of the pseudo-likelihood and, in turn, the pseudo-posterior is expected to be higher than that of the regular posterior.

For the proof we will use the following result.

Lemma 3.7. *For any cumulative distribution function, it holds*

$$\int_a^b F(w)\cdot 1dw = [F(w)w]_a^b - \int_a^b f(w)wdw = F(b)b - F(a)a - \int_a^b f(w)wdw.$$

Proof. Let us rewrite

$$\begin{aligned}
&\int_{-\infty}^{\infty} \Phi(V^{-1/2}(\theta-\theta_0))(1-\Phi(V^{-1/2}(\theta-\theta_0)))d\theta = \\
&= \int_{-\infty}^{\infty} [\Phi(V^{-1/2}(\theta-\theta_0)) - \Phi^2(V^{-1/2}(\theta-\theta_0)))]d\theta = \\
&= \int_{-\infty}^{\infty} \Phi(V^{-1/2}(\theta-\theta_0))d\theta - \int_{-\infty}^{\infty} \Phi^2(V^{-1/2}(\theta-\theta_0)))d\theta.
\end{aligned}$$

We recognize that $\Phi(\theta)^2$ is the cumulative distribution function of a Skew-Normal (SN) random variable (Azzalini, 1985) with shape equal to one. Thus, the mean of a Skew-Normal random variable, $\text{SN}(0, V, 1)$ is $(V/\pi)^{1/2}$. Using (3.7) we have

$$\int_{-\infty}^{\infty} \Phi(V^{-1/2}(\theta-\theta_0))(1-\Phi(V^{-1/2}(\theta-\theta_0))d\theta = (\infty - 0) - (\infty - \frac{V^{1/2}}{\sqrt{\pi}}) = \frac{V^{1/2}}{\sqrt{\pi}}.$$

□

Consistency. Under the assumption $V \underset{n\to\infty}{\to} 0$, and thanks to fist order asymptotic equivalence between $\tilde{\theta}$ and $\hat{\theta}$, we recognize that $\tilde{\theta}$ is consistent.

3.2.5 Example: Normal model

Consider a sample realization from a Normal model, $y_i \sim N(0, \theta)$, for $i = 1, \ldots, n$ with $\theta = 1$, a uniform prior $\pi(\theta) = \text{Uniform}(0, 6)$ and $t(y) = Var(y)$ as a summary

statistic, with $\delta = |t(y) - t^{\text{obs}}|$ distance function. We aim at comparing the Box-ABC pseudo-posterior to that of ABC with varying sample size $n = (10, 2000)$, fixing the tolerance $\varepsilon = 0.05$ and the number of proposals to 10^6. In this setting, the variance of the summary statistics changes with $1/\sqrt{n}$. The total number of accepted values, the fraction of accepted proposals on the total number of generations from the model (Acc/Gen), the average computational time necessary to obtain one accepted value (Time/Acc, with time expressed in seconds) are shown in table 3.1. The implementation was done in `R` programming language (R Core Team, 2015), and a single thread. A laptop CPU with a clock speed of 1.00 GHz was used. Note that the index (Acc/Gen) takes into account the fact that Box-ABC requires two model generations from the same parameter setting. The Resulting approximate posterior are compared to the true one in Figure 3.2. Note that as the sample size increases and, conversely, as the variance of the summary statistic decreases, for Box-ABC the acceptance probability decreases, thus the cost per simulation increases, but the discrepancy prior-posterior does not change. Conversely, without changing the prior/proposal, and with fixed ε, for ABC the larger the sample size, the higher is the discrepancy prior posterior agreement.

	n=10	Acc/Gen	Time/Acc	n=2000	Acc/Gen	Time/Acc
ABC	10969	0.01	$2.7\ 10^{-3}$	8268	0.008	0.02
Box-ABC	42950	0.02	$1.8\ 10^{-3}$	3006	0.002	0.10

TABLE 3.1: Synthesis of simulation results for the Normal model: absolute number of accepted proposals, proportion of accepted proposals on the number of generations from the model (Acc/Gen) and average computational time to accept once (Time/Acc), with time expressed in seconds.

3.3 Box-ABC: multivariate statistics

Let us consider a more general case, where the summary statistic is multidimensional, i.e. $t = (t_1, \ldots, t_d) \in \mathbb{R}^d,\ d > 1$. Some difficulties naturally arise for defining the "box", as there might be a dependence structure of the vector of summary statistics, and establishing a clear ordering in such case is not straightforward. If there is a positive correlation among all the components, it might be natural extending the order of the first component to the other components. Conversely, if the dependence is in the opposite direction, the order of the first component might be transferred inversely to the others. Exploring the dependence

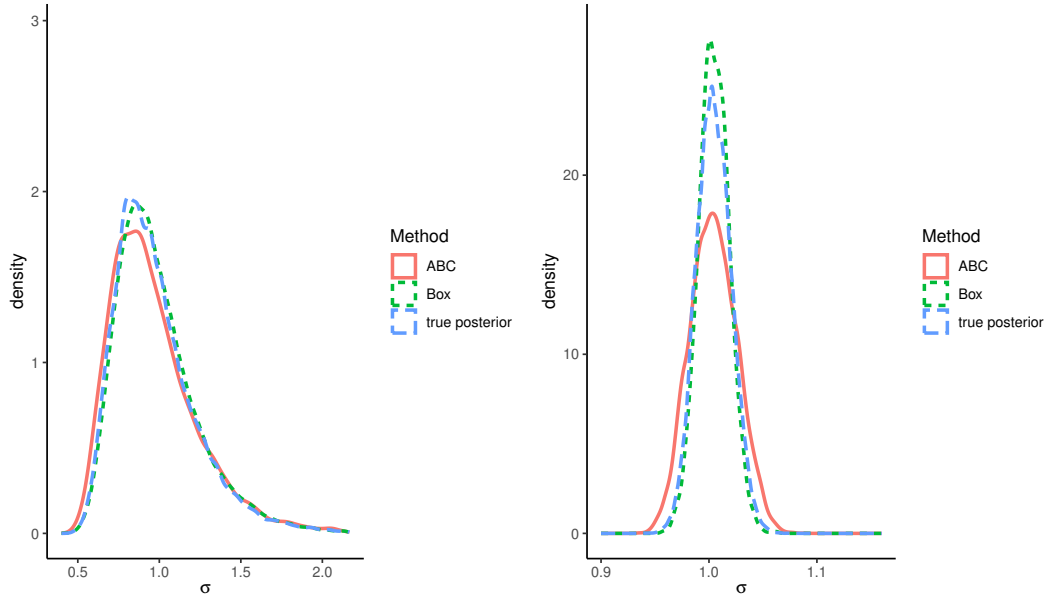

FIGURE 3.2: Normal model: approximate posteriors via ABC and Box-ABC compared to the true posteriorfor $n = 10$ (*left*) and $n = 2000$ (*right*).

structure, noting also that correlation might not be the most effective synthesis, is something we aim to avoid.

Hence, we consider and study the properties of a box built on the ordering of each component separately. The box in the multidimensional case is defined as

$$B = \times_{j=1}^{d} [t_j^{*(1)}, t_j^{*(2)}],$$

where $t_j^{*(1)}$ and $t_j^{*(2)}$ are the order statistics along the j-th coordinate. Equivalently, the parameter is accepted if

$$t_1^{(1)} < t_1^{\text{obs}} < t_1^{(2)} \;\&\; t_2^{(1)} < t_2^{\text{obs}} < t_2^{(2)} \;\&\; \ldots \;\&\; t_d^{(1)} < t_d^{\text{obs}} < t_d^{(2)}, \tag{3.5}$$

Thus, using the partitioning $t = (t_j, t_{[-j]})$, where $t_{[-j]} = (t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_d)$, $B = \cap_{j=1}^{d} B_j$, the corresponding pseudo-likelihood function can be written as

$$\mathcal{L}^{\text{box}}(\theta) \propto Pr(t^{\text{obs}} \in B|\theta) \propto \prod_{j}^{d} Pr(t_j \in B_j | t_{[-j]} \in B_{[-j]}, \theta). \tag{3.6}$$

3.3.1 Properties Box-ABC with multidimensional summaries

Consistency. Consider $t \in \mathbb{R}^d$ and $\theta \in \mathrm{R}$. To prove consistency in the multivariate case, we consider the following assumptions.

Assumption 2. Under the true θ_0 there exists τ s.t. $t \to \tau$ in probability.

Assumption 3. $Var_{\theta_0}[t(y)] \to 0$ as $n \to \infty$.

Theorem Under Assumptions 2 and 3, a CD is such that for $\theta \neq \theta_0$ $\lim_{n\to\infty} H_n(\theta)[1 - H_n(\theta)]| \to 0$. All the elements in the product (3.6) are conditional CDs, thus Box-ABC is consistent.

Median unbiasedness. To verify median unbiasedness, we write

$$Pr(t \in B) = Pr(t_1 \in B_1, t_2 \in B_2) = Pr(t_1 \in B_1 | t_2 \in B_2) Pr(t_2 \in B_2),$$

considering first a bivariate summary.

1. (Trivial case): if t_1 is independent of t_2 it becomes $Pr(t_1 \in B_1)Pr(t_2 \in B_2)$, where both are maximized for θ_0, hence the product is also maximized for θ_0.
2. If t_1 is not independent of t_2, alone $Pr(t_2 \in B_2)$ is maximized for θ_0. Denote the distribution of t_1 conditionally on the event $\{t_2 \in B_2\}$ with $F_{1,B_2}(\theta)$. We further assume that F_2 is symmetric. Then, there exists a region B_2^c symmetric in the sample space $T \in \mathbb{R}^2$ with the same probability mass, i.e. $Pr(B_2) = Pr(B_2^c)$ such that, each time a random interval B_2 contains the observed t_2, a random interval B_2^c is also formed by a set of two statistics defined a box and containing the observed t_2. The distribution of these statistics in the region $B_2 \cup B_2^c$ is symmetric and by construction its median coincides with the median of the marginal distribution.

3.3.2 Comparison to Data Depth approaches

Multivariate data, unlike univariate data, lacks a universally agreed-upon methodology for ordering. While order statistics theory has long been established for univariate data, extending this to multivariate cases remains challenging. The necessity for ordering multivariate observations spans diverse domains, for tasks such as estimating locations, identifying outliers, and enhancing visualization. Researchers have proposed various approaches, often using concepts of data depth to simplify the problem to a univariate context. Data depth functions (DD functions) are important tools that provide a measure of centrality within the multivariate sample space and guide the sequential ordering of points to ultimately delineate nested central regions. Although there is no unanimous consensus, as the multitude of depth notions leads to differing formulations of multivariate ordering, these techniques typically involve assigning depth values to data points in relation to their distribution, allowing ranking from the most distant outliers to the central

points. For example, the Simplicial depth (SD) method introduced by Liu (1990) determines the depth of a point by evaluating its presence within all combinations of simplex formed by the data points. Another instance of HD, also known as Tukey's depth, in one dimension is used to as the p-value for bilateral tests:

$$2\min\{Pr(Y \leq y^{\text{obs}} \mid \theta), Pr(Y \geq y^{\text{obs}} \mid \theta)\}.$$

The HD (Tukey's Depth) in the multivariate case requires the definition of a convex hull, which is the intersection of all halfspaces containing all sample points. The level sets of the HD are defined as the intersections of halfspaces containing $k < n$ sample points. Other DD functions are instead based on distance notions, as the Mahalanobis disance. For a comprehensive review, see Weller and Eddy (2015) Dungang *et al.* (2022) consider the concept of DD to define confidence intervals for multiparameter settings, called depth CDs, by ranking parameter values instead of data points. They propose to use the distribution of non-parametric bootstrap estimates to recover an approximate depth CD, motivated by the fact that efficient exact algorithms for computing half-space and simplicial depths in dimensions larger than 3 are not available. When examining the univariate counterpart of the SD, i.e. two independent observations drawn from a univariate cumulative distribution function, the SD is reduced to the form $\text{SD}_1(x) = 2F(x)[1F(x)]$, and the point that maximises $\text{SD}_1(x)$ corresponds to the median of the population. Note that the definition resembles that of the Box-ABC pseudo-likelihood function in the scalar case. In Box-ABC, the ordering performed on the sample space induces an ordering on the parameter space, similarly to the idea of the depth-CD of Dungang *et al.* (2022). In a multivariate setting, instead of verifying that one observation is central to all the obtainable simplexes, Box-ABC simulates hyper-rectangles and assigns a measure of centrality of the (fixed) observation via rejection sampling, to the parameter from which pseudo-observations are drawn. In particular, the complexity related to the reliance on the simplex is reduced with hyper-rectangles. While the CD furnishes a univariate ordering of the values of Θ, $\mathcal{L}^{\text{box}}(\theta)$ gives a natural ordering of $\theta \in \Theta$ from the center outward as a Data Depth (DD) function.

3.3.3 Simulation study

Here we study the median unbiasedness of the maximizer in a model where three summary statistics are used to resume information about a parameter of interest.

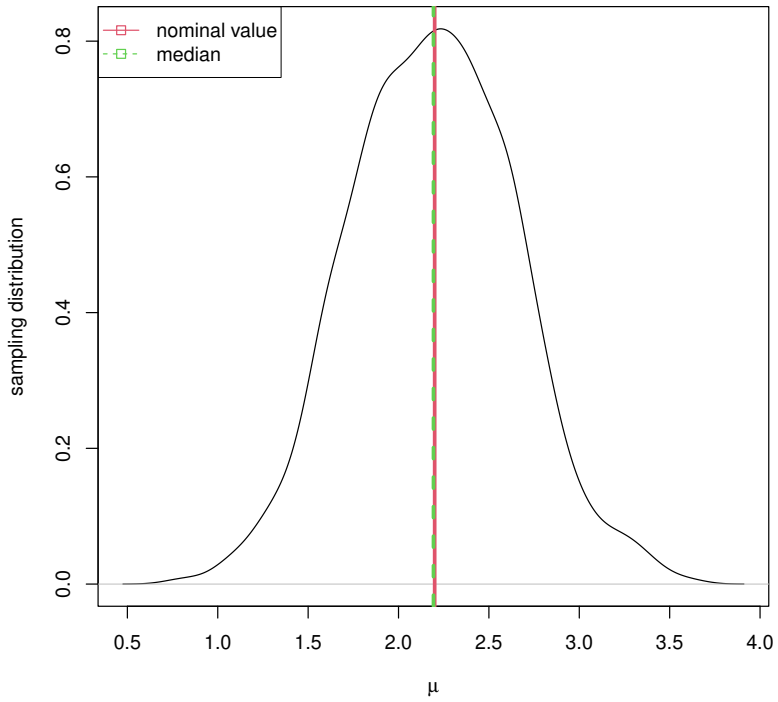

FIGURE 3.3: Distribution of maxima of the "box"-pseudo-likelihoods in a simulation study consisting of 1500 Replications from the multivariate correlated normal model.

We considered the following model:

$$y_i \sim \text{Normal}_2(\mu, \Sigma), \text{ with } \Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 2 \end{pmatrix}, \ i = 1, \ldots, 5.$$

We consider 1500 datasets obtained from this model. For each of them, we draw 10000 values from a uniform prior in $[0, 5]$ for the parameter of interest μ, whose nominal value was 2.2. For each dataset, we construct the Box-pseudo-likelihood using as summary statistics the empirical means of the three components of the normal, which are correlated. After a density estimation, we retrieve the maximizer of the likelihood. As shown in Figure 3.3 the estimator is median unbiased, as expected. Since the distribution of the summary statistics is symmetric, it is also unbiased. Finally, Figure 3.4 presents 10 replications of posteriors (shown as continuous lines) and corresponding approximations (depicted as dashed lines) based on Box-ABC, with each replication comprising $2 \cdot 10^4$ proposals. The color scheme remains consistent for each dataset.

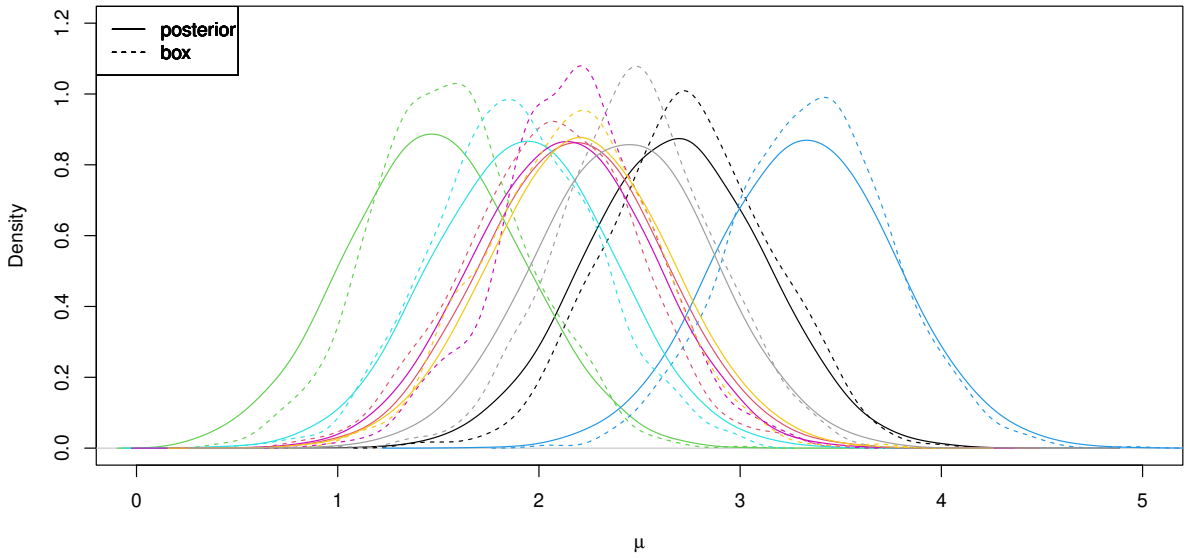

FIGURE 3.4: Posterior distributions (continuous lines) for the parameter μ across 10 datasets drawn from the multivariate correlated normal model, alongside "box"-approximations (dashed lines).

3.3.4 Example: Ricker's Model

We consider the Ricker's model (Ricker, 1954),which describes the evolution of the number of animals of a certain species according to the equation

$$\log(N(t)) = \log(r) + \log(N(t-1)) - N(t-1) + \sigma e(t),$$

where $N(t)$ is the unknown population at time t, $\log(r)$ is the logarithmic growth rate, σ is the standard deviation of innovation and $e(t)$ is an independent Gaussian error. Assuming $N(0) = 0$, the observed population at time t, y_t, is a Poisson random variable: $y_t \sim \text{Poisson}(\phi N(t))$ where ϕ is a scaling parameter. Suppose $\sigma = 0.3$, $\phi = 10$, and $\log(r) = 3.8$, where the latter parameter only is unknown. The size of the dataset considered is $t = 50$. Following Grazian and Fan (2020), the set of summary statistics used are

1. the number of observations greater than 10,
2. the median count,
3. the maximum count,
4. the quantile of level 0.75,
5. the sample mean of the observations greater than 1.

The prior distribution for the parameter of interest is $\log(r) \sim \text{Uniform}(0, 10)$. For ABC and Box-ABC two runs, with 10^5 and 10^6 simulations, respectively were performed and for ABC a series of thresholds were considered: $\varepsilon = 5.0, 10, 20, 30$ with distance as $\delta = ||t^* - t^{\text{box}}||^2$ was chosen to run to standard ABC. Results are displayed in figure 3.5, while table **??** summarizes number of accepted values, fraction of accepted proposals on number of simulations (Acc/Gen) from the model and average computational time in seconds on a laptop CPU with a clock speed of 1.00 GHz necessary to obtain one accepted value (Time/Acc).

	R=10^5	Acc/Gen	Time/Acc	R=10^6	Acc/Gen	Time/Acc
Box-ABC	303	0.15	0.467	2728	0.14	0.539
ABC $\varepsilon = 5$	46	0.05	4.884	113	0.02	5.549
ABC $\varepsilon = 10$	148	0.15	1.168	401	0.04	1.564
ABC $\varepsilon = 20$	204	0.20	0.363	1323	0.14	0.474
ABC $\varepsilon = 30$	258	0.26	0.208	2448	0.24	0.256

TABLE 3.2: Synthesis of simulation results for the Ricker's model with five summary statistics: absolute number of accepted proposals, proportion of accepted proposals on the number of generations from the model (Acc/Gen) and average computational time to accept once (Time/Acc), with time expressed in seconds.

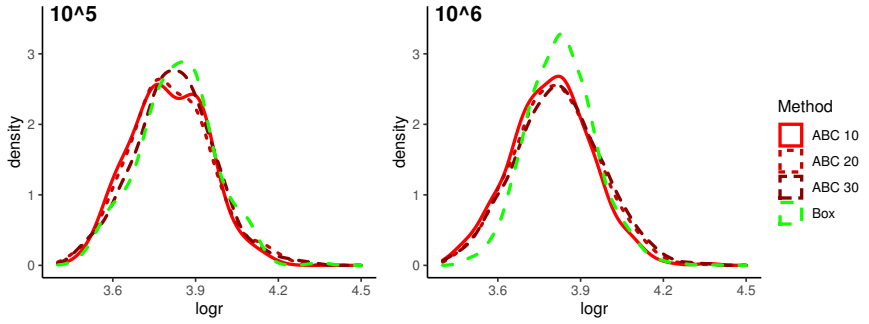

FIGURE 3.5: Results for the Ricker model: approximate posterior with five summary statistics.

3.4 Discussion

In this Chapter, we have examined a strategy for obtaining an approximate posterior distribution for performing inference in the absence of the likelihood function. The approximated likelihood is related to the concept of Confidence Distribution. We have focused on performing inference from many summary statistics without assuming a likelihood function for such summaries for a single parameter of interest. Missing from our discussion is the treatment of multidimensional parameters. Ideally, the method can

be extended to deal with many unknown parameters, by constructing a "box", incorporating all the summary statistics available, as done in Section 3.3.4, but proposing from a multidimensional parameter. Another idea is to consider partitions of summary statistics that are informative for blocks of parameters, similar to what has been studied by Clarté *et al.* (2021), thus working with the product of approximated full conditionals. The investigation of theoretical properties in the case of vector parameters still needs to be explored. Although the posterior is not recovered by the proposed method, since the likelihood is replaced by a transformation of a confidence distribution (or a depth CD) for one parameter at a time, we suggest that our proposal could be used as a means of monitoring the quality of the ABC approximation and as a means of tuning the tolerance parameters. This could be particularly beneficial for parameters where the challenge lies in the lack of concentration of the marginal posterior with a small ε range.

Chapter 4

Coupling of MCMC algorithms on manifolds

4.1 Distributions on submanifolds

Various problems in statistics lead to the problem of sampling from a probability distribution on a submanifold. This restricted space can be obtained through rather simple linear equations, or a more challenging set of functions, called constraints. Consider the space $\mathbb{R}^D$ and a constraint function $q(x) : \mathbb{R}^D \to \mathbb{R}^m$, $(D > m \geq 1)$. We consider distributions with support the submanifold $\{x \in \mathbb{R}^D | q(x) = 0_m\}$. In some models and scenarios, a reparametrization, such as polar coordinates, may aid in restoring the dimensionality of the space and ease sampling and performing inference. For example, a point on a three dimensional sphere can be represented using two angles, in the unconstrained space $[0, 2\pi] \times [0, 2\pi]$. However, for certain cases, such reparametrizations may not be available. Thus, samping distributions on such spaces might be challenging.

A commonly encountered example is the realm of positive semi-definite (PSD) matrices, which include covariance matrices. If the covariance matrix is of dimension $D \times D$, the effective space is $D(D+1)/2$ and the collection of PSD matrices forms a convex cone, known as the semidefinite cone in the space of symmetric matrices (see for instance Beskos and Kamatani, 2022). Other classical problems related to distributions on submanifolds can be found in the field of directional statistics (see e.g. Mardia *et al.* 2000), with numerous applications in astrostatistics and biology. Other instances come from the classical testing literature, where the constraints come from conditioning operations which allow to eliminate nuisance parameters and gaining power. For instance, consider an unknown parameter vector $\theta \in \mathbb{R}^p$, indexing a model $p(y|\theta)$, partitioned

into $\theta = (\psi, \lambda)$, where ψ is the scalar parameter of interest. If a sufficient statistic $s(y) = (V, U)$ is such that the model factorizes as $p_{U|V}(u|v, \psi)p_V(v|\lambda, \psi)$, then inference can be carried on the conditional model $p_{U|V}$, which does not depend on ψ. Often a closed form $p_{V|U}$ is not given, suggesting to resort to MCMC algorithms (see e.g. Davison 2003, Chapter 12 and Lindqvist *et al.* 2022). Diaconis *et al.*, 2013 describe how the problem arises in goodness-of-fit hypothesis testing, and more recently Barber and Janson (2022) consider the similar framewok of Co-sufficient Sampling (CSS), that addresses the challenge of testing a composite null hypothesis in the presence of unknown model parameters. CSS involves conditioning the analysis on the Maximum Likelihood Estimator (MLE) of these parameters, and simplifies subsequent statistical testing and analyses. In Bayesian statistics, conditioning is the key ingredient in the Bayes Theorem. Following the previous notation, let $\pi(\theta)$ be the prior distribution. The posterior is

$$\pi(\theta|y^{\text{obs}}) = \frac{\pi(\theta)p(y^{\text{obs}}|\theta)}{p(y^{\text{obs}})},$$

where the likelihood function $p(y^{\text{obs}}|\theta)$ is given by the density function of the model for fixed observed data y^{obs}. In this framework, when the data generating process can be represented as $y = g(\theta, u)$, with θ being the parameter vector, u a vector of random inputs of dimension $|u|$, following $\rho(u)$, and $g(\cdot, \cdot)$ a deterministic function, the likelihood function can be written as $P\{g(u, \theta) \in \mathcal{G}_\theta(y^{\text{obs}})|\theta\}$, which is

$$\int_{\mathcal{G}_\theta(y^{\text{obs}})} \frac{\rho(u)}{\det\left(\nabla_u g(u,\theta)\nabla_u g(u,\theta)^\top\right)^{1/2}}.$$

and $\mathcal{G}_\theta(y) = \{u \in U : g(u, \theta) = y\}$ (Liu *et al.*, 2022).

Graham and Storkey (2017) first identify this representation as the equation of a submanifold, of dimension $d - n$, where $d = p + |u|$, providing a solution for sampling posterior distributions with intractable likelihood functions, with strong links to Approximate Bayesian Computation (ABC) type of problems. Hannig *et al.* (2016) and Liu *et al.* (2022) identify a similar setting in the context of Generalized Fiducial Inference, where constraints are defined in terms of data generating equations and parameter estimates. Bornn *et al.* (2019) encounter similar problems when embedding moment conditions in Bayesian non-parametric priors, with applications e.g. to regression with instrumental variables, Gallant *et al.* (2022) list various cases in econometrics and Hartmann (2008) describes related problem in the field of statistical mechanics and in particular in simulation of molecular dynamics.

