Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN STATISTICA
CICLO XXXIII

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof. Paolo Salani

Approximate Bayesian Computation and Statistical Applications to Anonymized Data: an Information Theoretic Perspective

Settore Scientifico Disciplinare SECS-S/01

Dottoranda:
Cecilia Viscardi

Tutori
Prof. Fabio Corradi

Prof. Michele Boreale

Coordinatore
Prof. Paolo Salani

Anni 2017/2020

A Nonno Lino.

ABSTRACT

Realistic statistical modelling of complex phenomena often leads to considering several latent variables and nuisance parameters. In such cases, the Bayesian approach to inference requires the computation of challenging integrals or summations over high dimensional spaces. Monte Carlo methods are a class of widely used algorithms for performing simulated inference. In this thesis, we consider the problem of *sample degeneracy* in Monte Carlo methods focusing on Approximate Bayesian Computation (ABC), a class of likelihood-free algorithms allowing inference when the likelihood function is analytically intractable or computationally demanding to evaluate. In the ABC framework sample degeneracy arises when proposed values of the parameters, once given as input to the generative model, rarely lead to simulations resembling the observed data and are hence discarded. Such "poor" parameter proposals, i.e., parameter values having an (exponentially) small probability of producing simulation outcomes close to the observed data, do not contribute at all to the representation of the parameter's posterior distribution. This leads to a very large number of required simulations and/or a waste of computational resources, as well as to distortions in the computed posterior distribution. To mitigate this problem, we propose two algorithms, referred to as the Large Deviations Approximate Bayesian Computation algorithms (LD-ABC), where the ABC typical rejection step is avoided altogether. We adopt an information theoretic perspective resorting to the Method of Types formulation of Large Deviations, thus first restricting our attention to models for i.i.d. discrete random variables and then extending the method to parametric finite state Markov chains. We experimentally evaluate our method through proof-of-concept implementations.

Furthermore, we consider statistical applications to anonymized data. We adopt the point of view of an evaluator interested in publishing data about individuals in an ananonymized form that allows balancing the learner's utility against the risk posed by an attacker, potentially targeting individuals in the dataset. Accordingly, we present a unified Bayesian model applying to data anonymized employing group-based schemes and a related MCMC method to learn the population parameters. This allows relative threat analysis, i.e., an analysis of the risk for any individual in the dataset to be linked to a specific sensitive value beyond what is implied for the general population. Finally, we show the performance of the ABC methods in this setting and test LD-ABC at work on a real-world obfuscated dataset.

CONTENTS

I MONTE CARLO STRATEGIES FOR BAYESIAN SIMULATED INFERENCE 19

1 INTRODUCTION 21

2 BAYESIAN COMPUTATION WITH LIKELIHOOD EVALUATION 25
- 2.1 Rejection Sampling . . . 25
  - 2.1.1 Sampling from unormalized posterior distributions . . . 27
- 2.2 Importance Sampling . . . 28
  - 2.2.1 Effective Sample Size and Sample Degeneracy . . . 31
  - 2.2.2 Comparing RS and IS . . . 35
  - 2.2.3 Sampling from the posterior distribution via IS . . . 36
  - 2.2.4 Sampling from the tails of the distribution . . . 39
- 2.3 Markov Chain Monte Carlo . . . 40
  - 2.3.1 Metropolis Algorithm . . . 41
  - 2.3.2 Metropolis-Hastings Algorithm . . . 42
  - 2.3.3 Gibbs sampler . . . 44
  - 2.3.4 Metropolis within Gibbs . . . 47
  - 2.3.5 Convergence Diagnostics . . . 48
  - 2.3.6 MCMC for sampling from the posterior distribution . . . 51
- 2.4 Asymptotic approximations for Bayesian inference . . . 52
  - 2.4.1 First order approximation . . . 52
  - 2.4.2 Higher-order asymptotic approximations . . . 54
  - 2.4.3 HOTA sampling scheme . . . 55

3 APPROXIMATE BAYESIAN COMPUTATION 57
- 3.1 Summary statistics an tolerance threshold . . . 60
- 3.2 Some ABC sampling schemes . . . 62
  - 3.2.1 Rejection Sampling ABC . . . 62
  - 3.2.2 Importance Sampling ABC . . . 63
  - 3.2.3 Comparing Rejection Sampling and Importance Sampling ABC . . . 64
  - 3.2.4 Markov Chain Monte Carlo ABC . . . 66
- 3.3 Sample Degeneracy in ABC . . . 67

II IMPROVING ABC VIA LARGE DEVIATIONS THEORY 69

4 INTRODUCTION 71
- 4.1 Related work . . . 71

5 IMPROVING ABC VIA SANOV'S THEOREM 75
- 5.1 Large Deviations Theory via Method of Types . . . 75
- 5.2 Large Deviations Theory in ABC . . . 77
- 5.3 Large Deviations Approximate Bayesian Computation (LD-ABC) . . . 79
  - 5.3.1 Importance Sampling LD-ABC . . . 81
  - 5.3.2 Metropolis-Hastings LD-ABC . . . 83
- 5.4 A computational issue: the minimization of the KL-divergence . . . 84
- 5.5 Experiments . . . 85
  - 5.5.1 Example 1: Binomial distribution. . . . 86

5.5.2 Example 2: Mixture of binomial distributions . . . 89
5.5.3 Concluding remarks on the choice of the tolerance threshold . . 95
6 LD-ABC FOR FINITE STATE MARKOV CHAINS 99
6.1 Finite State Markov Chains and Doublet Probability Distributions: Set up and notation . . . 99
6.1.1 Entropy, Relative Entropy and Conditional Entropy . . . 100
6.2 The Method of Types for Markov chains . . . 103
6.2.1 Second order types as summary statistics . . . 104
6.3 ABC and Large Deviations for Markov Chains . . . 105
6.4 Experiments . . . 106
6.4.1 Example 1 . . . 109
6.4.2 Example 2 . . . 111

III SIMULATED INFERENCE FOR LEARNING FROM ANONYMIZED DATA 115
7 INTRODUCTION 117
8 RELATIVE PRIVACY THREATS AND LEARNING FROM ANONYMIZED DATA 119
8.1 Related works . . . 121
8.2 Group based anonymization schemes . . . 124
8.3 A unified probabilistic model . . . 125
8.3.1 Random variables . . . 126
8.3.2 Learner and attacker knowledge . . . 127
8.4 Measures of privacy threat and utility . . . 131
8.5 Learning from the obfuscated table by MCMC . . . 134
8.5.1 Definition and convergence of the Gibbs sampler . . . 135
8.5.2 Sampling from the full conditionals . . . 137
8.6 Experiments . . . 139
8.6.1 Horizontal schemes: k-anonymity . . . 139
8.6.2 Vertical schemes: Anatomy . . . 141
8.6.3 Discussion . . . 142
8.6.4 Assessing MCMC convergence . . . 142
8.7 Conclusions . . . 143
9 APPROXIMATE BAYESIAN INFERENCE FROM ANONYMIZED DATA: ABC VS LD-ABC 147
9.1 Learning from obfuscated data via ABC . . . 148
9.2 Comparing ABC and LD-ABC . . . 148

IV CONCLUSIONS, DISCUSSION AND FUTURE RESEARCH 151

Appendix 155
A INEQUALITIES AND CONVERGENCE 157
A.1 Laws of large numbers . . . 157
A.2 Central limit theorem . . . 157
A.3 Inequalities . . . 157
B APPROXIMATE METHODS 159
B.0.1 Expectation and Variance of a Ratio . . . 160
C MARKOV CHAINS 161
C.1 Definitions and main properties . . . 161

C.2 Classification of states . . . 162
C.3 The stationary distribution . . . 165
C.4 Some useful concepts about continuous state Markov chains . . . 166
D APPENDIX TO PART II 167
D.1 Theorems and proofs in Chapter 5 . . . 167
D.2 Theorems and proofs in Chapter 6 . . . 171
D.3 Additional results from the experiments . . . 174
D.3.1 Example 2: mixture of binomial distributions . . . 174
E APPENDIX TO PART III 177
E.1 Proofs . . . 177
E.2 An alternative group sampling method for vertical schemes . . . 177
E.3 Additional results from experiments . . . 178
BIBLIOGRAPHY 183

Figure 1 Right truncated exponential distribution (blue line) and standard normal distribution (red line). 40
Figure 2 Acceptance region, $\mathcal{B}_\epsilon$, types, T_{x^n} and T_{y^m}, and the probability distribution P_θ that generated y^m. Asymptotically (as $m \to \infty$) T_{y^m} converges to P_θ and the distance $D(\mathcal{B}_\epsilon \| T_{y^m})$ (red) converges to $D(\mathcal{B}_\epsilon \| P_\theta)$ (green). 79
Figure 3 Posterior distributions (LHS) and posterior cumulative density functions (RHS) provided employing IS and MCMC schemes with the uniform kernel (red lines) and with the LD kernel (blue lines). 87
Figure 4 Posterior cumulative density functions with 99% credible intervals derived from 100 reruns of IS (RHS) and MCMC algorithms (LHS) with the uniform kernel (red) and the LD kernel (blue). 87
Figure 5 Boxplots of the distributions of the sojourn time in the LD-MCMC-ABC and MCMC-ABC with four different thresholds: 0.55 (top-left), 0.6 (top-right), 0.75 (bottom-left), 0.8 (bottom-right). 88
Figure 6 Posterior distributions of θ approximated by LD-ABC (blue lines) and R-ABC (red lines). The black dashed lines represent the true $\mathtt{Beta}$ posterior distributions. Each panel corresponds to a different a value of N (3,4,5,6 and 6). 88
Figure 7 DAG representation of the finite mixture of binomials distributions. 91
Figure 8 Posterior distributions corresponding to four different pairs of tuning parameters (m, ϵ). Each panel refers to one of the three model parameters. Red lines represent the posterior density estimates provided via R-ABC. The blue lines represent the estimates provided via LD-ABC. The dashed black lines are the output of the Gibbs sampler. The gray dashed lines are the ratios $\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{x^n}) / \tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{x^n})$ providing a representation of the adjustment $\alpha_{\epsilon,m}$. 92
Figure 9 Posterior cumulative density functions for θ_2. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cumulative density function for a pair (m, ϵ). For $\theta_2 < 0.5$ both the cumulative density functions equal 0. 90% intervals over 100 reruns of each algorithm are also shown. 94

Figure 10 ESS of the two algorithms vs. the number of iterations. The dotted red line represents the $\widehat{\text{ESS}}$ achieved by R-ABC after $S = 100,000$ iterations. The gray dashed lines indicates the number of iterations (14,985) needed by the LD-ABC to get the same ESS as the R-ABC. 96
Figure 11 Posterior density functions of the three parameters of the Binomial mixture model in Section 5.5.2. The true posterior distributions are represented by the black dashed lines. The continuous red lines represent the posterior density estimates provided via R-ABC and the blue ones via LD-ABC. The dashed blue lines represent the density estimates achieved by the LD-ABC after 14,985 iterations. 97
Figure 12 Posterior distributions corresponding to $m = 120$ and $\epsilon = 0.005$. Each plot refers to one of the four parameters of the model. Red lines represent the posterior density estimates provided via R-ABC. The blue lines represent the estimates provided via LD-ABC. The dotted red lines are the estimates provided by the R-ABC using the first order type. The dashed gray lines are the true posterior distributions. 110
Figure 13 Posterior cumulative density functions for $\theta_1, \theta_2, \theta_3$ and λ. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cdf. 99% intervals over 100 rerun of each algorithm are also represented. 111
Figure 14 Observed time series $\mathbf{x}^{120} = x_1, ..., x_{120}$. 112
Figure 15 Approximate posterior distributions of each parameter with $S = 1,000,000$, $\epsilon = 0.01$ (solid lines) and $\epsilon = 0.05$ (dashed lines). 113
Figure 16 Sampling from $f(g|\theta, \mathbf{x}_{-i}, \mathbf{x}^*)$ ($g \in \mathcal{G}_i$) for horizontal schemes, across all the groups. 138
Figure 17 Results for k-anonymity. Top ($\ell = k = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = k = 4$. The skewness value (γ) represents the third standardized moment of the empirical distribution. Dark red areas show where the attacker performs better than the learner. 144
Figure 18 Results for Anatomy. Top ($\ell = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = 4$. The skewness value (γ) represents the third standardized moment of the empirical distribution. Dark red areas show where the attacker performs better than the learner. 145
Figure 19 Representation of the $r + h + s$ steps path between i and j. . . . 164
Figure 20 Representation of two transient states $(i, j) \in \mathcal{C}$. 164

Figure 21 Posterior cumulative density functions for θ_2. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cumulative density function for a pair (m, ϵ). For $\theta_1 < 0.5$ both the cumulative density functions equal to 0. The 90% intervals over 100 rerun of each algorithm are also shown. 174
Figure 22 Posterior cumulative density functions for λ. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cumulative density function for a pair (m, ϵ). For $\lambda > 0.5$ both the cumulative density functions equal 1. The 90% intervals over 100 rerun of each algorithm are also shown. 175
Figure 23 Sampling from $\psi(g|\theta, x^*)$ for vertical schemes. 178

LIST OF TABLES

Table 1 Summaries of the empirical distributions of the relative errors of the approximate distances. 85
Table 2 Squared errors, integrated squared errors and ESS averaged over 100 reruns, with $\epsilon = 0.01$, $m = 100$. 86
Table 3 $\widehat{ESS}$'s, running times and tolerance thresholds for each algorithm and dataset. 89
Table 4 Details for the simulation of the data-set and for the Gibbs implementation. 90
Table 5 Posterior estimates derived via Gibbs Sampling 90
Table 6 Squared errors and integrated squared errors averaged over 100 reruns. Each column contains results for one of the model parameters both for LD-ABC and R-ABC. 93
Table 7 $\widehat{ESS}$ and normalized perplexity averaged over 100 reruns for each pair of tuning parameters. 95
Table 8 Squared errors and integrated squared errors averaged over 100 reruns. Each column contains results for one of the parameters of the model both for LD-ABC and R-ABC. 110
Table 9 $\widehat{ESS}$ achieved by R-ABC and LD-ABC after $S = 1,000,000$ iterations with $\epsilon = 0.01$ and $\epsilon = 0.05$ 112
Table 10 A table (top) anonymized according to 2-anonymity via local recoding (bottom-left) and Anatomy (bottom-right). 120
Table 11 Summary of notation. 126
Table 12 Posterior distributions of diseases for a victim with $r_v = (M, 45501)$, for the anonymized x^* in Table 10(b). NB: figures affected by rounding errors. 129
Table 13 Summary of threat and faithfulness measures for anonymization according to k-anonymity and ℓ- diversity. 140
Table 14 Summary of threat and faithfulness measures for anonymization according to Anatomy. 141
Table 15 The leftmost table shows the Squared errors integrated over the 3-simplex and averaged over 100 reruns of ABC. Each column corresponds to an element of $\{\theta_{R|s} : s \in \{\text{Governament, Self-employed, Private, Without-pay}\}\}$. The rightmost table shows the Effective Sample Sizes acheved by R-ABC and LD-ABC averaged over 100 reruns. 149
Table 16 Squared errors averaged over 100 reruns of ABC. Each column corresponds to an element of $\{\theta_{R|s} : s \in \{\text{Governament, Self-employed, Private, Without-pay}\}\}$. 150
Table 17 Conditions for the existence and uniqueness of the stationary distribution. 165

Table 18 Posterior means via MCMC. Each column corresponds to the vector of posterior means for an element of $\left\{\theta_{R|s} : s \in \{\text{Governament, Self-employed, Private, Without-pay}\}\right\}$. 179

ACRONYMS

a.k.a	Also Known As
a.s.	Almost Surely
cdf	Cumulative Density function
gcd	Greatest Common Divisor
i.i.d	Independent and Identically Distributed
iff	If and only If
i.o.	Infinitely Often
w.r.t	With Respect To
pdf	Probability Density Function
pmf	Probability Mass function

OVERVIEW

Statistical inference is the area of science aimed at drawing conclusions about phenomena of interest through quantitative measures derived from observed data. Following the dictionary definition, the word *inference* means "a conclusion reached on the basis of *evidence* and/or *premises* ". In particular the expression *Bayesian Inference*, quoting Donald B. Rubin [129], refers to

> the method of statistical inference that draws conclusions by calculating conditional distributions of unknown quantities given a) known quantities and b) model specifications. Thus, in Bayesian Inference, known quantities are treated as observed values of random variables and unknown quantities are treated as unobserved random variables; the conditional distribution of unknowns given knowns follows from applying Bayes's theorem to the model specifying the joint distribution of known and unknown quantities.

According to the Bayesian paradigm the observed data represent the *evidence* and the *premises* are incorporated into a joint model specified through a *prior* distribution and a *likelihood* function. Usually the *parameters* are the unknown quantities and the prior distribution incorporates the prior knowledge about them. The likelihood function results from the probabilistic model assumed for the observable random variables and espresses how likely are the observed data given certain values of the parameters. The conclusions reached about the unknown quantities are represented by the *posterior distribution*, which is the conditional distribution of the unknown given the known quantities derived through the Bayes's Theorem. Unfortunately, only in few cases this derivation is straightforward. In particular, the analytical computation of the posterior distribution is feasible in the conjugate case [120], i.e., when the prior and the posterior distributions are in the same family of probability distributions. In many other cases the posterior computation requires approximation methods. Thus, the main challenges in the Bayesian framework are i) the formulation of a joint model capturing the key features of the scientific problem of interest; ii) the computation required to derive posterior quantities. Both the issues are in some sense related to the inclusion into the joint model of auxiliary random variables. Regarding the aspect i), one needs to specify the prior distribution and to model the relationship between the data and all the unknown quantities. Prior distributions can be elicited in a subjective manner or selected according to formal rules [71]. Modelling the structure of the data in a realistic and comprehensive manner often requires to involve other unknown quantities: nuisance parameters and latent variables that we are not directly interested in. Regarding ii), the computations required to perform Bayesian inference are mostly complex integrals (or summations in the discrete setting) over all the unknown quantities to derive the normalizing constant of the posterior distribution, the *marginal likelihood*. In such cases Monte Carlo techniques are well-known tools to approximate complex integrals via stochastic simulations. When the joint model is specified by introducing nuisance parameters or latent variables, further integrations are needed to derive the posterior distribution.

In the case of complex structure of the data, the introduction of latent variables may provide a way of modelling comprehensively the phenomena of interest. However, in such a case, the evaluation of the likelihood function may in turn involve high-dimensional integrals becoming analytically infeasible or computationally prohibitive. In such cases, an intuitive way of performing Bayesian inference is represented by the so called *likelihood-free* methods. The core of this thesis is a class of Monte Carlo likelihood-free methods known as Approximate Bayesian Computation (ABC). This class of algorithms dispenses from the evaluation of the likelihood function by resorting to comparisons between the observed data and the pseudo-data simulated from a generative model.

An interesting aspect is represented by the twofold role of the auxiliary random variables. On one hand they overcomplicate the joint model requiring the implementation of Monte Carlo strategies. On the other hand, ABC, as well as most of the Monte Carlo methods, relies on the simulation of instrumental random variables to enable complex computations.

MAIN CONTRIBUTIONS OF THE THESIS In the first part of this thesis we review some of the most important algorithms allowing simulated inference, starting from methods requiring the likelihood evaluation to conclude with ABC methods. During this excursus we provide formal comparisons among several Monte Carlo algorithms highlighting that most of them suffer from the sample degeneracy problem. In particular, we note that it becomes more serious in the ABC framework.

The original contributions of the thesis can be summarized as follows:

- The thesis presents a novel ABC method to mitigate the sample degeneracy problem. It is developed relaying on the "Method of Types" formulation of Large Deviations Theory and applies to discrete i.i.d. data. Two alternative ABC sampling schemes are presented and the performances are illustrated through several examples. Furthermore, formal guarantees of their convergence are provided.[1]

- The presented method is extended to finite state Markov chains by deepening the Method of Types and by considering more general results in Large Deviations Theory.[2]

- The last part of the thesis introduces a Bayesian probabilistic model and a related MCMC method for learning from anonymized. The method applies to data anonymized employing group-based anonymization schemes. Measures of (relative) privacy threats and utility deriving from publishing data in an obfuscated form are defined. The method is tested at work on real-world data.[3]

1 Part of the work presented in Chapter 5 has been accepted for publication in *Computational Statistics* as "Weighted Approximate Bayesian Computation via Sanov's Theorem". A brief synthesis of the chapter was published in the conference proceedings *Book of short papers SIS 2020* as "Improving ABC via large deviations theory".

2 A brief synthesis of Chapter 6 has been submitted for publication to the conference proceedings *Book of short papers SIS 2021* as "Inference on Markov chains parameters via Large Deviations ABC".

3 The work presented in Chapter 8 © 2020 IEEE was reprinted, with permission, from Michele Boreale, Fabio Corradi and Cecilia Viscardi, *Relative Privacy Threats and Learning From Anonymized Data*, IEEE Transactions on Information Forensics and Security, 2020.

- The use of ABC for inferring population parameters from obfuscated data is discussed. Finally, the novel ABC method is tested at work on an anonymized dataset.

Part I

MONTE CARLO STRATEGIES FOR BAYESIAN SIMULATED INFERENCE

"When one admits that nothing is certain one must, I think, also add that some things are more nearly certain than others. The longing for certainty... is in every human mind. But certainty is generally illusion. Doubt is not a pleasant condition but certainty is an absurd one. The only unchangeable certainty is that nothing is certain or unchangeable."

— Bertrand Russell

1 INTRODUCTION

Let $\mathbf{x} = x_1, ..., x_n$ be n observations drawn from the sample space $\mathcal{X}$ and assumed to be realizations of random variables $\mathbf{X} = X_1, ..., X_n$. Defining a *statistical model* on the sample space $\mathcal{X}$ corresponds to assuming a family of probability distributions indexed by the parameter θ which takes values on the *parameter space* Θ

$$\mathcal{F} \stackrel{\triangle}{=} \{p(\cdot|\theta) : \theta \in \Theta\}.$$

Throughout this thesis, we let $p(\cdot)$ and $p(\cdot|\cdot)$ denote respectively the probability density functions (pdf) and the conditional probability density functions w.r.t. suitable measures.

In Bayesian statistics the parameter θ is in turn modelled as a random variable and the assumed *prior distribution* on Θ is denoted by $\pi(\cdot)$. Thus, given the observed data, all the premises to the inference are incorporated into the assumed joint model

$$\pi(\theta) \cdot p(x|\theta).$$

The mathematical object of interest for the inference is the *posterior distribution* derived through Bayes's formula:

$$\pi(\theta|x) = \frac{\pi(\theta)p(x|\theta)}{\int_{\Theta} \pi(\theta)p(x|\theta)d\theta} \tag{1}$$

where $\int_{\Theta} \pi(\theta)p(x|\theta)d\theta = p(x)$ is the *marginal likelihood*.

In many applications the analytical evaluation of such integral is infeasible and requires a numerical approximation. Furthermore, the assumed family of probability distributions, $\mathcal{F}$, can be indexed by a vector of parameters θ. However, one can be interested in deriving the posterior distribution of only one component, say $\theta_1 \in \Theta_1$. In such cases the other components, called *nuisance parameters*, must be integrated out to derive the posterior distribution

$$\pi(\theta_1|x) = \frac{\pi(\theta_1)p(x|\theta_1)}{p(x)} = \frac{\int_{\Theta_{\setminus 1}} \pi(\theta)p(x|\theta_1)d\theta_{\setminus 1}}{\int_{\Theta_1}\int_{\Theta_{\setminus 1}} \pi(\theta)p(x|\theta)d\theta_{\setminus 1}d\theta_1}, \tag{2}$$

where $\theta_{\setminus 1}$ denotes all but the first parameter and $\Theta_{\setminus 1}$ denotes their joint space.

Other high-dimensional integrals are often required when is adopted a *missing data approach*. It represents a useful tool for modelling complex data structures in a realistic and comprhensive manner [85]. The general framework was developed to deal with problems in which the missingness of the data is not at random, meaning that the "missing mechanism" is not ignorable since it tells us something about the quantities we are interested in estimating [128]. In order to take into account various aspects of complex phenomena, many other difficulties can be treated as missing data problems. Examples are model involving latent variables, nuisance parameters and *latent*

class models, such as models for mixture data. In all these cases marginalizations as in (2) are needed. Further complications arise when the evaluation of the likelihood function involves in turn the computation of complex integrals making it *intractable*, meaning that its evaluation is infeasible or computationally demanding. Generally speaking, many methods for the fomputation of poterior quantities have been proposed. Here, we focus on Monte Carlo methods which represent a powerful tool for solving many inferential problems. In particular, the core of this thesis are the ABC methods which, when one is able to get samples from the assumed model, allow simulated posterior inference even when the evaluation of the likelihood is infeasible. However, besides Monte Carlo methods, other approximation methods are available. Among them we briefly review *asymptotic expansions* methods which probably represent the oldest solution to the discussed difficulties [82].

MONTE CARLO METHODS Monte Carlo (MC) techniques are a class of methods aimed at solving problems of optimization and inferential problems by means of *stochastic simulations*. The expression *stochastic simulations* refers to the execution of systems involving one or more random components in an environment controlled by the experimenter. The environment could be a smaller scale reproduction of the system (e.g. repeated tosses of a coin, repeated random drawing from an urn) or a computer. The original idea came from Enrico Fermi, the Italian physicist who first experimented it studying neutron diffusion in the 1930s. The formalization of the MC method as we know it, can be traced back to the 1940s, at the times when the construction of the first electronic computer, the ENIAC, had just been completed. It was formalized by Stanislaw Ulam, a Polish mathematician, and was implemented with John Von Neumann. The method was finally tested by Nicholas Metropolis who coined the name *Monte Carlo* [97].

Let us assume that $h(\cdot)$ is an integrable function of θ and that we are interested in evaluating the following integral

$$I = \mathbb{E}_\pi[h(\theta)] = \int_\Theta h(\theta)\pi(\theta|x)\,d\theta. \tag{3}$$

That integral corresponds to the expected value of $h(\theta)$ w.r.t. the posterior distribution, denoted as $\mathbb{E}_\pi$. Note that when the $h(\cdot)$ is the identity function, I represents the mean of the posterior distribution. Basically, MC methods approximate integrals by sample averages of observations obtained via stochastic simulations. Hence, a MC approximation of that integral in (3) can be got by drawing S random samples, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(S)}$, independent and identically distributed (i.i.d.) from $\pi(\cdot|x)$ and by computing

$$\hat{I} = \frac{1}{S}\sum_{s=1}^{s} h(\theta^{(s)}).$$

The *law of large numbers* (see A.1, Th.12) states that, as S goes to infinity, the approximation $\hat{I}$ converges to I almost surely. The rate at which this convergence occurs can be assessed by the *central limit theorem* (see A.2, Th.13) stating that

$$\sqrt{S}(\hat{I} - I) \xrightarrow{d} N(0, \sigma^2)$$

where $\xrightarrow{d}$ denotes convergence in distribution and $N(0, \sigma^2)$ denotes a Normal distribution with mean equal to zero and variance σ^2. It follows that MC approximations are based on an extensive use of simulations of random variables. Unfortunately, we are often not able to sample directly from the target distribution $\pi(\theta|x)$. In such cases we can implement other MC methods such as the Rejection Sampling, Importance Sampling or Markov Chain Monte Carlo techniques. All these methods require the ability of evaluating point-wise the unnormalized posterior distribution, thus avoiding the computation of the unavailable normalizing constant. As already pointed out, in many statistical applications the likelihood function is in turn intractable thus inhibiting also the evaluation of the unnormalized posterior distribution. In such cases ABC methods might be a solution. Such algorithms are based on the above-mentioned MC sampling schemes but requires the ability of simulating synthetic data from a simulator reproducing the stochastic process underlying the assumed probabilistic model.

This part of the thesis is organized as follows. In Chapter 2 we deal with approximation methods requiring the ability of evaluating point-wise the likelihood function. In particular, we introduce the Rejection Sampling, the Importance Sampling, the Markov Chain Monte Carlo methods and some comparisons among them. Furthermore, we detail how to perform Bayesian inference resorting to these MC methods and briefly review Bayesian approximation methods based on asymptotic expansions. In Chapter 3 we introduce ABC. Specifically, we consider the likelihood-free version of the algorithms dealt with in Chapter 2. Moreover, we emphasize the problem of sample degeneracy which will be the focus of Part II.

2

BAYESIAN COMPUTATION WITH LIKELIHOOD EVALUATION

In this chapter we introduce basic methods for the computations required to derive posterior quantities when the likelihood function is available. We focus on basic MC sampling schemes. The content of this chapter is mainly based on [85, 86, 125], with a focus on the the derivation of two Effective Sample Size estimates and the problem of the sample degeneracy.

A brief review of asymptotic approximations for posterior distributions is also given based on [15, 122].

2.1 REJECTION SAMPLING

The early methods for generating random variables were proposed in the same years as the MC approach. In 1947 Von Neumann outlined the Rejection Sampling (RS) [46] in a letter to Stanislaw Ulam and the seminal paper was published in 1951 [103]. The RS is a possible way of getting samples from a target distribution known up to a multiplicative constant. The key idea is to rely on the Fundamental Theorem of Simulation [125], thus resorting to an *instrumental* distribution from which is easier to get samples.

Theorem 1 (Fundamental Theorem of Simulation) *Let X be a random variable defined on $\mathcal{X}$ and distributed according $f(\cdot)$. Suppose that U is an uniform random variable defined on $\mathbb{R}^+$. Simulating $X \sim f(\cdot)$ is equivalent to simulating*

$$(X, U) \sim \mathtt{Unif}\{(x, u) : 0 < u < f(x)\}.$$

The above result suggests that the introduction of an *auxiliary* random variable U leading to the joint distribution

$$f_{X,U}(x, u) = \begin{cases} 1 & \text{if } 0 < u < f(x) \\ 0 & \text{otherwise} \end{cases}$$

allows retrieving the target distribution $f(\cdot)$ through marginalization being

$$f(x) = \int_0^{f_X(x)} du. \tag{4}$$

Thus, simulating pairs (x, u) from a superset of $\{(x, u) : 0 < u < f(x)\}$ and taking only pairs satisfying the constraint $0 < u < f(x)$ allows getting samples from $f(x)$ overcoming the problem of simulating from the target distribution just evaluating it pointwise. The superset mentioned above can be defined as

$$\{(x, u) : 0 < u < m(x)\}$$

under the constraint that $m(x) \geqslant f(x)$ for each $x \in \mathcal{X}$. Note that $m(x)$, the *envelope* function, cannot be a probability distribution unless the equality holds, however it can be the kernel of a probability distribution $g(\cdot)$ such that

$$m(x) = Mg(x) \quad \text{and} \quad M = \int_{\mathcal{X}} g(x)dx.$$

Accordingly, one can generate pairs (x, u) from $f_{X,U}(\cdot)$ as follows:

1. Draw $Y \sim g(\cdot)$
2. Draw $U|y \sim Unif(0, Mg(y))$
3. Accept (y, u) such that $0 < u < f(y)$.

Following this sampling procedure the retained y's represent realizations of the random variable X. In fact, the cumulative density function (cdf) of X can be retrieved as follows:

$$\Pr\{X \leqslant x\} = \Pr\{Y \leqslant x | U < f(Y)\} = \frac{\Pr\{Y \leqslant x, U < f(Y)\}}{\Pr\{U < f(Y)\}}.$$

The joint pdf derived from the first two steps of the outlined algorithm is

$$f_{Y,U}(y, u) = g(y)\frac{1}{Mg(y)} = \frac{1}{M}.$$

Thus,

$$\begin{aligned}
\Pr\{X \leqslant x\} &= \frac{\int_{-\infty}^{x} \int_0^{f(y)} \frac{1}{M} du dy}{\int_{\mathcal{X}} \int_0^{f(y)} \frac{1}{M} du dy} && (5) \\
&= \frac{\int_{-\infty}^{x} \int_0^{f(y)} du dy}{\int_{\mathcal{X}} \int_0^{f(y)} du dy} \\
&= \frac{\int_{-\infty}^{x} f(y) dy}{\int_{\mathcal{X}} f(y) dy} && (6) \\
&= \int_{-\infty}^{x} f(y) dy.
\end{aligned}$$

The described sampling procedure is often replaced by the equivalent procedure described in Algorithm 1. Looking at Algorithm 1, we note that the chosen instrumental distribution, $g(\cdot)$, must have tails thicker then those of the target $f(\cdot)$ in order to have

Algorithm 1 Rejection Sampling

for $s = 1, ..., S$ **do**
Draw $y^{(s)} \sim g(\cdot)$ and $u^{(s)} \sim \mathtt{Unif}[0, 1]$
Accept $X = y^{(s)}$ if $u^{(s)} \leqslant \dfrac{f(y^{(s)})}{Mg(y^{(s)})}$
end for

a bounded ratio $f(\cdot)/g(\cdot)$. Moreover, the probability of accepting $X = y^{(s)}$, at each iteration $s \in \{1, ..., S\}$, equals $\frac{1}{M}$:

$$\begin{aligned}
\Pr\left\{U \leqslant \frac{f(Y)}{Mg(Y)}\right\} &= \int_{\mathcal{X}} \Pr\left\{U \leqslant \frac{f(Y)}{Mg(Y)} | Y = y\right\} g(y)dy \\
&= \int_{\mathcal{X}} \frac{f(y)}{Mg(y)} g(y)dy \\
&= \frac{1}{M}.
\end{aligned}$$

Thus, M is the expected number of trials needed to obtain an acceptance. It follows that the efficiency of the algorithm, in terms of number of simulations needed to get an adequate sample size, depends on the size of M. In particular, when the instrumental distribution corresponds to the target

$$g(x) = f(x),$$

$M = 1$ and each proposed $y^{(s)}$ is accepted. Thus, when the instrumental distribution is more resembling the target, the algorithm is characterized by a greater efficiency. Accordingly, a possible strategy to maximize the efficiency is based on the choice of a parametric family for the instrumental distribution and the selection of the distribution $g(\cdot)$ satisfying the constraint $Mg(x) \geqslant f(x)$ with the minimum M in that family. Note that Algorithm 1 allows also sampling from distributions known up to the normalizing constant as shown in the following subsection.

2.1.1 *Sampling from unormalized posterior distributions*

Let us consider a Bayesian Inference scenario in which the target distribution is the posterior distribution of the parameter θ as defined in (1). Suppose that its analytical form is known up to the normalizing constant

$$K = \int_{\Theta} \pi(\theta)p(x|\theta)d\theta.$$

Samples from the posterior distribution can be got through Algorithm 1 as long as the unormalized posterior, hereafter denoted by $l(\theta) = \pi(\theta)p(x|\theta)$, can be easily computed. More precisely, Algorithm 1 can be implemented by a) replacing $f(x)$

with $l(\theta)$; b) resorting to the instrumental distribution $g(\cdot)$ defined on Θ. Following the same arguments as in equations from (5) to (6), it can be shown that

$$\begin{aligned}\Pr\{\theta \leqslant t\} &= \frac{\int_{-\infty}^{t} l(\theta)d\theta}{\int_{\Theta} l(\theta)d\theta} \\ &= \frac{\int_{-\infty}^{t} K\pi(\theta|x)d\theta}{\int_{\Theta} K\pi(\theta|x)d\theta} \\ &= \int_{-\infty}^{t} \pi(\theta|x)d\theta.\end{aligned}$$

proving that the described sampling procedure provides samples from $\pi(\theta|x)$. Note that the probability of accepting $\theta = t$, at each iteration, becomes:

$$\begin{aligned}\Pr\left\{u \leqslant \frac{l(\theta)}{Mg(\theta)}\right\} &= \int_{\Theta} \Pr\left\{u \leqslant \frac{l(\theta)}{Mg(\theta)}|\theta = t\right\} g(t)dt \\ &= \int_{\Theta} \frac{K \cdot \pi(t|x)}{Mg(t)} g(t)dt \\ &= \frac{K}{M}.\end{aligned}$$

The main drawback of this sampling scheme concerns the choice of the envelope function, $m(\cdot) = Mg(\cdot)$, since it requires a proper bound M and finding a value of M satisfying the constraint $Mg(\theta) \geqslant l(\theta)$ can be difficult, especially in a high-dimentional setting. Moreover, the acceptance probability depends on the constant M. Since the number of trials needed to get a sample of size S is a random variable distributed according to a Negative Binomial distribution [125], the time needed to approximate integrals such as (3) is in turn a random variable and cannot be evaluated a priori. However, the optimal choice of M leads to a maximization of the acceptance probability which minimizes the time and the computational cost. In contrast, a poor choice leads to a waste of computational efforts. A possible way to overcome this difficulty is the Importance Sampling described in the following section.

2.2 IMPORTANCE SAMPLING

Importance Sampling (IS) is a MC method which, as RS, allows approximating integrals as in (3) by drawing samples from an *easy to sample* instrumental distribution. Unlike RS, IS accepts all the proposed values, thus dispensing with the computation of the acceptance probability.

Let us consider the random variable X distributed according to $f(\cdot)$ over $\mathcal{X}$ and the general problem of approximating the integral

$$I = \mathbb{E}_f[h(x)] = \int_{\mathcal{X}} h(x)f(x)dx. \tag{7}$$

Algorithm 2 Importance Sampling

for $s = 1, ..., S$ **do**
  Draw $y^{(s)} \sim q(\cdot)$
  Assign to $y^{(s)}$ the *importance weight* $\omega^{(s)} = \frac{f(y^{(s)})}{q(y^{(s)})}$
end for

The basic idea of IS [94] is sampling from an *importance distribution* $q(\cdot)$, i.e., a probability distribution over $\mathcal{X}$ resembling $f(\cdot)$ but from which is easier to get samples, and then correcting the bias by weighting each sample as displayed in Algorithm 2. From the output of Algorithm 2, the integral in (7) can be approximated by the following weighted average

$$\hat{I} = \frac{1}{S} \sum_{s=1}^{S} \omega(y^{(s)}) h(y^{(s)}). \tag{8}$$

A formal justification for this approach comes from the *importance sampling fundamental identity*:

$$\begin{aligned} \mathbb{E}_f[h(x)] &= \int_{\mathcal{X}} h(x) f(x) dx \\ &= \int_{\mathcal{X}} h(x) \frac{f(x)}{q(x)} q(x) dx \\ &= \mathbb{E}_q[\omega(x) \cdot h(x)] \end{aligned}$$

where $\omega(x) = \frac{f(x)}{q(x)}$ is an importance weight.

Let $l(x) = f(x) \cdot K$ be the kernel of the target distribution. When the probability density function $f(\cdot)$ is known up to a normalizing constant, K, the importance weight can be computed at each iteration s as

$$\omega(y^{(s)}) = \frac{l(y^{(s)})}{q(y^{(s)})}.$$

It follows that the approximation of the integral in (7) becomes:

$$\frac{1}{S \cdot K} \sum_{s=1}^{S} \omega(y^{(s)}) h(y^{(s)}). \tag{9}$$

Being the normalizing constant unknown, an unbiased approximation [125] can be retrieved according to the importance sampling fundamental identity:

$$\begin{aligned} K &= \int_{\mathcal{X}} l(x) dx \\ &= \int_{\mathcal{X}} \omega(x) q(x) dx \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \omega(y^{(s)}). \end{aligned} \tag{10}$$

Thus, by substituting that unbiased approximation of the normalizing constant in (9) we obtain

$$\tilde{I} = \sum_{s=1}^{S} \tilde{\omega}(y^{(s)}) h(y^{(s)}) \tag{11}$$

where each $\tilde{\omega}(y^{(s)}) = \omega(y^{(s)}) / \sum_{s=1}^{S} \omega(y^{(s)})$ is a normalized weight. As long as $\mathrm{supp}(g) \supset \mathrm{supp}(f)$, whatever is the choice of the distribution $q(\cdot)$, both $\hat{I}$ and $\tilde{I}$ converge to (7) according to the Strong Low of Large Numbers.