The main difficulty in sampling from a target probability distribution supported

on a constrained set is that proposing values on the correct support is not immediate. Also constructing a standard Metropolis–Rosenbluth–Teller–Hastings (MRTH) algorithm becomes more challenging when dealing with constrained spaces compared to unconstrained scenarios. In the unconstrained case, the algorithm proposes new values in the neighborhood of the current chain value and accepts or rejects these values. However, when dealing with constrained spaces or submanifolds, determining how to propose moves within the submanifold becomes less straightforward. In particular, the submanifold may have intricate geometric properties, making it complex to navigate and ensuring that the sampled points satisfy the imposed nonlinear constraints adds complexity.

Despite the challenges, a substantial and longstanding body of research provides Markov Chain Monte Carlo (MCMC) algorithms tailored for sampling on constrained spaces (see Chapter 3 in Rousset *et al.*, 2010), encompassing various adaptations of discretized Langevin diffusions and Hamiltonian Monte Carlo techniques. The origins of this field trace back to early contributions such as Andersen (1983) and a stream of research stemming from the domain of molecular dynamics. Recently, a renewed attention and consideration in Mathematics and Statistics, especially from Zappa *et al.* (2018), highlighting the challenges and potentially proposing initial solutions. A more formal treatment and methodological extensions were subsequently introduced by Lelièvre *et al.* (2020), providing a framework for handling developments in this domain. These approaches have demonstrated utility in addressing the aforementioned examples.

In this chapter, we study couplings of the transition kernels of some archetypical MCMC algorithms designed to sample probability distributions on submanifolds, with the random walk proposal of Zappa *et al.* (2018) as the primary case. Our goal is to construct couplings that can be implemented to generate pairs of chains that coincide exactly after a random number of iterations, called faithful couplings in Rosenthal (1997).

Being able to generate faithful couplings of MCMC trajectories provides practical benefits for the MCMC practitioner: unbiased estimators that are easy to parallelize (Glynn and Rhee, 2014; Agapiou *et al.*, 2018; Jacob *et al.*, 2020), convergence diagnostics in the form of upper bounds on the distance to stationarity after a fixed number of iterations (Johnson, 1996, 1998; Biswas *et al.*, 2019), and asymptotic variance estimators that are useful to measure and compare the performance of MCMC algorithms (Douc *et al.*, 2022).

4.2 Random walk MCMC on submanifolds

We establish the notation following the formalism presented by Lelièvre *et al.* (2020) to describe the Random Walk Metropolis–Rosenbluth–Teller–Hastings algorithm introduced by Zappa *et al.* (2018). Consider the submanifold

$$\mathcal{S} = \{x \in \mathbb{R}^D \mid q(x) = 0 \in \mathbb{R}^m\}, \tag{4.1}$$

in the ambient space $\mathbb{R}^D$, where q is a $\mathbb{R}^m$-valued *constraint* function; i.e. $\mathcal{S}$ is the zero level set of $q : \mathbb{R}^D \to \mathbb{R}^m$ with $D > m$. We write $\nabla q(x)$ for the Jacobian matrix of q, that is the $D \times m$ matrix whose (i, j) entry is the derivative $\partial q_j(x)/\partial x_i$. We consider the goal of sampling from a probability distribution π, supported on $\mathcal{S}$,

$$\pi(dx) = \frac{1}{Z} \exp(-V(x))\sigma_{\mathcal{S}}(dx), \tag{4.2}$$

where $V : \mathbb{R}^D \to \mathbb{R}$ is a potential function, $\sigma_{\mathcal{S}}(dx)$ is the surface measure induced by the standard scalar product on $\mathbb{R}^D$, also called the Hausdorff measure on $\mathcal{S}$ and Z is the normalizing constant. We assume that we can evaluate the potential function V at all $x \in \mathcal{S}$. If the potential V is constant then π is the uniform distribution on $\mathcal{S}$. We introduce the same smoothness assumptions as Lelièvre *et al.* (2020) as we will rely on some of their results.

Assumption 4. The potential function V in (4.2) has two continuous derivatives, i.e. V is C^2.

We also make the following assumptions on the constraint function q.

Assumption 5. The function q is smooth (i.e. C^∞), and $\mathcal{S}$ is a compact subset of $\mathbb{R}^D$. For all $x \in \mathcal{S}$, $\nabla q(x)$ is of full rank.

Under Assumption 5, Theorem 5.12 of Lee (2012) states that $\mathcal{S}$ is of codimension m, i.e. is of dimension $d := D - m$. The tangent space of $\mathcal{S}$ at any $x \in \mathcal{S}$, denoted by $\mathcal{T}_x$, is a d-dimensional vector space (Proposition 3.10 in Lee 2012). Its orthogonal complement is denoted by $\mathcal{T}_x^\perp$ and is of dimension m. To obtain a orthonormal basis for $\mathcal{T}_x$, one numerically obtains a QR decomposition of $\nabla q(x)$:

$$\nabla q(x) = Q_x R_x = \begin{pmatrix} N_x & U_x \end{pmatrix} \begin{pmatrix} A_x \\ \hline 0 \end{pmatrix}, \tag{4.3}$$

where Q_x is a $D \times D$ matrix with orthonormal columns, and R_x is a $D \times m$ matrix made of an $m \times m$ upper triangular matrix A_x placed above $D - m$ rows of zeros. We denote

by N_x and U_x the matrices made of the first m and the last d columns of Q_x; then N_x forms a basis for $\mathcal{T}_x^\perp$ and U_x forms a basis for $\mathcal{T}_x$. The orthogonal projection matrix onto $\mathcal{T}_x$ can be written $P_x = I_D - N_x N_x^\top$, where I_D is the $D \times D$ identity matrix.

4.2.1 Random walk proposals on submanifolds

In the following we elaborate on the construction of an ergodic Markov chain taking values in $\mathcal{S}$ with stationary distribution π. The general scheme used by algorithms in the literature (Brubaker *et al.*, 2012; Zappa *et al.*, 2018; Lelièvre *et al.*, 2019, 2020) to explore the target space is as follows: if the chain is at point $x \in \mathcal{S}$, it moves along the tangent space to $\mathcal{S}$ at x; corrective steps are then taken to ensure that the proposal lies on the submanifold; the proposal is finally accepted or rejected.

4.2.2 Projections along the submanifold

A key ingredient to explore probability distributions on submanifolds is Newton's method to project a point $z \in \mathbb{R}^D$ onto $\mathcal{S}$ by following a direction $b \in \mathbb{R}^{D \times m}$, which in the present setting will be given by $\nabla q(x)$ for some $x \in \mathcal{S}$. The idea is to define $f : \mathbb{R}^m \to \mathbb{R}^m$, with $f(\alpha) = q(x + b\alpha)$, and to solve $f(\alpha) = 0$ by iterating $\alpha_t = \alpha_{t-1} - (\nabla f(\alpha_{t-1}))^{-1} f(\alpha_{t-1})$, see Algorithm 17.

Algorithm 17 Newton's method to project $z \in \mathbb{R}^D$ onto $\mathcal{S} = \{x \in \mathbb{R}^D : q(x) = 0\}$ along $b \in \mathbb{R}^{D \times m}$

function NEWTON'S METHOD (z, b, q, max iteration, tolerance)
  set $\alpha = (0, \ldots, 0) \in \mathbb{R}^m$
  set iteration $= 0$
  while iteration $\leq$ max iteration **and** $|q(z + b\alpha)| >$ tolerance **do**
    $\alpha = \alpha - (b^\top \nabla q(z + b\alpha))^{-1} q(z + b\alpha)$
    iteration $=$ iteration $+ 1$
  end while
return α
end function

Assume that a Markov chain is currently at state $x \in \mathcal{S}$. The idea of the random walk proposal in Zappa *et al.* (2018) is to sample a deviation from x on $\mathcal{T}_x$, of the form $x + U_x\nu \in \mathcal{T}_x$ for some $\mathbb{R}^d$-valued random variable ν with distribution p_ν. The point $x + U_x\nu$ can then be projected onto $\mathcal{S}$ along the direction of $\nabla q(x)$ using Newton's method. If successful, we obtain $\alpha \in \mathbb{R}^m$ such that $x + U_x\nu + \nabla q(x)\alpha \in \mathcal{S}$.

This projection step can fail for two reasons. Denote

$$\mathcal{F}_x(\nu) = \{y \in \mathcal{S} : \exists \alpha, y = x + U_x\nu + \nabla q(x)\alpha\}\,.$$

First, the set $\mathcal{F}_x(\nu)$ can be empty, in which case Newton's method (starting from $x+U_x\nu$ and with direction $\nabla q(x)$) is bound to fail. Second, $\mathcal{F}_x(\nu)$ can be non-empty and yet Newton's method fails, for example because the maximum number of iterations was reached before convergence. Denote by $\hat{\mathcal{F}}_x(\nu) \subseteq \mathcal{F}_x(\nu)$ the subset of points $y \in \mathcal{S}$ of the form $x + U_x\nu + \nabla q(x)\alpha$ for some α which can be found using Newton's method: it is possible that $\mathcal{F}_x(\nu) \neq \emptyset$ but $\hat{\mathcal{F}}_x(\nu) = \emptyset$. Lelièvre *et al.* (2020) entertain the possibility of multiple elements in $\hat{\mathcal{F}}_x(\nu)$, but here for simplicity we focus on the case where $\hat{\mathcal{F}}_x(\nu)$ is either empty or contains a single element. Effectively, we have restricted ourselves to a deterministic projection method: even if $\mathcal{F}_x(\nu)$ contains more than one element, only one element can be discovered by our implementation of Newton's algorithm, so $|\hat{\mathcal{F}}_x(\nu)| \leq 1$. Towards formalizing the proposal mechanism of Zappa *et al.* (2018), we follow Lelièvre *et al.* (2020) and make the following observation:

$$\forall x, y \in \mathcal{S}, \quad y \in \mathcal{F}_x(\nu) \Leftrightarrow \nu = G_x(y), \tag{4.4}$$

with $G_x : \mathcal{S} \to \mathbb{R}^d$ defined for all $x \in \mathcal{S}$ as

$$G_x(y) = U_x^\top (y - x). \tag{4.5}$$

Proposition 1 in Lelièvre *et al.* (2020) (under Assumptions 4-5, assumed throughout) shows that G_x is "locally" a C^1-diffeomorphism, for all $y \in \mathcal{S}$ except those such that $\det \nabla q(y)^\top \nabla q(x) = 0$. Introduce $\mathcal{C}_x = \{y \in \mathcal{S} : \det \nabla q(y)^\top \nabla q(x) = 0\}$. For a given x, this set could have non-zero mass with respect to $\sigma_{\mathcal{S}}$. However, the set $\mathcal{N}_x = G_x(\mathcal{C}_x)$ is of measure zero with respect to the Lebesgue measure on $\mathbb{R}^d$, according to Proposition 2 in Lelièvre *et al.* (2020). As a result, by sampling ν from a continuous measure on $\mathbb{R}^d$ and constructing $y = x + U_x\nu + \nabla q(x)\alpha$ for some α, as described above, we end up with $\mathbb{P}(y \in \mathcal{C}_x) = 0$.

4.2.3 Proposal distribution

We now define the proposal distribution $q(x, dy)$ of Zappa *et al.* (2018) in two ways: algorithmically and through its density. Algorithmically, from a current position $x \in \mathcal{S}$:

1. draw $\nu \sim p_\nu$ in $\mathbb{R}^d$, absolutely continuous with respect to the Lebesgue measure (for example $\nu \sim \text{Normal}(0, I)$;
2. run Newton's method (Algorithm 17) to project $x + U_x\nu$ onto $\mathcal{S}$ along $\nabla q(x)$ to obtain $\hat{\mathcal{F}}_x(\nu)$;
3. if $\hat{\mathcal{F}}_x(\nu) = \emptyset$, set $y = x$, otherwise set y as the unique element in $\hat{\mathcal{F}}_x(\nu)$.

In terms of density, the proposal distribution $q(x, dy)$ on $\mathcal{S}$ corresponding to the above mechanism is shown in Lemma 3 of Lelièvre *et al.* (2020) to be of the following form, for $x \in \mathcal{S}$,

$$q(x, dy) = \delta_x(dy)\, r(x) + |\det DG_x(y)| \cdot p_\nu(G_x(y)) \cdot 1(y \in \text{Im}\hat{\mathcal{F}}_x \setminus \mathcal{C}_x) \cdot \sigma_{\mathcal{S}}(dy). \quad (4.6)$$

Here we abuse notation and define

$$\text{Im}\hat{\mathcal{F}}_x = \bigcup_{\nu \in \mathbb{R}^d} \hat{\mathcal{F}}_x(\nu) = \left\{ y \in \mathcal{S} : \exists \nu \in \mathbb{R}^d \text{ such that } \hat{\mathcal{F}}_x(\nu) = \{y\} \right\},$$

the set of "Newton-reachable" points in $\mathcal{S}$ from x. The term $\det DG_x(y)$ represents the determinant of the differential of G_x at $y \in \mathcal{S}$ and can be computed as

$$\det DG_x(y) = \det U_x^\top U_y. \quad (4.7)$$

The Dirac mass at x in (4.6) corresponds to the cases where Newton's algorithm fails, i.e. $\hat{\mathcal{F}}_x(\nu) = \emptyset$. The associated probability can be written as

$$r(x) = 1 - \int_{\mathcal{S}} |\det DG_x(y)| \cdot p_\nu(G_x(y)) \cdot \mathbb{I}(y \in \text{Im}\hat{\mathcal{F}}_x \setminus \mathcal{C}_x)\sigma_{\mathcal{S}}(dy). \quad (4.8)$$

If $y \sim q(x, dy)$ is in fact equal to x, the chain remains at x. If $y \neq x$, the chain moves to y with a certain probability, following a standard Metropolis–Rosenbluth–Teller–Hastings (MRTH) scheme, with acceptance ratio of the form $\pi(y)q(y, x)/\pi(x)q(x, y)$. The reverse proposal density $q(y, x)$ features the indicator $1(x \in \text{Im}\hat{\mathcal{F}}_y \setminus C_y)$. To evaluate it, we need to check if x could have been reached from y; this is called the *reverse projection check* in Zappa *et al.* (2018). Note that we always have $x \in \mathcal{F}_y(\nu^\top)$ for $\nu^\top = G_y(x)$, however we might have $x \notin \hat{\mathcal{F}}_y(\nu^\top)$, for example because $\hat{\mathcal{F}}_y(\nu^\top)$ might be empty: in that case, $q(y, x) = 0$ and thus the proposal y is rejected. If the reversibility check goes through (i.e. $\hat{\mathcal{F}}_y(\nu^\top) = \{x\}$), y is accepted with probability

$$a(x, y) = \min\left(1, \frac{\exp(-V(y)) \cdot |\det DG_y(x)| \cdot p_\nu(G_y(x))}{\exp(-V(x)) \cdot |\det DG_x(y)| \cdot p_\nu(G_x(y))}\right), \quad (4.9)$$

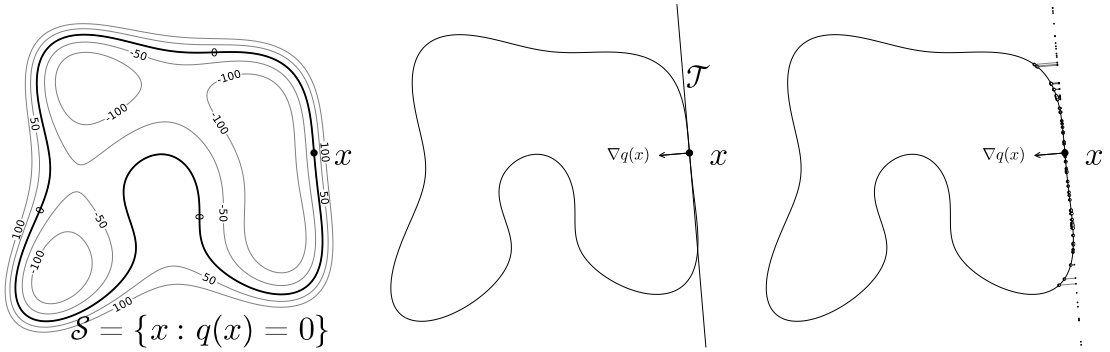

FIGURE 4.1: *Left*: submanifold $\mathcal{S}$ as a level set of a function q. *Middle*: $x \in \mathcal{S}$, its tangent space $\mathcal{T}_x$, and the direction $\nabla q(x)$. *Right*: proposals obtained by Newton-projecting points on $\mathcal{T}_x$ onto $\mathcal{S}$ following $\nabla q(x)$, with the possibility of failure.

where the ratio of determinants is equal to one because $\det DG_x(y) = \det U_x^\top U_y = \det DG_y(x)$. The overall transition kernel of random walk MRTH is denoted by $P(x, dy)$, and pseudo-code is in Algorithm 18. Figure 4.1 provides some visual description of the setup and the proposal mechanism.

Algorithm 18 MRTH kernel on a submanifold à la Zappa *et al.* (2018)

Input: $x \in \mathcal{S}$
Compute QR decomposition of $\nabla q(x)$, and obtain U_x, an orthonormal basis for $\mathcal{T}_x$.
Sample $\nu \sim p_\nu$.
Run Newton's method (Alg. 17) to find $\alpha \in \mathbb{R}^m$ such that $y := x + U_x\nu + \nabla q(x)\alpha \in \mathcal{S}$.
if fail **then return** x.
end if
Compute QR decomposition of $\nabla q(y)$, and obtain U_y, an orthonormal basis for $\mathcal{T}_y$.
Compute $\nu' = G_y(x) = U_y^\top(x - y)$.
Run Newton's method (Alg. 17) to find $\alpha' \in \mathbb{R}^m$ such that $x = y + U_y\nu' + \nabla q(y)\alpha'$.
if fail **then return** x.
end if
Draw $U \sim \text{Uniform}(0, 1)$.
if $U < \exp(-V(y)) \cdot p_\nu(G_y(x))/(\exp(-V(x)) \cdot p_\nu(G_x(y)))$ **then return** y.
else return x.
end if

Lelièvre *et al.* (2020) consider a "Langevin" generalization, where instead of considering $x + U_x\nu$ on $\mathcal{T}_x$ before projecting on $\mathcal{S}$, we consider $x - \lambda\nabla V(x) + U_x\nu$, where ∇V is the gradient of the potential in (4.2) and λ is a stepsize. This involves minimal changes to Algorithm 18, and our contributions can be easily adapted to this change.

4.2.4 Couplings of MCMC algorithms

The output of an MCMC algorithm is typically used to estimate integrals of the form $I = \int h(x)\pi(dx)$ for some test function h. The Markov Chain $(X_t)_{t\in\mathbb{N}}$ generated by iterating Algorithm 18 can be used to produce an estimator $\hat{I}_T = \frac{1}{T}\sum_{t=1}^{T} h(X_t)$ which verifies $\hat{I}_T \xrightarrow[T\to\infty]{Pr} I$. However, this estimator is biased after a finite number of iterations, since the elements (X_t) do not exactly follow π, so there are few guarantees for the quality of the result after a finite number of iterations. Also, as is the case for many MCMC algorithms, several tuning parameters must be chosen by the practitioner: the distribution and variance of the random step ν; the parameters of Newton's method; the total number of iterations. In practice, these are selected using an *ad hoc* measure of quality of the MCMC, but we lack a principled method to select these parameters or to compare different MCMC algorithms targeting the same distribution π; we also lack theoretical bounds on the quality of the MCMC output.

For the unconstrained case of a distribution with support in $\mathbb{R}^D$, a solution was proposed by Biswas *et al.* (2019) and Jacob *et al.* (2020), using pairs of lagged coupled Markov chains. We construct two Markov Chains $(X_t)_t$ and $(\tilde{X}_t)_t$. Marginally, each chain follows Algorithm 18; the chains are coupled so that they eventually meet: at some random time τ, $X_\tau = \tilde{X}_{\tau-\ell}$ for some user-specified lag ℓ. Once the chains have met, they stay together: $\forall t \geq \tau, X_t = \tilde{X}_{t-\ell}$.

With this framework, Glynn and Rhee (2014), and Jacob *et al.* (2020) show how to build an unbiased estimate of I. Biswas *et al.* (2019) show that the Total Variation distance between the target π and the distribution of the Markov Chain after t iterations can be bounded:

$$d_{TV}(\pi_t, \pi) \leq \mathbb{E}\left[\max\left(0, \left\lceil \frac{\tau - \ell - t}{\ell} \right\rceil\right)\right]. \tag{4.10}$$

In the remainder of this Chapter, we show how to build such coupled MCMC for distributions constrained to a submanifold, and demonstrate the utility of this approach on several examples.

4.3 Coupling random walk proposals on submanifolds

We wish to couple two chains, currently at position x, and $\tilde{x}$ respectively. Thus, we consider the problem of constructing proposal mechanisms, for $Y \sim q(x, dy)$ and $\tilde{Y} \sim q(\tilde{x}, dy)$ such that $\{Y = \tilde{Y}\}$ can occur, where $q(x, dy)$ is the proposal distribution of Zappa *et al.* (2018) described in Section 4.2.1. Here the projection on $\mathcal{S}$ can fail and

thus $q(x, \{x\})$ can be non-zero. For simplicity, we rewrite the proposal density in (4.6) as

$$q(x, dy) = r(x)\delta_x(dy) + q_s(x, y)\sigma_{\mathcal{S}}(dy). \tag{4.11}$$

We know how to evaluate $q_s(x, y)$ – the index s stands for "smooth" as it represents the smooth part of the proposal distribution $q(x, y)$ – and we do not necessarily know how to evaluate $r(x)$.

4.3.1 A first coupling

As described above, the problem is similar to the setting of Wang *et al.* (2021), who consider couplings of MRTH kernels where the accept/reject step also induces a Dirac mass at the current position and an extension of the $g-$coupling of Johnson (1998). The total variation (TV) distance between y and $\tilde{y}$ satisfies (e.g. Lemma 1 in Wang *et al.* 2021)

$$|q(x, dy) - q(\tilde{x}, dy)|_{\mathrm{TV}} = 1 - \int \min(q_s(x, y), q_s(\tilde{x}, y))\sigma_{\mathcal{S}}(dy),$$

and this is the maximal probability of the event $\{y = \tilde{y}\}$, achieved by so-called "maximal couplings". Algorithm 3 in Wang *et al.* (2021) samples from a maximal coupling of transition kernels of Markov Chains, $q(x, dy)$ and $q(\tilde{x}, dy)$, and is reproduced in Algorithm 19 here. The idea of the procedure is as follows: the next state of the x-chain is proposed. If the proposal (y) is not rejected after a Metropolis-Hastings ratio check (the transition kernel does not output a Dirac) and an attempt to set the position of second chain equal to the first is successful, the same state is used as next position of the chain. Otherwise, if one of the conditions is not met, it is checked whether the proposal $(\tilde{y})$ of the $\tilde{x}$-chain is such that a meeting is not possible, either because the proposal is a Dirac or because an attempt to couple the proposals would fail. The validity of that algorithm and its maximality are established in Proposition 2, Appendix A.1 of Wang *et al.* (2021). In words, our coupling strategy consists in dealing with the rejections in the proposal kernel as done for rejections of a transition kernel as Wang *et al.* (2021). To draw $Y \sim q(x, dy)$, we draw $\nu \sim p_\nu$ and compute $\hat{\mathcal{F}}_x$. If $\hat{\mathcal{F}}_x = \emptyset$, then $Y = x$ and we cannot hope to output $\tilde{Y} = Y$; else let y be the single element of $\hat{\mathcal{F}}_x$, and we have $Y = y$, which is a realization of $q_s(x, dy)$. We now attempt to propose $\tilde{Y} = y$ by using a maximal coupling of $q_s(x, dy)$ and $q_s(\tilde{x}, dy)$. We project y onto $\mathcal{T}_{\tilde{x}}$, giving the unique value $\tilde{\nu}$ such that $y \in \mathcal{F}_{\tilde{x}}$, namely

$$\tilde{\nu} = U_{\tilde{x}}^{\top}(y - \tilde{x}). \tag{4.12}$$

If $\hat{\mathcal{F}}_{\tilde{x}} = \{y\}$, then $q_s(\tilde{x}, y) > 0$ and the two chains may meet; else the chains cannot couple at this step.

Algorithm 19 Max coupling of two kernels with point mass of form (4.11)

function SAMPLE FROM MAX COUPLING($q(x, dy)$, $q(\tilde{x}, dy)$)
  Draw $Y \sim q(x, dy)$ and $U \sim \text{Uniform}(0, 1)$.
  if $Y \neq x$ and $U \leq q_s(\tilde{x}, Y)/q_s(x, Y)$ **then**
    return (Y, Y) (identical states).
  else
    while true **do**
      Draw $\tilde{Y} \sim q(\tilde{x}, dy)$.
      if $\tilde{Y} = \tilde{x}$ **then**
        return $(Y, \tilde{Y})$.
      else
        Draw $V \sim \text{Uniform}(0, 1)$.
        if $V > q_s(x, \tilde{Y})/q_s(\tilde{x}, \tilde{Y})$ **then**
          return $(Y, \tilde{Y})$
        end if
      end if
    end while
  end if
end function

Note that to implement Algorithm 19 we need to evaluate ratios of the form $q_s(x, y)/q_s(\tilde{x}, y)$, which involves the computation of $\det DG_x(y) = \det U_x^\top U_y$ and $\det DG_{\tilde{x}}(y) = \det U_{\tilde{x}}^\top U_y$, which is normally not required when running a single chain. Once we have a way of coupling the proposals, we can use a common random uniform variable to accept/reject Y for the first chain and $\tilde{Y}$ for the second chain. Alternatively, we may employ a strategy to couple the MRTH kernels described in Algorithm 5 of Wang *et al.* (2021).

4.3.2 Scaling and reflection couplings

The above strategy makes it possible to obtain exact meetings, but scaling with dimension is not expected to work well, since it is a direct adaptation of a standard coupling of random walk MRTH (Johnson, 1998), which does not scale well with dimension, even in an unconstrained space, as shown experimentally in Jacob *et al.* (2020). More precisely, average meeting times of coupled chains are expected to increase rapidly with dimension, even using an optimal scaling of the stepsize under which the mixing time of the chain increases linearly.

In Jacob *et al.* (2020) it was shown experimentally that reflection couplings (e.g. Bou-Rabee *et al.*, 2018) of Normal proposals lead to shorter meeting times. Papp and

Sherlock (2022) explain this phenomenon for spherical Normal targets and propose another coupling that provide better (in a sense, optimal) performance for more general targets. The key point is that meeting is only possible in high dimensions if the chains are close, therefore short meeting times can only achieved by couplings that induce a contraction between the chains. The naive maximal coupling described above acts as an independent coupling when the chains are not close, and thus fails to induce contraction. Below we first describe reflection couplings of Normal distributions and then describe how they can be used in the submanifold setting.

Consider two chains at positions x and $\tilde{x}$, in the generic, unconstrained space $\mathbb{R}^D$. First, observe that proposals variables $x^\star$ and $\tilde{x}^\star$, Normal(x, Σ_a) and Normal$(\tilde{x}, \Sigma_a)$, can be drawn as two rescaled standard normals, centered at the original positions x and $\tilde{x}$, respectively. Hence, from the first chain a random perturbation is chosen, and the proposed point is set to $y = x + \Sigma_a^{1/2} z$, with $\Sigma_a^{1/2}$ being the lower-triangular matrix obtained by Cholesky decomposition of Σ and z standard Normal. While, for the second chain, $\tilde{z}$ is chosen to point towards the opposite direction of z with respect to a vector passing through x and y. This is summarized in Algorithm 20.