One of the main drawbacks of IS is that its efficiency strongly depends on the choice of the importance distribution $q(\cdot)$. In fact, a poor selection of the importance distribution can lead to an estimator with an infinite variance. However, a properly selected importance distribution can lead to an estimator more efficient than the one provided by directly sampling from the target distribution.

The following theorem by Rubistain (see [131]) states that, for each target distribution $f(\cdot)$ there exists an optimal importance distribution depending on the integrable function $h(\cdot)$.

Theorem 2 *Given a target distribution* $f(\cdot)$ *and a function* $h(\cdot)$*, the importance distribution* $q^*(\cdot)$ *minimizing the variance of the estimator in* (8) *is*

$$q^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(x)|f(x)dx}. \tag{12}$$

Proof The variance of the importance sampling estimator in (8) can be expressed, up to a multiplicative constant, as:

$$\mathbb{V}\mathrm{ar}_q\left[\frac{h(X)f(X)}{q(X)}\right] = \mathbb{E}_q\left[\left(\frac{h(X)f(X)}{q(X)}\right)^2\right] - \left\{\mathbb{E}_q\left[\frac{h(X)f(X)}{q(X)}\right]\right\}^2,$$

By noting that

$$\mathbb{E}_q\left[\frac{h(X)f(X)}{q(X)}\right] = \int_{\mathcal{X}} \frac{h(x)f(x)}{q(x)} q(x)dx$$

does not depend on $q(\cdot)$, to minimize the variance it is sufficient to minimize the first term. From Jensen's inequality (see Theorem 14 in Appendix A.3) follows that, whatever is $q(\cdot)$, the first term is bounded below:

$$\mathbb{E}_q\left[\left(\frac{h(X)f(X)}{q(X)}\right)^2\right] \geqslant \left\{\mathbb{E}_q\left[\frac{|h(X)|f(X)}{q(X)}\right]\right\}^2 = \left(\int_{\mathcal{X}} |h(x)|f(x)dx\right)^2.$$

It is easy to proof that this lower bound is achieved by $q^*(\cdot)$:

$$\begin{aligned}\mathbb{E}_{q^*}\left[\left(\frac{h(X)f(X)}{q^*(X)}\right)^2\right] &= \int_{\mathcal{X}} \frac{\big(h(x)f(x)\big)^2}{q^*(x)}dx \\ &= \left(\int_{\mathcal{X}} |h(x)|f(x)dx\right)^2,\end{aligned}$$

where the last equality follows from the definition of $q^*(\cdot)$ in (12). □

Note that the optimal importance distribution depends on the integral I. However, by considering that importance distribution in (12), the estimator becomes

$$\tilde{I} = \frac{\sum_{s=1}^{S} f(y^{(s)})|h(y^{(s)})|^{-1}}{\sum_{s=1}^{S} |h(y^{(s)})|^{-1}}$$

which does not depend on the intractable integral. Hence, Theorem 2 suggests to sample $y^{(1)}, ..., y^{(S)}$ from an importance distribution proportional to $f \cdot |h|$. Note that the importance distribution still depends on the function $h(\cdot)$, hence this result is of limited applicability since in many applications we are interested in using the same sample to approximate many different quantities.

In order to evaluate the efficiency of an estimator derived through IS a practical rule of thumb is the evaluation of the Effective Sample Size (ESS).

2.2.1 *Effective Sample Size and Sample Degeneracy*

Conceptually the ESS represents the number of samples from the target distribution $f(\cdot)$, if it would be available, needed to get an estimator

$$\bar{I} = \frac{1}{S} \sum_{s=1}^{S} h(y^{(s)})$$

with the same variance as the IS estimator $\tilde{I}$ defined in (11). More formally, the ESS is defined as

$$\text{ESS} \triangleq S \frac{\mathbb{V}ar_f[\bar{I}]}{\mathbb{V}ar_q[\tilde{I}]}.$$

Unfortunately, both the variances are not analytically available hence the exact computation is infeasible. Indeed, in the MC literature the ESS is usually approximated as proposed in [76]:

$$\widehat{\text{ESS}} = \frac{\left(\sum_{s=1}^{S} \omega_s\right)^2}{\sum_{s=1}^{S} \omega_s^2}$$

where the importance weights $\omega(y^{(s)})$ are referred to as ω_s for short. This approximation is derived in [85] and [47] by approximating the variance:

$$\mathbb{V}ar_q[\tilde{I}] = \mathbb{V}ar_q\left[\frac{\sum_{s=1}^{S} \omega(Y^{(s)})h(Y^{(s)})}{\sum_{s=1}^{S} \omega(Y^{(s)})}\right]. \tag{13}$$

VARIANCE APPROXIMATION Following [47], the variance in (13) can be approximated by resorting to the Delta Method described in Appendix B.0.1. In particular, denoting $\omega(Y)$ by W and $h(Y)$ by H, from (157) we obtain

$$\mathbb{V}ar_q[\tilde{I}] \approx \frac{1}{S}\left[\mathbb{V}ar_q[W]\frac{\mathbb{E}_q^2[WH]}{\mathbb{E}_q^4[W]} + \frac{\mathbb{V}ar_q[WH]}{\mathbb{E}_q^2[W]} - 2\frac{\mathbb{E}_q[WH]}{\mathbb{E}_q^2[W]}\mathbb{C}ov_q[W, WH]\right]$$

where the symbol $\approx$ denotes the approximation following to the second order Taylor expansion. For the sake of simplicity hereafter we assume that the importance weights are computed involving the normalized target distribution, $f(\cdot)$. This assumption will be relaxed in the final approximation.

By noting that

$$\begin{aligned}\mathbb{E}_q[W] &= \int_{\mathcal{X}} \omega(y)q(y)dy \\ &= \int_{\mathcal{X}} \frac{f(y)}{q(y)}q(y)dy = 1\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_q[WH] &= \int_{\mathcal{X}} \omega(y)h(y)q(y)dy \\ &= \int_{\mathcal{X}} \frac{f(y)}{q(y)}h(y)q(y)dy \\ &= \int_{\mathcal{X}} h(y)f(y)dy = I,\end{aligned}$$

we obtain

$$\mathbb{V}ar_q[\tilde{I}] \approx \frac{1}{S}\left[\mathbb{V}ar_q[W]I^2 + \mathbb{V}ar_q[WH] - 2I\mathbb{C}ov_q[W, WH]\right]. \tag{14}$$

Let us expand

$$\begin{aligned}\mathbb{C}ov_q[W, WH] =& \mathbb{E}_q[W^2H] - \mathbb{E}_q[WH]\mathbb{E}_q[W] \\ =& \mathbb{E}_q[W^2H] - I \\ =& \mathbb{E}_f[WH] - I \qquad (15)\\ =& \mathbb{C}ov_f[WH] + \mathbb{E}_f[W]\mathbb{E}_f[H] - I \qquad (16)\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}ar_q[WH] =& \mathbb{E}_q[W^2H^2] - \mathbb{E}_q^2[WH] \\ =& \mathbb{E}_f[WH^2] - I^2. \qquad (17)\end{aligned}$$

where (15) follows from

$$\begin{aligned}\mathbb{E}_q[W^2H] &= \int_{\mathcal{X}} \left(\frac{f(x)}{q(x)}\right)^2 h(x)q(x)dx \\ &= \int_{\mathcal{X}} \frac{f(x)}{q(x)}h(x)f(x)dx \\ &= \mathbb{E}_f[WH]\end{aligned}$$

and (17) follows from

$$\begin{aligned}\mathbb{E}_q[W^2H^2] &= \int_{\mathcal{X}} \left(\frac{f(x)}{q(x)}\right)^2 \big(h(x)\big)^2 q(x)\,dx \\ &= \int_{\mathcal{X}} \frac{f(x)}{q(x)} \big(h(x)\big)^2 f(x)\,dx \\ &= \mathbb{E}_f[WH^2].\end{aligned}$$

By applying the Delta Method to $\mathbb{E}_f[WH^2]$ (see (156)) follows that

$$\mathbb{E}_f[WH^2] \approx \mathbb{E}_f[W]\mathbb{E}_f^2[H] + \frac{1}{2}\mathbb{V}ar_f[H] \cdot 2\mathbb{E}_f[W] + \mathbb{C}ov_f[WH] \cdot 2\mathbb{E}_f[H]. \tag{18}$$

Thus,

$$\begin{aligned}\mathbb{V}ar_q[WH] =& \mathbb{E}_f[WH^2] - I^2 \\ =& \mathbb{E}_f[W]I^2 + \mathbb{V}ar_f[H]\mathbb{E}_f[W] + 2I\mathbb{C}ov_f[WH].\end{aligned} \tag{19}$$

By substituting (16) and (19) into (14) follows

$$\mathbb{V}ar_q[\tilde{I}] \approx \frac{1}{S}\Big[Var_f[H]\mathbb{E}_f[W] + I^2\big(\mathbb{V}ar_q[W] - \mathbb{E}_f[W] + 1\big)\Big]. \tag{20}$$

Since

$$\mathbb{V}ar_f[\bar{I}] = \frac{1}{S}\mathbb{V}ar_f[H]$$

and

$$\begin{aligned}\mathbb{E}_f[W] &= \int_{\mathcal{X}} \frac{f(x)}{q(x)} f(x)\,dx \\ &= \int_{\mathcal{X}} \left(\frac{f(x)}{q(x)}\right)^2 q(x)\,dx &(21)\\ &= \mathbb{E}_q[W^2] = 1 + \mathbb{V}ar_q[W], &(22)\end{aligned}$$

the variance in (20) can be written as

$$\mathbb{V}ar_q[\tilde{I}] \approx \mathbb{V}ar_f[\bar{I}]\big(1 + \mathbb{V}ar_q[W]\big).$$

The ESS resulting from the variance approximation derived above is

$$ESS \approx \frac{S}{1 + \mathbb{V}ar_q[W]}.$$

When the target distribution can be evaluated only up to a normalizing constant, the ESS can be adapted a posteriori as

$$\begin{aligned}ESS \approx& \frac{S}{1 + \frac{\mathbb{V}ar_q[W]}{K^2}} &(23)\\ =& \frac{SK^2}{K^2 + \mathbb{V}ar_q[W]} \\ =& \frac{SK^2}{\mathbb{E}_q[W^2]}. &(24)\end{aligned}$$

Thus, by resorting to the approximation of the normalized constant in (10)

$$\begin{aligned}\widehat{\text{ESS}} &= S\frac{\left(\frac{1}{S}\sum_{s=1}^{S}\omega_s\right)^2}{\frac{1}{S}\sum_{s=1}^{S}\omega_s^2} \\ &= \frac{\left(\sum_{s=1}^{S}\omega_s\right)^2}{\sum_{s=1}^{S}\omega_s^2}.\end{aligned} \tag{25}$$

Looking at (23), it is apparent that highly variable importance weights correspond to a small ESS. Thus, an importance distribution leading to highly variable weights results in an inefficient estimator $\tilde{I}$. Furthermore, it can be shown that $\mathbb{V}ar_q[W]$ is a measure of distance between the importance and the target distribution [86]:

$$\begin{aligned}Var_q[W] =& \mathbb{E}_q[W^2] - \mathbb{E}_q^2[W] \\ =& \mathbb{E}_q[W^2] - 1 \\ =& \int_{\mathcal{X}} \left(\frac{f(x)}{q(x)}\right)^2 q(x)dx - 2 + 1 \\ =& \int_{\mathcal{X}} \frac{\left(f(x)\right)^2}{q(x)}dx - 2\int_{\mathcal{X}} f(x)dx + \int_{\mathcal{X}} q(x)dx \\ =& \int_{\mathcal{X}} \frac{\left(f(x)\right)^2}{q(x)}dx - 2\int_{\mathcal{X}} \frac{f(x)q(x)}{q(x)}dx + \int_{\mathcal{X}} \frac{\left(q(x)\right)^2}{q(x)}dx \\ =& \int_{\mathcal{X}} \frac{[f(x) - q(x)]^2}{q(x)}dx \\ =& \chi^2(f, q)\end{aligned}$$

where $\chi^2(f, q)$ is the Pearson divergence between the target and the importance distribution. This result suggests that an importance distribution close to the target distribution leads to an efficient estimator and to an adequate ESS. On the other hand, an importance distribution far from the target leads to one of the main drawbacks in IS methods: the sample degeneracy.

SAMPLE DEGENERACY The problem of sample degeneracy in IS techniques arises when only a small fraction of the importance weights has relative high weights. Generally speaking, this problem is due to an importance distribution far from the target. As already shown, a great distance between the two distributions corresponds to highly variable weights. It follows that when sample degeneracy occurs the IS gives inefficient estimators and a very large number of simulations is required to get adequate estimates. A common measure of the degree of sample degeneracy is the ESS. As will be discussed in the following sections this problem becomes more serious when one resort to RS or to a Random Weights Importance Sampling.

2.2.2 *Comparing* RS *and* IS

In order to formally compare the performances of RS and IS we refer to Chen [23] proving that the RS estimator is always less efficient than the one based on IS with importance distribution equal to $g(\cdot)$. In particular, RS can be seen as a special IS, that relaying on the following target and importance distributions defined on the augmented space $\mathcal{X} \times [0, 1]$:

$$f^*(x, y) = \begin{cases} Mg(x) & \text{for } x \in \mathcal{X} \quad y \in \left[0, \frac{f(x)}{Mg(x)}\right] \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$q^*(x, y) = \begin{cases} g(x) & \text{for } x \in \mathcal{X} \quad y \in \left[0, \frac{f(x)}{Mg(x)}\right] \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Note that the desired target distribution $f(\cdot)$ can be retrieved by marginalizing out the auxiliary variable y:

$$\int_0^1 f^*(x, y)dy = \int_0^{\frac{f(x)}{Mg(x)}} Mg(x)dy = \frac{f(x)}{Mg(x)} Mg(x) = f(x).$$

From (26) and (27) follows that the importance weights are

$$\omega^*(x, y) = \begin{cases} M & \text{for } x \in \mathcal{X} \quad y \in \left[0, \frac{f(x)}{Mg(x)}\right] \\ 0 & \text{otherwise,} \end{cases}$$

meaning that the rejection step is replaced by a weighting strategy leading to an equivalent estimator. In fact, by denoting with $\mathcal{A}$ the set of iterations at which is assigned a strictly positive weight and assuming $|\mathcal{A}| = S^*$, the resulting estimator, $\tilde{I}_R$, is equivalent to the RS estimator:

$$\tilde{I}_R = \sum_{s=1}^{S} \omega_s^* \frac{h(Y^{(s)})}{\sum_{s=1}^{S} \omega_s^*} \sum_{s=1}^{S} \omega_s^* \frac{h(Y^{(s)})}{S^* M} = \frac{1}{S^*} \sum_{s \in \mathcal{A}} h(Y^{(s)}).$$

Theorem 3 *The Pearson distance between the target* $f(\cdot)$ *and the proposal* $q(\cdot)$ *is less than or equal to the one between the target* $f^*(\cdot)$ *and the proposal* $q^*(\cdot)$*. In other words:*

$$\mathbb{V}ar_q\left[\frac{f(X)}{q(X)}\right] \leqslant \mathbb{V}ar_{q^*}\left[\frac{f^*(X, Y)}{q^*(X, Y)}\right]$$

Proof From (22) follows that

$$\mathbb{V}ar_q[W] = \mathbb{E}_q[W^2] - 1.$$

Thus,

$$
\begin{aligned}
1 + \mathbb{V}ar_{q^*}\left[\frac{f^*(X,Y)}{q^*(X,Y)}\right] &= \int_{\mathcal{X}} \int_{0}^{\frac{f(x)}{Mg(x)}} \left(\frac{f^*(X,Y)}{g^*(X,Y)}\right)^2 q^*(x,y)\,dy\,dx \\
&= \int_{\mathcal{X}} \int_{0}^{\frac{f(x)}{Mg(x)}} M^2 g(x)\,dy\,dx \\
&= \int_{\mathcal{X}} Mf(x)\,dx \\
&= M
\end{aligned}
$$

and

$$
\begin{aligned}
1 + \mathbb{V}ar_{q}\left[\frac{f(X)}{q(X)}\right] &= \int_{\mathcal{X}} \left(\frac{f(x)}{q(x)}\right)^2 q(x)\,dx \\
&= \int_{\mathcal{X}} \frac{f(x)}{g(x)} f(x)\,dx \\
&\leqslant \int_{\mathcal{X}} \frac{Mg(x)}{g(x)} f(x)\,dx \\
&= M
\end{aligned}
$$

where the inequality follows from the envelope function definition.

Thus,

$$
\mathbb{V}ar_{q}\left[\frac{f(X)}{q(X)}\right] \leqslant M - 1 = \mathbb{V}ar_{q^*}\left[\frac{f^*(X,Y)}{q^*(X,Y)}\right].
$$

□

Summing up, the envelope function involved in RS can be used as the importance distribution in an equivalent IS based on a target and an importance distribution defined on an augmented space and whose marginal distributions are $f(\cdot)$ and $q(\cdot)$, respectively. Conducting IS directly on the marginal distribution is at least efficient as the RS.

2.2.3 *Sampling from the posterior distribution via* IS

Recalling that one of the main problems of Bayesian inference is the computation of intractable marginal likelihood, one can resort to IS to address it [58]. In such a case, a sufficient condition is the availability of the analytical form of the kernel of the posterior distribution:

$$
l(\theta) = \pi(\theta)p(x|\theta).
$$

In fact, by sampling $\theta^{(1)}, ..., \theta^{(S)}$ i.i.d from an importance distribution, $q(\cdot)$, defined on the parameter space, Θ, and by computing the importance weights

$$
\omega_s = \frac{l(\theta^{(s)})}{q(\theta^{(s)})} = \frac{\pi(\theta^{(s)})p(x|\theta^{(s)})}{q(\theta^{(s)})} \quad \forall s \in \{1, ..., S\},
$$

Algorithm 3 IS^2

for $s = 1, ..., S$ **do**
  Draw $\theta^{(s)} \sim q(\cdot)$
  Run an IS to draw $y^{(1)}, ..., y^{(n)}$ and compute $\hat{p}_N(x|\theta^{(s)})$
  Assign to $\theta^{(s)}$ the *importance weight* $\hat{\omega}^{(s)} = \frac{\pi(\theta^{(s)})\hat{p}_N(x|\theta^{(s)})}{q(\theta^{(s)})}$
end for

the integral in (3) is approximated by

$$\tilde{I}_{IS} = \sum_{s=1}^{S} \frac{\omega_s h(y^{(s)})}{\sum_{s=1}^{S} \omega_s}.$$

Unfortunately, in many cases the likelihood $p(x|\theta)$ is intractable and cannot be involved in the computation of the importance weights. Note that likelihood-free methods overcoming this difficulty will be discussed in the next chapter, however here we introduce an IS technique based on an estimate of the likelihood function in turn retrieved employing IS .

2.2.3.1 *Random Weights Importance Sampling*

The Random Weights Importance Sampling (RW-IS) is an IS algorithm in which the evaluation of the likelihood function is replaced by a random estimate. In particular in [151] it is referred to as IS^2, being the likelihood estimate in turn derived employing IS. For example, suppose that the random variable X is associated to a latent variable Y defined on $\mathcal{Y}$ and that the joint probability $p(x, y|\theta)$ is analytically available. When the likelihood function

$$p(x|\theta) = \int_{\mathcal{Y}} p(x, y|\theta)dy = \int_{\mathcal{Y}} p(x|y, \theta)p(y|\theta)dy \tag{28}$$

is analytically intractable, one of the possible ways of approximating quantities such as (3) is to approximate (28) employing IS and then resort to Algorithm 3. Note that the importance weights $\hat{\omega}^{(s)}$ depend on the random estimate $\hat{p}_N(x|\theta^{(s)})$. More precisely, let $q(\cdot|x, \theta)$ be an importance distribution for Y an let

$$\omega(y^{(n)}, \theta) = \frac{p(x, y^{(n)}|\theta)}{q(y^{(n)}|x, \theta)} = \frac{p(x|y^{(n)}, \theta)p(y^{(n)}|\theta)}{q(y^{(n)}|x, \theta)}$$

be the corresponding importance weights. The density $p(x|\theta)$ is estimated by

$$\hat{p}_N(x|\theta) = \frac{1}{N} \sum_{n=1}^{N} \omega(y^{(n)}, \theta). \tag{29}$$

A formal justification of Algorithm 3 is provided in [151] and is based on the following arguments:

- the estimator $\hat{p}_N(x|\theta)$ can be rewritten as

$$p(x|\theta)e^z,$$

where $Z \stackrel{\triangle}{=} \log \hat{p}_N(x|\theta) - \log p(x|\theta)$ is a random variable whose randomness is induced by the randomness occurring in the estimator $\hat{p}_N(x|\theta)$ and whose probability density is denoted by $q_N(z|\theta)$;

- IS^2 can be considered as an IS scheme aimed at sampling from the joint posterior distribution

$$\tilde{\pi}(\theta, z|x) \stackrel{\triangle}{=} \pi(\theta|x) e^z q_N(z|\theta) = \frac{\pi(\theta)\hat{p}_N(x|\theta)}{K} q_N(z|\theta)$$

on the extended space $\Theta \times \mathbb{R}$.

Assuming that $\hat{p}_N(x|\theta)$ is an unbiased estimator, the following equality holds:

$$\mathbb{E}_{q_N}[e^z] = \int_{\mathbb{R}} e^z q_N(z|\theta) dz = 1.$$

Thus, the marginal posterior equals the target distribution

$$\int_{\mathbb{R}} \tilde{\pi}(\theta|x) dz = \pi(\theta|x) \int_{\mathbb{R}} e^z q_N(z|\theta) dz = \pi(\theta|x).$$

It follows that the integral in (3) can be rewritten as

$$\begin{aligned}\mathbb{E}_\pi[h(\theta)] &= \int_\Theta \int_{\mathbb{R}} h(\theta)\pi(\theta|x) e^z q_N(z|\theta) dz d\theta \\ &= \int_\Theta \int_{\mathbb{R}} h(\theta) \frac{\pi(\theta|x) e^z q_N(z|\theta)}{q(\theta, z)} q(\theta, z) dz d\theta\end{aligned}$$

where $q(\theta, z) = q_N(z|\theta) q(\theta)$ is the importance distribution over $\Theta \times \mathbb{R}$. Accordingly, by computing unnormalized weights

$$\begin{aligned}\hat{\omega}(\theta) &= \frac{l(\theta) e^z q_N(z|\theta)}{q(\theta, z)} \\ &= \frac{\pi(\theta) p(x|\theta)}{q_N(z|\theta) q(\theta)} e^z q_N(z|\theta) \\ &= \frac{\pi(\theta)\hat{p}_N(x|\theta)}{q(\theta)}\end{aligned}$$

from Algorithm 3 one can estimate (3) by computing

$$\tilde{I}_{IS^2} = \frac{\sum_{s=1}^{S} \hat{\omega}^{(s)} h(\theta^{(s)})}{\sum_{s=1}^{S} \hat{\omega}^{(s)}}.$$

In [151], they prove that $\tilde{I}_{IS^2}$ converges almost surely to $\mathbb{E}_\pi[h(\theta)]$. They also show that, under mild conditions, the estimator $\tilde{I}_{IS^2}$ has a variance retrieved as

$$\mathbb{V}ar[\tilde{I}_{IS^2}] = C \cdot \tilde{I}_{IS},$$

where $C > 1$ is a constant depending on the variance of the stochastic term Z. It follows that $\tilde{I}_{IS^2}$ is less efficient then $\tilde{I}_{IS}$, meaning that when the likelihood is replaced by a random estimate, the ESS is smaller and the problem of sample degeneracy becomes more serious.

2.2.4 *Sampling from the tails of the distribution*

As already shown, the variability of the importance weights depends on the importance distribution which, if properly chosen, may reduce the sample degeneracy leading to an adequate sample size. The choice of the importance distribution is also crucial when we are interested in approximating quantities involving an evaluation of the probability in the tails of the target distribution. to show this fact we discuss an example based on the Example 3.5. from [125].

Example 1. Approximating the Normal cumulative density function Suppose that the we are interested in evaluating the cumulative density function (cdf) of a random variable normally distributed. Since the cdf, denoted by $\Phi(\cdot)$, cannot be written in an explicit form, one can resort to a MC estimate. For example, the approximation of

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

can be got by sampling $y^{(1)}, ..., y^{(S)}$ i.i.d. from a standard normal distribution and by computing

$$\hat{\Phi}(0) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\{y^{(s)} \leqslant 0\}. \tag{30}$$

Note that each $\mathbb{1}\{y^{(s)} \leqslant 0\}$ is a Bernoulli random variable with success probability $\Phi(0)$. Accordingly, the exact variance of (30) is

$$\frac{\Phi(0)(1-\Phi(0))}{S}.$$

Since in this simple example we know that $\Phi(0) = 0.5$, we are able to compute the number of simulations S needed to achieve the desired precision. For example achieving a precision of four decimals with a confidence level of $\approx 99.5\%$ requires

$$S \approx (\sqrt{2} \cdot 10^4)^2 = 2 \cdot 10^8.$$

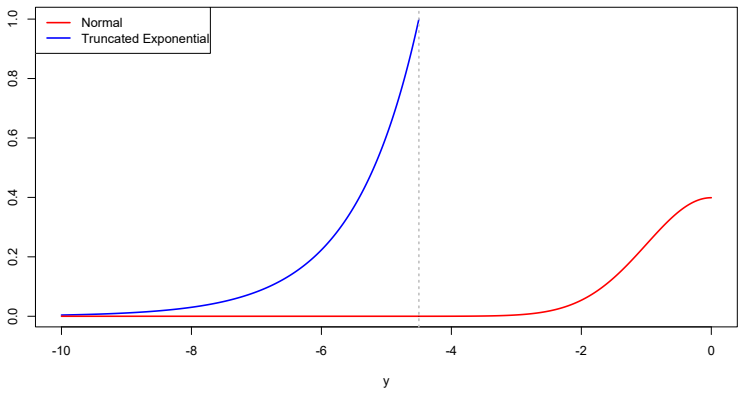
Now suppose that we are interested in approximating $\Phi(-4.5)$ through the following MC estimator

$$\hat{\Phi}(-4.5) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\{y^{(s)} \leqslant -4.5\}. \tag{31}$$

Since $\Phi(-4.5) = 3.3977 \cdot 10^{-6}$, to get an accurate estimate we need a precision of at least seven decimals corresponding to $\approx 13 \cdot 10^8$ simulations. Otherwise, small value of S usually produce all zeros of the indicator function since we are approximating the probability of a very rare event. In [125], the Example 3.8 shows that IS may improve the accuracy by relying on a clever importance distribution. In particular, they suggest to resort to a right truncated exponential distribution with density

$$q(y) = e^{4.5+y} \mathbb{1}\{y \leqslant -4.5\}. \tag{32}$$

Figure 1: Right truncated exponential distribution (blue line) and standard normal distribution (red line).



As shown in Figure 1, sampling values smaller than -4.5 from a truncated exponential distribution (blue line) is more likely than from the standard normal distribution (red line). It follows that denoting by $f(\cdot)$ the standard normal pdf, by simulating from $q(\cdot)$ we obtain the following approximation with $S = 10,000$

$$\hat{\Phi}(-4.5) = \frac{1}{S}\sum_{s=1}^{S}\frac{f(y^{(s)})}{q(y^{(s)})}\mathbb{1}\{y^{(s)} \leqslant -4.5\} = 3.4139 \cdot 10^{-6}. \tag{33}$$

2.3 MARKOV CHAIN MONTE CARLO

As shown in the previous section, IS is a possible solution for approximating integrals such as (7) without sampling directly from the target distribution $f(\cdot)$. Markov Chain Monte Carlo (MCMC) methods represent another strategy for getting samples, $X_1, \ldots X_S$, approximately distributed according to $f(\cdot)$. In particular, a MCMC method is any method producing an ergodic Markov chain, $\{X^{(s)}\}_{s=1}^{S}$, whose stationary distribution is $f(\cdot)$ [125]. For details on Markov chains and their properties see Appendix C. Under certain conditions the sequence produced by an MCMC method can be employed to approximate integrals via sample averages, as already shown in the case of i.i.d. samples.

The first MCMC method was developed by Nicholas Metropolis et al. in 1953 [98]. The algorithm was proposed as a modified MC integration method aimed at investigating the properties of substances consisting of interacting individual molecules. In 1970, Hastings generalized the Metropolis algorithm presenting the Metropolis–Hastings algorithm [63]. Another widespread MCMC method is the Gibbs Sampler introduced by Stuart Geman and Donald Geman in 1985 [56]. In 1990 Gelfand and Smith [52] shed light on the usefulness of the Gibbs sampler for calculating Bayesian posterior distributions. Afterwords, Tierney [149] showed how the other MCMC methods could be used for exploring intractable posterior distributions.

Algorithm 4 Metropolis Algorithm

Initialize $x^{(0)}$
for $s = 1, ..., S$ **do**
  Draw $y \sim \tilde{q}(x^{(s-1)}, \cdot)$ and $u^{(s)} \sim \mathcal{U}nif[0, 1]$
  Compute $\alpha(x^{(s-1)}, y) = \min\left\{1, \frac{f(y)}{f(x^{(s-1)})}\right\}$
  if $u^{(s)} \leqslant \alpha(x^{(s-1)}, y)$ **then**
    Set $X^{(s)} = y$
  else
    $X^{(s)} = x^{(s-1)}$
  end if
end for

2.3.1 *Metropolis Algorithm*

The Metropolis algorithm can be seen as a generalized RS [86]. In fact, it allows sampling from a *proposal* distribution, $\tilde{q}(\cdot, \cdot)$, different from the target and resorts to a rejection step to ensure the convergence to the target distribution. The main difference between the two algorithms is that in the Metropolis algorithm, at each step, both the proposal distribution and the acceptance probability depend on the value accepted at the previous iteration. Thus, samples are no longer independent of one another but rather form a Markov chain. It follows that to ensure the convergence the algorithm must generate a Markov chain characterized by *transition probabilities*, denoted by $q(\cdot, \cdot)$, satisfying the following equation

$$\int_{\mathcal{X}} f(x)q(x, y)dx = f(y), \tag{34}$$

where x denotes the current value and y denotes the proposed value of the chain. Equation (34) states that the target distribution is the stationary distribution of the Markov chain characterized by that transition probabilities (see Appendix C.3 and C.4). However, the *reversibility* (see Definition 16) of the target distribution is often an easier-to-check sufficient (although not necessary) condition for the stationarity. In fact, $f(\cdot)$ is reversible whenever the Detailed Balance condition holds:

$$f(x)q(x, y) = f(y)q(y, x). \tag{35}$$

From (35) follows

$$\begin{aligned}\int_{\mathcal{X}} f(x)q(x, y)dx &= \int_{\mathcal{X}} f(y)q(y, x)dx \\ &= f(y)\int_{\mathcal{X}} q(y, x)dx \\ &= f(y).\end{aligned}$$

Thus, the Metropolis algorithm is based on the definition of a transition kernel satisfying this condition. In particular, the transition probabilities following from the implementation of the Algorithm 4 are of the form

$$q(x, y) = \tilde{q}(x, y)\alpha(x, y)$$

Algorithm 5 Metropolis–Hastings Algorithm

Initialize $x^{(0)}$
for $s = 1, ..., S$ **do**
  Draw $y \sim \tilde{q}(x^{(s-1)}, \cdot)$ and $u^{(s)} \sim \mathcal{U}nif[0, 1]$
  Compute $\alpha(x^{(s-1)}, y) = \min\left\{1, \frac{f(y)\tilde{q}(y, x^{(s-1)})}{f(x^{(s-1)})\tilde{q}(x^{(s-1)}, y)}\right\}$
  if $u^{(s)} \leqslant \alpha(x^{(s-1)}, y)$ **then**
    Set $X^{(s)} = y$
  else
    $X^{(s)} = x^{(s-1)}$
  end if
end for

where $\tilde{q}(x, y)$ is a symmetric proposal function and

$$\alpha(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

In other terms

$$q(x, y) = \begin{cases} \tilde{q}(x, y)\frac{f(y)}{f(x)} & \text{if } f(x) > f(y) \\ \tilde{q}(x, y) & \text{otherwise.} \end{cases} \tag{36}$$

For example, suppose that $f(x) > f(y)$, equation (35) becomes

$$\begin{aligned} f(x)q(x, y) &= f(x)\tilde{q}(x, y)\frac{f(y)}{f(x)} \\ &= f(y)\tilde{q}(x, y) \\ &= f(y)\tilde{q}(y, x) \\ &= f(y)q(y, x) \end{aligned} \tag{37}$$

where (37) follows from symmetry.

2.3.2 *Metropolis-Hastings Algorithm*

The Metropolis algorithm was generalized by Hastings as displayed in Algorithm 5. This generalization, known as Metropolis–Hastings (MH) algorithm, allows overcoming the requirement for a symmetric proposal distribution. Accordingly, the transition probabilities in (36) become

$$q(x, y) = \begin{cases} \tilde{q}(x, y)\frac{f(y)\tilde{q}(y, x)}{f(x)\tilde{q}(x, y)} & \text{if } f(x) > f(y) \\ \tilde{q}(x, y) & \text{otherwise.} \end{cases} \tag{38}$$

It is straightforward to show that the Detailed Balance condition is still satisfied without symmetry:

$$\begin{aligned} f(x)q(x,y) &= f(x)\tilde{q}(x,y)\frac{f(y)}{f(x)}\frac{\tilde{q}(y,x)}{\tilde{q}(x,y)} \\ &= f(y)\tilde{q}(y,x) \\ &= f(y)q(y,x). \end{aligned}$$

Thus, MH algorithm admits the use of almost any pdf as proposal distribution. Here we focus on two of the most used family of proposal distributions.

INDEPENDENT MH : At each iteration the proposal distribution is independent from the value accepted at the previous iteration:

$$\tilde{q}(x,y) = \tilde{q}(y).$$

It follows that the acceptance probability becomes:

$$\alpha(x,y) = \min\left\{1, \frac{f(y)\tilde{q}(x)}{f(x)\tilde{q}(y)}\right\}.$$

Note that, even though the proposal distribution is independent from the current state of the chain, the acceptance ratio still depends on it thus building a Markov chain.

This kind of proposal encourages the exploration of the sample space but can be a poor choice for complex target distributions.

RANDOM WALK METROPOLIS : At each iteration the proposal distribution is centred on the value accepted at the previous iteration, say x. Thus, the proposed candidate can be expressed as

$$y = x + \epsilon \qquad \epsilon \sim q(\cdot|\mu = 0, \sigma)$$

where μ is a centrality parameter and σ is a scale parameter.

Note that, being the transition probability symmetric, the acceptance probability reduces to

$$\alpha(x,y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

Thus, this kind of proposal leads to the implementation of a Metropolis algorithm and encourages a local exploration around the value accepted at the previous step.

Obviously the choice of the proposal distribution and its scale parameter strongly influences the acceptance rate. In particular, a proposal distribution with a high variance usually leads to a low acceptance rate, often implying that the chain gets stuck because of the large number of rejections. On the other hand, a low variance leads to proponing local moves around the last accepted value thus resulting in a high acceptance rate. However, a too high acceptance rate typically corresponds to a chain failing to explore the entire sample space. Ideally the proposal distribution involved in a Random Walk Metropolis should be chosen in such a way that the acceptance rate is approximativly 40% for univariate random variables and declines to about 23% as the size of the variable increases [53].

2.3.2.1 *Convergence*

MH algorithm satisfies by construction the Detailed Balance condition, thus ensuring that the target distribution, $f(\cdot)$, represents the stationary distribution of the induced Markov chain. However, to properly approximate integrals such as (7) through sample averages, $\{X^{(s)}\}_{s=1}^{S}$ must be an *ergodic* chain. Recalling that a Markov chain is said to be *ergodic* when it is *irreducible*, *aperiodic* and *positive recurrent* (see Appendix C.2), the structurally properties required to an appropriate convergence are the *irreducibility* and *aperiodicity*. As shown in [125, Ch 7], the following two conditions ensures aperiodicity and irreducibility, respectively:

- A sufficient condition for the aperiodicity is that the algorithm allows events such as $\{X^{(s+1)} = X^{(s)}\}$, i.e., there is a strictly positive probability of staying in the same state at two consecutive steps, meaning that there exists a strictly positive probability that

  $$\alpha(X, Y) = \min\left\{1, \frac{f(Y)\tilde{q}(Y, X)}{f(X)\tilde{q}(X, Y)}\right\} < 1.$$

  It follows that a sufficient condition for aperiodicity is that:

  $$\Pr\left\{\frac{f(Y)\tilde{q}(Y, X)}{f(X)\tilde{q}(X, Y)} \geqslant 1\right\} = \Pr\left\{f(Y)\tilde{q}(Y, X) \geqslant f(X)\tilde{q}(X, Y)\right\} < 1.$$

  Note that this condition does not contradict the Detailed Balance condition since the proposal distribution $\tilde{q}(\cdot, \cdot)$ does not correspond to the transition probability $q(\cdot, \cdot)$.

- The irreducibility follows from the *positivity* of the chain $\{X^{(s)}\}_{s=1}^{S}$, that is

  $$\tilde{q}(x, y) > 0 \quad \forall (x, y) \in \mathcal{X}^2$$

  meaning that at each iteration every subset of $\mathcal{X}$ can be reached in a single step.

Since a MH irreducible chain is *Harris recurrent* (17), as proved in [125, Th. 7.3], the conditions described above ensure that the following equality holds:

$$\lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} h(x^{(s)}) = \int_{\mathcal{X}} h(x) f(x) dx \qquad \text{a.s.}$$

where $h(\cdot)$ must be a Lebesgue integrable function. Furthermore, the distribution of $\{X^{(s)}\}_{s=1}^{S}$ converges in *total variation norm* to the stationary distribution. For a formal proof we refer the reader to [125, Th. 7.4].

2.3.3 *Gibbs sampler*

The Gibbs sampler is another MCMC method which can be considered as a special case of the more general MH algorithm. It allows sampling from the joint distribution and approximating the marginal distributions when dealing with a multivariate random variable.