Algorithm 20 Reflection coupling of two Normal distributions with common variance.

function REFLECTION COUPLING(x, $\tilde{x}$, Σ)
  Compute $e = \Sigma^{-1/2}(x - \tilde{x})$.
  Compute $\bar{e} = e/|e|_2^2$ where $|u|_2^2 = \sum_{i=1}^D u_i^2$.
  Draw $z \sim \text{Normal}(0, I_D)$.
  Compute $\tilde{z} = z - 2(\bar{e}^\top z)\bar{e}$.
return $x^\star := x + \Sigma^{1/2} z \sim \text{Normal}(x, \Sigma)$ and $\tilde{x}^\star := \tilde{x} + \Sigma^{1/2}\tilde{z} \sim \text{Normal}(\tilde{x}, \Sigma)$.
end function

Back to the submanifold setting, we focus on a default choice of distribution for $\nu \sim p_\nu$, which is a d-dimensional centered Normal$(0, \Sigma_{\mathcal{T}})$ distribution. The immediate difficulty in using reflection couplings with Algorithm 18 is that the proposals (pre-projection) $x + U_x\nu$ and $\tilde{x} + U_{\tilde{x}}\tilde{\nu}$ are supported on different spaces. It is simpler to couple directly ν with $\tilde{\nu}$ on $\mathbb{R}^d$, but it is not directly clear how the current locations x and $\tilde{x}$ can be used to design the coupling of ν with $\tilde{\nu}$.

We observe that the variable $x+U_x\nu$ has the same distribution as the vector $x+P_xQ_x\xi$ obtained by applying the orthogonal projector $P_x = I_D - N_x N_x^\top$ to the Normal variable $Q_x\xi$ where $\xi \sim \text{Normal}(0, \Sigma_a)$, where Q_x is the rotation matrix obtained by the QR

decomposition of $\nabla q(x)$ as in (4.3), and with

$$\Sigma_a = \begin{pmatrix} \Sigma^* & C \\ C^\top & \Sigma_{\mathcal{T}} \end{pmatrix}, \tag{4.13}$$

for any choice of matrices Σ^* and C such that Σ_a is symmetric positive definite.

For any $z \in \mathbb{R}^D$, we can see $Q_x z$ as $N_x z_n + U_x z_t$, where $z = (z_n, z_t)$. Then multiplying on the left by $P_x = I_D - N_x N_x^\top$, the orthogonal projector on $\mathcal{T}_x$, we obtain $x + P_x Q_x \xi = x + U_x Z_t$ where $Z_t \sim \text{Normal}(0, \Sigma_{\mathcal{T}})$. The matrices Σ^* and C play no role here, apart from making Σ_a symmetric positive definite.

Therefore we can think of the proposal mechanism as 1) sampling $\xi \sim \text{Normal}(0, \Sigma_a)$ in the ambient space, with a covariance matrix that does not depend on x, 2) left-multiplying by Q_x i.e. applying a rotation, 3) orthogonally projecting $x + Q_x \xi$ onto $\mathcal{T}_x$, and 4) Newton-projecting $x + P_x Q_x \xi$ onto $\mathcal{S}$ following the direction of $\nabla q(x)$.

The advantage of the above view is that the first step, sampling $\xi \sim \text{Normal}(0, \Sigma_a)$ in the ambient space, is directly amenable to a reflection coupling: indeed we can obtain ξ and $\tilde{\xi}$ by reflection coupling of $\text{Normal}(0, \Sigma_a)$ and $\text{Normal}(0, \Sigma_a)$.

If $\Sigma_{\mathcal{T}}$ is simply of the form $s^2 I_d$ with $s > 0$, then we can define $\Sigma_a = s^2 I_D$, and then $\xi \sim \text{Normal}(0, s^2 I_D)$, with a covariance matrix that does not depend on x. So in that case, we can directly define a reflection coupling between $\xi \sim \text{Normal}(x, s^2 I_D)$ and $\tilde{\xi} \sim \text{Normal}(\tilde{x}, s^2 I_D)$. For an intuitive depiction of such reflection coupling, see Figure 4.2.

If the reflection coupling strategy successfully induces contraction between the chains, it can be used in a two-scale strategy, analogously to what it is presented in Papp and Sherlock (2022):

1. if $|x - \tilde{x}| >$ threshold, employ a contractive coupling.
2. if $|x - \tilde{x}| \leq$ threshold, employ a coupling that induces exact meetings.

4.3.3 Computational complexity

Algorithm 18 involves computing the QR decomposition of the matrix $\nabla q(x)$, which is of order $O(m^2(D-m))$ operations. This step is performed twice to find the bases of $\nabla q(x)$ and $\nabla q(y)$. Computing points in the tangent space $x + U_x \nu \in \mathcal{T}_x$ and $y + U_y \nu' \in \mathcal{T}_y$ has a cost of order $O(Dd)$. One iteration of the Newton's solver costs $O(m^2 D)$,while finding the reverse projections $\nu' = U_y^\top (x - y)$ is $O(D)$. Overall, the cost is $O(m^2 D)$. When running two chains the additional cost to perform doubled number of operations comes from finding $\tilde{\nu} = U_{\tilde{x}}^\top (\tilde{x} - y)$, of order $O(Dd)$, and evaluating differentials. For

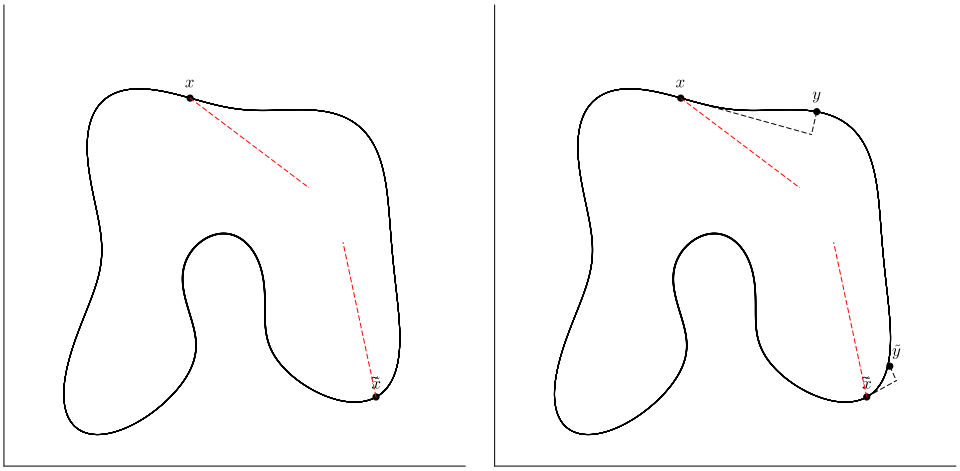

FIGURE 4.2: Depiction of a reflection coupling (*left*) in the augmented space, for x and $\tilde{x}$ on a submanifold. The red segments represent $\xi - x$ and $\tilde{\xi} - \tilde{x}$, where $\xi \sim \text{Normal}(x, s^2 I_D)$ and $\tilde{\xi} \sim \text{Normal}(\tilde{x}, s^2 I_D)$ can be obtained by Algorithm 20. Projections on the tangent space are represented as dashes segments tangent to the submanifold (*r*ight) and the points y and $\tilde{y}$ represent the points projected through Newton's method.

the latter, the cost of computing the inner product, $U_x^\top U_y$, is $O(d^2 D)$, while computing the determinant of the resulting $d \times d$ matrix is $O(d^3)$. Hence, when the dimension of the submanifold is greater than the codimension, i.e. $d > m$, there is an additional cost in running the maximal coupling of the proposals.

The complexity of computing reflections instead is $O(Dm)$, due to computation of the projection matrix P_x. If Σ is diagonal with all elements equal, instead one can directly obtain the proposal from the second chain, as $\tilde{\nu} = U_{\tilde{x}}^\top U_x \nu$, which is of complexity $O(Dd)$.

For this reason, it will be practical to use reflection couplings, which allow to move the chains closer and minimize the attempts to use the maximum coupling, which is computationally more costly, especially in higher dimensions.

4.3.4 Sequence of hyperspheres

To illustrate how the maximal coupling algorithm scales with the dimension of the space, we consider the problem of sampling the Uniform distribution on a sequence of Hyperspheres, $\mathcal{HS}^d = \{x \in \mathbb{R}^D : \sum_{i=1}^D x_i^2 = 1\}$, with $d = D - 1$ in $\{5, 10, 15, 20\}$. We choose the standard deviation for the proposals in the tangent space in order to have comparable acceptance probabilities across all the dimensions. Note that since the radius of the hypersphere is 1, we need $\|\nu\|_2^2 = \sum_{i=1}^d \nu_i^2 \leq 1$ for the orthogonal projections of $x + U_x \nu$ to exist at any point in $\mathcal{HS}^d$. Note that if $\nu \sim \text{Normal}(0, I_d)$, $\|\nu\|_2^2 \sim \chi_d^2$ with

$\mathbb{E}(\|\nu\|_2^2) = d$, while rescaling $\nu \sim \text{Normal}(0, I_d/d)$, ensures that $\mathbb{E}\|\nu\|_2^2 = 1$ and $\mathbb{P}(\|\nu\|_2^2 > 1) \leq 0.5$, with equality for $d \to \infty$, as the Chi-square distribution becomes symmetric. The expected proportion of accepted values, computed as $\mathbb{P}(\chi_d^2 < d)$ is shown in Figure 4.3 together with the empirical proportion of accepted proposals. As expected, the probability that the projection is successful is above 0.5 for any d. Note that the fraction of accepted proposals in this example coincides with the fraction of successful reverse projections since the distribution on the manifold is uniform and the sphere symmetric, thus the Metropolis-Hastings ratio is always equal to 1, after reersibility is checked. We also verified that the average number of iterations used to obtain successful projections by Newton's methods was weakly increasing, but comparable: 5.23, 5.52, 5.72, 5.85 for the hyperspheres of dimension 5, 10, 15 and 20 respectively.

When running pairs of chains, initialized from opposite points respect to the origin of the hypersphere, and with lag $\ell = 50$, we observe that expected meeting times increase linearly with the dimension of the space when the maximal coupling algorithm is used at each iteration. Conversely, when combining maximal couplings with reflection couplings, for chains at least σ apart, i.e. $\|x - \tilde{x}\|_2^2 > \sigma = 1/\sqrt{d}$, average meeting times are constant, see Figure 4.4. This suggests interestingly that the mixing properties of the algorithm, for this target distribution don't depend on the dimension of the space when the variance of the proposal distribution is set equal to $1/d$.

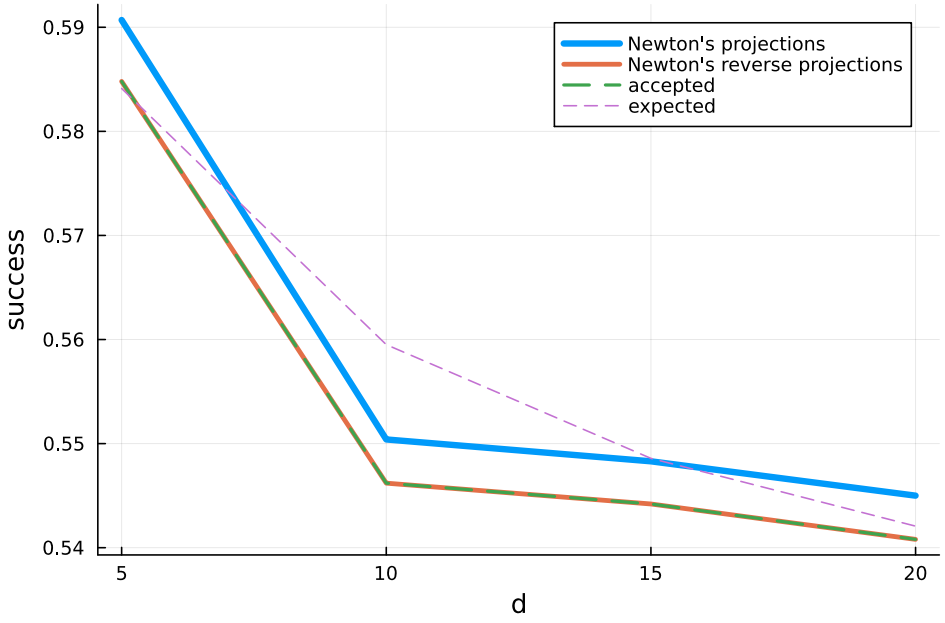

FIGURE 4.3: Monitoring the fraction of successful proposals in different dimensions, computed on chains of length 10^4.

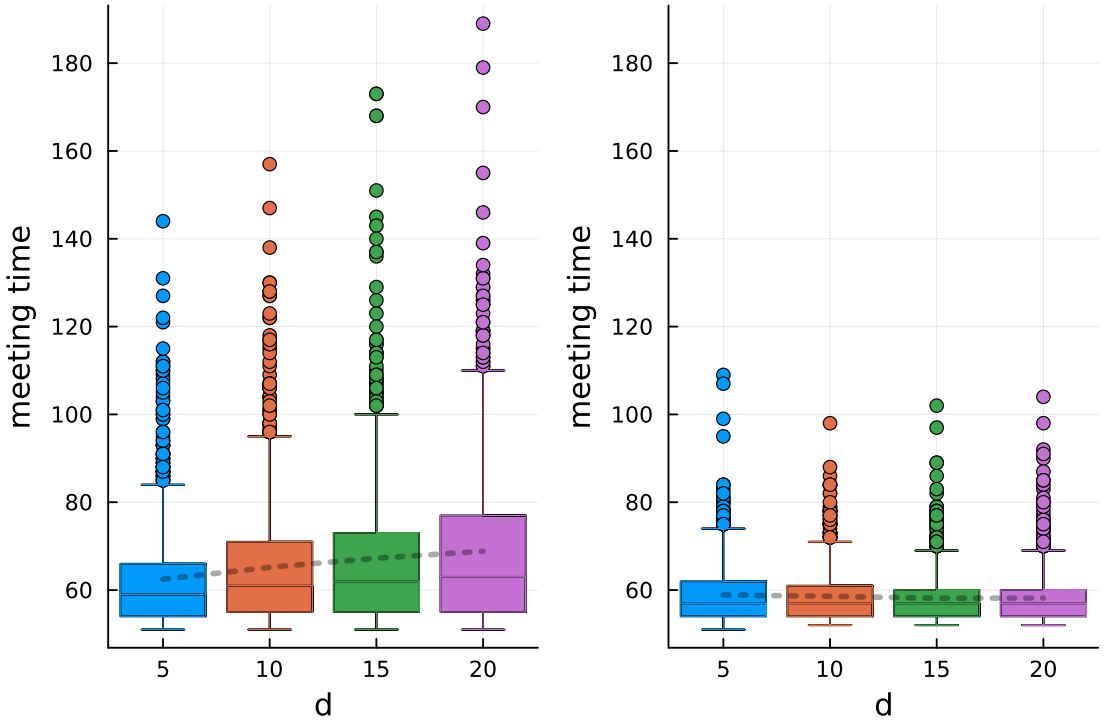

FIGURE 4.4: Meeting times using maximal couplings (left) and maximal couplings combined with reflection couplings (right), based on 10^3 parallel chains. Dotted lines represent average meeting times.

4.3.5 Goodness of fit example

A typical application of sampling over constrained spaces arises in the testing literature, where conditioning on sufficient statistics for the model under the null hypothesis provides greater power for tests to the unconditional counterparts. In this context, sampling over the constrained space is required to retrieve samples generated under the null hypothesis. Diaconis *et al.* (2013) and Lindqvist *et al.* (2022) consider a goodness-of-fit test to the Gamma distribution, under the sum and product condition ($S(x) = \sum_{i=1}^{n} x_i$, $P(x) = \prod_{i=1}^{n} x_i$, with $x \in \mathbb{R}_+^n$). In revisiting this example, we consider a sample x^{obs} of $D = 20$ data points, and define the submanifold as

$$\mathcal{G} = \{x \in \mathbb{R}_+^{20} | \sum_{j}^{20} x_j = S(x^{\text{obs}}), \prod_{j=1}^{20} x_j = P(x^{\text{obs}})\},$$

where $S(x^{\text{obs}}) = 34.8$, $P(x^{\text{obs}}) = 2199.4$. For the Zappa *et al.* (2018) *ZHG* random walk, we study the impact of the maximum number of iterations in Newton's method (4,5,10) and the choice of standard deviation for the proposal in the tangent space ($\sigma = 1/\sqrt{D} \approx$ 0.22,0.3,0.4,0.5), we thus have 12 setups to compare. Figure 4.5 is based on single runs of chains, each of length 10^4, but with different configurations of tuning parameters, and

shows the fraction of cases where the algorithm *ZHG* proposes a valid point $y \in \mathcal{G}$ (left), finds the current state with the reverse projection (center), and accepts the proposal y as a new state (right). The last step involves verifying that $y_j > 0$, $\forall j$. Note that as σ increases, i.e. with larger proposals, Newton's method tends to fail in finding points on $\mathcal{G}$. When the number of allowed iterations for the projection step is 4, this probability decreases quite rapidly. When the number of iterations is 10 (green line), Newton's method yields higher rates of proposals satisfying the constraints, even with fairly large values of σ, but a good fraction of these proposals are discarded in the acceptance phases (about 20% of them with $\sigma = 0.5$, right panel). The further question we aim to answer using meeting times drawn with coupled chains is whether a lower acceptance probability depending on larger proposals is compensated by a fast convergence. We used maximal couplings when the chains were close, i.e. $\|x - \tilde{x}\|_2^2 < \sigma$, otherwise reflections couplings, as explained in Section 4.3.2 and for each setting we considered 20 parallel chains. In Figure 4.6 we show meeting times obtained from coupled chains initialized on the observed set of points x^{obs} at lag $L = 5000$. Our experiments show that choosing large proposals, corresponding to acceptance probability of about 0.5, short meeting times can be obtained, once an adequate number iterations for Newtons's method are used (> 5, green boxplots). In particular, using 5 iterations in Newton's algorithm produces results only slightly less good than with 10 steps, at half the computational cost. Instead, when Newton's method fails (blue boxplots, 4 iterations used), even if the final acceptance probability is higher than 0.5, the meeting times are generally larger. Also, in Figure 4.6 we show the asymptotic variance estimation for the chains, which confirms that when using only 4 iterations provide much worse results that when considering 5 or 10. Finally, using coupled chains, we can compare the performance of the *ZHG* random walk with that of an algorithm originally proposed by Diaconis *et al.* (2013), referred to as *DHS*. In the latter, $D - 2$ data points $(x_3, \ldots, x_D)$ are updated in each iteration using a random walk proposal, and the remaining ones are updated using

$$x_{1,2} = \frac{(S - s) \pm \sqrt{(S - s)^2 - 4P/p}}{2}, \tag{4.14}$$

with $s = \sum_{j=3}^{D} x_j$, $p = \prod_{j=3}^{D} x_j$, provided $(S - s)^2 > 4P/p$ and a random assignment to x_1, x_2. The algorithm can be easily coupled by using a Thorisson maximal coupling of the $D - 2$ dimensional unconstrained proposal, obtaining the remaining data points by (4.14) and common random numbers for randomising the choice of $x_1 > x_2$ or $x_1 < x_2$. When the chains are far apart, i.e., $\|x_{3:D} - \tilde{x}_{3:D}\|_2^2 > \delta_{DHS}$, reflection couplings can also be implemented, following Jacob *et al.* (2020). Chains are initialized with a lag ℓ of

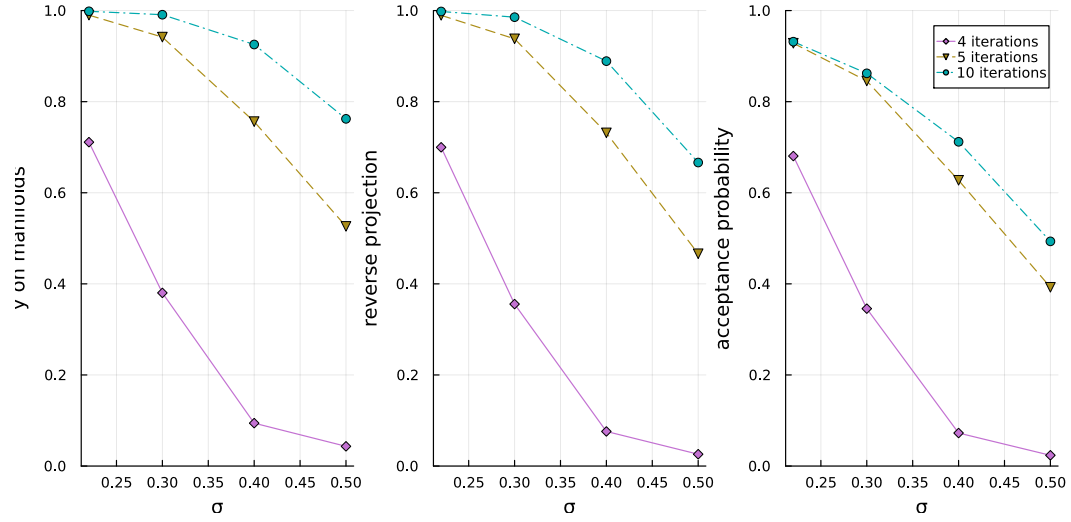

FIGURE 4.5: Proportion of successful proposals (left), reverse projections (middle) and acceptance rate (right) in the Random walk *ZHG* on $\mathcal{G}$ with different standard deviations (σ) for the proposal on tangent space and maximum number of iterations allowed for Newton's method.

5000 steps. After tuning the *DHS* random walk, we found that choosing a proposal Normal$(0, 0.7 \cdot I)$, and $\delta_{DHS} = 2\sigma$ provided the lowest observed average meeting time. In Figure 4.7 the corresponding bounds in total variation from stationarity, compared to those of *ZHG* algorithm with $\sigma = 0.5$ and 10 Newton's iterations. In this specific problem, and dimension $D = 20$, we obtained better convergence guarantees for the *ZHG* algorithm, for which about 500 iterations are required against 150000.

4.4 A note on maximal coupling of composite proposals

The proposal mechanism described in Algorithm 18 can be cast into the form of a composite proposal

$$y = \Phi(x, \epsilon), \tag{4.15}$$

where x and $*yY$ are the current and the proposed state respectively, ϵ indicates a random perturbation and Φ encodes a deterministic transformation. In particular, can we recognize in place of ϵ the proposal ν, while the deterministic function is given by the projection steps with Newton's method, once fixed the number of iterations. Also, we can write explicitly the deterministic function Φ in place of (4.15), as

$$y = G_x(\nu)^{-1}\hat{\mathcal{F}}_x(\nu)\mathbb{I}_{\{\hat{\mathcal{F}}_x(\nu)\neq\emptyset\}} + x1_{\{\hat{\mathcal{F}}_x(\nu)=\emptyset\}}, \tag{4.16}$$

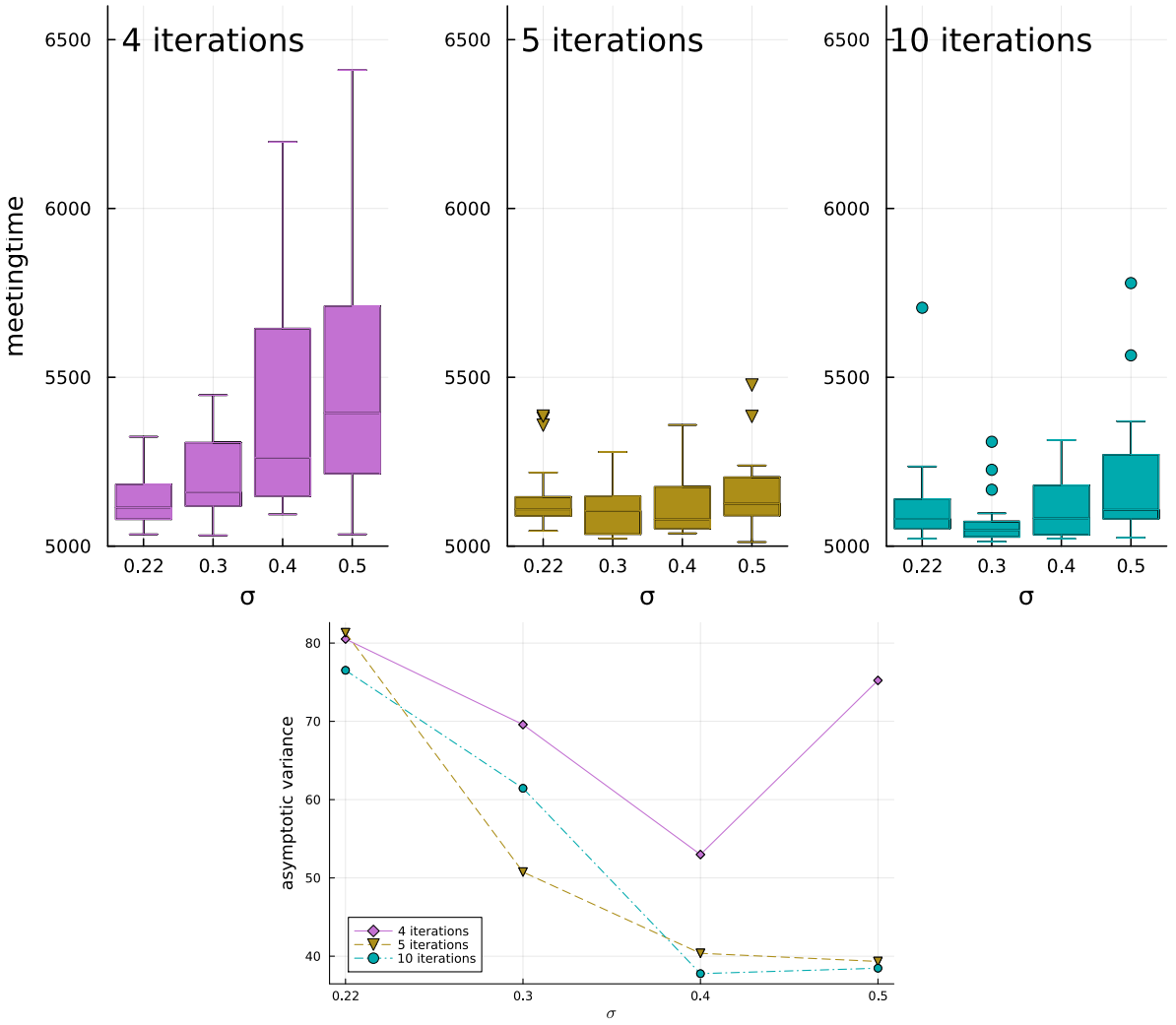

FIGURE 4.6: Meeting times (above) and estimate of asymptotic variance (below) obtained by couplings of *ZHG* algorithm for different standard deviations (σ) for the proposal on tangent space and maximum number of iterations allowed for Newton's method.

with $\hat{\mathcal{F}}_x(\nu)$ defined in Section 4.2.