Algorithm 6 Gibbs sampler

Initialize $\mathbf{x}^{(0)}$
for $s = 1, ..., S$ **do**
  Draw $x_1^{(s)} \sim f_{X_1}(x_1|x_2^{(s-1)}, ..., x_d^{(s-1)})$
  Draw $x_2^{(s)} \sim f_{X_2}(x_2|x_1^{(s)}, x_3^{(s-1)}, ..., x_d^{(s-1)})$
  $\vdots$
  Draw $x_d^{(s)} \sim f_{X_d}(x_d|x_1^{(s)}, ..., x_{d-1}^{(s)})$
end for

Let us consider the d-dimensional random variable $\mathbf{X} = (X_1, ..., X_d)$ taking values in $\mathcal{D}$ and distributed according to the joint density $f_{\mathbf{X}}(\cdot)$. Suppose that we are interested in obtaining some features of the marginal density

$$f_{X_i}(x_i) = \int f(x_1, ..., x_d) d\mathbf{x}_{\setminus i}, \tag{39}$$

where $\mathbf{x}_{\setminus i}$ denotes $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$. As already stressed, this kind of computation is often infeasible. The Gibbs sampler allows generating samples distributed according to $f_{X_i}(\cdot)$, for each $i \in \{1, ..., d\}$, without drawing directly from it. In fact, as displayed in Algorithm 6, the Gibbs sampler resorts to what is called *full conditional* distributions, in order to get samples $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(S)}$ from the joint distribution. A full conditional distribution is the probability distribution of a single component of $\mathbf{X}$ given all the others and is denoted by

$$f(x_i|\mathbf{x}_{\setminus i}) = f(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$$

for each component $i \in \{1, ..., d\}$. In practice, due to the conditional dependence structure between random variables established by the assumed model, the full conditional simplifies. In fact, each component x_i can be sampled from its probability distribution given the *Markov blanket*, [109] which represents the subset of components of $\mathbf{X}$ conditioned on which X_i is independent from all the others.

In addition to Algorithm 6, also known as *Systematic scan*, many alternative approaches to determine the order of the coordinates to be sampled can be devised. Two well-known methods are the *Symmetric scan* [125] and the *Random scan* [84].

SYMMETRIC SCAN The coordinates are sampled from their full conditionals first in an ascending order and then in a descending order.

RANDOM SCAN The coordinates are sampled from their full conditionals at each iteration in a different random order.

Unlike the chain induced by Algorithm 6, both the above methods induce a reversible Markov chain. Generally speaking, despite the fact that the chain induced by a Gibbs sampler is possibly reversible, it is straightforward to show that the target distribution is a stationary distribution.

Let us denote by $q(\mathbf{x}, \mathbf{y})$ the transition probability from the state $\mathbf{x} \in \mathcal{D}$ to the state $\mathbf{y} \in \mathcal{D}$, from the sampling scheme outlined in Algorithm 6 follows that

$$q(\mathbf{x}, \mathbf{y}) = f_{X_1}(y_1|x_2, ..., x_d) f_{X_2}(y_2|y_1, x_3, ..., x_d) ... f_{X_d}(y_d|y_1, ..., y_{d-1}).$$

Let $f^i_X(\cdot)$ denote the marginal density obtained by integrating out the i-th component, then

$$\begin{aligned}
&\int f_X(x) q(x, y) dx \\
&= \int f_X(x) f_{X_1}(y_1|x_2, ..., x_d) f_{X_2}(y_2|y_1, x_3, ..., x_d) ... f_{X_d}(y_d|y_1, ..., y_{d-1}) dx \\
&= \int [f^1_X(x_2, ..., x_d) f_{X_1}(x_1|x_2, ..., x_d)] f_{X_1}(y_1|x_2, ..., x_d) ... f_{X_d}(y_d|y_1, ..., y_{d-1}) dx_1, ..., dx_d \\
&= \int f_X(y_1, x_2, ..., x_d) f_{X_2}(y_2|y_1, x_3, ..., x_d) ... f_{X_d}(y_d|y_1, ..., y_{d-1}) dx_2, ..., dx_d \qquad (40)
\end{aligned}$$

where (40) follows by integrating out x_1 and combining $f^1_X(x_2, ..., x_d)$ with $f_{X_1}(y_1|x_2, ..., x_d)$. Doing the same for the other components from x_2 to x_{d-1} we obtain:

$$\begin{aligned}
&\int f_X(y_1, ..., y_{d-1}, x_d) f_{X_d}(y_d|y_1, ..., y_{d-1}) dx_d \\
=&f^d_X(y_1, ..., y_{d-1}) f_{X_d}(y_d|y_1, ..., y_{d-1}) \\
=&f_X(y).
\end{aligned}$$

Thus, the target distribution satisfies the condition of stationarity.

2.3.3.1 *Convergence*

As already discussed for MH, we need to investigate the irreducibility and aperiodicity of the chain to ensure that the simulated Markov chain converges appropriately. Roberts and Smith in 1994 [126] provided simple conditions for the converge of the Gibbs sampler and Metropolis–Hastings algorithms. In particular, for the Gibbs sampler they showed that when 1) the target $f_X(\cdot)$ is lower semi-continuous at 0; 2) f^i_X is locally bounded for $i \in \{1, ..., d\}$ and 3) $\mathcal{D}$ is connected,

$$\lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} h(x^{(s)}) = \int_{\mathcal{D}} h(x) f_X(x) dx \qquad \text{a.s.}$$

and the distribution of $\{X^{(s)}\}_{s=1}^S$ converges in *total variation* to the stationary distribution. However, this level of generality may not be necessary when $\mathcal{D}$ is a product set. In such a case, irreduciblity and aperiodicity follow from the well-definedness of the full conditional distributions and from the Fubini's theorem (see the discussion following Corollary 1 of [126]). Since Harris recurrence follows from irreducibility, the induced Markov chain is ensured to be ergodic. For further details about covergence see e.g. [125, Ch. 10]

2.3.3.2 *Comparing Gibbs sampler and Metropolis Hastings*

As already mentioned, the Gibbs sampler can be considered as a special case of MH algorithm. In particular, regarding Algorithm 6 the following theorem [125, Th. 10.13] holds.

Theorem 4 *The Gibbs sampler method is equivalent to the composition of* d *Metropolis–Hastings algorithms, with acceptance probability uniformly equal to* 1.

Proof Let us consider the Gibbs sampler as the composition of d MH algorithms. Let us denote by $\mathbf{x}^i = (y_1, ...y_{i-1}, x_i, ..., x_d)$ and by $\mathbf{y}^i = (y_1, ..., y_i, x_{i+1}, ..., x_d)$ respectively the current and the proposed state at the i-th step of the current iteration. Then, the proposal distribution is given by

$$q_i(\mathbf{x}^i, \mathbf{y}^i) = \delta_{X_{\setminus i}}(y_1, ..., y_{i-1}, x_{i+1}, ..., x_d) f_{X_i}(y_i | y_1, ..., y_{i-1}, x_{i+1}, x_d)$$

where $\delta(\cdot)$ denotes the Dirac delta function. It follows that the acceptance probability is

$$\begin{aligned}
&\frac{f_{\mathbf{X}}(\mathbf{y}^i) q_i(\mathbf{y}^i, \mathbf{x}^i)}{f_{\mathbf{X}}(\mathbf{x}^i) q_i(\mathbf{x}^i, \mathbf{y}^i)} \\
&= \frac{f_{\mathbf{X}}(\mathbf{y}^i) f_{X_i}(x_i | y_1, ..., y_{i-1}, x_{i+1}, ..., x_d)}{f_{\mathbf{X}}(\mathbf{x}^i) f_{X_i}(y_i | y_1, ..., y_{i-1}, x_{i+1}, ..., x_d)} \\
&= \frac{f_{X_i}(y_i | y_1, ..., y_{i-1}, x_{i+1}, ..., x_d) f_{X_i}(x_i | y_1, ..., y_{i-1}, x_{i+1}, x_d)}{f_{X_i}(x_i | y_1, ..., y_{i-1}, x_{i+1}, ..., x_d) f_{X_i}(y_i | y_1, ..., y_{i-1}, x_{i+1}, ..., x_d)} \\
&= 1. \qquad (41)
\end{aligned}$$

□

Thus, from this theorem the Gibbs sampler seems to be more efficient than MH. In fact, it avoids wastes of computational resources accepting all the proposed values. Furthermore, it dispenses from the choice of the proposal distribution by deriving the conditional distributions from the target $f_{\mathbf{X}}(\cdot)$. However, it is worth norting that the acceptance probability in (41) concerns the transition from the state x_i to the state y_i only for the i-th component of the multivariate random variable $\mathbf{X}$. A complete step of the Gibbs sampler results from the compositions of d MH steps, one for each comptonent, and even though the acceptance probability equals 1 for each MH step, the global acceptance probability is usually different from 1 (see Example 10.14 in [125]). Note also that the d MH steps must be considerate jointly in order to build a convergent Markov chain on the joint space $\mathcal{D}$. In fact, although the sequence $\{\mathbf{X}^{(s)}\}_{s=1}^{S}$ forms a Markov chain enjoying the convergence properties described in Section 2.3.3.1, considering each MH step as a different state of the chain does not produce an irreducible Markov chain. In fact, at each MH step $i \in 1, ..., d$, the random variable $\mathbf{X}$ can take values in a sample space constrained by the values taken by the components $X_1, .., X_{i-1}, X_{i+1}, ..., X_d$.

2.3.4 *Metropolis within Gibbs*

The implementation of a Gibbs sampler requires the ability of sampling from the univariate full conditional distributions. In many cases these conditional distributions do not have a standard analytical form and their normalizing constants can be in turn intractable. In such cases, sampling from the full conditionals is infeasible unless one resorts to the *Metropolis within Gibbs* (MwG) algorithm. MwG is a *hybrid* MCMC algorithm, meaning that it uses simultaneously both Gibbs sampler and MH steps. In particular, MH steps should be used within the Gibbs scheme to sample from intractable full conditional distributions exploiting the fact that in the MH acceptance ratio the intractable normalizing constants cancel out.

Algorithm 7 Metropolis within Gibbs sampler

Initialize $x^{(0)}$
for $s = 1, ..., S$ **do**
  for $i = 1, ..., d$ **do**
    Draw $y_i \sim \tilde{q}(x_i^{(s-1)}, \cdot | x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_d^{(s-1)})$
    $u^{(s)} \sim \mathrm{Unif}[0, 1]$
    Compute

$$\alpha(x_i^{(s-1)}, y_i) = \min\left\{1, \frac{f_{X_i}(y_i | x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_d^{(s-1)}) \tilde{q}(y_i, x_i^{(s-1)} | x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_d^{(s-1)})}{f_{X_i}(x_i^{(s-1)} | x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_d^{(s-1)}) \tilde{q}(x_i^{(s-1)}, y_i | x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_d^{(s-1)})}\right\}$$

    if $u^{(s)} \leqslant \alpha(x_i^{(s-1)}, y_i)$ **then**
      Set $X_i^{(s)} = y_i$
    else
      $X_i^{(s)} = x_i^{(s-1)}$
    end if
  end for
end for

A general MwG scheme is displayed in Algorithm 7 in which is adopted a Gibbs scheme but samples from each full conditional distribution are got by means of a MH step. Note that the internal MH algorithm should be used only for those components that cannot be sampled directly from the full conditional.

2.3.5 *Convergence Diagnostics*

Theoretical foundations of MCMC algorithms ensure the ergodicity of the induced Markov chain and thus the convergence of the algorithm. However, from a practical point of view, it is often difficult to establish a *stopping rule* deciding when it is reasonable to consider samples as representative of the underlying stationary distribution. The notion of *convergence* related to MCMC methods is different from the one of other iterative methods: the output of the algorithm is not a single number or a probability distribution but rather represents a sample form an unknown probability distribution. Furthermore, the Markov nature of the sampling procedure slows the algorithm in its attempt to get samples from the target distribution. Accordingly, there are three different types of convergence to be assessed:

- convergence to the stationary distribution;
- convergence of averages;
- convergence to i.i.d. samples.

In practice, a lot of convergence diagnostics have been developed. Among all the diagnostics assessing the three types of convergence, we can distinguish between methods involving the simulation of a single chain and methods involving the simulation of multiple parallel chains.

The first type of convergence, the convergence to the stationary distribution, seems to be a minimal requirement since MCMC algorithms are supposed to get samples from the target distribution. However, despite the fact that from a theoretical point of view the target distribution corresponds to the stationary distribution, it is only a limiting distribution. Indeed, any inference from a finite number of simulations is an approximation and the quality of the approximation may also still be affected by the starting point. Accordingly, it is a standard practice to discard the initial iterations not providing good information about the target distribution. Thus, following this *burn-in* idea, according to [16, Ch. 6] the first part of the simulated sequences should be discarded. An early attempt of checking the convergence to the stationary distribution was presented by Gelfand and Smith [52]. The method is based on the comparison between the empirical distributions evaluated at nearly consecutive iterations and concludes that the convergence is achieved when the difference between the two is negligible. This strategy usually relies on a graphical comparison and requires the simulation of a single chain. A more formal comparison is based on nonparametric tests of stationarity. Examples are tests involving the Kolmogorov–Smirnov statistic [79] in the comparison between the first and the second halves of a single chain. Since this kind of nonparametric tests relies on the assumption of independence and identical distribution, in this framework one needs to correct the statistic by discarding batches of consecutive simulations thus leading to the construction of two quasi-independent subsamples (see e.g. [125, Sec. 12.2.2]).

As already noted, MCMC methods are often implemented to approximate integrals such as (7) via sample averages. Thus, we are often interested in evaluating the convergence of averages. A widespread diagnostic for the convergence of averages was presented by Gelman and Rubin [54]. This method is composed of two steps:

Step 1 Obtain an overdispersed estimate of the target distribution. Generate from it M different values to use as starting points for M parallel chains of length S.

Step 2 For the scalar quantity of interest computed from each simulated chain, say $\psi = h(X)$, compute the *between-chain* and *within-chain* variances, denoted by B and W, respectively.

The variance of the quantities of interest can be estimated by the following weighted average

$$\widehat{\mathbb{V}\mathrm{ar}}[\psi] = \frac{S-1}{S} W + \frac{1}{S} B. \tag{42}$$

To monitor the converge, one can evaluate the estimate of the *potential scale reduction*

$$\hat{R} = \sqrt{\frac{\widehat{\mathbb{V}\mathrm{ar}}[\psi]}{W}}. \tag{43}$$

As the length of each chain goes to infinity, $\hat{R}$ coverages to 1. Thus, further simulations may improve the quality of the approximation as long as the potential scale reduction is higher than 1. Note that the approximation in (43) is derived in [55] and differs from the one originally proposed in [54].

Another well-known diagnostic for the convergence of averages is the one introduced by Geweke [59]. Geweke's diagnostic is calculated by splitting a single chain

in two subsequences and by comparing the sample averages of the quantity of interest. The evaluation of the convergence is based on an adapted two-samples test on the means. The asymptotic standard errors involved in the computation of the test statistic are computed from the spectral density estimates for the two subsequences. A similar approach for assessing the convergence when the random variables $X^{(s)}$ take values in a discrete state space is based on a χ^2 test adjusted for the autocorrelation of the chain. Again, the method proceeds by partitioning the chain in subsequences and by testing the homogeneity of the empirical distribution of each subsequence and the empirical distribution of the whole chain [38]. For a more comprehensive review of the diagnostic methods we refer the reader to [28] and [125, Ch. 12].

A useful measure of convergence to an i.i.d sample is the ESS. It represents the effective number of independent samples thus indicating how close to be i.i.d the considered sample is.

2.3.5.1 *Effective Sample Size*

Consider the case in which we resort to MCMC algorithms to simulate a Markov chain $\{X^{(s)}\}_{s=1}^{S}$ for approximating the integral in (7). Let us denote by $\hat{I}_{MCMC} = 1/S \sum_{s=1}^{S} h(X^{(s)})$ the resulting sample average. Note that if the random variables $X^{(1)}, ..., X^{(S)}$ were sampled employing an i.i.d. MC sampling scheme, the variance of the resulting estimator would be

$$\mathbb{V}ar[\hat{I}_{MC}] = \frac{1}{S^2} \sum_{s=1}^{S} \mathbb{V}ar[h(X^{(s)})] = \frac{1}{S} \mathbb{V}ar[h(X^{(s)})]. \tag{44}$$

However, due to the autocorrelation in MCMC algorithms, the variance in (44) underestimates the desired variance. In fact, the asymptotic variance of the sum of correlated random is derived by

$$\begin{aligned} \lim_{S\to\infty} S\mathbb{V}ar[\hat{I}_{MCMC}] &= \mathbb{V}ar[h(X^{(s)})] + 2 \sum_{t=1}^{\infty} \mathbb{C}ov(h(X^{(s)}), h(X^{(s+t)})) \\ &= \mathbb{V}ar[h(X^{(s)})] + 2 \sum_{t=1}^{\infty} \rho_t \mathbb{V}ar[h(X^{(s)})] \\ &= \mathbb{V}ar[h(X^{(s)})](1 + 2 \sum_{t=1}^{\infty} \rho_t), \end{aligned} \tag{45}$$

where ρ_t is the autocorrelation at lag t. As already shown in Section 2.2.1 for IS , the ESS depends on the ratio between the variances of the two estimators:

$$ESS \triangleq \frac{S\mathbb{V}ar[\hat{I}_{MC}]}{\mathbb{V}ar[\hat{I}_{MCMC}]}. \tag{46}$$

Thus, by substituting (44) and (45) into (46) we obtain the following well-known equality [55, Section 11.5] [125, Section 12.3.5]:

$$ESS = \frac{S}{1 + 2 \sum_{t=1}^{\infty} \rho_t}. \tag{47}$$

However, in order to compute the ESS we need an approximation of the sum of the autocorrelations. Here, we give some details on the derivation of a computable approximation of the ESS reported in [55, Section 11.5].

First of all we define the *variogram*, V_t, which describes the degree of dependence between pairs of random variables at lag t, say $h(X^{(s)})$ and $h(X^{(s-t)})$:

$$\begin{aligned} V_t &\triangleq \mathbb{V}ar\big[h(X^{(s)}) - h(X^{(s-t)})\big] \\ &= \frac{1}{S}\sum_{s=t+1}^{S}\big(h(X^{(s)}) - h(X^{(s-t)})\big)^2, \end{aligned} \tag{48}$$

where (48) follows form the fact that each variable in the sequence has the same expected value. The variogram is related to the autocorrelation by the following equation:

$$\begin{aligned} V_t &= \mathbb{V}ar[h(X^{(s)})] + \mathbb{V}ar[h(X^{(s-t)})] - 2\mathbb{C}ov[h(X^{(s)}), h(X^{(s-t)})] \\ &= 2\mathbb{V}ar[h(X^{(s)})] - 2\rho_t\mathbb{V}ar[h(X^{(s)})] \\ &= 2\mathbb{V}ar[h(X^{(s)})]\big(1-\rho_t\big). \end{aligned}$$

Thus, by approximating $\mathbb{V}ar[h(X^{(s)})]$ as in (42), we obtain the estimate for the autocorrelation:

$$\hat{\rho}_t = 1 - \frac{V_t}{2\widehat{\mathbb{V}ar}[h(X)]}. \tag{49}$$

It follows that the ESS can be computed as

$$\mathrm{ESS} \approx \frac{S}{1 + 2\sum_{t=1}^{S}\hat{\rho}_t}. \tag{50}$$

Thus, an highly correlated Markov chain leads to a small ESS and poorly approximates an i.i.d. sample.

SAMPLE DEGENERACY As already noted, in MCMC methods the choice of the proposal distribution plays a key role since the acceptance ratio depends from it. When an MCMC algorithm leads to a low acceptance ratio the resulting Markov chain is highly autocorrelated. An high autocorrelation leads to highly variable estimates (see (45)) and to a low ESS. As already discussed, this problem is known as sample degeneracy.