Similarly, the representation in (4.15) can be used to describe a broader class of MCMC algorithms, with a primary example being Hamiltonian Monte Carlo with leapfrog integrator. Hence, we observe that the corresponding coupling strategy explained in Section 4.3, can be extended to such family of proposals. In particular, when considering two chains X, $\tilde{X}$ evolving with the same scheme, if the point y is proposed from the first one, X, in order to obtain exact meetings, the random perturbation $\tilde{\epsilon}$ from the second chain can be chosen in such a way that $y = \Phi(\tilde{X}, \tilde{\epsilon})$. Denoting by $\Phi_a^{-1}(\cdot)$ the inverse map with the first argument fixed to a, we can write $\tilde{\epsilon} = \Phi_{\tilde{x}}^{-1}(y)$ to determine the necessary perturbation that applied to the secondary chain delivers y. In some cases, Φ^{-1} has a closed form solution, as (4.12) in the random walk on submanifolds,

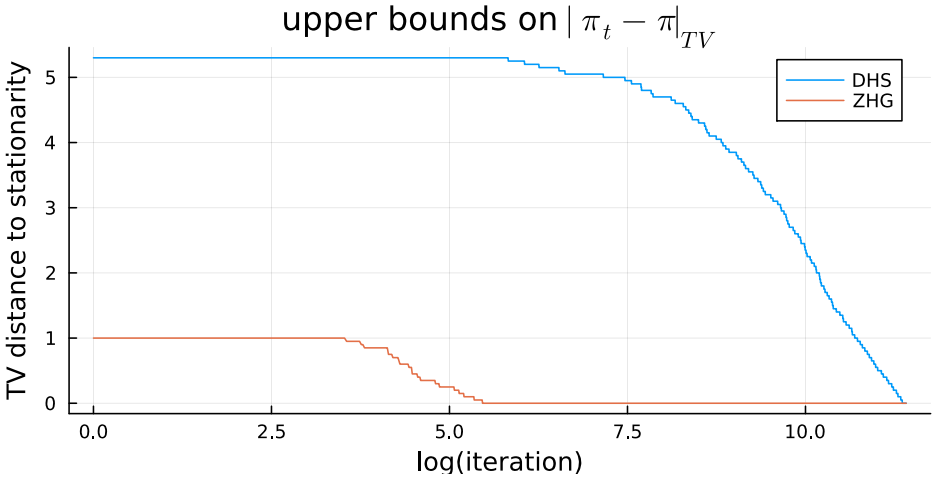

FIGURE 4.7: Comparison of upper bounds on distance from stationarity of tuned *DHS* and *ZHG* random walks on the submanifold $\mathcal{G}$.

in other cases, where an explicit expression is not available, the necessary perturbation that drives the chain to y can be obtained by running an optimization routine, finding

$$\tilde{\epsilon} = \arg\min_{\epsilon} \|\Phi(\tilde{x}, \epsilon) - y\|^2, \tag{4.17}$$

where quasi-Newton or Nelder-Mead, implemented in statistical software (e.g., Mogensen and Riseth 2018), can be utilized. The minimum of the target function in (4.17) is zero. Since as previously observed, numerical methods can fail in finding it, to the end of providing a coupling strategy that is valid, in the following we elaborate on this possibility. We assume we have access to the result of the numerical optimizer in terms of a function $y \mapsto \widehat{\Phi}_x^{-1}(y)$ that returns either "`fail`", in case the minimum is different than zero, or the exact value $\Phi_x^{-1}(y)$. The function $\Phi_x^{-1} : \mathbb{R}^D \to \mathbb{R}^D$ does not have an explicit form and needs to be approximated numerically. For fixed x and y, we define the optimization program

$$\min_{v'} |y - \Phi_x(v')|^2. \tag{4.18}$$

Without direct access to Φ_x^{-1}, we can use numerical optimization methods to solve (4.18).

4.4.1 Coupling of Hamiltonian Monte Carlo kernel

To illustrate the concept, consider (unconstrained) Hamiltonian Monte Carlo on $\mathbb{R}^D$, The Hamiltonian Monte Carlo (HMC) algorithm is a MCMC method introduced to Bayesian computation by Neal (1993) and extensively used since then (Neal, 2011;

Betancourt *et al.*, 2017). The proposals for the next state of the chain are generated by simulating the evolution of the position x and momentum v in response to the total energy of the system represented by the Hamilton function $H(x, v)$,

$$\frac{dx}{dt} = \frac{\partial H}{\partial v} \tag{4.19}$$

$$\frac{dv}{dt} = -\frac{\partial H}{\partial x}. \tag{4.20}$$

Each iteration of the MCMC scheme involves sampling a momentum variable v from a Gaussian distribution with zero mean and covariance matrix M and run a numerical integrator for n steps, that approximates the evolution given by (4.19) and (4.20), leading to a proposal y. The numerical integrator, for assigned initial position x, number of steps n and momentum v is deterministic. The mechanism for generating a proposal, using the leapfrog integrator, is summarized in Algorithm 5. The endpoint of the Hamiltonian trajectory is then accepted or rejected with a Metropolis-Hastings step based on the ratio of the target densities at the current and proposed states.

In the context of the current discussion, let us imagine a pair of chains running in parallel. After we have proposed a new position y of one of them according to algorithm 5, the momentum of the second one can be chosen such that the chain is led to the same point y as the first chain after the deterministic transformation induced by the leapfrog integrator for given initial conditions. A similar argument is presented in Bou-Rabee and Eberle (2023). In particular, writing the leapfrog integrator with initial conditions as $\mathcal{L}(x, v)$, a meeting would be possible for

$$\tilde{v} = \arg\min_{v} \|\mathcal{L}(\tilde{x}, v) - y\|^2 \text{ with } \mathcal{L}(\tilde{x}, \tilde{v}) - y = 0,$$

where $\tilde{v}$ is the numerical solution of an optimization routine.

Previous research presented as Heng and Jacob (2019) investigating couplings for HMC has used contractive couplings that utilise joint momentum updating. However, to obtain exact meetings these kernels are combined to Metropolis Random walk through mixture. Similarly, in Xu *et al.* (2021), contractive and optimal transport couplings, which are able to bring HMC chains closer together, are combined with a Metropolis kernel. In contrast, the strategy proposed here uses the exact same HMC kernel. The advantages of using a coupling that brings about the exact meetings are multiple. Firstly, it could make it possible to possibly obtain the meetings faster and bounds in total variation. Second, unlike other work that relies on a mixture of kernels, it does not require additional parameters to be set. Thirdly, and most importantly, this design

Algorithm 21 Coupled HMC Kernels

Require: current states $x, \tilde{x} \in \mathbb{R}^D$, numerical precision δ, maxit,

function CoupledHMCKernel($x, \tilde{x}$)
  identical←FALSE
  draw $v \sim N(0, M)$
  $y, v_K \leftarrow$Leapfrog$(x, v, K,\ \eta)$
  $\tilde{v} \leftarrow \hat{\Phi}^{-1}(y, \tilde{x})$
  $v' \leftarrow \hat{\Phi}^{-1}(y, x)$
  $\tilde{y}, \tilde{v}_K \leftarrow$Leapfrog$(\tilde{x}, \tilde{v}, K,\ \eta)$
  if $y = \tilde{y}$ and $v' = v$ **then**
    $D = |\det(\nabla_v \text{Leapfrog}(x, v))|$
    $\tilde{D} = |\det(\nabla_{\tilde{v}} \text{Leapfrog}(\tilde{x}, \tilde{v})|$
    sample $W \sim$ Uniform(0,1)
    if $W < \frac{p(\tilde{v})D}{p(v)\tilde{D}}$ **then**
      identical←TRUE
      return (y, v_K) and $(y, \tilde{v}_K)$
    end if
  else
    done←FALSE
    while done←FALSE **do**
      draw $\tilde{v} \sim N(0, M)$
      $\tilde{y}, \tilde{v}_K \leftarrow$Leapfrog$(\tilde{x}, \tilde{v}, K,\ \eta)$
      $v \leftarrow \hat{\Phi}^{-1}(\tilde{y}, x)$
      $v' \leftarrow \hat{\Phi}^{-1}(\tilde{y}, \tilde{x})$
      $y, v_K \leftarrow$Leapfrog$(x, v, K,\ \eta)$
      if $\tilde{y} \neq y$ or $v' \neq \tilde{v}$ **then**
        done←TRUE
        return (y, v_K) and $(\tilde{y}, \tilde{v}_K)$
      else
        $\tilde{D} = |\det(\nabla_{\tilde{v}} \text{Leapfrog}(\tilde{x}, \tilde{v})|$
        $D = |\det(\nabla_v \text{Leapfrog}(x, v))|$
        sample $\tilde{W} \sim$ Uniform(0,1)
        done← $\tilde{W} > \frac{p(v)\tilde{D}}{p(\tilde{v})D}$
      end if
    end while
    return (y, v_K) and $(\tilde{y}, \tilde{v}_K)$
  end if
end function

is based exactly on the transition kernel of HMC, thus performance measures and convergence diagnostics such as the distance from stationarity and the asymptotic variance of the chain obtained by the coincidence of times reflect the properties of the original HMC scheme -not to a mixture of kernels- and provide guidelines for tuning the original HMC. One drawback is that in practice, the numerical solver may be chosen to have machine precision and take more time depending on the difficulty of the coupling. And in higher dimensions, the computational effort required to solve the optimisation problem becomes demanding. Finally, the function Φ must be invertible. The invertibility of $v \mapsto \Phi_x(v)$ is defined everywhere under a smoothing condition in Lemma 4 of Chen and Gatmiry (2023), restated below.

Lemma 4.1 (Lemma 4 in Chen and Gatmiry 2023)**.** *Suppose that $V : \mathbb{R}^D \to \mathbb{R}$ is M-smooth: $\nabla^2 V(x) \preceq M I_D$, where I_D is the $D \times D$ identity matrix, i.e. $M I_D - \nabla^2 V(x)$ is positive semi-definite. Choose K and η such that $K\eta M^{1/2} \leq 1/4$. Then the map $v \mapsto \Phi_x(v)$ is invertible for all $x \in \mathbb{R}^D$.*

In words, invertibility of Φ_x means that for any x and y, there exists a unique velocity v such that the leapfrog integrator starting from (x, v) yields the location y after K steps with stepsize η. In the sequel, we assume that Φ_x is invertible for all x. In order to avoid checking the assumptions at each iteration of the chain, one practical possibility is using in combination of contractive couplings and a local one-shot coupling when chains are closed or when the loglikelihood is nearly the same. This strategy is broadly applicable for distributions whose log-density is non-globally gradient Lipschitz or non-globally concave.

Consider two chains in positions denoted by x and $\tilde{x}$ in $\mathbb{R}^D$. We assume that the function $v \mapsto \Phi_x(v)$ is invertible with inverse $y \mapsto \Phi_x^{-1}(y)$. Then the change-of-variable formula gives

$$q(x, y) = \text{Normal}(\Phi_x^{-1}(y); 0, I_D) |\det D\Phi_x^{-1}(y)|. \tag{4.21}$$

To evaluate the determinant we can use the equivalence, for v, y such that $y = \Phi_x(v)$,

$$|\det D\Phi_x^{-1}(y)| = |\det D\Phi_x(v)|^{-1}, \tag{4.22}$$

so that it becomes easy to evaluate these determinants with an implementation of leapfrog integration that supports automatic differentiation.

If we can evaluate (4.21) for all x, y, we can implement the coupling described by Gerber and Lee (2020), recalled below.

Define, for some $C < 1$, and for the current positions x, $\tilde{x}$,

$$\phi : y \mapsto \min(C \cdot 1(\widehat{\Phi}_x^{-1}(y) \neq \texttt{fail} \;\&\; \widehat{\Phi}_{\tilde{x}}^{-1}(y) \neq \texttt{fail}), w(y)). \tag{4.23}$$

This is zero if the numerical optimizer fails to invert either $\Phi_x(y)$ or $\Phi_{\tilde{x}}(y)$. Otherwise, $\phi(y)$ equals $\min(C, w(y))$, and $w(y)$ involves evaluations of both $\Phi_x^{-1}(y)$ and $\Phi_{\tilde{x}}^{-1}(y)$. With this in place, the coupling algorithm of $y \sim q(x, \cdot)$ and $\tilde{y} \sim q(\tilde{x}, \cdot)$ can be obtained as follows:

1. Sample $y \sim q(x, \cdot)$. With probability $\phi(y)$, set $\tilde{y} = y$ and stop,

   otherwise go to step 2.

2. Sample $\tilde{y} \sim q(\tilde{x}, \cdot)$ and stop with probability $1 - \phi(\tilde{y})/w(\tilde{y})$, otherwise repeat.

As explained in Gerber and Lee (2020) the algorithm produces a valid coupling as long as $\phi \leq w$, but is not necessarily maximal (e.g. if $\phi = 0$, it reverts to an independent coupling). Gerber and Lee (2020) consider $\phi : y \mapsto \min(C, w(y))$ for $C < 1$ as a way of controlling the variance of the computing cost of the algorithm. Recall that, when using $C = 1$, the cost has a variance that goes to infinity as $|q(x, \cdot) - q(\tilde{x}, \cdot)|_{\mathrm{TV}}$ goes to zero.

4.4.2 Example: Banana-shaped distribution

We consider the nonconvex potential of a banana-shaped distribution on $\mathbb{R}^2$ (top panel of Figure 4.8), given by the Rosembrock function, $V(x_1, x_2) = (1 - x_1)^2 + 10(x_2 - x_1^2)^2$. The aim of this example is to illustrate the validity of the meeting-inducing coupling scheme proposed for HMC and the gain in the estimation of TV bounds from stationarity against the mixture of two couplings.

Figure displays, for two fixed current states, points obtained by running the leapfrog integrator for 15 steps from a single chain (blue dots), a collection of points obtained from a second chain coupled with the first one (red triangles) and a second chain which is not coupled with the first one. Possible meeting points are marked as stars. In the third panel of Figure 4.8 we display the marginal distribution of proposals obtained by coupling and not coupling the second chain with the first one, to prove the validity of the evolution scheme of the coupled chain.

In order to evaluate the performance of the coupling strategy proposed, we compare it to the mixture of HMC and Random walk kernel of Heng and Jacob (2019), which is $K_{\mathrm{mix}} = 0.95 q_{\mathrm{HMC}} + 0.05 q_{\mathrm{RW}}$. Specifically, the number of leapfrog steps in this example were set to 500, and their size to $1/500$. For the RW propsal, a Gaussian with standard deviation equal to 0.01 was used.

Furthermore, to obtain bounds in total variation from stationarity, we run 500 parallel chains. Specifically, these were initiated from a Uniform distribution in the interval $[-5, 5]^2$, and with a lag of 20 iterations. TV upper bounds obtained by meeting times of tuned coupled HMC employing meeting-inducing couplings and tuned mixture of HMC and random walk are displayed in the bottom panel of Figure 4.9, with the first method exhibiting faster meeting times and providing lower bounds. In fact, the average convergence time of a pair of chains across 500 runs required about 19 HMC iterations against 158 iterations of the mixture of kernels. By employing an implementation in the Julia programming language this took about 1.6 seconds for each pair of chains.

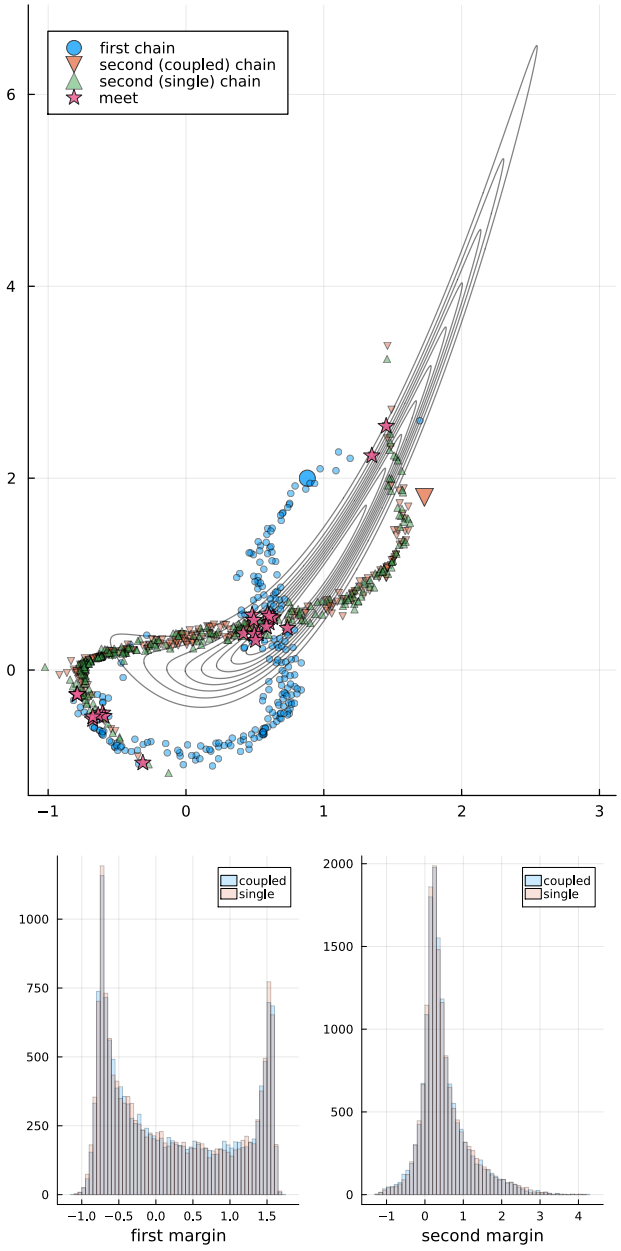

FIGURE 4.8: Comparison of joint and marginal proposal distribution from single chains and coupled chains.

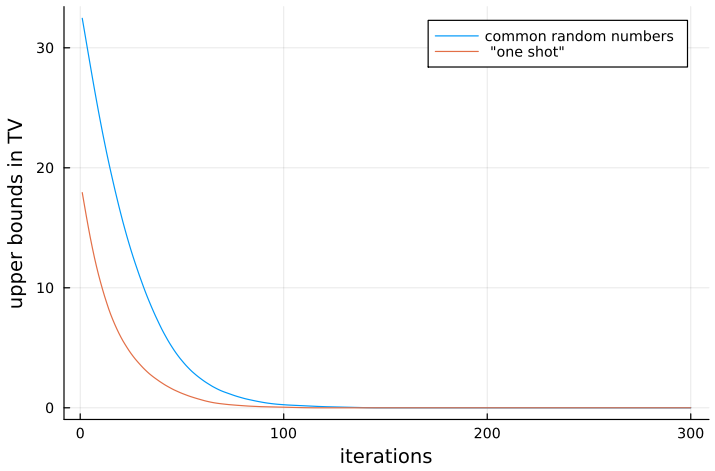

FIGURE 4.9: Bounds in total variation from stationarity for the banana-shaped distribution by meeting-inducing couplings for HMC and mixture of HMC and RW.

4.5 Discussion and future developements

We mention a possible extension of coupling schemes for random walk on submanifolds to Langevin diffusion and HMC kernels on submanifolds. Indeed, one immediate generalization of the random walk MRTH algorithm is the Metropolis Adjusted Langevin algorithm (MALA), based on the discretised Langevin diffusion. Denoting by γ the discretisation step, and by $\nabla(-V(x))$ the gradient of the log-target, the proposal on a submanifold can be written as

$$x + \gamma\nabla(-V(x)) + U_x\nu, \tag{4.24}$$

followed by a projection step using Newton's method. Note that the point defined by (4.24) is not on the tangent space of the submanifold at x, hence, for coupling this variant is sufficient to consider a proposal centered at $x + \gamma\nabla(-V(x))$ and $\tilde{x} + \gamma\nabla(-V(\tilde{x}))$ instead of in x and $\tilde{x}$ respectively. Lelièvre *et al.* (2019) consider a more sophisticated algorithm, called Generalized Hamiltonian Monte Carlo (GHMC), where a constrained integrator, RATTLE, is employed and at each MCMC iteration only a partial refreshment of the momentum is performed. For such case, if only one integration step is considered, the expression of the momentum can be found to be Langevin diffusion with partial refreshment in closed form Lelièvre *et al.* (2019). If more than one integration step is done, one strategy could be to couple this in the same way that we couple unconstrained HMC in Section 4.4.1, i.e. using numerical inversion. Since this is expected to be computationally intensive, following Heng and Jacob (2019), another solution is proposing same momenta for the pair of chains, and when the chains are close, using

the meeting inducing coupling of random walks.

In the study of couplings for HMC, the invertibility condition seems to be the most difficult to handle. One potential solution involves adapting the Hamiltonian Monte Carlo (HMC) method by constraining the flow when multiple momenta guide the chain to the same point, selecting randomly, only one of these trajectories to proceed, similarly to ideas in Lelièvre *et al.* (2020). In alternative, the choice can be deterministic, by introducing additional constraints, as for instance choosing the shortest path, or the smallest momentum, or rather identifying the closest velocity $\tilde{v}$ to the one proposed v. But also, there may be not velocity $\tilde{v}$ such that $y = \Phi(\tilde{x}, \tilde{v})$. Also, the complexity of the optimization routines could become high in multidimensional settings, and tailored optimization routines could be needed. These aspects could be object of future investigations.

Chapter 5

Manifold-Based Sampling from Generic Target Distributions

5.1 Introduction

There is a growing range of sampling methods designed for target distributions that are defined on a d-dimensional submanifold $\mathcal{S}$ of $\mathbb{R}^D$ where $d < D$, defined as a level set of a smooth function $q : \mathbb{R}^D \to \mathbb{R}^m$, known as the constraint. These techniques have been primarily introduced to tackle sampling problems where a submanifold arises naturally. For example, in Statistical mechanics, there is a need to generate representative samples tied to a function linked with the energy level of a system, under fixed constraints, termed in that context *reaction coordinates* (see e.g. Andersen, 1983; Rousset *et al.*, 2010). In Statistics, constrained sampling becomes valuable when handling hypothesis testing or over-specified models (see for instance Diaconis *et al.* 2013; Bornn *et al.* 2019; Graham and Storkey 2017; Liu *et al.* 2022).

However, the ability to sample from submanifolds extends beyond these specific applications and proves beneficial in generic sampling problems within unconstrained spaces ($\mathbb{R}^d$). Recent work by Au *et al.* (2023), Graham *et al.* (2022) proposes the artificial introduction of a submanifold embedded in a higher dimensional space to facilitate sampling from a target distribution originally defined on unconstrained state spaces. Similarly, in this Chapter, we identify two ways to introduce a submanifold related to the global shape of the target distribution, by augmenting the state space or by conditioning on relevant directions of the function of interest. The ideas presented aim to utilize intrinsic geometric information of the target distribution to efficiently explore the space by devising convenient proposal mechanisms.

5.1.1 Use of geometric information in MCMC

Markov chain Monte Carlo (MCMC) methods stand as the bedrock of Bayesian analysis, wielding the machinery necessary for precise uncertainty quantification (Metropolis *et al.*, 1953; Hastings, 1970; Tierney, 1994; Roberts and Rosenthal, 2004; Robert and Casella, 1999; Liu, 2008). Careful design of proposal for the sampling mechanisms is central in dictating the efficiency of MCMC algorithms, enabling to explore complex parameter spaces and ease converge to the desired posterior distribution. For fast exploring the target distribution it is crucial to enforce the alignment of proposals with relevant direction, especially when the size of the step is large, and successively accept the new states with high probability, Green *et al.* (2015). Often, performing such types of steps is possible if the proposal is tailored to mirror the specific characteristics of the target distribution, such as local shape and scale (Rosenthal *et al.*, 2011; Andrieu and Thoms, 2008). A general approach to achieve this is by defining the negative log-posterior distribution as a potential energy function within a fictitious physical system. The dynamics of the MCMC chains can be then guided by discretized trajectories following the motion equation along relevant directions of the potential energy function as in Hamiltonian Monte Carlo (Duane *et al.*, 1987; Neal, 1999, 2011). More advanced use of geometric information has further proven successful, for instance, in Hamiltonian Monte Carlo, Girolami and Calderhead (2011) introduce the Fisher information matrix as the metric tensor in the Hamiltonian dynamic, serving to align the geometry of the target distribution with the coordinate system and implicitly perform tuning. More recently, for certain classes of models where the posterior distribution is marked by strong anisotropy, Au *et al.* (2023) show how it is feasible to sample from the target introducing a distribution supported on a manifold embedded within a higher-dimensional space, demonstrating a comparable efficiency to a well-preconditioned Hamiltonian Monte Carlo while offering computational advantages.

With the aim of sampling from a highly anisotropic posterior distribution for a vector parameter θ, which is typical of certain models with additive noise terms, Au *et al.* (2023) pioneered an elegant solution that transforms the latter distribution, originally defined on $\mathbb{R}^d$ onto a lifted distribution on a submanifold embedded within a higher-dimensional space. In the augmented space, n Gaussian latent variables, denoted as η, are introduced, where n is of the same number of the observations from the model, y^{obs}, and the submanifold has the form

$$\mathcal{S} = \{(\theta, \eta) : y^{\text{obs}} = F(\theta) + \sigma(\theta)\eta\}, \tag{5.1}$$

where $F(\theta)$ is a function that depends on the model and σ is a scale parameter. The auxiliary distribution can be written as

$$\bar{\pi}(\theta, \eta|Y) \propto \pi(\theta)\text{Normal}(\eta; 0, 1)1_{\{Y=F(\theta)+\sigma\eta\}}(\det J(\theta, \eta))^{-1/2}(\theta, \eta), \tag{5.2}$$

where $\pi(\theta)$ is the prior and the correction term, $(\det J(\theta, \eta))^{-1/2}$ is obtained from the co-area formula, with $J(\theta, \eta) = \nabla q(\theta, \eta)^\top \nabla q(\theta, \eta)$ and ∇q the Jacobian matrix of $q(\theta, \eta) = Y - F(\theta) - \sigma\eta$. The θ-marginal of (5.2) corresponds to the original posterior distribution. When the diffusion of the target decreases, i.e. as $\sigma \to 0$, the peakedness of the manifold increases. Consequently, a constant step size with which the Markov chain moves along $\mathcal{S}$ adapts seamlessly without the need for it to be adjusted even in scenarios of extreme anisotropy.

Similarly, but for a broader class of models, in this chapter we present some strategies for enhancing the exploration of general target distributions by implementing constrained moves based on intrinsic geometric information based on the construction of an artificial submanifold. The idea revolves around two core concepts: 1) proposing moves that align with the distribution's shape and 2) incorporating substantial step sizes for effective exploration.