2.3.6 MCMC *for sampling from the posterior distribution*

The described MCMC methods represent also a possible way of getting samples from an intractable posterior distribution. In particular, properly defined the proposal distribution on the parameter space, both the Metropolis and MH algorithm only require the ability of evaluating the unnormalized posterior distribution $l(\theta)$. In fact, looking

at the acceptance ratio we can see that the intractable normalizing constant cancels out:

$$\begin{aligned}\alpha(\theta^{(s)},\theta^{(s+1)}) &= \min\left\{1, \frac{\pi(\theta^{(s+1)}|x)q(\theta^{(s+1)},\theta^{(s)})}{\pi(\theta^{(s)}|x)q(\theta^{(s)},\theta^{(s+1)})}\right\} \\ &= \min\left\{1, \frac{l(\theta^{(s+1)})q(\theta^{(s+1)},\theta^{(s)})}{l(\theta^{(s)})q(\theta^{(s)},\theta^{(s+1)})}\right\} \\ &= \min\left\{1, \frac{\pi(\theta^{(s+1)})p(x|\theta^{(s+1)})q(\theta^{(s+1)},\theta^{(s)})}{\pi(\theta^{(s)})p(x|\theta^{(s+1)})q(\theta^{(s)},\theta^{(s+1)})}\right\}. \end{aligned} \tag{51}$$

In many cases the target of the inference is a vector of parameters, θ. In such cases, may be difficult (or impossible) to sample directly from the joint posterior distribution. The Gibbs sampler allows sampling from the target distribution by getting samples from the full conditional distribution of each unknown quantity. However, when sampling from one or more full conditional distributions is in turn prohibitive, one can resort to a MH step.

MCMC methods, as well as the other methods described in this chapter, requires the ability of evaluating pointwise the likelihood function. In many contexts they require also a great computational effort to achieve convergence, while the methods described in the next section are based on analytical asymptotic approximation.

2.4 ASYMPTOTIC APPROXIMATIONS FOR BAYESIAN INFERENCE

In this section we review some important asymptotic results for approximating the posterior distribution. These approximation methods are based on Taylor expansions and integrations over kernels of normal pdf. The key idea of these methods is that as the sample size increases the likelihood function becomes roughly normal and dominated by a unique mode. In contrast to MC methods, asymptotic methods rely on analytical approximations, thus requiring only differentiation and maximization procedures. Although the computation of Bayesian approximations is quite straightforward, they do not appear to be much used in the literature on applications of Bayesian methods, preference being given to MC methods [122].

2.4.1 *First order approximation*

The main asymptotic result in likelihood-based Bayesian inference is that the posterior distribution is asymptotically normal (see [82, Ch. 7 Th. 1]). To make this statement more formal, we introduce the following notation:

- $\mathbf{X} = X_1, ..., X_n$ is a vector of i.i.d. random variables distributed according to $p(\cdot|\theta)$;
- $\mathbf{x} = x_1, ..., x_n$ is a realization of $\mathbf{X}$;
- $p(\mathbf{x}|\theta)$ denotes the joint pdf evaluated at $\mathbf{x}$;
- $\mathcal{L}(\theta;\mathbf{x}) \stackrel{\triangle}{=} c(\mathbf{x})p(\mathbf{x}|\theta)$ is the likelihood function which equals the joint pdf up to a normalizing constant $c(\mathbf{x})$;

- $\ell(\theta;\mathbf{x}) \stackrel{\triangle}{=} \log \mathcal{L}(\theta;\mathbf{x})$ is the log-likelihood function;
- $\hat{\theta}$ is the maximum likelihood estimate (MLE);
- $j(\hat{\theta}) = -d^2\ell(\theta;\mathbf{x})/d\theta^2|_{\theta=\hat{\theta}}$ is the observed Fisher information;
- $W(\theta) = 2\{\ell(\hat{\theta};\mathbf{x}) - \ell(\theta;\mathbf{x})\}$ is the log-likelihood ratio.

More formally, the asymptotic result mentioned above can be written as

$$\lim_{n\to\infty} \int_{a_n}^{b_n} \pi(\theta|\mathbf{x})d\theta = \Phi(b) - \Phi(a),$$

where $a_n = a \cdot j(\hat{\theta})^{-1/2} + \hat{\theta}$ and $b_n = b \cdot j(\hat{\theta})^{-1/2} + \hat{\theta}$. Under regularity conditions for the prior distribution and conditions required for the normality of the maximum likelihood estimator, the asymptotic normality of the posterior distribution can be motivated as follows. By considering the Taylor expansion of the log-likelihood about the MLE we get

$$\begin{aligned}\ell(\theta;\mathbf{x}) &= \ell(\hat{\theta};\mathbf{x}) + \frac{d\ell(\theta;\mathbf{x})}{d\theta}\Big|_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + \frac{1}{2}\frac{d^2\ell(\theta;\mathbf{x})}{d\theta^2}\Big|_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2 + R \\ &\approx \ell(\hat{\theta};\mathbf{x}) - \frac{1}{2}j(\hat{\theta})(\theta-\hat{\theta})^2, \end{aligned} \tag{52}$$

where (52) follows from the fact that $\hat{\theta}$ maximizes the likelihood and R represents the reminder of order $n^{-1/2}$. Accordingly, the Bayes' formula can be written as

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta)\mathcal{L}(\theta;\mathbf{x}) = \exp\big[\log\pi(\theta) + \ell(\theta;\mathbf{x})\big]. \tag{53}$$

Accordingly, by substituting (52) into (53), the posterior distribution can be rewritten as

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \exp\big[\log\pi(\theta)\big]\exp\big[\ell(\hat{\theta};\mathbf{x}) - \frac{1}{2}j(\hat{\theta})(\theta-\hat{\theta})^2 + R\big] \\ &\propto \exp\big[\log\pi(\theta)\big]\exp\big[-\frac{1}{2}j(\hat{\theta})(\theta-\hat{\theta})^2 + R\big] \\ &\approx \exp\big[-\frac{1}{2}j(\hat{\theta})(\theta-\hat{\theta})^2\big], \end{aligned} \tag{54}$$

where, since $\ell(\theta;\mathbf{x})$ increases as n goes to infinity and $\log\pi(\theta)$ remains constant, this latter can be ignored when considering an asymptotic approximation. Note that (54) represents the kernel function of a Normal distribution, meaning that $\theta|\mathbf{x}$ is approximately distributed according to a $N(\hat{\theta}, j(\hat{\theta})^{-1})$. For a rigorous statement of this asymptotic result, a more detailed discussion of the arguments reported above and of the extensions to the case of a vector of parameters θ see e.g. [82, Ch 7].

The first order approximation allows approximating the marginal likelihood and integrals such as (3) without resorting to the simulation methods described until now. However, the quality of the approximation depends on the features of the posterior distributions and can be very low when the posterior distribution is asymmetric or skewed. In such cases, higher-order asymptotic approximations are needed to provide improvement.

2.4.2 *Higher-order asymptotic approximations*

One appealing feature of higher-order approximations is that they may be applied at little additional computational cost over simple first-order approximations and, at the same time, they allow avoiding the implementation of MCMC methods. In fact, their main advantage over the MCMC approach is that higher-order approximations can be achieved through independent samples, so that much less computational time is needed.

An higher-order approximation of the posterior distribution can be derived by resorting to the Laplace approximation of the marginal likelihood [150]. In particular, by considering the Taylor expansion of the integrand function about $\hat{\theta}$, one can reach an approximation of the posterior distribution of order n^{-1} in regions of the parameter space in which $|\theta - \hat{\theta}| < \delta/\sqrt{n}$, $\delta > 0$ [155]:

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \frac{\pi(\theta)\mathcal{L}(\theta;\mathbf{x})}{\int_{\Theta} \pi(\hat{\theta}) \exp\left[\ell(\hat{\theta};\mathbf{x}) - \frac{1}{2} j(\hat{\theta})(\theta-\hat{\theta})^2\right] d\theta} \\ &= \frac{1}{\sqrt{2\pi\, j(\hat{\theta})^{-1}}} \frac{\pi(\theta)}{\pi(\hat{\theta})} \exp\left[\ell(\theta;\mathbf{x}) - \ell(\hat{\theta};\mathbf{x})\right]. \end{aligned} \tag{55}$$

It follows that the intractable integrals required to compute the probability in the tails of the posterior distribution can be approximated to the same order according to the following arguments. Starting from the log-likelihood ratio, $W(\theta)$, we can rewrite the exponent in (55) as $\exp[-1/2 \cdot W(\theta)]$. By considering the *likelihood root* $r(\theta) \triangleq \text{sign}(\hat{\theta} - \theta) W(\theta)^{\frac{1}{2}}$, the tail area probability can be written as

$$\int_{\theta_0}^{\infty} \pi(\theta|\mathbf{x}) d\theta \approx \int_{\theta_0}^{\infty} \frac{1}{\sqrt{2\pi j(\hat{\theta})^{-1}}} \frac{\pi(\theta)}{\pi(\hat{\theta}) \exp\left[-\frac{1}{2} r(\theta)^2\right]} d\theta.$$

After two changes of the variable of integration (for further details see e.g. [155]), first from θ to $r(\theta)$ and then to $r^*(\theta) = r(\theta) - \frac{1}{r(\theta)} \log\left[|j(\hat{\theta})|^{1/2} \big(\pi(\theta)/\pi(\hat{\theta})\big) \big(r(\theta)/\ell'(\theta)\big)\right]$, we get

$$\int_{\theta_0}^{\infty} \pi(\theta|\mathbf{x}) d\theta \approx \int_{-\infty}^{r_0^*} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(r^*)^2\right] dr^* = \Phi(r_0^*),$$

where $r_0^* = r(\theta_0) + \frac{1}{r(\theta_0)} \log\left[|j(\hat{\theta})|^{-1/2} \big(\pi(\hat{\theta})/\pi(\theta)\big) \big(\ell'(\theta)/r(\theta_0)\big)\right]$. This result gives an approximation accurate to order $O(n^{-3/2})$ in regions where $|\theta - \hat{\theta}| < \delta/\sqrt{n}$ [137, Ch 2]. This approximation offers an alternative solution for the evaluation of probabilities involving the computation of intractable integrals. Following similar arguments, one can retrieve the analogous result when the posterior computation requires the marginalization w.r.t. nuisance parameters, as in (4). In particular, denoting by $\theta = (\psi, \lambda)$ the vector of parameters, where ψ is the parameter of interest and $\lambda \in \Lambda$ is the nuisance parameter (or a vector thereof), the marginal posterior distribution can be approximated through the Laplace approximation of the following integral

$$\pi(\psi|\mathbf{x}) = \int_{\Lambda} \pi(\psi, \lambda|\mathbf{x}) d\lambda,$$

leading to

$$\pi(\psi|\mathbf{x}) \approx \frac{1}{\sqrt{2\pi j_p(\hat{\psi})^{-1}}} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})} \exp\left[\ell_p(\psi; \mathbf{x}) - \ell_p(\hat{\psi}; \mathbf{x})\right] \frac{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}$$

where the subscript p denotes that the observed Fisher information and the log-likelihood are computed from the *profile likelihood* (see [29, Ch 3]) and $\hat{\lambda}_\psi$ is the constrained MLE of λ given ψ. Note that, when θ is a d-dimensional vector, the Fisher information computed from the full log-likelihood is given by a $d \times d$ matrix and $j_{\lambda,\lambda}$ denotes its (λ, λ)-block. Following arguments similar to the ones reported above, the probabilities in tail areas are approximated by (see [155])

$$\int_{\psi_0}^{\infty} \pi(\psi|\mathbf{x}) d\psi \approx \Phi\left(r_p^*(\psi_0)\right), \tag{56}$$

where $r^*(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)} \log\left[\frac{1}{r_p(\psi)} \ell_p'(\psi) |j_p(\hat{\psi})|^{-1/2} \frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi, \hat{\lambda}_\psi)} \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}\right]$.

2.4.3 *HOTA sampling scheme*

The first order approximation represents an analytical approximation of the posterior distribution and requires no simulation for evaluating the marginal likelihood. Moreover, expectation w.r.t. the posterior distribution, as in (3), can be computed at a little computational cost. Higher-order tail area approximations (HOTA) [132] require additional computational effort but still lead to a gain with respect to IS or MCMC methods. In fact, from (56) one can define a simple HOTA sampling scheme for getting a sample from $\pi(\psi|\mathbf{x})$. It is essentially an inverse sampling method [125, Ch 2] and can be summarized as follows:

1. Draw $z^{(s)} \sim N(0, 1)$;
2. Set $\psi^{(s)} = \psi^*$, with ψ^* such that $r_p^*(\psi^*) = z^{(s)}$

for $s \in \{1, ..., S\}$. Meaning that only a numerical procedure for solving the equation $r_p^*(\psi^*) = z^{(s)}$ is needed.

This method is often implemented to perform sensitivity analysis since it gives the same MC error also with different models and prior distributions (see e.g. [70, 123]). However, HOTA sampling schemes are affected both from the MC error and the asymptotic error. Stated otherwise, the quality of the approximation depends also on the sample size. Furthermore, the approximations are only available in models enjoying regularity conditions such as those characterized by posterior densities having a unique mode and being differentiable in regions with moderate deviations from the mode (see [69], for further details).

Finally, asymptotic approximations, as well as all the MC methods in the previous sections, require the ability of evaluating the likelihood function and cannot be implemented when the assumed model is intractable or computationally intensive to evaluate. In such a case, Bayesian inference can be conducted by resorting to other methods such as pseudo-likelihoods, composite-likelihoods [154] or ABC. This latter will be introduced in the following chapter and will be the focus of this thesis.

3

APPROXIMATE BAYESIAN COMPUTATION

ABC is a broad class of methods allowing Bayesian inference on parameters governing complex models. Since for such models the likelihood evaluation is typically infeasible, either analytically or numerically, the methods described in the previous chapter do not apply. In fact, all of them require a pointwise evaluation of the likelihood function to determine acceptance probabilities or importance weights. ABC methods dispense with exact likelihood computation and only require the ability of simulating pseudo-data by sampling observations from the assumed model employing a *generative model* a.k.a. *simulator*. A simulator can be thought as a probabilistic computer program taking as input a parameter value (or a vector thereof) $\theta \in \Theta$ and returning a sample from the distribution $p(\cdot|\theta)$. In general, no knowledge of the analytical form of the likelihood is necessary to write down such a program.

The original intuition comes from the interpretation of Bayes' Theorem provided by Rubin [129]: samples from the prior distribution are converted into samples from the posterior by collecting only those values that, when given as input to the generative model, produce pseudo-data exactly matching the observed data. More specifically, in the primal rejection ABC sampling scheme, whose origins can be traced back to Tavaré et al. [148] and Pritchard et al. [116], the following actions are taken:

1. $S \geqslant 1$ parameter values from the prior distribution $\pi(\cdot)$ are generated;
2. for each $s \in \{1, ..., S\}$, given the parameter proposal $\theta^{(s)}$ as input, the simulator generates a realization of a random variable $\mathbf{Y} \in \mathcal{X}^n$ distributed according to $p(\cdot|\theta^{(s)})$;
3. only parameter values leading to pseudo-data equal to the observed data are accepted, thereby samples from the exact posterior are derived by conditioning on the event $\{\mathbf{Y} = \mathbf{x}\}$.

However, the probability that the simulated and the observed data are identical is zero in the continuous setting and may be extremely small when $\mathcal{X}^n$ is a large discrete sample space. Thus, even in the discrete case, a very large number of simulations may be required to get an appreciable number of accepted parameters. Introducing a twofold approximation scheme, as illustrated in Algorithm 8, might increase the efficiency of the algorithm outlined above. First, one introduces a summary statistic, $s(\cdot)$, which is a function from the sample space $\mathcal{X}^n \subseteq \mathbb{R}^n$ to a lower-dimensional space $\mathcal{S} \subset \mathbb{R}^k$, with $k \ll n$. Second, exact matching of the simulated and the observed data is relaxed to similarity, expressed in terms of a predefined distance function $d(\cdot, \cdot)$ and tolerance threshold $\epsilon > 0$.

The underlying idea of ABC methods relies on the introduction of an auxiliary latent variable: the pseudo-data, $\mathbf{Y}$. As always, the aim of the method is getting samples from the posterior distribution in order to approximate quantities such as the integral in (3). Since sampling directly from

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta)p(\mathbf{x}|\theta)$$

Algorithm 8 R-ABC

for $s = 1, ..., S$ **do**
  Draw $\theta^{(s)} \sim \pi(\cdot)$
  Generate $\mathbf{y}^{(s)} \sim p(\cdot|\theta^{(s)})$ from the simulator
  Accept the pair $\left(\theta^{(s)}, s(\mathbf{y}^{(s)})\right)$ if $d\big(s(\mathbf{y}^{(s)}), s(\mathbf{x})\big) \leqslant \epsilon$
end for

is infeasible, samples from the desired posterior distribution are obtained by drawing from a joint posterior distribution defined on a augmented space and from which is easier to get samples:

$$\pi(\theta, \mathbf{y}|\mathbf{x}) \propto \pi(\theta)p(\mathbf{y}|\theta)\mathbb{1}\{\mathbf{y} = \mathbf{x}\}, \tag{57}$$

where $\mathbb{1}\{\mathbf{y} = \mathbf{x}\}$ is the indicator function assuming the value 1 if the pseudo-data equals the observed data and 0 otherwise.

The primal rejection scheme is just a way of sampling pairs $(\theta, \mathbf{y})$ from $\pi(\theta, \mathbf{y}|\mathbf{x})$. It follows that marginalizing out $\mathbf{y}$ in (57), that is ignoring the simulated data $\mathbf{y}$, the output of the algorithm becomes a sample from the *exact* posterior distribution

$$\pi(\theta|\mathbf{x}) = \int_{\mathcal{X}^n} \pi(\theta, \mathbf{y}|\mathbf{x}) = \int_{\mathcal{X}^n} \pi(\theta)\, p(\mathbf{y}|\theta)\, \mathbb{1}\{\mathbf{y} = \mathbf{x}\}\, d\mathbf{y}.$$

Abbreviating $s(\mathbf{y})$ by $s_\mathbf{y}$ and $s(\mathbf{x})$ by $s_\mathbf{x}$, the output of the Algorithm 8 is a sample of pairs $(\theta, s_\mathbf{y})$ from an *approximate* joint posterior distribution due to the data compression in summary statistics and the relaxation of the equality constraint. In particular, Algorithm 8 provides samples from the following joint approximate posterior distribution:

$$\tilde{\pi}_{\epsilon,d}(\theta, s_\mathbf{y}|s_\mathbf{x}) \propto \pi(\theta)\, p(s_\mathbf{y}|\theta)\, \mathbb{1}\{d(s_\mathbf{y}, s_\mathbf{x}) \leqslant \epsilon\}. \tag{58}$$

Marginalizing out $s_\mathbf{y}$ in (58), that is, ignoring the simulated summary statistics, the output of Algorithm 8 becomes a sample from the following *approximate* marginal posterior distribution

$$\tilde{\pi}_{\epsilon,d}(\theta|s_\mathbf{x}) \propto \int_S \pi(\theta)\, p(s_\mathbf{y}|\theta)\, \mathbb{1}\{d(s_\mathbf{y}, s_\mathbf{x}) \leqslant \epsilon\}\, ds_\mathbf{y}. \tag{59}$$

Note that the approximate posterior distribution can be also written as

$$\begin{aligned}\tilde{\pi}_{\epsilon,d}(\theta|s_\mathbf{x}) &\propto \pi(\theta) \int_S p(s_\mathbf{y}|\theta)\, \mathbb{1}\{d(s_\mathbf{y}, s_\mathbf{x}) \leqslant \epsilon\}\, ds_\mathbf{y} \\ &= \pi(\theta) \cdot \Pr\left(d(s_\mathbf{Y}, s_\mathbf{x}) \leqslant \epsilon|\theta\right).\end{aligned} \tag{60}$$

The probability $\Pr\left(d(s_\mathbf{Y}, s_\mathbf{x}) \leqslant \epsilon|\,\theta\right)$, where $s(\mathbf{Y})$ is referred to as $s_\mathbf{Y}$ for short, is the probability of accepting a simulation given the parameter value and approximates the likelihood. In fact, from (60) follows that the ABC *approximate likelihood* can be written as

$$\tilde{\mathcal{L}}_{\epsilon,d}(\theta; s_\mathbf{x}) = \int_S p(s_\mathbf{y}|\theta)\, \mathbb{1}\{d(s_\mathbf{y}, s_\mathbf{x}) \leqslant \epsilon\}\, ds_\mathbf{y}.$$

Note that as $\epsilon \to 0$, the approximate posterior distribution converges to the true posterior distribution obteined conditiong on the observed summary statistics s_x [140]:

$$\lim_{\epsilon \to 0} \pi(\theta) \int_{\mathcal{S}} p(s_y|\theta)\, \mathbb{1}\{d(s_y, s_x) \leqslant \epsilon\}\, ds_y = \pi(\theta) \int_{\mathcal{S}} \delta_{s_x}(s_y) p(s_y|\theta)\, ds_y \propto \pi(\theta|s_x).$$

See also [112, Appendix A, p. 832] for a proof of the ABC likelihood convergence.