5.2 Sampling on the graph of a function

The first way to introduce such artificial submanifold for sampling from a generic d-dimensional target π, is by defining a distribution on the graph of a function $f : \mathbb{R}^d \to \mathbb{R}$. For the function f, various representations can be considered, such as $\pi(\cdot)$ up to a multiplicative constant, or $\log(\pi(\cdot))$ up to an additive constant, or other smooth transformations, as tempered versions of π used in sequential Monte Carlo (see e.g. Dai *et al.* 2022). Denoting by $G(f)$ the "graph map" function (Simon, 2014):

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d \quad G(f)(x) = (x, f(x)), \tag{5.3}$$

we can introduce an auxiliary variable $x_D \in \mathbb{R} := f(x)$ and define a submanifold as

$$\mathcal{S} := G(f)(\mathbb{R}^d) = \{(x, x_D) : x \in \mathbb{R}^d, x_D \in \mathbb{R}^1 | x_D = f(x)\}. \tag{5.4}$$

Thus, $\mathcal{S}$ is a submanifold embedded in a $D = d + 1$-dimensional ambient space. The auxiliary variable in (5.4) resembles that used in slice sampling (Neal, 2003). We use this representation to propose a new strategy to sample from π.

Defining the Jacobian matrix as

$$J_{G(f)}(x) = \sqrt{\det(\nabla G(f)(x)^\top \nabla G(f)(x))} = \sqrt{1 + |\nabla f(x)|^2}, \tag{5.5}$$

where $\|\nabla f(x)\|^2$ is the squared 2-norm of the gradient of f at x, i.e. $\nabla f(x)^\top \nabla f(x)$, and using the co-area formula, the density on $G(f)(X)$ with respect to the surface measure on $\mathcal{S}$ is given by

$$\pi_{G(f)}(x, x_D) \propto \pi(x)(1 + |\nabla f(x)|^2)^{-1/2}. \tag{5.6}$$

5.2.1 Effects of moving on the graph

Let us assume that we run the algorithm of Zappa *et al.* (2018) on $\mathcal{S}$ with $f = \log(\pi)$, discussed in Section 4.2. The constraint function is $q(x, x_D) = f(x) - x_D$, with $\nabla q(x, x_D) = (-\nabla f(x), 1)'$. We write z for the pair (x, x_D). To simplify the reasoning, we consider the case $d = 1$ and propose steps $\nu \sim \text{Normal}(0, \sigma^2)$. Note first that multiplying a tangential step ν by the basis of the tangent space U_z of $\mathcal{S}$ at z rescales the proposal according to the magnitude of the absolute value of the gradient of the instrumental function f. As a limiting case, when the function f is flat, the proposal on the $x-$ coordinates (d dimensional) coincides with the proposal of $\mathcal{T}_z$. Conversely, if the gradient of f is steep, only a small part of the motion along U_z is conserved on the d dimensional support of the target distribution. Then, the new state proposal is obtained by Newton projection of $z + U_z\nu$ onto $\mathcal{S}$. Since $d = 1$, the vector orthogonal to $\nabla q(z)$ is $(-\nabla_2 q(z), \nabla_1 q(z))$, and U_x is obtained by normalization of that vector. If the function f is linear, then the proposal x^* coincides with $z + U_z\nu$, since $\alpha = 0_m$, in notation of Newton's projections (Algorithm 17). In this case

$$x^* = x - \sigma\epsilon/\sqrt{1 + |\nabla f(x)|^2},$$

in particular, if the function is steep, (either with positive or negative gradient), the increment on the state space (x coordinate) after one step will be small. Conversely, if the function is flat, the proposal on the $x-$ coordinates (d dimensional) coincides with the proposal on $\mathcal{T}_z$. When the function f is not linear, and $\alpha \neq 0$, the value of α can depend on the curvature of the function. In particular, assuming α is obtained by one iteration of Newton's method, then it will satisfy

$$\alpha = -\frac{f(x - \sigma/\sqrt{1 + |\nabla f(x)|^2}\epsilon) - f(x) + \sigma\epsilon\nabla f(x)/\sqrt{1 + |\nabla f(x)|^2}}{\nabla f(x)\nabla f(x - \sigma\epsilon/\sqrt{1 + |\nabla f(x)|^2}) + 1}. \tag{5.7}$$

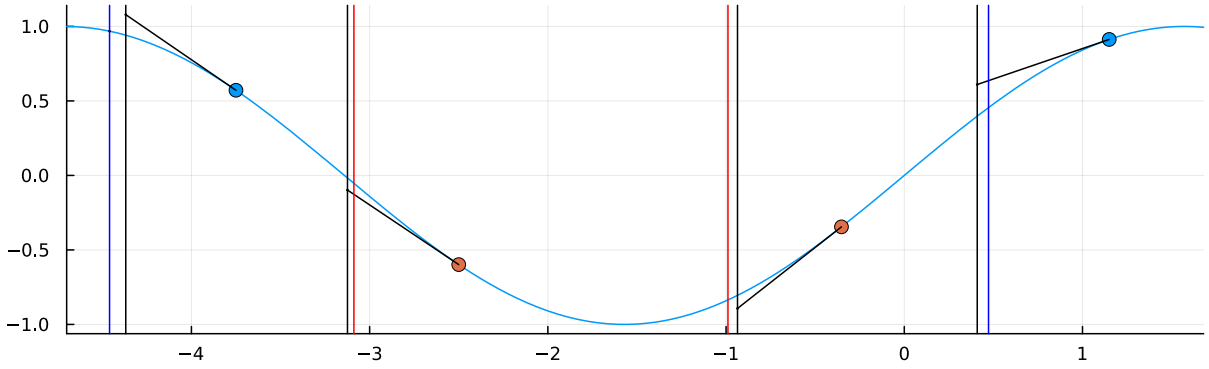

FIGURE 5.1: Graph of the sine function $(x, \sin(x))$. Blue markers represent points at which the function is locally convex, red markers represent points at which the function is locally concave. The black lines represent the coordinates of x before the projection steps, while blue and red lines those after the projection steps.

. In (5.2.1) α increases with $\sigma\epsilon$.

Assuming $\nabla f(x+\epsilon)\nabla f(x) > 0$, i.e. there is not a critical point between the original state and the x-coordinate after the displacement, for fixed ϵ, and initial value x, let us rewrite

$$\alpha^* = -\frac{(f(x-\sigma/\sqrt{1+|\nabla f(x)|^2}\epsilon) - f(x))/\nabla f(x) + \sigma\epsilon 1/\sqrt{1+|\nabla f(x)|^2}}{\nabla f(x-\sigma/\sqrt{1+|\nabla f(x)|^2}\epsilon + 1}$$

note that the sign of α depends on the sign of

$$f(x-\sigma/\sqrt{1+|\nabla f(x)|^2}\epsilon) - f(x). \tag{5.8}$$

Finally, for fixed sign of (5.8), and value x, the sign of α changes with the sign of $\nabla f(x)$. Thus, the sign of α is determined by the position of the tangent space at the proposal point with respect to the function f. This is also intuitively shown in Figure 5.1. There, the function is locally convex if evaluated at points corresponding to blue markers, while on red markers the function is locally concave. From each point a proposal in the tangent space is drawn, either climbing the local mode or descending from it. The vertical black line shows the increments on the state space before the projection step, while blue and red lines shows the arrival x position after finding projections. In summary, the position of the chain after one iteration depends on the peakiness of the auxiliary function f and at the same time on its curvature.

5.2.2 Example: Multimodal target

We consider a mixture of four normal random variables in $\mathbb{R}^2$, centered at $(-5, 5)$, $(-2, -2)$, $(1, 1)$ and $(4, -4)$ and with weights of 0.3, 0.2, 0.2 and 0.3 respectively. The covariance matrices are diagonal with non-isotropic components and scales smaller with respect to the distances between the means, which are reported in Table 5.1. The target distribution is therefore multimodal, as shown in the first panel of Figure 5.2. The remaining panels in Figure 5.2 show samples obtained with standard MRTH, MALA and graph-based sampling based on 50000 iterations. In a simple random walk, to move between modes, one must choose a step size of the same order of magnitude as the distance between modes to move between all regions, but at the same time the acceptance rate becomes negligible. To obtain the results shown in Figure 5.2, the step size was chosen equal to $\sigma = 2.3$ and the acceptance probability was 0.02. When such a step size is chosen, the sampler jumps between modes and fails in exploring the mode centered at $(-2, -2)$. Similarly, MALA struggles to visit all the modes and also has vanishing acceptance rate when choosing relatively large stepsize. Moving on the graph with a step size of 1.5, allows to move through the four modes with an acceptance probability of 0.10, exploring the modes locally, especially the mode centred at $(-2, -2)$, for which it is difficult to construct a good proposal due to anisotropy. For the sample obtained by the latter strategy, the percentage of draws such that $x_2 > 2.5$, thus are compatible with the first component of the mixture was as expected 0.3.

	(-5,5)	(-2,-2)	(1,1)	(4,-4)
(-5,5)	-	7.62	7.21	12.73
(-2,-2)	7.62	-	4.24	6.32
(1,1)	7.21	4.24	-	5.83
(4,-4)	12.73	6.32	5.83	-

TABLE 5.1: Mixture of four Normals: distance between the four modes.

While graph-based moves seems helping in this example, extensions to multimodal targets in higher dimensional settings might be less convincing by only using the log-target distribution as graph map.

5.3 Improving MCMC by walking on level sets

In some challenging sampling scenarios—where the target distribution exibits strong anisotropy, significant inter-component correlations, or within high-dimensional settings— a strategy for devising proposal moves is aligning with the shape of the target

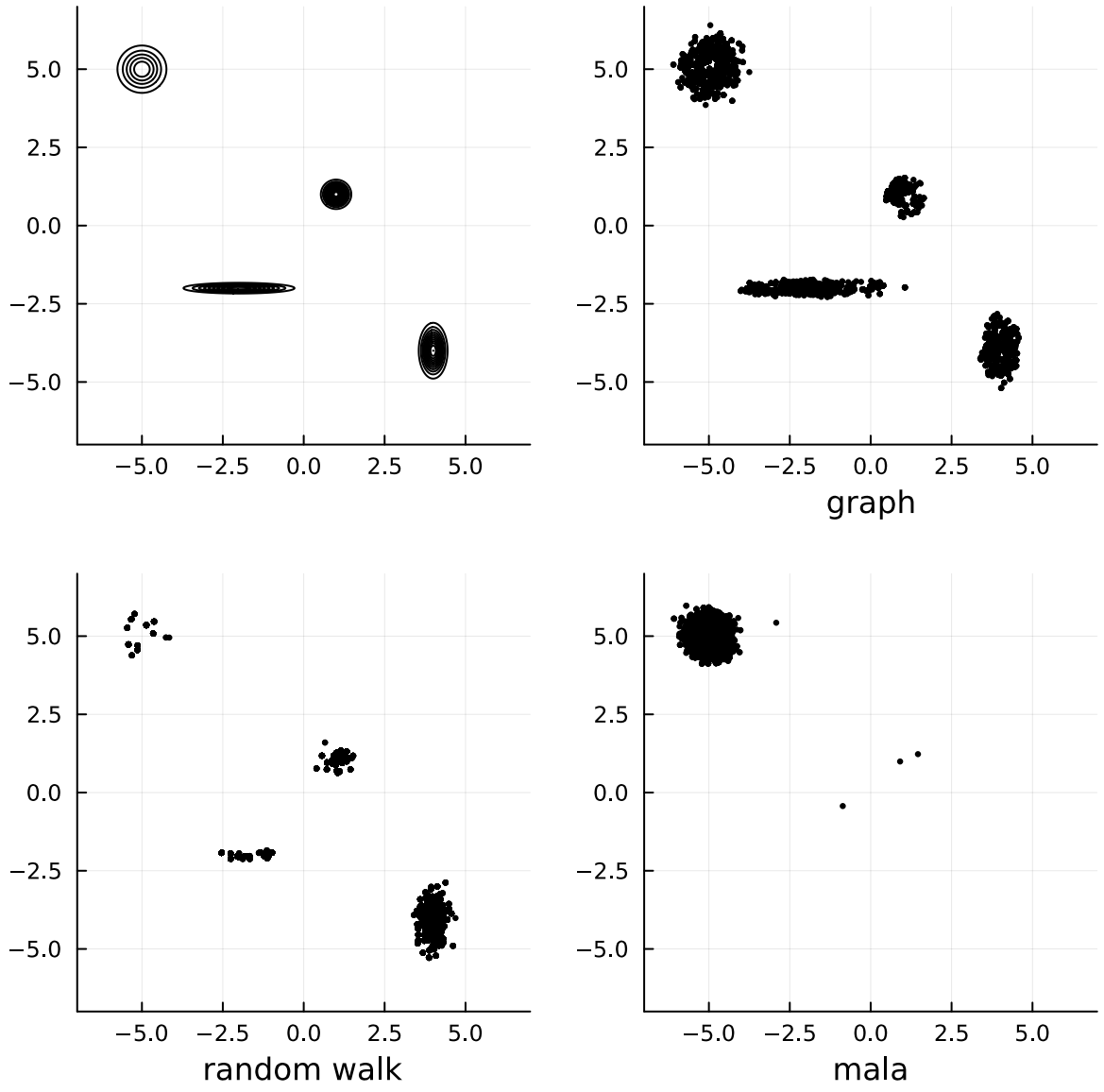

FIGURE 5.2: Mixture of four Normals: contour plot, and draws of length 50000 obtained by sampling on the graph, random walk, MALA.

function. Methods that use geometric information about the target as gradients, exploring specific directions, with Hamiltonian Monte Carlo (HMC) among others, have demonstrated to be successful in this context, being effective when global reparametrizations are not sufficient to meaningfully reshape the target. Instead of relying on gradient based methods and introducing a phase space, as in HMC, we propose to improve the exploration of the target space by introducing constrained moves defined upon intrinsic geometric information of the target. We propose to achieve such alignment between proposal and shape of the target distribution by interweaving standard MCMC moves with moves designed for targets supported on the level sets of the target distribution.

This is achieved by conditioning on the value of a nonlinear function q:

$$\pi(x|q(x) = 0) \propto \pi(x)(\det J(x))^{-1/2} 1\{(q(x) = 0\}, \tag{5.9}$$

where $J(x) = \nabla q(x) \nabla q(x)^\top$ is the Gram matrix, see (4) in Au *et al.* (2023). In practice, it is equivalent to use $q(x) = \log \pi(x)$ or directly $q(x) = \pi(x)$ assuming smoothness conditions on $\pi(x)$. In this case, the submanifold is defined only upon one constraint ($m = 1$).

Note that by construction along the same contour, the acceptance probability only depends on the ratio of the proposals $p_\nu(\nu')/p_\nu(\nu)$ and on the ratio of jacobian-related terms $J(x)$. If the contour levels were spheres (for symmetric distributions) the ratio would be equal to one, independently on the dimension of the space.

In order to explore fully the target π, then these moves need to be combined with moves that enable to change between level sets. The algorithm operates alternating moves on a fixed contour set and free steps, that can be performed with any MCMC kernel that leaves π invariant. A schematic description of the algorithm is in Algorithm 22.

Algorithm 22 Alternate kernel

Markov chain currently at state X_t, target distribution π, unconstrained reversible Markov kernel K_U

1: Compute $q_0 = \log \pi(X_t)$.
2: Define $q(x) = \log \pi(x) - q_0$
3: Call algorithm 18 with distribution (5.9) on $\mathcal{S}$
4: Perform an unconstrained move with kernel K_U to obtain X_{t+1}.

In particular, for the unconstrained kernel, we can consider several options, as standard random walk, or MALA. Alternatively, one could encourage moves between contours that are exactly or approximately orthogonal to the current contour set.

A similar idea is considered by Ludkin and Sherlock (2023), who introduce *Hug and Hop*, an Algorithm that interweaves a series of moves along fixed contour sets (Hug-kernel) using the Bouncy Particle Sampler (see e.g. Bouchard-Côté *et al.* 2018) and across contour levels (Hop-kernel), using a mechanism that resembles a preconditioned MALA, to promote the transition of maximizing the change in log-posterior value. This can be obtained by proposing a new point via a proposal distribution with a higher variance in the direction of the gradient and a lower variance in the directions perpendicular to it. In particular, if the target distribution is symmetric, the Hug-kernel maintains successive steps of the chain along a fixed contour set as the constrained moves, while in non-symmetric targets Hug-moves only approximately follow such sets. The results

of a single constrained move would be similar of that obtained with several hug steps of small size.

5.3.1 Example: multivariate normals

We consider a sequence of multivariate normals of dimension $d = (5, 10, 15, 20, 25)$ centered at the origin 0_d and with covariance matrix is given by $\Sigma = I/d$. We want to compare the performance of HMC, random walk and alternating moves on level sets. In particular, we consider alternating contour moves and random walk as well as contour moves and moves along the gradient. To choose the hyperparameters HMC, we set the number of leapfrog steps to 10 and change the step size to obtain an acceptance rate of 0.90. For the random walk kernel, used alone or in combination with contour moves, we choose independent proposals with a variance of $2.38^2/d^2$, which allows us to achieve an acceptance rate of 0.25 for each d. For moves on contour planes, we choose independent $d-1$-dimensional proposals with a covariance matrix $0.5 \cdot I/d$, which allows to obtain acceptance rate equal to 1, as d increases, and for movements along the gradient direction, we choose a one-dimensional proposal with variance equal to $1/d$. The chains were initialized from the stationary distribution and run for 5000 iterations. For the comparison, we consider the effective sample size (ESS) for the function $h(x) = x^2$ calculated on each marginal draw from the target. Figure 5.3 shows the minimum over the d components of the ESS computed. Orange and purple lines, associated to constrained moves on level sets, corresponding to the second and the third lines from the bottom, seem to stabilize as d increases, while the green line (HMC), with the highest ESS values, and yellow line (random walk), with lowest values seem to depend more on the value d. With an implementation in `Julia` programming language (Bezanson *et al.*, 2017) on a laptop CPU with a clock speed of 1.00 GHz, the average cost of performing one iteration (combining contour moves and moves between the level sets) was 2.9 milliseconds when RW was used in combination versus 21 milliseconds when moves along the gradient direction were used.

5.3.2 Example: Funnel distribution

To illustrate the benefits of leveraging the geometric properties of contour levels, and of sampling on the graph, we consider Neal's two-dimensional funnel distribution, defined by the potential function:

$$V(x_1, x_2) = 0.5 \left(\frac{x_2^2}{9} + \frac{x_1^2}{e^{x_2}} + \log(2\pi e^{x_2}) \right).$$

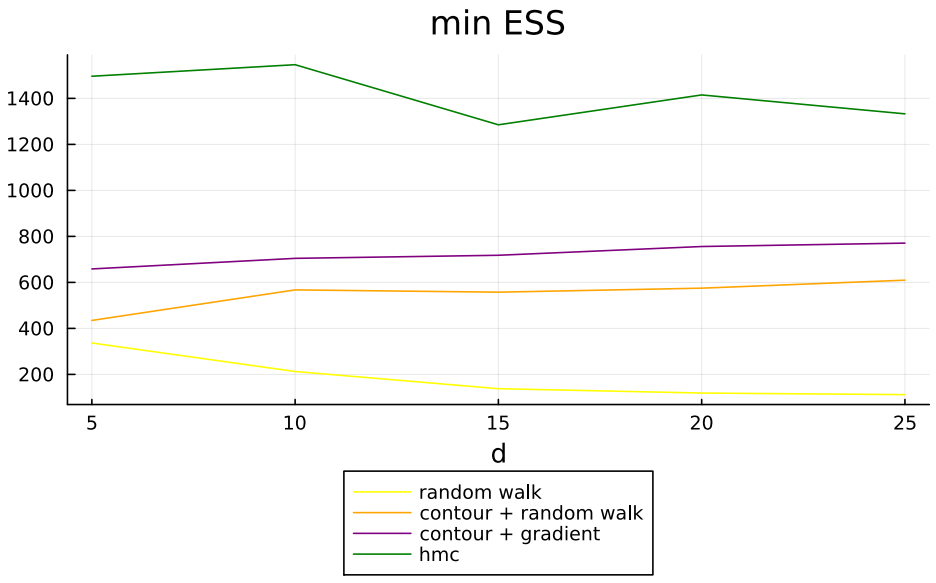

FIGURE 5.3: Comparison of minimum ESS for the function $h(x_j) = x_j^2$, over all the components $j = 1, \ldots, d$ on the multivariate normal example.

This distribution is renowned in the literature due to its complex shape. It exhibits on one side a sharply constrained peak, collapsing on a line while the distribution remains diffuse across $\mathbb{R}^2$ on the other side.

Figure 5.4 shows a sample of length 50000 obtained from random walk with independent increments and standard deviation equal to 0.05, HMC with stepsize 0.05 and 10 timesteps, alternated random walk and contour walk moves and by sampling on the graph with 2.5 stepsize. All the algorithms were allowed to perform small steps and achieve hight acceptance probabilities, respectively 0.80 (random walk), 0.99 (HMC), 0.35 (contour walk), 0.42 (graph), while Effective Sample Size was 555 (random walk), 607 (HMC), 1662 (contour walk), 2500 (graph). Even if the acceptance probability is globally higher than the latter two methods, both HMC and random walk display difficulties in exploring the target. In particular, the minimum value reached by the HMC algorithm over the second coordinate in this run was above -5. On the contrary, moving on the contour sets or on the graph of the function seems more effective to explore the space, even in the region $[-7, -5]$. The average computational time based on `Julia` implementation on a laptop CPU with a clock speed of 1.00 GHz to perform one MCMC iteration with RW was about 1 microsecond, with HMC was 20 microseconds, by sampling on the graph was of 58 microseconds, with alternate moves (contour walk) 159 microseconds.

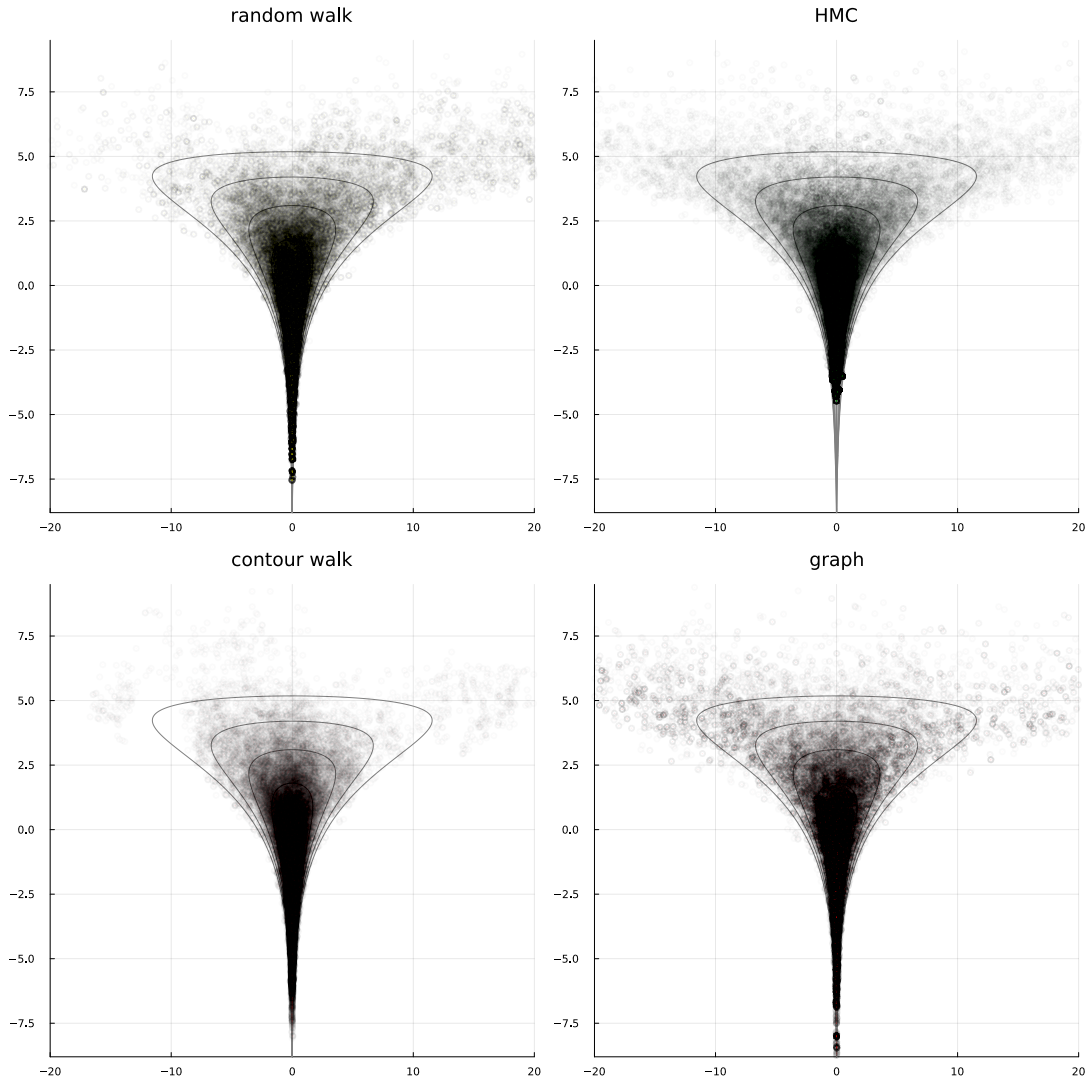

FIGURE 5.4: Comparison of random walk, HMC, contour moves and graph moves one long run with the funnel distribution as target.

5.4 Discussion and future extensions

Inspired by the concepts presented in Au *et al.* (2023), we developed two strategies for sampling from a target distribution by formulating an artificial submanifold sampling problem. We'll revisit and highlight novelties and advantages provided by such ideas, some of which are under investigation.

Crucially, these strategies exhibit generality and are not confined to any specific class of additive models. Also, the artificial constraints are always one dimensional, and the augmented space is either of the same dimension of the original problem of requires the introduction of one auxiliary variable. Finding the initialization point of the chain is not more complex than a standard MCMC. In both the approaches presented, one needs to fix an initial state x_0 in the original $d-$dimensional space and

either compute the value $f(x_0)$ of a one-dimensional auxiliary variable or fix the level set as $q(x_0)$. Conversely, in general finding one initial point satisfying a set of potentially high dimensional constraints of type 5.1 is challenging. Also, the complexity of the algorithm naturally won't depend on the number of observations from the model (n) more than a standard gradient-based algorithm, as HMC, but only on the dimension of the space d. As well, the adaptation of the proposal is automatically refined through the use of projections. When sampling on the graph, the step size is contingent upon the graph's peakiness. In both graph-type and contour-type moves, the ultimate step size is determined by the projection step, which offers substantial freedom as it enables movement of any magnitude.

Finally, defining coupling strategies, especially for measuring the performance of the algorithms introduced would represent a progressive stride in development of the methodology.

Chapter 6

Objective priors with invariance properties for e-value computation

6.1 Introduction

In Bayesian testing, nuisance parameters are additional parameters introduced into the model to enhance its flexibility and realism, even though the main focus of inference usually revolves around a specific parameter of interest. Dealing with nuisance parameters often entails the cumbersome task of eliciting information about these components and performing multidimensional integration, which can significantly increase computational complexity. This Chapter introduces novel methodologies for statistical hypothesis testing within the framework of Bayesian inference, specifically focusing on the computation of e-values and on the Full Bayesian Significance Test (FBST) theory. It begins by presenting asymptotic expansions of the posterior distribution, which serve to significantly reduce computational costs for the e-value calculation, particularly in models with nuisance parameters. These expansions are highlighted for their ability to provide fast-converging numerical approximations. Furthermore, combined with matching priors, the proposed approach offers the advantage of eliminating the need for eliciting information on the nuisance components and for conducting multidimensional integration, and it produces invariant $e-$values in the presence of nuisance parameters. Recognizing the challenges posed by the intractability of deterministic approximations of the posterior, the Chapter also provides computational strategies to enabling practical implementation of the proposed methodologies.