The accuracy of the approximation of the posterior distribution depends both on how much information about the parameters is preserved by the summary statistics and on the magnitude of the threshold ϵ. In fact, as long as sufficient summary statistics for θ are chosen, the approximate posterior distribution $\tilde{\pi}(\cdot|s_x)$ converges to the true posterior $\pi(\cdot|x)$ (see [140, Ch. 1]). On the other hand, as $\epsilon \to \infty$, the probability $\Pr\big(d(s_Y, s_x) \leqslant \epsilon|\theta\big)$ approaches to 1 and samples are generated from the prior distribution. This establishes a trade-off between the statistical bias and the computational efficiency [83]: as the tolerance level ϵ decreases, the error of the approximation of the ABC posterior vs. the true posterior decreases at the cost of higher computational effort.

Remark 1 (Computational cost) *The evaluation of the computational efficiency depends on two alternative termination criteria: 1) stop when S values have been proposed; 2) stop when S parameter proposals have been accepted. In the first case the running time does not depend on ϵ but an evaluation of the computational efficiency may be done looking at the resulting sample size. Otherwise the running time and the number of proposed values are indicative of the algorithm computational cost. For a complete analysis of the asymptotic effects of ϵ both on the computational cost and on the bias, see [156, Ch. 1, 2].*

A discussion about the choice of the threshold ϵ and the summary statistics is provided in the next section.

As pointed out in [140, Ch. 1], the use of the indicator function does not enable one to discriminate between whether the pseudo-data y coincides with the observed data and whether y is just close enough. This may lead to a waste of information. For this reason, the indicator function in (58) is often replaced by

$$K_\epsilon\big(d(s_y, s_x)\big) = \begin{cases} \kappa\big(d(s_y, s_x)\big) & \text{if } d(s_y, s_x) \leqslant \epsilon \\ 0 & \text{if } d(s_y, s_x) > \epsilon \end{cases} \tag{61}$$

where $\kappa(\,\cdot\,)$ is a kernel function (e.g., triangular, Epanechnikov, Gaussian, etc.) defined on a compact support and decaying continuously from 1 to 0 (see e.g. [6]).

Now the ABC approximate likelihood becomes the convolution of the true likelihood with the kernel K_ϵ [112]:

$$\tilde{\mathcal{L}}_{\epsilon,d,K}(\theta; s_x) = \int_{\mathcal{S}} p(s_y|\theta)\, K_\epsilon(\, d(s_y, s_x)\,)\, ds_y \tag{62}$$

leading to the general approximate posterior distribution

$$\tilde{\pi}_{\epsilon,d,K}(\theta|s_x) \propto \pi(\theta)\tilde{\mathcal{L}}_{\epsilon,d,K}(\theta; s_x). \tag{63}$$

Note that this general setting encompasses also the case of R-ABC employing the uniform kernel. For the sake of an easier notation, hereafter we omit the indexes of the sources of approximation and denotes the approximate posterior by $\tilde{\pi}(\cdot|s_x)$.

In the literature, a variety of methods for sampling from (63) have been proposed. Some of them are known as *Marginal samplers* (e.g., [93, 141], etc.) since they allow directly sampling from the approximate marginal posterior distribution $\tilde{\pi}(\theta|s_x)$. The key idea is that $\tilde{\pi}(\theta|s_x)$ can be estimated pointwise as

$$\pi(\theta^{(s)}) \cdot \frac{1}{M} \sum_{i=1}^{M} K_\epsilon\{d(s_y^{(i)}, s_x) \leqslant \epsilon\} \qquad \forall s \in \{1, ..., S\} \tag{64}$$

by simulating M pseudo-datasets from $p(\cdot|\theta^{(s)})$ and computing $s_y^{(i)}$ for $i \in 1, ..., M$ at each iteration s. As is apparent, the second term in (64) provides a Monte Carlo estimate of the ABC approximate likelihood in (62). Instead marginalizing the output of Algorithm 8 corresponds to the implementation of a marginal sampler with $M = 1$. In such case the indicator function represents a crude Monte Carlo estimate of the probability $\Pr\left(d(s_Y, s_x) \leqslant \epsilon|\theta\right)$.

3.1 SUMMARY STATISTICS AN TOLERANCE THRESHOLD

A crucial point in the implementation of ABC algorithms is the choice of an adequate threshold and suitable summary statistics. Regarding the tolerance threshold, a practical rule is to choose the α-th quantile of the empirical distribution of the distances between the observed and the simulated summary statistics [6].

The quality of the posterior distribution depends also on the amount of preserved information about the parameters. However, low-dimensional sufficient summary statistics are often unavailable and preserving a great amount of information by involving many non-sufficient summaries does not represent a clever solution. Indeed, the opportunities for random discrepancies between s_y and s_x increase as the size of the summaries increases. This is a weak point in ABC methods as stated by Beaumont et al. [6]:

> A crucial limitation of the...method is that only a small number of summary statistics can usually be handled. Otherwise, either acceptance rates become prohibitively low or the tolerance...must be increased, which can distort the approximation.

It follows that a good choice of ABC summary statistics must strike a balance between low dimension and informativeness. The summary selection methods have been extensively discussed in the literature and, according to the overview given in [140, Ch. 5], can be classified in three categories: i) subset selection ii) projection and iii) auxiliary likelihood.

Methods based on subset selection start from a vector of candidate summary statistics, say $z = (z_1, ..., z_k)$, and try to select an informative minimal subset. Some of them (e.g. [4, 67, 105]) require to run ABC with different possible subsets in order to find the best among them according to a specific criterion. Joice and Marjoram [67] presented a stepwise selection based on an *approximate sufficiency test* criterion. The key

idea is that, given the minimum size subset of sufficient statistics, adding other summaries does not affect the approximation of the posterior distribution. Accordingly, this approach consists in testing whether changing the subset lead to a significantly different approximation of the posterior distribution. A different approach was proposed by Nunes and Balding [105]. They suggested to measure the informativeness of the ABC posterior distribution by computing its entropy and to select the subset of z which minimises it. However, as pointed out by Blum et al. [11], given a particularly precise prior the correct posterior may be more diffuse. In such cases, the ABC posterior having the smaller entropy does not correspond to more accurate inference. Barnes et al. [4] found out that a null Kullback–Leibler divergence between $\pi(\theta|s_y)$ and $\pi(\theta|x)$ represents a necessary condition for sufficiency of s_y. Accordingly, their stepwise selection method adds statistics to the actual subset one-by-one until the Kullback–Leibler is lower than a predefined positive threshold. Sedki and Pudlo [136] and Blum et al. [11] proposed a method that, when compared to the methods mentioned above, has the advantage of not requiring to run ABC many different times. It represents a regularisation procedure based on the idea of fitting a linear regression with response θ and covariates z based on training data and performs a variable selection to find an informative subset of z.

Projection methods start with a vector of summaries z as well. However, here the idea is to find an informative lower-dimensional projection of z, e.g., through a linear transformation. Among others, Fearnhead and Prangle [48] proposed to fit the following linear model to the training data:

$$\theta \sim N(Az + b, \Sigma).$$

The resulting vector of parameter estimates, $\hat{\theta} = \hat{A}z + \hat{b}$, is used as ABC summary statistics.

A completely different perspective is that of the methods based on auxiliary likelihoods. This approach is similar in spirit to another likelihood-free method: the *indirect inference* approach [62]. In fact, the key idea is to specify an auxiliary tractable model and to derive summary statistics from it. A possibility is to use the MLE of the auxiliary model as summary statistic [43, 60]. This idea is supported by the fact that the MLE is typically asymptotically sufficient for the auxiliary model and the same holds for the intractable model whether this latter is nested in the auxiliary model. Despite the fact that this kind of tractable auxiliary model are often unavailable, this approach still reasonable even without asymptotic sufficiency since Bayesian consistency can be attained [95]. Soubeyrandand and Haon-Lasportes [144] discussed the properties of the posterior distributions conditional on MLE estimates obtained by maximizing pseudo-likelihood functions built by ignoring some dependence structures in the data. Gleim and Pigorsch [60] suggested to use the score function of the auxiliary model as summary statistic. Ruli et al. [133] showed that the ABC posterior distribution, when based on a suitably rescaled score function derived form the true likelihood, converges to the true posterior as $\epsilon \to 0$. Moreover, the size of the summary statistics equals the number of the parameters and the resulting posterior distribution is also invariant to re-parametrization. However, being the true likelihood function intractable, they propose the cs-ABC in which a rescaled score function is derived from a *composite likelihood* (see [154], for a review on composite likelihood methods). In [134], the cs-ABC was included in a more general framework in which they proposed

Algorithm 9 General Rejection Sampling ABC

for $s = 1, ..., S$ **do**
Draw $\theta^{(s)} \sim g(\cdot)$
Generate $\mathbf{y}^{(s)} \sim p(\cdot|\theta^{(s)})$ from the simulator
Accept the pair $(\theta^{(s)}, s_{\mathbf{y}}^{(s)})$ with probability

$$\frac{\pi(\theta^{(s)}) K_\epsilon(d(s_{\mathbf{y}}^{(s)}, s_{\mathbf{x}}))}{M g(\theta^{(s)})}$$

end for

to get a robust ABC inference involving robust estimating functions (e.g., the score function or the composite score function) as summaries. They also showed that the approach based on the estimating functions use the same information as the method based on the MLE but resorting to different distance metrics. Hence, as $\epsilon \to 0$ both methods converge to the posterior conditional on the MLE.

3.2 SOME ABC SAMPLING SCHEMES

For almost all the existing MC methods a likelihood-free version for sampling from the approximate posterior distribution have been implemented. Here we focus on the ABC version of the sampling schemes described in the previous chapter. For further details on ABC sampling schemes we refer the reader to [140, Ch. 4].

3.2.1 *Rejection Sampling* ABC

The Rejection Sampling ABC is a sampling scheme allowing to get samples from the approximate posterior distribution by resorting to an easier-to-sample distribution. In particular, in order to get samples form $\tilde{\pi}(\theta, s_{\mathbf{y}}|s_{\mathbf{x}})$, we define the following proposal distribution on the joint space $\Theta \times \mathcal{S}$:

$$g(\theta, s) = g(\theta) p(s|\theta).$$

As described in Section 2.1 for the standard RS, also the likelihood-free version of RS requires an envelope function, $m(\cdot, \cdot)$, satisfying

$$m(\theta, s) = M g(\theta, s) \geqslant \tilde{\pi}(\theta, s|s_{\mathbf{x}})$$

for each pair $(\theta, s) \in \Theta \times \mathcal{S}$. Accordingly, M must be

$$\begin{aligned} M \geqslant & \max_{\theta, s} \frac{\pi(\theta) p(s|\theta) K_\epsilon(d(s, s_{\mathbf{x}}))}{g(\theta) p(s|\theta)} \\ = & K_\epsilon(0) \max_{\theta} \frac{\pi(\theta)}{g(\theta)}. \end{aligned} \tag{65}$$

Algorithm 10 IS-ABC

for $s = 1, ..., S$ **do**
Draw $\theta^{(s)} \sim q(\cdot)$
Generate $y^{(s)} \sim p(\cdot|\theta^{(s)})$ from the simulator
Set the IS weight for $(\theta^{(s)}, s_y^{(s)})$ to $\omega_s = K_\epsilon\big(d(s_y^{(s)}, s_x)\big) \cdot \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$.
end for

Following the same arguments as in Section 2.1, the likelihood-free RS can be implemented as displayed in Algorithm 9, where the intractable likelihood cancels out in the computation of the acceptance probability:

$$\begin{aligned} AP \triangleq \frac{\tilde{\pi}(\theta^{(s)}, s_y^{(s)}|s_x)}{g(\theta^{(s)}, s_y^{(s)})} &= \frac{\pi(\theta^{(s)})p(s_y^{(s)}|\theta^{(s)})K_\epsilon(d(s_y^{(s)}, s_x))}{Mg(\theta^{(s)})p(s_y^{(s)}|\theta^{(s)})} \\ &= \frac{\pi(\theta^{(s)})K_\epsilon(d(s_y^{(s)}, s_x))}{Mg(\theta^{(s)})}. \end{aligned} \tag{66}$$

Note that RS in Algorithm 9 encompasses also R-ABC scheme in Algorithm 8 where 1) the proposal distribution $g(\theta)$ corresponds to the prior distribution; 2) the kernel function $K_\epsilon(\cdot)$ corresponds to the indicator function $\mathbb{1}\{d(s_y, s_x) \leqslant \epsilon\}$; 3) $M = 1$ according to (65) .

3.2.2 *Importance Sampling* ABC

The efficiency of Rejection Sampling ABC, as the standard RS, depends on the choice of the optimal envelope function $m(\theta, s) = Mg(\theta, s)$. An inadequate envelope function can lead to a waste of computational effort due to a large number of rejections. To overcome this problem one can resort to an Importance Sampling scheme in which is assigned an importance weight to any proposal, thus avoiding rejections.

Following the standard Importance Sampling scheme introduced in Section 2.2, the IS-ABC displayed in Algorithm 10 consists of sampling pairs (θ, s_y) from an importance distribution, $q(\cdot, \cdot)$, and of weighting each pair avoiding the computation of the acceptance probabilities. In the ABC framework, the importance distribution can be set as

$$q(\theta, s) = q(\theta)p(s|\theta),$$

thus the parameter proposals are drawn from the importance distribution on the parameter space, $q(\cdot)$, and the pseudo-data is generated from the simulator. This implies that the importance weights do not depend from the intractable likelihood. In fact, denoted by Z the normalizing constant of the joint posterior, the resulting importance weights $\bar{\omega}(\theta^{(s)}, s_y^{(s)})$, referred to as $\bar{\omega}_s$ for short, are

$$\begin{aligned} \bar{\omega}_s &= \frac{\pi(\theta^{(s)})\ p(s_y^{(s)}|\theta^{(s)})\ K_\epsilon\big(d(s_y^{(s)}, s_x)\big)}{Z\, q(\theta^{(s)})\ p(s_y^{(s)}|\theta^{(s)})} \\ &= \frac{K_\epsilon\big(d(s_y^{(s)}, s_x)\big)}{Z} \cdot \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})} \qquad \forall s \in \{1, ..., S\}. \end{aligned}$$

By computing, at each iteration s, the following unnormalized weight

$$\omega_s = K_\epsilon\left(d(s_y^{(s)}, s_x)\right) \cdot \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})},$$

an approximation of the constant Z is obtained as

$$\begin{aligned} Z &= \int_\Theta \int_S \pi(\theta)\ \ p(s_y|\theta)\, K_\epsilon\left(d(s_y, s_x)\right) ds_y\, d\theta \\ &= \int_\Theta \int_S \omega(\theta, s_y)\, q(\theta, s_y)\, ds_y\, d\theta \approx \frac{1}{S}\sum_{s=1}^{S} \omega_s, \end{aligned}$$

where the second equality is got by multiplying and dividing by $q(\theta, s_y)$. It follows that, given an integrable function $h : \Theta \to \mathbb{R}$, the output of Algorithm 10 allows estimating posterior quantities such as

$$\mathbb{E}_{\tilde{\pi}}[h(\theta)] = \int_\Theta h(\theta)\tilde{\pi}(\theta|s_x)d\theta$$

by computing the sample average

$$\frac{1}{S\,Z}\sum_{s=1}^{S} \omega_s\, h(\theta^{(s)}) \approx \sum_{s=1}^{S} \tilde{\omega}_s\, h(\theta^{(s)}) \tag{67}$$

where each $\tilde{\omega}_s = \omega_s / \sum_{r=1}^{S} \omega_r$ is a normalized weight.

At each iteration, the importance weight depends on the distance $d(s_Y, s_x)$, hence on the random variable $s(\mathbf{Y})$ [140]. To wrap up, IS-ABC can be considered equivalent to a RW-IS in which $s(\mathbf{Y})$ represents the auxiliary latent variable and $K_\epsilon\left(d(s_y^{(s)}, s_x)\right)$ is a random estimate of the intractable likelihood. Accordingly, we can say that IS-ABC replaces the random estimate $\hat{p}_N(x|\theta)$ derived via IS in Section 2.2.3.1 with a crude MC estimate. Thus, as already noted in Section 2.2.3.1, being the likelihood replaced by a random estimate, IS-ABC leads to highly variable estimates and IS-ABC estimator results less efficient than the IS estimator one would obtain knowing the likelihood. Furthermore, being the crude MC estimate derived employing the evaluation of a kernel function defined on a compact support, it often assumes the value 0 leading to a null weight corresponding to an implicit rejection. As a result, in ABC framework the sample degeneracy problem becomes more serious. In fact, since IS-ABC leads to estimators characterized by a higher variance, the ESS value becomes smaller.

3.2.3 *Comparing Rejection Sampling and Importance Sampling* ABC

In order to compare Importance Sampling ABC and Rejection Sampling ABC, it is straightforward to show that also in the ABC framework the same arguments as in Section 2.2.2 hold. Thus, by properly defining a target and an importance distribution on an augmented space, one can show that the general Rejection Sampling ABC is a special case of the IS-ABC.

In particular, by introducing an auxiliary variable $u \in [0,1]$, as in Section 2.2.2, we can define the following target and instrumental distribution on the augmented space $\Theta \times \mathcal{S} \times [0,1]$:

$$\tilde{\pi}^*\big((\theta, s_y), u\big) = \begin{cases} M g(\theta, s_y) & \text{for } (\theta, s_y) \in \Theta \times \mathcal{S}\,, u \in \big[0, \mathrm{AP}\big] \\ 0 & \text{otherwise} \end{cases}$$

$$q^*\big((\theta, s_y), u\big) = \begin{cases} g(\theta, s_y) & \text{for } (\theta, s_y) \in \Theta \times \mathcal{S}\,, u \in \big[0, \mathrm{AP}\big] \\ 0 & \text{otherwise} \end{cases}.$$

It follows that the importance weights are

$$\omega^*\big((\theta, s_y), u\big) = \begin{cases} M & \text{for } (\theta, s_y) \in \Theta \times \mathcal{S}\,, u \in \big[0, \mathrm{AP}\big] \\ 0 & \text{otherwise} \end{cases}$$

where AP is the acceptance probability as defined in (66). Again the target distribution $\tilde{\pi}(\theta, s_y|s_x)$ is obtained by marginalizing $\tilde{\pi}^*\big((\theta, s_y), u\big)$ w.r.t. u:

$$\begin{aligned} \int_0^{\mathrm{AP}} M\, g(\theta, s_y)\, du &= M\, g(\theta, s_y) \int_0^{\mathrm{AP}} du \\ &= M\, g(\theta) p(s_y, \theta) \frac{\pi(\theta) K_\epsilon\big(d(s_y, s_x)\big)}{M\, g(\theta)} \\ &\propto \tilde{\pi}(\theta, s_y|s_x) \end{aligned}$$

and the resulting estimators are equivalent to the one of the Rejection Sampling ABC in Algorithm 9. Hence, according to Theorem 3 the IS-ABC involving $M g(\theta, s_y)$ as instrumental distribution is at least efficient as Algorithm 9.