The parametric framework that we consider can be described as follows. Consider a random sample y^{obs} of size n from a random variable Y with parametric model $p(y;\theta)$, indexed by a paramater θ, with $\theta \in \Theta \subseteq^d$. Given a prior $\pi(\theta)$ on θ, Bayesian inference

for θ is based on the posterior density

$$\pi(\theta|y^{\text{obs}}) \propto \pi(\theta)L(\theta), \tag{6.1}$$

where $L(\theta)$ represents the likelihood function based on $p(y^{\text{obs}};\theta)$. Interest is in particular in the situation in which $\theta = (\psi, \lambda)$, where ψ is a scalar parameter for which inference is required and λ represents the remaining $(d-1)$ nuisance parameters. In such case, Bayesian inference for ψ is based on the marginal posterior density

$$\pi_m(\psi|y^{\text{obs}}) = \int \pi(\psi, \lambda|y^{\text{obs}})\, d\lambda \propto \int \pi(\psi, \lambda)L(\psi, \lambda)\, d\lambda, \tag{6.2}$$

which for its computation requires both elicitation on the nuisance parameter λ and multidimensional integration.

Asymptotic arguments are widely used in Bayesian inference through (6.1) and (6.2), based on developments of so-called higher-order asymptotics (see, e.g., Reid 2003, Brazzale *et al.* 2007, Ventura *et al.* 2013, Ventura and Reid 2014 and Cabras *et al.* 2015). Indeed, the theory of asymptotic expansions provides very accurate approximations to posterior distributions, and to various summary quantities of interest, including tail areas, credible regions and for the Full Bayesian Significance Test (see, e.g., Pereira and Stern 1999 and Madruga *et al.* 2003). Moreover, they are particularly useful for sensitivity analyses (see Kass *et al.* 1989, Reid and Sun 2010 and Ruli *et al.* 2014) and also for the derivation of matching priors (see Datta and Mukerjee 2004, and references therein). For instance, focusing on the presence of nuisance parameters, the Laplace approximation to (6.2) provides

$$\pi_m(\psi|y^{\text{obs}}) \doteqdot \frac{1}{\sqrt{2\pi}}|j_p(\hat{\psi})|^{1/2}\exp\{\ell_p(\psi) - \ell_p(\hat{\psi})\}\frac{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}\frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})}, \tag{6.3}$$

where $\ell_p(\psi) = \log L_p(\psi) = \log L(\psi, \hat{\lambda}_\psi)$ is the profile log-likelihood for ψ, with $\hat{\lambda}_\psi$ the constrained maximum likelihood estimate (MLE) of λ given ψ, $(\hat{\psi}, \hat{\lambda})$ is the full MLE, and $j_p(\psi) = -\partial^2\ell_p(\psi)/\partial\psi^2$ is the profile observed information. Moreover, $j_{\lambda\lambda}(\psi, \lambda)$ is the $(\lambda, \lambda)-$block of the observed information from the full log-likelihood $\ell(\psi, \lambda) = \log L(\psi, \lambda)$, and the notation $\doteqdot$ indicates that the approximation is accurate to order $O(n^{-3/2})$ in moderate deviation regions (see, e.g., Severini (2000), Chapter 2). One appealing feature of higher-order approximations like (6.3) is that they may routinely be applied in practical Bayesian inference, since they require little more than standard likelihood quantities for their implementation, and hence they may be available at little additional computational cost over simple first-order approximations.

In the presence of nuisance parameters, starting from approximation (6.3) it is possible to define a general posterior distribution for ψ of the form

$$\pi^*(\psi|y^{\text{obs}}) \propto \pi^*(\psi)L_p(\psi), \tag{6.4}$$

where $\pi^*(\psi)$ is now a prior distribution on ψ only. Bayesian inference based on pseudo-likelihood functions – i.e. functions of ψ only and of the data y^{obs} with properties similar to those of a genuine likelihood function, such as the profile likelihood – has been widely used and discussed in the recent statistical literature. Moreover, it has been theoretically motivated in several papers (see, for instance, Ventura and Racugno 2016, Giummolé *et al.* 2019, Leisen *et al.* 2020, Miller 2021, and references therein), also focusing on the derivation of suitable objective priors. Especially when the dimension of λ is large, there are two advantages in using (6.4) instead of the marginal posterior distribution (6.2). First, the elicitation over λ is not necessary and, second, the computation of the integrals in (6.2) is circumvented.

Focusing on (6.4), we are interested in testing the precise (or sharp) null hypothesis

$$H_0 : \psi = \psi_0 \quad \text{against} \quad H_1 : \psi \neq \psi_0 \tag{6.5}$$

using the measure of evidence for the Full Bayesian Significance Test (see, e.g., Pereira and Stern 1999 and Madruga *et al.* 2003). The Full Bayesian Significance Test (FBST) quantifies evidence by considering the posterior probability associated with the least probable points in the parameter space under H_0. Higher-order asymptotic computation of the FBST for precise null hypotheses in the presence of nuisance parameters has been discussed in Cabras *et al.* (2015).

The original measure of evidence for the FBST is not invariant under suitable transformations of the parameter, a property which has been reached however in the more recent definition of the $e-$value (see Pereira and Stern 2022 and Diniz *et al.* 2020, and references therein). Neverthless, when working on a scalar parameter of interest, in the presence of nuisance parameters, the $e-$value is not invariant with respect to marginalisations of the nuisance parameter and it must be used in the full dimensionality of the parameter space. This requires elicitation on the complete parameters, numerical optimization and numerical integration, that can be computationally heavy especially when the dimension of λ is large. The aim of this contribution is to consider the $e-$value in the context of the pseudo-posterior distribution $\pi^*(\psi|y^{\text{obs}})$, suggesting in this respect a suitable objective prior $\pi^*(\psi)$ to be used in (6.4). More precisely, focus is on a particular matching prior, which ensure invariance of the posterior mode of the pseudo-posterior

distribution. As a consequence also Highest Probability Density credible (HPD) sets are invariant, as well as the $e-$value.

6.2 The FBST measure of evidence

Suppose that we need to decide between two hypotheses: the null H_0 and the alternative H_1. The usual Bayesian testing procedure is based on the well-known Bayes factor (BF), defined as the ratio of the posterior odds to the prior odds in favor of the null hypothesis. A high BF or its logarithm suggest evidence in favor of H_0. However, it is well-known that, when improper priors are used, the BF can be undetermined and, when the null hypothesis is precise, as specified in (6.5), the BF can lead to the so-called Jeffreys-Lindley's paradox (see, e.g. Robert 2014). Moreover, the BF is not calibrated, i.e. its finite sampling distribution is unknown and it may depend on the nuisance parameter.

To avoid these drawbacks, in recent years an alternative Bayesian procedure, called FBST, has been introduced by Pereira and Stern (1999) in case of sharp hypothesis H_0 identified by the null set Θ_0, a submanifold of Θ of lower dimension. The FBST quantifies evidence by considering the posterior probability associated with the least probable points in the parameter space Θ_0. When this probability is high, it favors the null hypothesis, providing a clear and interpretable measure of support for H_0 (see, e.g. Madruga *et al.* 2001, Madruga *et al.* 2003 and Pereira and Stern 2022, and references therein). The FBST is based on a specific loss function, and thus the decision made under this procedure is the action that minimizes the corresponding posterior risk.

The FBST operates by determining the $e-$value, a representation of Bayesian evidence associated to H_0. To construct the $e-$value, the authors introduced the *posterior surprise function* and its supremum given, respectively, by

$$\pi_s(\theta|y^{\text{obs}}) = \frac{\pi(\theta|y^{\text{obs}})}{r(\theta)} \quad \text{and} \quad s^* = \pi_s(\theta^*|y^{\text{obs}}) = \sup_{\theta\in\Theta_0} \pi_s(\theta|y^{\text{obs}}),$$

where $r(\theta)$ is a suitable *reference function* to be chosen. Then, they introduce the *tangential set* $T_y(\theta^*)$ defined as the set of parameter values for which the posterior surprise function exceeds the supremum s^*, that is

$$T_y(\theta^*) = \{\theta \in \Theta : \pi_s(\theta|y) > s^*\}.$$

This set, often referred to as the Highest Relative Surprise Set, includes parameter values

with higher surprise than those within the null set Θ_0. The $e-$value is then computed as

$$ev = 1 - \int_{T_y(\theta^*)} \pi_s(\theta|y^{\text{obs}})\, d\theta,$$

and H_0 is rejected for *small* values of ev.

The original FBST, as proposed by Pereira and Stern (1999) and Pereira and Stern (2001), relies on a flat reference function $r(\theta) \propto 1$, so that this first version involved the determination of the tangential set $T_y(\theta)$ starting only from the posterior distribution $\pi(\theta|y^{\text{obs}})$. However, this initial version lacked invariance under reparameterizations. Subsequent refinements of the FBST introduced the importance of reference density functions, making the $e-$value explicitly invariant under appropriate transformations of the parameter. Common choices for the reference function include uninformative priors, like the uniform distribution, maximum entropy densities, or Jeffreys' invariant prior. In Druilhet and Marin (2007), the use of the Jeffreys' prior, $\pi(\theta) \propto |i(\theta)|^{1/2}$, where $i(\theta)$ is the Fisher information derived from $L(\theta)$, is discussed as the reference function to derive invariant HPD sets and Maximum A Posteriori (MAP) estimators that are invariant under reparameterizations. Note that the ev uses the full dimensionality of the parameter space. Moreover, this measure is not invariant with respect to transformations of the nuisance parameters and the use of high posterior densities to construct credible sets may produce inconsistencies.

Concerning the asymptotic behavior of the ev it can be proven that, under suitable regularity conditions as the sample size increases, with θ_0 representing the true parameter value (see Pereira and Stern 2022), it holds:

- If H_0 is false, i.e. $\theta_0 \notin H_0$, then ev converges in probability to 1.
- If H_0 is true, i.e. $\theta_0 \in H_0$, then, denoting by $V(c) = Pr(ev \leq c)$ the cumulative distribution function of ev, we have that $V(c) \approx Q(d - h, Q^{-1}(d, c))$, with $d = dim(\Theta)$, $h = dim(\Theta_0)$ and $Q(k, x)$ the cumulative chi$-$square distribution with k degrees of freedom.

In practice, the computation of ev is performed in two steps: (a) a numerical optimization and (b) a numerical integration. The numerical optimization step consists in finding the maximizer θ^* of $\pi_s(\theta|y)$ under the null hypothesis. The numerical integration step consists of integrating the posterior surprise function over the region where it is greater than $\pi_s(\theta^*|y)$, to obtain the $e-$value. These computational steps make the FBST a computationally intensive procedure. Despite efficient computational algorithms for local and global optimization, as well as numerical integration, obtaining precise results for hypotheses like (6.5) is highly demanding, especially with large nuisance parameter

dimensions. Numerical integration can be tackled by resorting to higher-order tail area approximations, as reviewed in the Bayesian framework in Ventura and Reid (2014). An application of asymptotic approximation to the FBST in its first formulation, i.e. with reference function $r(\theta) \propto 1$, has been discussed in Cabras *et al.* (2015).

6.2.1 Asymptotic approximations for the $e-value$

A first-order approximation for the $e-$value, when testing (6.5), is simply given by (see, e.g., Pereira *et al.* 2008, Diniz *et al.* 2012)

$$ev \doteq 2\left(1-\Phi\left(\left|\frac{\psi_0-\hat{\psi}}{\sqrt{j_p(\hat{\psi})^{-1}}}\right|\right)\right), \tag{6.6}$$

where the symbol "$\doteq$" indicates that the approximation is accurate to $O(n^{-1/2})$ and $\Phi(\cdot)$ is the standard normal distribution function. Thus, to first-order, ev agrees with the $p-$value based on the profile Wald statistic

$$w_p(\psi) = \frac{(\hat{\psi}-\psi_0)}{\sqrt{j_p(\hat{\psi})^{-1}}}. \tag{6.7}$$

In practice, the approximation (6.6) of ev may be inaccurate, in particular when the dimension of λ is large with respect to the sample size, because it forces the marginal posterior distribution to be symmetric.

The practical computation of ev requires the evaluation of integrals of the marginal posterior distribution. In order to have more accurate evaluations of ev, it may be useful to resort to higher-order asymptotics based on tail area approximations (see, e.g., Reid 2003, Ventura and Reid 2014, and references therein). Indeed, the measure of evidence involves integrals of the marginal surprise posterior density $\pi_{ms}(\psi|y^{\text{obs}})$. In particular, extending the application of the tail area argument to the marginal surprise posterior density, we can derive a $O(n^{-3/2})$ approximation to the marginal surprise posterior tail area probability, given by

$$\int_{\psi_0}^{\infty} \pi_{ms}(\psi|y^{\text{obs}})\, d\psi \doteqdot \Phi(r_B^*(\psi_0)), \tag{6.8}$$

where

$$r_B^*(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)}\log\frac{q_B(\psi)}{r_p(\psi)},$$

with

$$r_p(\psi) = \text{sign}(\hat{\psi} - \psi)[2(\ell_p(\hat{\psi}) - \ell_p(\psi))]^{1/2}$$

profile likelihood root and

$$q_B(\psi) = \ell_p'(\psi)|j_p(\hat{\psi})|^{-1/2}\frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}\frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi, \hat{\lambda}_\psi)}\frac{r(\psi, \hat{\lambda}_\psi)}{r(\hat{\psi}, \hat{\lambda})}.$$

In the expression of $q_B(\psi)$, $\ell_p'(\psi) = \partial\ell_p(\psi)/\partial\psi$ is the profile score function.

Using the tail area approximation (6.8), a third-order approximation of the measure of evidence ev can be derived. The approximation, assuming without loss of generality that ψ_0 is smaller than the MAP of $\pi_{ms}(\psi|y^{\text{obs}})$, is given by

$$ev(\psi) \doteqdot 1 - \Phi(r_B^*(\psi_0)) + \Phi(r_B^*(\psi_0^*)), \tag{6.9}$$

with ψ_0^* the value of the parameter such that $\pi_{ms}(\psi_0^*|y^{\text{obs}}) = \pi_{ms}(\psi_0|y)$. Note that

$$\Phi(r_B^*(\psi_0)) - \Phi(r_B^*(\psi_0^*)) \doteqdot \int_{\psi_0^*}^{\psi_0} \pi_{ms}(\psi|y^{\text{obs}})\, d\psi = 1 - ev$$

in (6.9) gives the posterior probability of the HPD credible interval (ψ_0, ψ_0^*). Note also that the higher-order approximation (6.9) does not call for any condition on the prior $\pi(\psi, \lambda)$, i.e. it can be also improper. Finally, when $\pi_{ms}(\psi|y^{\text{obs}})$ is symmetric, Equation (6.9) reduces to $ev \doteqdot 2(1 - \Phi(r_B^*(\psi_0)))$.

While tail area approximations require little more than standard likelihood quantities for their implementation and, in this respect, they are available at little additional computational cost over the first-order approximation, they require elicitation on the complete parameter θ and to choose the reference function $r(\theta)$.

6.3 An invariant objective prior

The aim of this section is to derive a default prior $\pi^*(\psi)$ to be used in (6.4). To this end, following Datta and Mukerjee (2004) we use the shrinkage argument, which is a crucial procedure in the development of matching priors, i.e. priors that ensure, up to the desired order of asymptotics, an agreement between Bayesian and frequentist procedures. Examples of matching priors are (see Datta and Mukerjee 2004) for posterior quantiles, for credible regions and for prediction. Here, we focus on a specific matching prior, that ensures the invariance of the posterior mode in the posterior distribution

(6.4). As a consequence, the invariance extends to HPDs, as well as the $e-$value, achieved incorporating the reference function within the prior.

The proposed choice of the prior $\pi^*(\psi)$, that makes the MAP and thus also HPDs and the $e-$value invariant under 1-1 reparameterization, will depend on the log-likelihood $\ell(\theta)$ and on its derivatives. In regular parametric estimation problems, both the MLE and the score estimating function exhibit an asymptotically symmetric distribution centered at the true parameter value and at zero, respectively. However, these asymptotic behaviours may poorly reflect exact sampling distributions particularly in cases with small or moderate sample information, sparse data, or complex models. Several proposals have been developed to correct the estimate or the estimating function. Most available methods are aimed at approximate bias adjustment, either of the MLE or of the profile score function, also when nuisance parameters are present (see Kosmidis 2014 for a review of bias reduction for the MLE.the median modification of the score, or profile score, does not rely on finiteness of the MLE, thereby effectively preventing infinite estimates.

In practice, to derive the median matching prior prior $\pi^*(\psi)$, we impose that the MAP of $\pi^*(\psi|y)$ coincides with a refined version of the MLE, obtained as the solution of the median modified score function (Kenne Pagui *et al.*, 2017). To introduce this new invariant prior, we initially explore the scenario without nuisance parameters and then the situation in which nuisance parameters are present.

6.3.1 No nuisance parameters

Let's explore first the scenario where θ is scalar. In order to obtain median bias reduction of the MLE, it is possible to resort to a modified version of the score function of the form

$$t(\theta) = \ell_\theta(\theta) + m(\theta), \tag{6.10}$$

where $\ell_\theta(\theta) = \ell_\theta(\theta; y^{\text{obs}}) = \partial\ell(\theta; y)/\partial\theta$ is the score function and $m(\theta)$ a suitable correction term of order $O(1)$. In particular, the median modified score function assumes for $m(\theta)$ the expression

$$m(\theta) = -\frac{E(\ell_\theta(\theta)^3)}{6\, i(\theta)}.$$

The solution $\tilde{\theta}$ to the equation $t(\theta) = 0$ not only upholds equivariance under componentwise monotone reparameterizations but also approximates median unbiasedness (Kenne Pagui *et al.*, 2017). Note that likelihood inference based on (6.10) does not

depend explicitely on the MLE. Indeed, the modified score function has been found to overcome infinite estimate problems. Likewise the MLE, also $\tilde{\theta}$ is asymptotically $N(\theta, i(\theta)^{-1})$, so that the Wald-type statistics only differ in location.

Since Bayes' theorem is a statement of adittivity on the log scale $\log \pi(\theta|y) = \log \pi(\theta) + \log L(\theta)$+constant, we observe that in the Bayesian framework $m(\theta)$ can be interpreted as the derivative of the logarithm of a prior, that is $m(\theta) = \partial \log \pi(\theta)/\partial\theta$. We are thus looking for a matching prior $\pi^*(\theta)$ such that

$$\frac{\partial \log \pi^*(\theta)}{\partial \theta} = -\frac{E(\ell_\theta(\theta)^3)}{6\, i(\theta)}.$$

In the scalar parameter case, it is straightforward to show that the proposed *median matching prior* takes the form

$$\begin{aligned} \pi^*(\theta) &\propto \exp\left(-\frac{1}{6}\int i(\theta)^{-1} E(\ell_\theta(\theta)^3)\, d\theta\right) \\ &\propto \exp\left(\frac{1}{6}\int i(\theta)^{-1}(3E(\ell_{\theta\theta}(\theta)\ell_\theta(\theta)) + E(\ell_{\theta\theta\theta}(\theta)))\, d\theta\right), \end{aligned}$$

where $\ell_{\theta\theta}(\theta) = \partial\ell_\theta(\theta)/\partial\theta$ and $\ell_{\theta\theta\theta}(\theta) = \partial\ell_{\theta\theta}(\theta)/\partial\theta$. The posterior based on the median matching prior is thus

$$\pi^*(\theta|y^{\text{obs}}) \propto \exp\left(\ell(\theta) - \frac{1}{6}\int i(\theta)^{-1} E(\ell_\theta(\theta)^3)\, d\theta\right).$$

A first-order approximation for the $e-$value, when testing $H_0 : \theta = \theta_0$, is simply given by

$$ev \doteq 2\left(1 - \Phi\left(\left|\frac{\theta_0 - \tilde{\theta}}{\sqrt{i(\theta_0)^{-1}}}\right|\right)\right), \tag{6.11}$$

which differs in location with respect to the classical first-order approximation for the $e-$value based on the MLE. A second approximation for the $e-$value, when testing $H_0 : \theta = \theta_0$, can be obtained from the asymptotic distribution of the modified score function (6.10), that is

$$ev \doteq 2\left(1 - \Phi\left(\left|\frac{t(\theta_0)}{\sqrt{i(\theta_0)}}\right|\right)\right). \tag{6.12}$$

Although the first-order equivalence between (6.11) and (6.12), note that (6.11) is based on an easily understandable comparison between estimated value and hypothetical value,

taking estimation error into account, and is widely used in applications, but does not satisfy the principle of parameterization invariance. On the other hand, $t(\theta)/\sqrt{i(\theta)}$ is parameterization invariant.

Note that, when using a *predictive matching prior*, i.e. a prior ensuring asymptotic equivalence of higher-order frequentist and Bayesian predictive densities (see, e.g., Datta and Mukerjee 2004), the term $m(\theta)$ in (6.10) corresponds to the Firth's adjustment

$$m_F(\theta) = -\frac{(E(\ell_\theta(\theta)^3) + E(\ell_{\theta\theta}(\theta)\ell_\theta(\theta)))}{2i(\theta)}.$$

In view of this, for general regular models, Firth's estimate coincides with the mode of the posterior distribution obtained using the default predictive matching prior. However, lack of invariance affects this kind of adjustment (Kosmidis, 2014), unless dealing with linear transformations.

Example 1: One parameter exponential family. For a one parameter exponential family with canonical parameter θ, i.e. with density

$$f(y;\theta) = \exp\{\theta a(y) - K(\theta)\}b(y),$$

the median modified score function has the form

$$t(\theta) = \ell_\theta(\theta) + \frac{K_{\theta\theta\theta}}{6K_{\theta\theta}},$$

where $K_{\theta\theta\theta} = \partial^3 K(\theta)/\partial\theta^3$ and $K_{\theta\theta} = \partial^2 K(\theta)/\partial\theta^2 = i(\theta)$. In this parameterization, $t(\theta)$ can be seen as the first derivative of the log-posterior

$$\log \pi(\theta|y) = \ell(\theta) + \log i(\theta)/6.$$

On the other hand, Firth's modified score takes the form $t_F(\theta) = \ell_\theta(\theta) + K_{\theta\theta\theta}/(2K_{\theta\theta})$. The effect of the median modification is to consider the median matching prior $\pi^*(\theta) \propto i(\theta)^{1/6}$, while $t_F(\theta)$ implies a Jeffreys' prior $\pi_J(\theta) \propto i(\theta)^{1/2}$. Note that, for a one parameter exponential family with canonical parameter θ, both $\pi^*(\theta)$and $\pi_J(\theta)$ belong to the family of invariant priors discussed in Hartigan (1964) and Hartigan (1965).

Example 2: Scale model. Consider the scale model $f(y;\theta) = (1/\theta)p_0(y/\theta)$, where $p_0(\cdot)$ is a given function. Let $g(\cdot) = -\log p_0(\cdot)$. We have $E(\ell_\theta^3) = c_1/\theta^3$, $E(\ell_\theta^3\ell_\theta) = c_2/\theta^3$ and $i(\theta) = c_3/\theta^2$, with $c_1 = \int(y^3g'''(y)+6y^2g''(y)+6yg'(y)-2)p_0(y)dy$, $c_2 = \int(3yg'(y)+y^2g''(y)-2y^2g'(y)^2-y^3g'(y)g''(y)-1)p_0(y)dy$ and $c_3 = \int(2yg'(y)+y^2g''(y)-1)p_0(y)dy$.

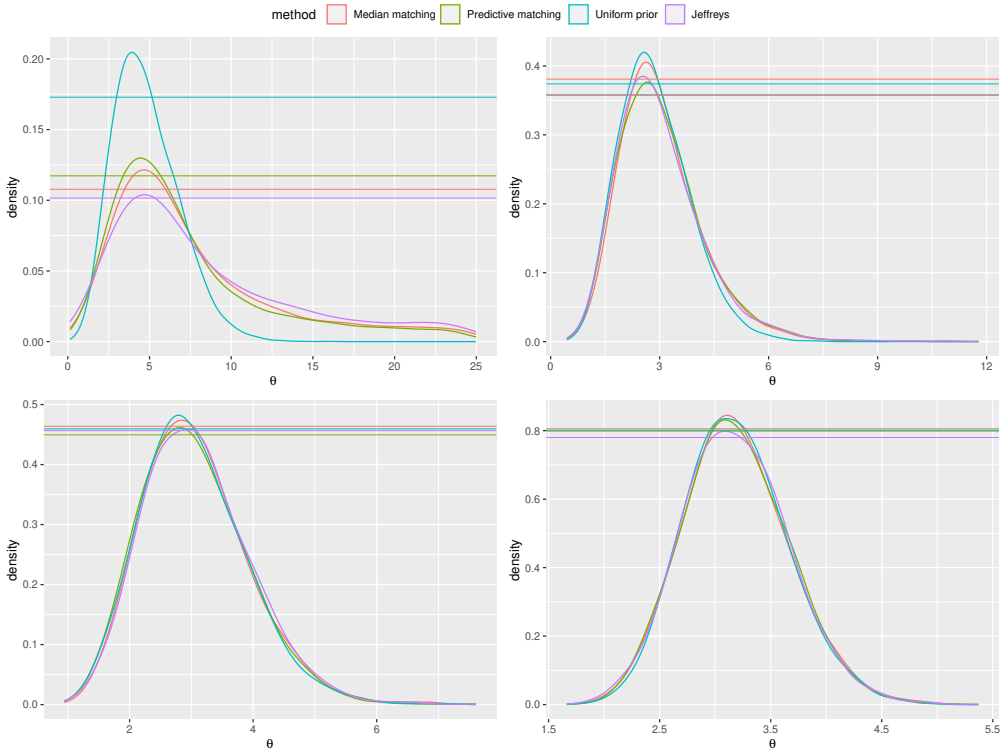

FIGURE 6.1: Inference for the scalar parameter θ of the skew-normal model with sample sizes $n = 20, 30, 50, 200$. The red line is used for the posterior obtained from the median matching prior, the green one for the predictive matching prior, the violet one for the Jeffreys' prior and the blue one from an improper flat prior. The horizontal lines identify tangential sets associated to the hypothesis $H_0 : \theta = 3$.

The median matching prior is thus $\pi^*(\theta) \propto \theta^{-c_1/6c_3}$, while the Jeffreys' prior for a one-parameter scale model is $\pi_J(\theta) \propto \theta^{-1}$ and the prior associated to the Firth's adjustment is $\pi_F(\theta) \propto \theta^{-(c_1+c_2)/2c_3}$.

Example 3: Skew-normal distribution. Consider a skew-normal distribution, with shape parameter $\theta \in$, and density $f(y;\theta) = 2\phi(y)\Phi(y\theta)$, where $\phi(\cdot)$ is the standard normal probability density function and $\Phi(\cdot)$ is its cumulative density function. The median correction term for the score function associated to the median matching prior is (see Sartori 2006,Kenne Pagui *et al.* 2017)

$$m(\theta) = \frac{E(y^3\phi(y\theta)^3/\Phi(y\theta)^3)}{6E(y^2\phi(y\theta)^2/\Phi(y\theta)^2)}.$$

In this setting, numerical integration must be performed to obtain the expected values involved in $m(\theta)$.

In order to illustrate the proposed prior, we consider draws from the skew-normal model with true parameter $\theta_0 = 3$ and increasing sample sizes ($n = 20, 30, 50, 200$, top-left, top-right, bottom-left and bottom-right panels of Figure 6.1, respectively). The

H_0	n	**Flat prior**	**Median m. prior**	**Predictive m. prior**	**Jeffreys' prior**
$\theta = 3$	20	0.59	0.62	0.56	0.65
	30	0.65	0.70	0.73	0.64
	50	0.81	0.84	0.82	0.91
	200	0.79	0.79	0.81	0.82
$\theta = 4$	20	0.82	0.91	0.98	0.84
	30	0.17	0.22	0.22	0.22
	50	0.20	0.20	0.20	0.21
	200	0.07	0.08	0.08	0.09

TABLE 6.1: Skew-normal: e-values for hypotheses $H_0 : \theta = 3$ and $H_0 : \theta = 4$.

posterior distributions are obtained with the method by Ruli *et al.* (2020), i.e. drawing 10^5 values and accepting the best 5%. The $e-$values associated to the null (true) hypothesis $H_0 : \theta = 3$ and the (false) hypothesis $H_0 : \theta = 4$ are reported in Table 6.1. For comparison, also the Jeffreys' prior (Liseo and Loperfido 2006), the flat prior, with uniform reference function, and the predictive matching prior (Sartori 2006) are considered. Progressive agreement among evidence values obtained with the proposed median matching prior and the other priors for larger sample size is shown. Also, as expected, progressively increasing n, the evidence values indicate agreement with the true hypothesis $\mathcal{H} : \theta_0 = 3$ and disagreement with the $\mathcal{H} : \theta_0 = 4$ for all the priors used. Anyway, note that the posterior distribution obtained with a flat prior, and a uniform reference function, is proportional to the likelihood function that can be monotone. In view of this, while the MAPs of the posterior based on the default priors are always finite, in some samples the MAP of the posterior with the non-informative prior may be infinite. An example of this effect is illustrated in Figure 6.2.

The properties of first-order approximations of the $e-$values have been investigated by a simulation study, again with sample sizes $n = 20, 30, 50, 200$. Results are displayed in Figure 6.3. Distributions of the $e-$value from the posterior based on the median matching prior are better in terms of convergence to the Uniform distribution both for small and moderate sample sizes. Moreover, score-type $e-$values are also preferable than Wald-type $e-$values. For the results with the posterior distribution obtained with a flat prior, we found $4.3\%, 4.2\%, 0.9\%, 0\%$ of infinite estimates for the sample sizes considered in the simulation study, respectively and in these cases the $e-$value was considered as 0.

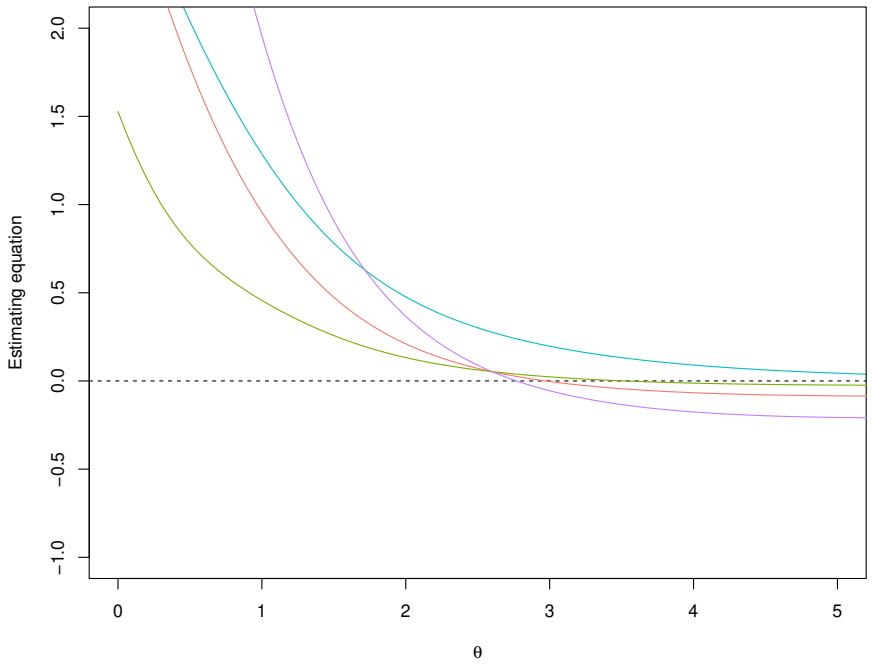

FIGURE 6.2: Skew-normal model: an example of $\partial \log \pi(\theta|y)/\partial\theta$ (estimating equation) with a flat prior (blue line), the median matching prior (red line) the predictive matching prior (green line) and the Jeffreys' prior (violet line) in a sample where all the observations are positive.

6.3.2 Presence of nuisance parameters

In the presence of nuisance parameters, in order to obtain median bias reduction of the MLE, it is possible to resort to a modified version of the profile score function of the form

$$t_p(\psi) = \ell'_p(\psi) + m(\psi, \hat{\lambda}_\psi), \tag{6.13}$$

where $m(\psi, \lambda)$ a suitable correction term of order $O(1)$. In particular, for the median modified profile score function, the adjustment $m(\psi, \lambda)$ assumes the expression

$$m(\psi, \lambda) = -\kappa_{1\psi} + \frac{\kappa_{3\psi}}{6\kappa_{2\psi}},$$

where $\kappa_{1\psi}$, $\kappa_{2\psi}$ and $\kappa_{3\psi}$ are the first three cumulants of $\ell'_p(\psi)$ (see Kenne Pagui *et al.* 2017, Section 2.2, for their expression). For the estimator $\tilde{\psi}_p$, defined as the solution of $t_p(\psi) = 0$, parametrization equivariance holds under interest respecting reparametrizations (Kenne Pagui *et al.* 2017).

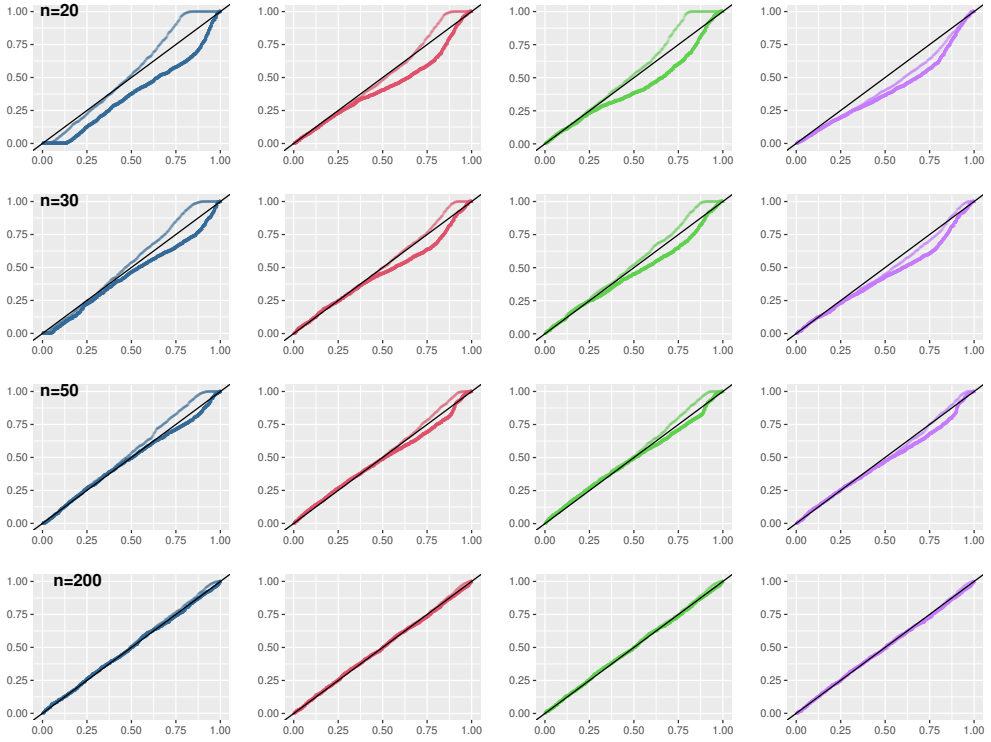

FIGURE 6.3: Skew-normal model: distribution of $e-$values from a simulation study under the null hypothesis $H_0 : \theta = 3$, using a flat prior (blue line), the median matching prior (red line), the predictive matching prior (green line), and the Jeffreys' prior (violet line). The darker line is used for the approximation (6.11) while the lighter for that based on (6.12).

Note that, also the context of nuisance parameters, we are in the situation in which the proposed prior $\pi^*(\psi)$ is known through its first derivative; this is typically the situation with matching priors (see, e.g., Datta and Mukerjee 2004). Since the parameter of interest is scalar, the posterior based on the median matching prior can be written as

$$\pi^*(\psi|y^{\text{obs}}) \propto \exp\left(\ell_p(\psi) + \int m(\psi, \hat{\lambda}_\psi)\, d\psi\right). \tag{6.14}$$

A simple analytical way of approximating to first-order the posterior distribution (6.14) based on the median matching prior is to resort to a quadratic form of $t_p(\psi)$. In particular, the approximate posterior distribution for ψ takes the form

$$\pi^*(\psi|y^{\text{obs}}) \dot{\propto} \exp\left(-\frac{1}{2}s_p(\psi; y^{\text{obs}})\right), \tag{6.15}$$

where $s_p(\psi) = t_p(\psi)^2 j_p(\psi)^{-1}$ is a Rao score-type statistic based on (6.13) and the symbol $\dot{\propto}$ means asymptotic proportionality at first-order. In this case, a first-order

approximation of the $e-$value, when testing $H_0 : \psi = \psi_0$, is given by

$$ev \doteq 2\left(1 - \Phi\left(\left|\frac{t_p(\psi_0)}{\sqrt{j_p(\psi_0)}}\right|\right)\right). \tag{6.16}$$

In this case, an higher order approximation via (6.3) would be impractical since a closed form prior is not available. As an alternative, simulation-based approaches may be used to derive the implied posterior distribution (6.14) based on the median matching prior. The first one relies on Approximate Bayesian Computation (ABC) techniques, using $\tilde{\psi}_p$ or the modified profile score function $t_p(\psi)$ as summary statistics; see Bortolato and Ventura (2023) for the modification of the algorithm of Ruli *et al.* (2020) by using a profile estimating equation. This first method introduces an approximation at the level of the posterior estimation. The second one still relies on (6.13) but considers use of Manifold MCMC methods (see ,e.g., Brubaker *et al.* 2012) to conditioning exactly on the profile equation and not up to a tolerance level, as in ABC (see also Lewis *et al.* 2021 and Graham and Storkey 2017 for similar ideas). The algorithm moves on the constrained space $\{(y, \psi) \in \mathcal{Y} \times \ominus | t_p(\tilde{\psi}_p) = 0\}$, where $\tilde{\psi}_p$ is the solution of the estimating equation on the original data. For the latter method, we need minimal regularity assumptions on $m(\psi, \lambda)$, which is assumed to be continuous, differentiable and available in closed form expression. Note, for instance, that in the skew-normal example in Section 3.1 these conditions are not met.

Example 4: Exponential family. If $f(y; \theta)$ is an exponential family of order d with canonical parameter (ψ, λ), i.e. $f(y; \psi, \lambda) = \exp\{\psi t(y) + \lambda^\top s(y) - K(\psi, \lambda)\}h(y)$, quantities involved in $m(\psi, \lambda)$ are simply obtained from derivatives of $K(\psi, \lambda)$ (Kenne Pagui *et al.*, 2017). Note that, in this framework, $\ell'_p(\psi) - \kappa_{1\psi}$ is an approximation with error of order $O(n^{-1})$ of the score for ψ in the conditional model given $s(y)$. Then, in the continuous case, the MAP $\tilde{\psi}_p$ is an approximation of the optimal conditional median unbiased estimator, and $\pi^*(\psi | y^{\text{obs}})$ is related to the conditional likelihood for ψ given by $L_c(\psi) = \exp(\psi t(y) - K_s(\psi))$; see Severini (1999) for a Bayesian interpretation of such pseudo-likelihoods.

Example 5: Multivariate regression model. Consider a regression model of the form

$$Y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, 2,$$

where it is assumed that $\epsilon_i \sim N_2(0, \Sigma)$, with $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ positive definite matrix, with $\sigma > 0$, $-1 < \rho < 1$ and $(\beta_0, \beta_1, \beta_2, \sigma^2, \rho)$ are unknown parameters. This model

is widely used for instance in time series analysis, where the past observables are often used as regression covariates. We focus on the problem of testing hypothesis on the correlation coefficient ρ. For obtaining the posterior (6.14) for ρ we first compute the MAPs with the median matching prior and also, for comparison, with the predictive matching prior, which are respectively 0.953 and 0.92. Note that the expression of the predictive matching prior for (ψ, λ) corresponds to the Firth's adjustment to the score function. The expressions of the modified profile estimating functions $t_p(\psi)$ and $t_F(\psi)$ are obtained from Bortolato and Kenne Pagui (2023) and are available in closed form expressions. Hence, the Manifold MCMC method can be used to obtain the implied posteriors, whose approximation is comparable to that of any MCMC sampler. In particular we used 20000 iterations.

We compare the posterior distribution obtained from a draw with true parameter $\rho_0 = 0.95$ based on the proposed median matching prior, with those obtained with the predictive matching prior and with an inverse-Wishart prior for the covariance matrix Σ with one degree of freedom and identity position, and uniform prior on the regression parameters. The resulting posterior distributions are displayed in Figure 6.4. The hypothesis of interest is $H_0 : \rho = 0.9$, and a smaller e-value indicating disagreement with the hypothesis should be preferable. The $e-$values are 0.25 with the median matching prior, 0.36 for the predictive matching prior and 0.60 with the inverse-Wishart prior. Note that the $e-$value based on the inverse-Wishart prior involves the constrained maximization and multidimensional integration, thus is not directly readable in Figure 6.4. Indeed, one crucial difference is that the original $e-$value formulation links the evidence of the null hypothesis to the evidence of a more refined hypothesis, choosing the MAP under the null hypothesis for all the nuisance parameters, while in the alternative (tangential) set all values are used, and integration is performed on the full dimensionality of the space. On the contrary, in the proposed posterior based on the median matching prior, the maximizer of nuisance parameters are taken both in the null and non null sets.

Finally, for the posterior based on the inverse-Wishart prior, we also computed the e-value based on high-order tail area approximation (6.9) of the marginal surprise posterior, which is equal to 0.27. This procedure still avoids the multidimensional integration but the result is not invariant to changes of parametrization.

Example 6: Logistic regression model. Let y_i, $i = 1, \ldots, n$, be independent realizations of binary random variables with success probability π_i. We indicate with $\log(\pi_i/(1-\pi_i)) = \eta_i = x_i\beta$ the linear predictor, where $x_i = (x_{i1}, \ldots, x_{ip})$ is a row vector of covariates. For such a model, we assume that a generic scalar component of β is of

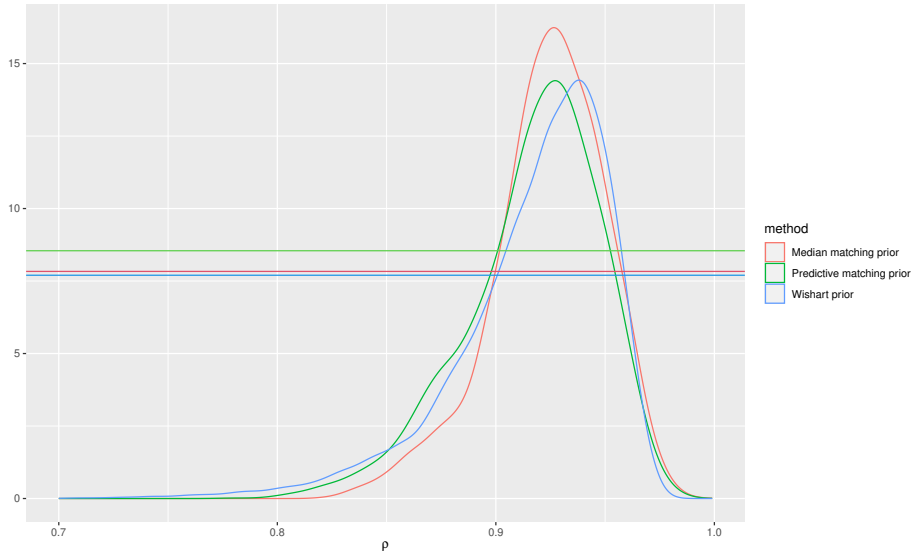

FIGURE 6.4: Posterior distributions for the correlation parameter ρ of the bivariate regression model obtained from MCMC draws and the three different priors.

interest and we treat the remaining components as nuisance parameters.

As an example, we consider the *endometrial cancer grade* dataset analyzed, among others, in Agresti (2015). The aim of the clinical study was to evaluate the relationship between the histology of the endometrium (HG) (encoded as a binary response variable) of $n = 79$ patients and three risk factors: 1. Neovascularization (NV), that indicates the presence or extent of new blood vessel formation; 2. Pulsatility Index (PI), that measures blood flow resistance in the endometrium; 3. Endometrium Height (EH), that indicates the thickness or height of the endometrium. A logistic model for HG, including an intercept and using all the covariates (NV, PI, EH), has been fitted, but maximum likelihood leads to infinite MLE of the coefficient β_2 related to NV, due to quasi-complete separation. This phenomenon prohibits the use of diffuse priors for β_2, since the corresponding posterior wouldn't concentrate. Moreover, the $e-$value with non-informative priors cannot be obtained also for any hypothesis concerning parameters different from β_2.

If we consider β_2 as the parameter of interest, while the remaining regression coefficients are treated as nuisance parameters, the analysis with the median matching prior allows to obtain a global proper posterior, with MAP equal to 3.86, open to interpretation both in the original scale and in terms of odds ratios. Similarly, the posterior based on the predictive matching prior, which in this model coincides with Jeffreys' prior $\pi(\beta) \propto |i(\beta)|^{1/2}$ is proper, with the MAP set at 2.92. The latter suffers from lack of interpretability on different scales, since a different parametrization in estimation phase would affect the results.

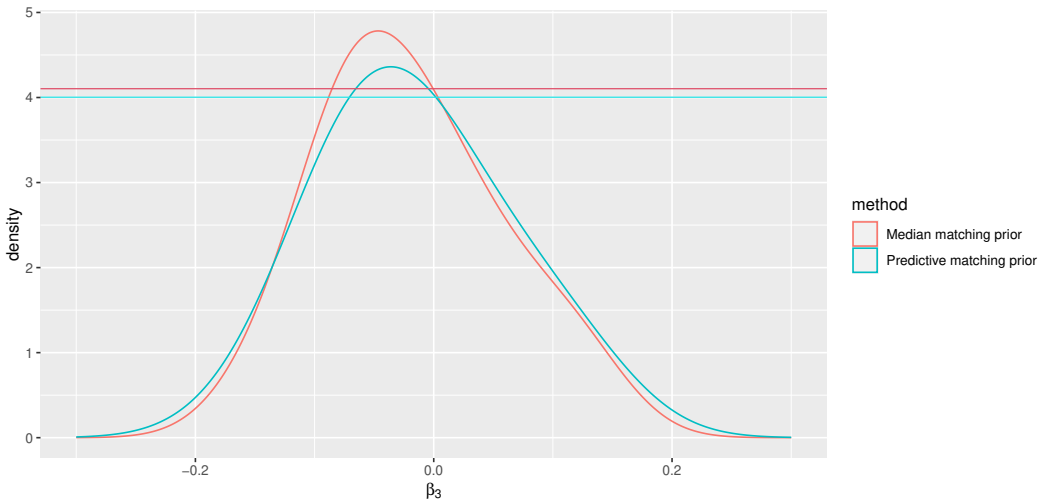

FIGURE 6.5: Median matching posterior distribution and predictive matching posterior distribution for β_3 in the logistic regression model.

If we consider β_3 as the parameter of interest, related to the risk factor PI, the MAPs are -0.038 when using the median matching prior and -0.035 when using the predictive matching prior. The $e-$values for the hypothesis $\mathcal{H}_0 : \beta_3 = 0$ are 0.60 and 0.55, respectively (see Figure 6.5). Likewise, the interpretation of e-values remains consistent and independent of parametrization solely in the first case.

6.4 Discussion and remarks

Although (6.14) cannot always be considered orthodox in a Bayesian setting, the use of alternative likelihoods is nowadays widespread, and several works focus on the Bayesian application of some well-known pseudo-likelihoods. In particular, the proposed posterior $\pi^*(\psi|y)$ has the advantage of avoiding the elicitation on the nuisance parameter λ and the computation of multidimensional integrals. Moreover, it provides invariant MAPs, HPDs and $e-$values, without the adoption of a reference function. Finally, we remark that frequentist properties of the MAP of the posterior based of the proposed median matching prior in comparison with the MAP of the posterior based of the predictive matching prior have been investigated in Kenne Pagui *et al.* (2017) and Bortolato and Kenne Pagui (2023) for some of the examples discussed in this Chapter.

For inference on a full vector parameter θ, with $d > 1$ components, a direct extension of the rationale leading to (6.10) does not seem to be practicable due to lack of a manageable definition of multivariate median. Actually, in Kosmidis *et al.* (2020) it is shown how the method can be extended to a vector parameter of interest, in

the presence of nuisance parameters, by simultaneously solving median bias corrected score equations for all parameter components. This leads to componentwise third-order median unbiasedness and parameterization equivariance. Moreover, the use of default priors involving all parameter components, also the nuisance, becomes necessary to regularize likelihoods in case of monotonicity. We note that among the possible objective priors that ensures invariance of the posterior, we did not focus on the Jeffreys' in the multidimensional case, since it often exhibits poor convergence properties. Conversely, the default matching priors considered in this Chapter are easily generalizable to the multidimensional case Kosmidis *et al.* (2020) preserving good convergence properties.

As a final remark, we highlight that this investigation opens to several topics of future research. In particular, from a computational point of view, it could be of interest: to develop a library of computational routines exploring the methods proposed in this Chapter for a wide range of statistical models of interest, together with semi-automated procedures for further expanding this library, as done for point estimation for Generalized Linear Models in the `R` package `brglm2` Kosmidis (2023). From a theoretical point of view, it could be of interest to further explore the theoretical connections between the $e-$value invariance properties and matching priors, to explore the existence of similar connections in other classes of pseudo-likelihoods. Also, it would be to apply and extend the methodology to consider other objective priors used in Bayesian inference, such as those obtained from scoring rules, as proposed by Leisen *et al.* (2020), as solutions of differential equations or in the context of context of empirical and profile empirical likelihoods, with a large number of nuisance parameters (see e.g. Bedoui and Lazar 2020).

Conclusions

Discussion

This dissertation focuses on two main research lines. The first line investigates the computation of confidence distributions within the frequentist framework. This includes special attention to treating nuisance parameters, extensions to possible use of multivariate statistics, applications to non-regular models and to likelihood-free setups. The second line of research explores advanced methods for Markov chain Monte Carlo (MCMC) algorithms on submanifolds, including coupling for MCMC algorithms on submanifolds, with the aim of providing convergence diagnostics, parallelizing computation and measuring the precision of MCMC methods. We have also proposed new sampling schemes that are widely applicable for a range of problems in Bayesian statistics, introducing artificial submanifold-based MCMC strategies. Finally, another methodological application and original contribution related to this line of research concerns the computation of posterior distributions with non tractable priors.

Future directions of research

The investigation of the topics presented in this work paves the way for numerous future research projects. These include the exploration of connections between the topics examined as well as the deepening of the investigations carried out.

In Chapter 2, for example, we did not discuss elimination of nuisance parameters through conditioning. We point out that it may be possible to combine constrained simulations based on manifold-MCMC methods to produce conditional CDs, as well as to use coupling methods to retrieve unbiased uncertainty quantification. Another point to be explored is the connection and differences between simulation-based CDs when using estimating equations and the constrained bootstrap, especially with respect to the treatment of nuisance parameters. An interesting application could be the extension of this framework to compute CD for lasso regressions (Tibshirani, 1996) working in high

dimensional framework. From a computational point of view, all the algorithms used in Chapter 2 and Chapter 3 are based on rejection sampling, but adaptive or iterative schemes could be developed as well.

In Chapter 3, we did not examine the frequentist properties of the data depth functions derived from the Box ABC method to derive valid confidence intervals. This certainly represents an interesting area of research, with the possibility of combining inferences from multiple summaries, as also mentioned in Dungang *et al.* (2022). Another topic that should be further investigated is the convergence properties of the algorithm in presence of many unknown parameters.

In the context of MCMC on manifolds, one line of research that is of particular interest is the elaboration of meeting-inducing couplings for constrained Hamiltonian Monte Carlo (see for example Lelièvre *et al.* 2019, Graham and Storkey 2017), by further elaborating the strategy for HMC in the unconstrained case presented in Chapter 4.

For the MCMC algorithms presented in Chapter 5, where artificial submanifolds are introduced to solve general sampling problems, a research direction not yet explored is the use of tempered versions or other suitable, possibly simpler functions, instead of the graph map. And also the consideration of other types of conditioning or smooth surfaces, when considering alternate moves.

Finally, the use of estimating equations presented in Chapter 6, to deal with intractable prior distributions can be extended to perform posterior inference in many cases where the prior distribution emerges as the solution of partial differential equation, as those derived from scoring rules (see for instance Leisen *et al.*, 2020). An open question is whether the use of estimating equations can be extended in the context of Generalized Fiducial Inference, replacing the data generating equations with the estimating equations, for reducing the dimensionality of the sampling problem.

Bibliography

Agapiou, S., Roberts, G. O. and Vollmer, S. J. (2018) Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models. *Bernoulli* **24**(3), 1726–1786.

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Ahsanullah, M., Nevzorov, V. B. and Shakil, M. (2013) *An introduction to order statistics*. Volume 8. Springer.

Albers, W., Bickel, P. and van Zwet, W. (1978) Correction to "Asymptotic Expansions for the Power of Distribution free Tests in the One-sample problem". *Annals of Statistics* **6**(1), 1170–1171.

An, Z., Nott, D. J. and Drovandi, C. (2020) Robust Bayesian synthetic likelihood via a semi-parametric approach. *Statistics and Computing* **30**(3), 543–557.

Andersen, H. C. (1983) Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of computational Physics* **52**(1), 24–34.

Anderson, M. J. and Robinson, J. (2001) Permutation tests for linear models. *Australian & New Zealand Journal of Statistics* **43**(1), 75–88.

Andrieu, C., Lee, A. and Livingstone, S. (2020) A general perspective on the Metropolis-Hastings kernel. *arXiv preprint arXiv:2012.14881* .

Andrieu, C. and Thoms, J. (2008) A tutorial on adaptive MCMC. *Statistics and Computing* **18**(4), 343–373.

Anonymous (1978) Influenza in a boarding school. *British Medical Journal* **1**, 587–587.

Au, K. X., Graham, M. M. and Thiery, A. H. (2023) Manifold lifting: scaling Markov chain Monte Carlo to the vanishing noise regime. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **85**(3), 757–782.

Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**(2), 171–178.

Barber, R. F. and Janson, L. (2022) Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *The Annals of Statistics* **50**(5), 2514–2544.

Barker, A. A. (1965) Monte Carlo calculations of the radial distribution functions for a proton electron plasma. *Australian Journal of Physics* **18**(2), 119–134.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and asymptotics*. Volume 13. Springer.

Basu, D. (1975) Statistical information and likelihood. *Sankhyā: The Indian Journal of Statistics, Series A (with discussion)* pp. 1–71.

Beaumont, M., Zhang, W. and Balding, D. (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025.

Bedoui, A. and Lazar, N. A. (2020) Bayesian empirical likelihood for Ridge and Lasso regressions. *Computational Statistics & Data Analysis* **145**, 106917.

Bee, M., Benedetti, R. and Espa, G. (2017) Approximate maximum likelihood estimation of the Bingham distribution. *Computational Statistics & Data Analysis* **108**, 84–96.

Bernton, E., Jacob, P. E., Gerber, M. and Robert, C. P. (2019) Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(2), 235–269.

Besag, J. and Clifford, P. (1989) Generalized Monte Carlo significance tests. *Biometrika* **76**(4), 633–642.

Beskos, A. and Kamatani, K. (2022) MCMC algorithms for posteriors on matrix spaces. *Journal of Computational and Graphical Statistics* **31**(3), 721–738.

Betancourt, M., Byrne, S., Livingstone, S. and Girolami, M. (2017) The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* **23**(4A), 2257–2298.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017) Julia: A fresh approach to numerical computing. *SIAM review* **59**(1), 65–98.

Biau, G., Cérou, F., Guyader, A. and ALTRI (2015) New insights into approximate Bayesian computation. *Annales de l'I.H.P. Probabilités et statistiques* **51**(1), 376–403.

Biswas, N., Jacob, P. E. and Vanetti, P. (2019) Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, pp. 7389–7399.

Blum, M. G., Nunes, M. A., Prangle, D. and Sisson, S. A. (2013) A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science* **28**(2), 189–208.

Bornn, L., Shephard, N. and Solgi, R. (2019) Moment conditions and Bayesian nonparametrics. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **81**(1), 5–43.

Bortolato, E. and Kenne Pagui, E. (2023) Bias reduction and robustness in Gaussian longitudinal data analysis. *Journal of Statistical Computation and Simulation* .

Bortolato, E. and Ventura, L. (2023) On approximate robust confidence distributions. *Econometrics and Statistics* .

Bou-Rabee, N. and Eberle, A. (2023) Mixing time guarantees for unadjusted Hamiltonian Monte Carlo. *Bernoulli* **29**(1), 75–104.

Bou-Rabee, N., Eberle, A. and Zimmer, R. (2018) Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability* **30**(3), 1209–1250.

Bouchard-Côté, A., Vollmer, S. J. and Doucet, A. (2018) The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association* **113**(522), 855–867.

Brazzale, A., Davison, A. and Reid, N. (2007) *Applied Asymptotics: Case Studies in Small Sample Statistics*. Cambridge University Press.

Brubaker, M., Salzmann, M. and Urtasun, R. (2012) A family of MCMC methods on implicitly defined manifolds. In *Artificial intelligence and statistics*, pp. 161–172.

Bunke, H. (1975) Statistical inference: Fiducial and structural vs. likelihood. *Mathematische Operationsforschung und Statistik* **6**(5), 667–676.

Cabras, S., Racugno, W. and Ventura, L. (2015) Higher-order asymptotic computation of Bayesian significance tests for precise null hypotheses in the presence of nuisance parameters. *Journal of Statistical Computation and Simulation* **85**, 2989–3001.

Carhart-Harris, R., Giribaldi, B., Watts, R., Baker-Jones, M., Murphy-Beiner, A., Murphy, R., Martell, J., Blemings, A., Erritzoe, D. and Nutt, D. J. (2021) Trial of psilocybin versus escitalopram for depression. *New England Journal of Medicine* **384**(15), 1402–1411.

Carlin, B. P. and Gelfand, A. E. (1991) An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statistics and Computing* **1**, 119–128.

Chen, Y. and Gatmiry, K. (2023) When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm? *arXiv preprint arXiv:2304.04724* .

Chopin, N. (2002) A sequential particle filter method for static models. *Biometrika* **89**, 539–552.

Chopin, N. and Papaspiliopoulos, O. (2020) *An introduction to sequential Monte Carlo*. Volume 4. Springer.

Clarté, G., Robert, C. P., Ryder, R. J. and Stoehr, J. (2021) Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika* **108**(3), 591–607.

Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I. (2018) Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis* **13**(2), 627 – 679.

Cunen, C. and Hjort, N. L. (2022) Combining information across diverse sources: the II-CC-FF paradigm. *Scandinavian Journal of Statistics* **49**(2), 625–656.

D'Agostino Sr, R. B., Massaro, J. M. and Sullivan, L. M. (2003) Non-inferiority trials: design concepts and issues–the encounters of academic consultants in statistics. *Statistics in Medicine* **22**(2), 169–186.

Dai, C., Heng, J., Jacob, P. E. and Whiteley, N. (2022) An invitation to sequential Monte Carlo samplers. *Journal of the American Statistical Association* **117**(539), 1587–1600.

Dalmasso, N., Zhao, D., Izbicki, R. and Lee, A. B. (2021) Likelihood-free frequentist inference: Bridging classical statistics and machine learning in simulation and uncertainty quantification. *arXiv preprint arXiv:2107.03920* .

Datta, G. and Mukerjee, R. (2004) *Probability Matching Priors: Higher-Order Asymptotics*. Lecture Notes in Statistics. Springer.

Davison, A. C. (2003) *Statistical models*. Volume 11. Cambridge university press.

Dawid, A. P., Musio, M. and Ventura, L. (2016) Minimum scoring rule inference. *Scandinavian Journal of Statistics* **43**(1), 123–138.

Dawid, A. P. and Stone, M. (1982) The functional-model basis of fiducial inference. *The Annals of Statistics* pp. 1054–1067.

Del Moral, P., Doucet, A. and Jasra, A. (2007) Sequential Monte Carlo for Bayesian computation. In *Bayesian statistics 8: proceedings of the eighth Valencia International Meeting, June 2-6, 2006*, eds J. Bernardo, M. Bayarri, J. Degroot, A. Dawid, D. Heckerman, A. Smith and M. West, volume 8, p. 115.

Del Moral, P., Doucet, A. and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and computing* **22**, 1009–1020.

Devroye, L. (1985) *Non-uniform random variate generation*. Springer-Verlag, New York.

Diaconis, P., Holmes, S., Shahshahani, M. *et al.* (2013) Sampling from a manifold. *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton* **10**, 102–125.

DiCiccio, T. J. and Efron, B. (1996) Bootstrap confidence intervals. *Statistical Science* **11**(3), 189–228.

Diciccio, T. J., Martin, M. A. and Stern, S. E. (2001) Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canadian Journal of Statistics* **29**(1), 67–76.

DiCiccio, T. J., Martin, M. A. and Young, G. A. (1993) Analytical Approximations to Conditional Distribution Functions. *Biometrika* pp. 781–790.

Diciccio, T. J. and Romano, J. P. (1988) A review of Bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **50**(3), 338–354.

Diniz, M., Pereira, C. and Stern, J. M. (2020) Cointegration and unit root tests: A fully Bayesian approach. *Entropy* **22**, 968.

Diniz, M. A., Pereira, C., Polpo, A., Stern, J. and Wechsler, S. (2012) Relationship between Bayesian and frequentist significance indices. *Int. J. Uncertain. Quantif.* **2**, 161–172.

Doll, J. and Dion, D. (1976) Generalized Langevin equation approach for atom/solid–surface scattering: Numerical techniques for Gaussian generalized Langevin dynamics. *The Journal of Chemical Physics* **65**(9), 3762–3766.

Douc, R., Jacob, P. E., Lee, A. and Vats, D. (2022) Solving the Poisson equation using coupled Markov chains. *arXiv preprint arXiv:2206.05691* .

Drovandi, C. and Frazier, D. T. (2022) A comparison of likelihood-free methods with and without summary statistics. *Statistics and Computing* **32**(3), 42.

Druilhet, P. and Marin, J. (2007) Invariant HPD credible sets and map estimators. *Bayesian Analysis* **2**, 681–692.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Physics letters B* **195**(2), 216–222.

Dungang, L., Regina Y., L. and Xie, M.-G. (2022) Nonparametric fusion learning for multiparameters: Synthesize inferences from diverse sources using Data Depth and Confidence Distribution. *Journal of the American Statistical Association* **117**(540), 2086–2104.

Efron, B. (1979) Computers and the theory of statistics: thinking the unthinkable. *SIAM review* **21**(4), 460–480.

Efron, B. (1993) Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**(1), 3–26.

Efron, B. (2003) Second thoughts on the Bootstrap. *Statistical Science* pp. 135–140.

Ermak, D. L. (1975) A computer simulation of charged particles in solution. technique and equilibrium properties. *The Journal of Chemical Physics* **62**(10), 4189–4196.

Farcomeni, A. and Ventura, L. (2012) An overview of robust methods in medical research. *Statistical Methods in Medical Research* **21**(2), 111–133.

Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for Approximate Bayesian computation: semi-automatic ABC (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Field, C. A. and Ronchetti, E. (1990) Small sample asymptotics.

Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38.

Fisher, R. A. (1936) Design of experiments. *British Medical Journal* **1**(3923), 554.

Fraser, D. A. (1961) The fiducial method and invariance. *Biometrika* **48**(3/4), 261–280.

Fraser, D. A. S. (2004) Structural inference. *Encyclopedia of Statistical Sciences* **13**.

Gallant, A. R., Hong, H., Leung, M. P. and Li, J. (2022) Constrained estimation using penalization and MCMC. *Journal of Econometrics* **228**(1), 85–106.

Gallant, A. R. and Tauchen, G. (1996) Which moments to match? *Econometric theory* **12**(4), 657–681.

Garcia-Angulo, A. C. and Claeskens, G. (2022) Exact uniformly most powerful postselection confidence distributions. *Scandinavian Journal of Statistics* .

Garrett, A. D. (2003) Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine* **22**(5), 741–762.

Genton, M. G. and Ronchetti, E. (2003) Robust indirect inference. *Journal of the American Statistical Association* **98**(461), 67–76.

Gerber, M. and Lee, A. (2020) Discussion on the paper by Jacob, O'Leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**(3), 584–585.

Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. *Technical report, University of Minnesota, School of Statistics* .

Ghosh, A. and Basu, A. (2013) Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression .

Girolami, M. and Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(2), 123–214.

Giummolé, F., Mameli, V., Ruli, E. and Ventura, L. (2019) Objective Bayesian inference with proper scoring rules. *Test* **28**, 728–755.

Glynn, P. W. and Rhee, C.-H. (2014) Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability* **51**(A), 377–389.

Gonçalves, L., de Oliveira, M. R., Pascoal, C. and Pires, A. (2012) Sample size for estimating a binomial proportion: comparison of different methods. *Journal of Applied Statistics* **39**(11), 2453–2473.

Good, P. I. (2004) *Permutation, Parametric and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling Methods for Testing Hypotheses*. Third edition. Springer Series in Statistics. Springer.

Gourieroux, C., Monfort, A. and Renault, E. (1993) Indirect inference. *Journal of applied econometrics* **8**(S1), S85–S118.

Graham, M. and Storkey, A. (2017) Asymptotically exact inference in differentiable generative models. In *Artificial Intelligence and Statistics*, pp. 499–508.

Graham, M. M., Thiery, A. H. and Beskos, A. (2022) Manifold Markov chain Monte Carlo methods for Bayesian inference in diffusion models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(4), 1229–1256.

Grazian, C. and Fan, Y. (2020) A review of approximate Bayesian computation methods via density estimation: Inference for simulator-models. *Wiley Interdisciplinary Reviews: Computational Statistics* **12**(4), e1486.

Greco, L., Racugno, W. and Ventura, L. (2008) Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference* **138**(5), 1258–1270.

Green, P. J., Łatuszyński, K., Pereyra, M. and Robert, C. P. (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing* **25**(4), 835–862.

Hannig, J. (2009) On Generalized Fiducial inference. *Statistica Sinica* pp. 491–544.

Hannig, J., Iyer, H., Lai, R. C. and Lee, T. C. (2016) Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association* **111**(515), 1346–1361.

Hartigan, J. (1964) Invariant prior densities. *Ann.Math.Statist.* **35**, 836–845.

Hartigan, J. (1965) The asymptotically unbiased density. *Ann. Math. Statist.* **36**, 1137–1152.

Hartmann, C. (2008) An ergodic sampling scheme for constrained Hamiltonian systems with applications to molecular dynamics. *Journal of Statistical Physics* **130**, 687–711.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109.

Heng, J. and Jacob, P. E. (2019) Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* **106**(2), 287–302.

Heritier, S. and Ronchetti, E. (1994) Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association* **89**(427), 897–904.

Hjort, N. L. and Schweder, T. (2018) Confidence distributions and related themes.

Ising, E. (1924) *Beitrag zur theorie des ferro-und paramagnetismus*. Ph.D. thesis, Grefe & Tiedemann Hamburg, Germany.

Jacob, P. E., O'Leary, J. and Atchadé, Y. F. (2020) Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society Series B (with discussion)* **82**(3), 543–600.

Johnson, V. E. (1996) Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association* **91**(433), 154–166.

Johnson, V. E. (1998) A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association* **93**(441), 238–248.

Kahn, H. and Marshall, A. W. (1953) Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America* **1**(5), 263–278.

Karabatsos, G. (2021) Automatic tolerance selection for approximate Bayesian computation. *Machine Learning eJournal* .

Kass, R. E., Tierney, L. and Kadane, J. B. (1989) Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**(4), 663–674.

Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017) Median bias reduction of maximum likelihood estimates. *Biometrika* **104**(4), 923–938.

Kent, J. T. (1982) The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **44**(1), 71–80.

Kosmidis, I. (2014) Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics* **6**(3), 185–196.

Kosmidis, I. (2023) *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.9.2.

Kosmidis, I., Kenne Pagui, E. C. and Sartori, N. (2020) Mean and median bias reduction in generalized linear models. *Statistics and Computing* **30**(1), 43–59.

Lee, J. M. (2012) *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer.

Lee, S. M. and Young, G. A. (2005) Parametric bootstrapping with nuisance parameters. *Statistics & probability letters* **71**(2), 143–153.

Legramanti, S., Durante, D. and Alquier, P. (2022) Concentration and robustness of discrepancy-based ABC via Rademacher complexity. *arXiv preprint arXiv:2206.06991* .

Leisen, F., Villa, C. and Walker, S. G. (2020) On a class of objective priors from scoring rules (with discussion). *Bayesian Analysis* **15**(4), 1345–1423.

Lelièvre, T., Rousset, M. and Stoltz, G. (2019) Hybrid Monte Carlo methods for sampling probability measures on submanifolds. *Numerische Mathematik* **143**(2), 379–421.

Lelièvre, T., Stoltz, G. and Zhang, W. (2020) Multiple projection MCMC algorithms on submanifolds. *arXiv preprint arXiv:2003.09402* .

Lewis, J. R., MacEachern, S. N. and Lee, Y. (2021) Bayesian restricted likelihood methods: Conditioning on insufficient statistics in Bayesian regression (with discussion). *Bayesian Analysis* **16**(4), 1393–1462.

Li, W. and Fearnhead, P. (2018) On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika* **105**(2), 285–299.

Lindqvist, B. H., Erlemann, R. and Taraldsen, G. (2022) Conditional Monte Carlo revisited. *Scandinavian Journal of Statistics* **49**(3), 943–968.

Liseo, B. and Loperfido, N. (2006) A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical planning and inference* **136**(2), 373–389.

Liu, D., Liu, R. Y. and Xie, M.-G. (2014) Exact meta-analysis approach for discrete data and its application to 2× 2 tables with rare events. *Journal of the American Statistical Association* **109**(508), 1450–1465.

Liu, J. S. (2008) *MonteCarlo strategies in scientific computing*. Springer Science & Business Media.

Liu, R. Y. (1990) On a notion of Data Depth based on random simplices. *The Annals of Statistics* pp. 405–414.

Liu, Y., Hannig, J. and Murph, A. C. (2022) A Geometric Perspective on Bayesian and Generalized fiducial inference. *arXiv preprint arXiv:2210.05462* .

Ludkin, M. and Sherlock, C. (2023) Hug and hop: a discrete-time, nonreversible Markov chain Monte Carlo algorithm. *Biometrika* **110**(2), 301–318.

Madruga, M. R., Esteves, L. G. and Wechsler, S. (2001) On the Bayesianity of Pereira-Stern tests. *Test* **10**, 291–299.

Madruga, M. R., Pereira, C. d. B. and Stern, J. M. (2003) Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference* **117**(2), 185–198.

Mardia, K. V., Jupp, P. E. and Mardia, K. (2000) *Directional statistics*. Volume 2. Wiley Online Library.

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statistics and Computing* **22**(6), 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**(26), 15324–15328.

Martin, G. M., Frazier, D. T. and Robert, C. P. (2023) Computing Bayes: From then 'til now. *Statistical Science* **1**(1), 1–17.

Masserano, L., Dorigo, T., Izbicki, R., Kuusela, M. and Lee, A. B. (2022) Simulation-based inference with Waldo: Perfectly calibrated confidence regions using any prediction or posterior estimation algorithm. *arXiv preprint arXiv:2205.15680* .

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**(6), 1087–1092.

Miller, J. W. (2021) Asymptotic normality, concentration, and coverage of generalized posteriors. *The Journal of Machine Learning Research* **22**(1), 7598–7650.

Mogensen, P. and Riseth, A. (2018) Optim: A mathematical optimization package for julia. *Journal of Open Source Software* **3**(24).

Müller, A. (1997) Integral probability metrics and their generating classes of functions. *Advances in applied probability* **29**(2), 429–443.

Nayak, S. M., Bari, B. A., Yaden, D. B., Spriggs, M. J., Rosas, F. E., Peill, J. M., Giribaldi, B., Erritzoe, D., Nutt, D. J. and Carhart-Harris, R. (2023) A Bayesian reanalysis of a trial of psilocybin versus escitalopram for depression. *Psychedelic Medicine* **1**(1), 18–26.

Neal, R. M. (1993) Bayesian learning via stochastic dynamics. *Advances in neural information processing systems* pp. 475–475.

Neal, R. M. (1999) Circularly-coupled Markov chain sampling. Technical report, Department of Statistics, University of Toronto.

Neal, R. M. (2003) Slice sampling. *The annals of statistics* **31**(3), 705–767.

Neal, R. M. (2011) MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo* **2**(11).

Papp, T. P. and Sherlock, C. (2022) A new and asymptotically optimally contracting coupling for the random walk Metropolis. *arXiv preprint arXiv:2211.12585* .

Pereira, C. A. and Stern, J. M. (2001) Model selection: full Bayesian approach. *Environmetrics: The official journal of the International Environmetrics Society* **12**(6), 559–568.

Pereira, C. A. d. B. and Stern, J. M. (1999) Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy* **1**(4), 99–110.

Pereira, C. A. d. B. and Stern, J. M. (2022) The e-value: a fully Bayesian significance measure for precise statistical hypotheses and its research program. *São Paulo Journal of Mathematical Sciences* **16**(1), 566–584.

Pereira, C. A. d. B., Stern, J. M. and Wechsler, S. (2008) Can a significance test be genuinely Bayesian? .

Pesarin, F. and Salmaso, L. (2010) The permutation testing approach: a review. *Statistica* **70**(4), 481–509.

Pitman, E. J. (1937) Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society* **4**(1), 119–130.

Price, L. F., Drovandi, C. C., Lee, A. and Nott, D. J. (2018) Bayesian Synthetic likelihood. *Journal of Computational and Graphical Statistics* **27**(1), 1–11.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (2003) Asymptotics and the theory of inference. *The Annals of Statistics* **31**(6), 1695–2095.

Reid, N. and Sun, Y. (2010) Assessing sensitivity to priors using higher order approximations. *Communications in Statistics - Theory and Methods* **39**(8-9), 1373–1386.

Ricker, W. E. (1954) Stock and recruitment. *Journal of the Fisheries Board of Canada* **11**(5), 559–623.

Robert, C. (1994) *The Bayesian choice*. Springer-Verlag, New York.

Robert, C. P. (2014) On the Jeffreys-Lindley paradox. *Philosophy of Science* **81**(2), 216–232.

Robert, C. P. *et al.* (2007) *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Volume 2. Springer.

Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. First edition. Springer-Verlag, New York.

Roberts, G. O. and Rosenthal, J. S. (2004) General state space Markov chains and MCMC algorithms. *Probability Surveys* **1**, 20–71.

Roberts, G. O. and Tweedie, R. L. (1996) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* pp. 341–363.

Ronchetti, E. and Ventura, L. (2001) Between stability and higher-order asymptotics. *Statistics and Computing* **11**, 67–73.

Ronchetti, E. M. and Huber, P. J. (2009) *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA.

Rosenthal, J. S. (1997) Faithful couplings of Markov chains: now equals forever. *Advances in Applied Mathematics* **18**(3), 372 – 381.

Rosenthal, J. S. *et al.* (2011) Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain MonteCarlo* **4**(10.1201).

Rothmann, M. D., Wiens, B. L. and Chan, I. S. (2011) *Design and analysis of non-inferiority trials*. CRC press.

Rousset, M., Stoltz, G. and Lelievre, T. (2010) *Free energy computations: a mathematical perspective*. World Scientific.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* pp. 1151–1172.

Rubio, F. and Johansen, A. M. (2013) A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics* **7**, 1632–1654.

Ruli, E., Sartori, N. and Ventura, L. (2014) Marginal Posterior Simulation via Higher-order Tail Area Approximations. *Bayesian Analysis* **9**(1), 129 – 146.

Ruli, E., Sartori, N. and Ventura, L. (2016) Approximate Bayesian computation with composite score functions. *Statistics and Computing* **26**(3), 679–692.

Ruli, E., Sartori, N. and Ventura, L. (2020) Robust approximate Bayesian inference. *Journal of Statistical Planning and Inference* **205**, 10–22.

Ruli, E. and Ventura, L. (2021) Can Bayesian, confidence distribution and frequentist inference agree? *Statistical Methods & Applications* **30**, 359–373.

Ruli, E., Ventura, L. and Musio, M. (2022) Robust confidence distributions from proper scoring rules. *Statistics* **56**(2), 455–478.

Sartori, N. (2006) Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference* **136**(12), 4259–4275.

Schweder, T. and Hjort, N. L. (2016) *Confidence, likelihood, probability*. Volume 41. Cambridge University Press.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2022) Bayes and big data: The consensus Monte Carlo algorithm. In *Big Data and Information Theory*, pp. 8–18. Routledge.

Severini, T. A. (1999) On the relationship between bayesian and non-bayesian elimination of nuisance parameters. *Statistica Sinica* pp. 713–724.

Severini, T. A. (2000) *Likelihood methods in statistics*. Oxford University Press.

Severini, T. A. and Wong, W. H. (1992) Profile likelihood and conditionally parametric models. *The Annals of statistics* pp. 1768–1802.

Simola, U., Cisewski-Kehe, J., Gutmann, M. U. and Corander, J. (2021) Adaptive approximate Bayesian computation tolerance selection. *Bayesian analysis* **16**(2), 397–423.

Simon, L. (2014) *Introduction to geometric measure theory*. Second edition.

Singh, K., Xie, M.-G. and Strawderman, W. E. (2005) Combining information from independent sources through Confidence Distributions .

Sisson, S. A., Fan, Y., Beaumont, M. and Altri (2018) *Handbook of Approximate Bayesian Computation*. CRC Press.

Smith Jr, A. A. (1993) Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* **8**(S1), S63–S84.

Soubeyrand, S. and Haon-Lasportes, E. (2015) Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC. *Statistics & Probability Letters* **107**, 84–92.

Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016) Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**(5), 702–712.

Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescence times from dna sequence data. *Genetics* **145**(2), 505–518.

Thornton, S., Li, W. and Xie, M.-G. (2022) Approximate confidence distribution computing. *arXiv preprint arXiv:2206.01707* .

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1786.

Tierney, L. (1998) A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability* pp. 1–9.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6**(31), 187.

Tsallis, C. (1988) Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics* **52**, 479–487.

Varin, C., Reid, N. M. and Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica* **21**(1), 5–42.

Ventura, L. and Racugno, W. (2016) Pseudo-likelihoods for Bayesian inference. In *Topics on methodological and applied statistical inference*, pp. 205–220.

Ventura, L. and Reid, N. (2014) Approximate Bayesian computation with modified log-likelihood ratios. *Metron* **72**, 231–245.

Ventura, L., Sartori, N. and Racugno, W. (2013) Objective Bayesian higher-order asymptotics in models with nuisance parameters. *Computational Statistics Data Analysis* **60**, 90–96.

Verdinelli, I. and Wasserman, L. (1991) Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing* **1**, 105–117.

G. de Vries, T. Hillen, M. L. J. M. B. S. (2006) *A course in mathematical biology : Quantitative modeling with mathematical and computational methods*. Siam edition.

Wang, G., O'Leary, J. and Jacob, P. (2021) Maximal couplings of the Metropolis–Hastings algorithm. In *International Conference on Artificial Intelligence and Statistics*, pp. 1225–1233.

Wang, P., Xie, M.-G. and Zhang, L. (2022a) Finite-and large-sample inference for model and coefficients in high-dimensional linear regression with repro samples. *arXiv preprint arXiv:2209.09299* .

Wang, X. and Dunson, D. B. (2013) Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605* .

Wang, Y., Kaji, T. and Rockova, V. (2022b) Approximate Bayesian Computation via Classification. *J. Mach. Learn. Res.* **23**(1).

Weller, G. B. and Eddy, W. F. (2015) Multivariate order statistics: Theory and application. *Annual Review of Statistics and Its Application* **2**, 237–257.

Wood, S. N. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**(7310), 1102–1104.

Xie, M.-G. and Singh, K. (2013) Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review* **81**(1), 3–39.

Xie, M.-G. and Wang, P. (2022) Repro samples method for finite- and large-sample inferences. *arXiv preprint arXiv:2206.06421* .

Xu, K., Fjelde, T. E., Sutton, C. and Ge, H. (2021) Couplings for Multinomial Hamiltonian Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, pp. 3646–3654.

Zappa, E., Holmes-Cerfon, M. and Goodman, J. (2018) Monte Carlo on manifolds: sampling densities and integrating functions. *Communications on Pure and Applied Mathematics* **71**(12), 2609–2647.