Note also that R-ABC in Algorithm 8 corresponds to an IS on the augmented space with

$$\tilde{\pi}^*\big((\theta, s_y), u\big) = \begin{cases} \pi(\theta) p(s_y|\theta) & \text{for } (\theta, s_y) \in \Theta \times \mathcal{S}\,, u \in \big[0, \mathbb{1}\{d(s_y, s_x) \leqslant \epsilon\}\big] \\ 0 & \text{otherwise} \end{cases}$$

and

$$q^*\big((\theta, s_y), u\big) = \begin{cases} \pi(\theta) p(s_y|\theta) & \text{for } (\theta, s_y) \in \Theta \times \mathcal{S}\,, u \in \big[0, \mathbb{1}\{d(s_y, s_x) \leqslant \epsilon\}\big] \\ 0 & \text{otherwise} \end{cases}.$$

It follows that the importance weights are

$$\omega^*\big((\theta, s_y), u\big) = \begin{cases} 1 & \text{for } (\theta, s_y) \in \Theta \times \mathcal{S}\,, u \in \big[0, \mathbb{1}\{d(s_y, s_x) \leqslant \epsilon\}\big] \\ 0 & \text{otherwise} \end{cases},$$

meaning that they are equal to 1 whenever one gets $d(s_y, s_x) \leqslant \epsilon$, since u lies in $[0,1]$ by definition. On the other side, $\omega^*\big((\theta, s_y), u\big) = 0$ when $d(s_y, s_x) > \epsilon$. Summing up the primal Rejection Sampling is a special case of the IS-ABC with 1) the prior as instrumental distribution on the parametric space, i.e., $q(\theta) = \pi(\theta)$; 2) the indicator function $\mathbb{1}\{d(s_y, s_x) \leqslant \epsilon\}$ as kernel function; 3) the importance weights equal to 0 or 1 depending on the acceptance or rejection.

3.2.4 *Markov Chain Monte Carlo* ABC

Likelihood-free MCMC algorithms were introduced by Marjoram et al. in [93] as an answer to the inefficiency of the R-ABC in Algorithm 8. In fact, with non-informative priors, it does not take into account the information provided by the data at the proposal stage. This leads to proponing values located in region corresponding to low posterior probabiliies [92]. Marjoram et al. presented a MH algorithm defined as a marginal sampler: the Markov chain is defined on the parameter space Θ and the likelihood involved in the acceptance ratio is approximated by means of a MC estimate as in (64). Here, for the sake of consistency with the other sampling schemes, we adopt the same approach as in [140, Ch. 4] in which the MCMC-ABC provides samples from the approximated joint posterior distribution $\tilde{\pi}(\theta, s_y|s_x)$ by defining a Markov chain on the joint space $\mathcal{S} \times \Theta$. More specifically, Algorithm 11 requires the definition of a proposal distribution $\tilde{q}\big((\theta, s_y), (t, s)\big)$, where (θ, s_y) represents the state of the chain at the current iteration and (t, s) is the proposed pair. In the ABC framework the proposal is set to be

$$\tilde{q}\big((\theta, s_y), (t, s)\big) = \tilde{q}(\theta, t)p(s|t).$$

In fact, by resorting to this proposal distribution the acceptance ratio simplifies as follows

$$\begin{aligned} r((\theta, s_y), (t, s)) &\stackrel{\triangle}{=} \frac{\tilde{\pi}(t, s|s_x)\tilde{q}((t, s), (\theta, s_y))}{\tilde{\pi}(\theta, s_y|s_x)\tilde{q}((\theta, s_y), (t, s))} \\ &= \frac{\pi(t)p(s|t)K_\epsilon\big(d(s, s_x)\big)\tilde{q}(t, \theta)p(s_y|\theta)}{\pi(\theta)p(s_y|\theta)K_\epsilon\big(d(s_y, s_x)\big)\tilde{q}(\theta, t)p(s|t)} \\ &= \frac{\pi(t)K_\epsilon\big(d(s, s_x)\big)\tilde{q}(t, \theta)}{\pi(\theta)K_\epsilon\big(d(s_y, s_x)\big)\tilde{q}(\theta, t)} \end{aligned} \tag{68}$$

and does not depend on the intractable likelihood.

Accordingly, Algorithm 11 builds a Markov chain with the following transition kernel

$$q((\theta, s_y), (t, s)) = \begin{cases} \tilde{q}((\theta, s_y), (t, s))r((\theta, s_y), (t, s)) & \text{if } \tilde{\pi}(t, s|s_x) > \tilde{\pi}(\theta, s_y|s_x) \\ \tilde{q}((\theta, s_y), (t, s)) & \text{otherwise.} \end{cases}$$

Thus, the Detailed Balance condition w.r.t. to the target distribution, $\tilde{\pi}(\theta, s_y|s_x)$, is satisfied:

$$\begin{aligned} &\tilde{\pi}(\theta, s_y|s_x)q((\theta, s_y), (t, s)) \\ &= \tilde{\pi}(\theta, s_y|s_x)\tilde{q}((\theta, s_y), (t, s))r((\theta, s_y), (t, s)) \\ &= \tilde{\pi}(\theta, s_y|s_x)\tilde{q}((\theta, s_y), (t, s))\frac{\tilde{\pi}(t, s|s_x)\tilde{q}((t, s), (\theta, s_y))}{\tilde{\pi}(\theta, s_y|s_x)\tilde{q}((\theta, s_y), (t, s))} \\ &= \tilde{\pi}(t, s|s_x)q((t, s), (\theta, s_y)). \end{aligned}$$

It follows that the conclusions about the convergence drawn in Section 2.3 are still valid in the ABC framework.

An interesting aspect is that the acceptance ratio in (68) is equal to the one in (51) except that the likelihood is approximated by the kernel function defined on a

Algorithm 11 MCMC-ABC

Initialize $\theta^{(0)}$ and generates $\mathbf{y}^{(0)} \sim p(\cdot|\theta^{(0)})$
for $s = 1, ..., S$ **do**
  Draw $\theta^* \sim \tilde{q}(\theta^{(s-1)}, \cdot)$
  Draw $\mathbf{y}^* \sim p(\cdot|\theta^*)$ from the simulator and compute $s^*_{\mathbf{y}}$
  Compute

$$\alpha\big((\theta^{(s-1)}, s_{\mathbf{y}}^{(s-1)}), (\theta^*, s^*_{\mathbf{y}})\big) = \min\left\{1, \frac{\pi(\theta^*) K_\epsilon\big(d(s^*_{\mathbf{y}}, s_{\mathbf{x}})\big)\tilde{q}(\theta^*, \theta^{(s-1)})}{\pi(\theta^{(s-1)}) K_\epsilon\big(d(s_{\mathbf{y}}^{(s-1)}, s_{\mathbf{x}})\big)\tilde{q}(\theta^{(s-1)}, \theta^*)}\right\}$$

  Draw $u^{(s)} \sim \mathtt{Unif}[0, 1]$
  if $u^{(s)} \leqslant \alpha\big((\theta^{(s-1)}, s_{\mathbf{y}}^{(s-1)}), (\theta^*, s^*_{\mathbf{y}})\big)$ **then**
    Set $(\theta^{(s)}, s_{\mathbf{y}}^{(s)}) = (\theta^*, s^*_{\mathbf{y}})$
  else
    $(\theta^{(s)}, s_{\mathbf{y}}^{(s)}) = (\theta^{(s-1)}, s_{\mathbf{y}}^{(s-1)})$
  end if
end for

compact support. This means that in the ABC framework the acceptance probability can be equal to 0. In particular, it is equal to 0 whenever the distance between simulated and observed data exceeds the threshold. This leads to an implicit rejection step based on the comparison between the simulated and the observed data.

3.3 SAMPLE DEGENERACY IN ABC

As already pointed out, in the ABC framework the sample degeneracy problem becomes more serious. In fact, all the described sampling schemes involve an implicit rejection step relying on kernel functions defined on a compact support to approximate the likelihood.

Remark 2 (Likelihood approximation) *Recalling that the* ABC *approximate likelihood is defined as*

$$\tilde{\mathcal{L}}(\theta; s_{\mathbf{x}}) = \int_{\mathcal{S}} p(s_{\mathbf{y}}|\theta)\, K_\epsilon(\, d(s_{\mathbf{y}}, s_{\mathbf{x}})\,)\, ds_{\mathbf{y}}, \tag{69}$$

we note that its evaluation involves the computation of an integral on the space of the summary statistics. Replacing that integral with a random estimate corresponds to adopt the same strategy as in Section 2.2.3.1. The main difference is that in the ABC *framework we are able to get samples from the true likelihood* $p(\cdot|\theta)$ *by means of a simulator. This fact can be exploited to derive a* MC *estimate for such integral by sampling from* $p(\cdot|\theta)$ *and computing the following average:*

$$\frac{1}{M}\sum_{i=1}^{M} K_\epsilon(d(s_{\mathbf{y}}^{(i)}, s_{\mathbf{x}}) \leqslant \epsilon).$$

This is the estimate of the likelihood function involved in the marginal samplers. It can be proved that the more efficient choice is $M = 1$ *[13], thus* ABC *algorithms are usually*

implemented resorting to a crude MC *estimate. However, this choice implies that all the* ABC *algorithms implicitly involve a rejection step based on the comparison between a single pseudo-dataset and the observed data.*

The consequences of sample degeneracy in ABC methods are discussed in the next part of this thesis in which we propose a novel ABC method based on Large Deviations Theory.

Part II

IMPROVING ABC VIA LARGE DEVIATIONS THEORY

Ab assuetis non fit passio.

4

INTRODUCTION

In the literature, a variety of ABC methods have been proposed, see Sisson, Fan, and Beaumont [140, Ch. 4], and for recent reviews [68, 83]. In the vast majority of these methods, the approximate likelihood function takes positive values only when the distance between the simulated and the observed data is lower than a predefined threshold. In other words, most ABC schemes involve — implicitly or explicitly — a rejection step, which often leads to discarding a very large number of proposals. This results in a waste of computational resources and/or in an inadequate sample size, that is, in sample degeneracy. Sample degeneracy may also cause serious distortions in the form of the approximate posterior distribution. Indeed, accepting poor parameter proposals, i.e., those producing simulated data very rarely resembling the observed data, is a rare event. In the lack of accepted values, the posterior probability of such proposals will be approximated just as zero, in turn resulting in a distortion in the tails. This may be especially problematic for posterior distributions with long tails.

Our idea is to mitigate the problem of sample degeneracy by improving the approximation of the likelihood function. In particular, we speculate that taking into account the positive, however small, probability of rare events, i.e., poor proposals leading to simulated data resembling the observed data, allows avoiding the rejection step altogether; we instead weight all parameter proposals. To this end, we resort to Large Deviations Theory (LDT). Our aim is to show how LDT provides a convenient way to define an approximate likelihood, as well as guarantees of its convergence to the true likelihood as the size of the pseudo-dataset goes to infinity. In order to make the incorporation of LDT into ABC as smooth as possible, we rely on one of the less general formulations of Sanov's theorem and on its extension to finite state Markov chains. Accordingly, we only consider models for discrete random variables which, despite their apparent simplicity, will be shown to be of interest in several applications of ABC. This allows adopting a straightforward information theoretic formulation of LDT known as the *Method of Types* [27, 31]. Here, a type is basically an empirical distribution and, as an additional benefit of this approach, is a natural candidate for the summary statistics.

4.1 RELATED WORK

Generally speaking, inefficiency in ABC originates from the low probability of accepting certain parameter proposals. Thus, getting a proper sample size may require considerable computational effort. In the literature, there have been many proposals aimed at improving the computational efficiency of basic ABC. Prangle [111] proposed *Lazy* ABC, which saves computing time by abandoning simulations likely to lead to a poor match between the simulated and the observed data. To this end, at each iteration, the simulation is given up with a probability depending on its acceptance and on the expected required time for its completion.

Unlike our method, Lazy ABC does not avoid rejection, but rather accelerates the process leading to discarding a proposal.

Another way to improve computational efficiency is to consider proposal distributions closer to the posterior on the parameter space, employing sophisticated sampling methods, such as MCMC [93], Population Monte Carlo [7] and Sequential Monte Carlo [37]. In the same vein, Chiachio et al. [24] proposed a sequential way of achieving computational efficiency by overcoming the difficulties in getting samples resembling the observed data. This was the first attempt to improve the acceptance rate by adopting a rare-event approach. In particular, in [24], they combined the ABC scheme with a rare-event sampler that draws conditional samples from a nested sequence of subdomains. However, even this method cannot completely avoid rejections, and only partially mitigates the sample degeneracy problem.

In order to tackle the problem more systematically, clever proposal distributions should be combined with better approximations to the likelihood. Accordingly, Prangle [112] also resorted to a sequential approach, but explicitly considering a likelihood estimate that takes into account the probability of rare events. As a comparison, our method evaluates the probabilities of rare events are based on theoretical results (LDT), rather than on MC estimates of tail probabilities. Moreover, they focus on continuous data by showing that extensions to discrete data can be challenging and require application-specific solutions; in contrast the Method of Types provides a natural way of dealing with discrete random variables by summarizing data via empirical distributions, thus avoiding the common practice of summarizing data by selecting ad hoc summary statistics.

Other methods have been proposed avoiding the selection of summary statistics and relying on empirical distributions. In particular, Park et al. [108] rely on the maximum mean discrepancy between the embeddings of the simulated and the observed empirical distributions. They avoid rejection by weighting each parameter proposal by means of a kernel function defined on a non-compact support. Other interesting methods involve the Wasserstein distance [8] or the Kullback–Leibler divergence [65] as measure of discrepacy between the observed and the simulated data. In particular, Jiang [65] approximates the likelihood by means of an estimator of the Kullback–Leibler divergence between the unknown distribution of the data given the true parameter, and given the parameter sampled at the current iteration. Exploiting the fact that the maximum likelihood estimator is the one minimizing that Kullback–Leibler divergence, they prove that their approximate posterior distribution converges to a restriction of the prior distribution on the region in which the above mentioned divergence is smaller than a predefined threshold. Even though most of the above mentioned methods apply to continuous data, we note that ABC applications to discrete data appear frequently in population genetics, epidemiology, ecology and system biology (see [5] for an overview of the applications of ABC in these fields). In particular, in population genetics, discrete (possibly i.i.d.) data representing the genotyping at a few loci of different (unrelated) individuals have often been summarized through their empirical distributions (see [17, 93] among others).

A very different way of bypassing the selection of summary statistics relies on the random forest method [121]. Here, regression random forests are trained by using a training-set composed of a large number of parameter proposals and pseudo-data sampled from the prior distribution and the generative model, respectively. Since all

the summary statistics are involved as covariates, summary selection is avoided. The output of the algorithm is the predicted expected value of an arbitrary function of interest on the parameter space, conditional on the observed data.

Going beyond ABC, other methods addressing the intractability of the likelihood are the quasi-likelihood approximation [18], the Bayesian computation via empirical likelihood [96] and the Bayesian synthetic likelihood [115], aimed at reducing the number of simulations from the generative model [140, Ch 12]. The Bayesian empirical likelihood approach is based on the maximization of a likelihood built empirically under constraints on the moments. The quasi-likelihood approximation and the Bayesian synthetic likelihood can be considered as direct competitors of ABC, being simulator-based methods. As the empirical Likelihood approach, the quasi-likelihood approximation replaces the true likelihood with an estimate based on unbiased estimating functions, however here such a function is derived via simulations from the generative model as in the Bayesian synthetic likelihood approach. This latter derives a parametric approximated likelihood function from a normal density estimate for the summary statistics, with plug-in mean and covariance matrix obtained byMC simulations from the model. Thanks to the parametric approximation, this algorithm scales better as the dimension of summary statistics increases. However, it is still inherently dependent on the number of simulations required by the plugged-in mean and variance estimates. As involving a parametric auxiliary model, Bayesian synthetic likelihood is a member of the so called parametric Bayesian Indirect Likelihood (pBIL) class of methods [44]. ABC can instead be viewed as a member of the nonparametric Bayesian Indirect Likelihood (npBIL) class of methods.

Referring to the classification outlined in [44], our method can be placed in the class of nonparametric Bayesian Indirect Likelihood methods as well. In particular, it can be considered belonging to the class denoted as "npdBIL", being the non-parametric auxiliary model applied to the full dataset instead of summary statistics.

STRUCTURE OF PART II This part of the thesis is structured as follows. In Chapter 5 we introduce our method restricting our attention to i.i.d discrete data. In particular, in Section 5.1 we introduce LDT in the i.i.d. case by adopting the Method of Types. In Section 5.2 we show how LDT allows poor parameters proposals to contribute to the representation of the approximate posterior distribution. For the sake of an easy introduction, we firstly consider the case of a basic R-ABC, then we give two LD-ABC algorithms in Section 5.3. In Section 5.4 we give some details on the resolution of a practical computational difficulty and in Section 5.5 we illustrate the results obtained from several toy examples.

In Chapter 6 we extend the theory developed in the i.i.d case to finite state Markov chains. In Section 6.1 we introduce the notation and some preliminary concepts necessary for the extension of the Method of Types to finite state Markov chains, which is introduced in Section 6.2. The LD-ABC algorithms for Markov chains are given in Section 6.3 and tested at work in Section 6.4.

5

IMPROVING ABC VIA SANOV'S THEOREM

In the ABC sampling schemes described in Chapter 3, at each iteration s, the indicator function represents a crude estimate for the ABC approximate likelihood based on the estimate of the acceptance probability (see Remark 2). A possible approach to mitigate sample degeneracy is to provide a finer estimate for the ABC likelihood by evaluating that probability. In order to deal with rare events, we resort to LDT, which studies the exponential decay of the probabilities of such events. In particular, in this chapter we consider i.i.d. data and rely on one of the major results in LDT: Sanov's Theorem.

We speculate that taking into account the positive probability of large deviations events allows one to avoid rejection at all. This might provide a higher ESS, thus making the algorithm more efficient.

SET UP AND NOTATION For the sake of a smooth introduction of LDT into ABC from now on we will confine our attention to discrete random variables, and adopt an information theoretic point of view based on the Method of Types [27, 31]. In particular, we will assume that $\mathcal{X} = \{r_1, ..., r_{|\mathcal{X}|}\}$ is a finite, nonempty set. Moreover, $\mathcal{F} \stackrel{\triangle}{=} \{P(\cdot|\theta) : \theta \in \Theta\}$ is a family of probability mass function (pmf) on $\mathcal{X}$, where each $P(\cdot|\theta) = P_\theta$ has full support: $\text{supp}(P(\cdot|\theta)) \stackrel{\triangle}{=} \{r : P(r|\theta) > 0\} = \mathcal{X}$ for each $\theta \in \Theta$. We will let $\mathbf{X}^n = \{X_i\}_{i=1}^n$, $\mathbf{Y}^m = \{Y_i\}_{i=1}^m$ and so on denote sequences of i.i.d. random variables, distributed according to an (intractable) probability distribution $P_\theta \in \mathcal{F}$. Note also that $\log(\cdot)$ will denote the logarithm to base 2.

5.1 LARGE DEVIATIONS THEORY VIA METHOD OF TYPES

The Method of Types was fully developed by Csiszár and Körner [32], who derived the main theorems of information theory from this viewpoint. Key applications range from hypothesis testing to LDT. This powerful tool moves the focus from a sequence of observation, say $\mathbf{x}^n$, to its empirical distribution, the *type*.

More formally the type is defined as follows.

Definition 1 (Type) *Let $\mathbf{x}^n = (x_1, ..., x_n) \in \mathcal{X}^n$. The* type *of $\mathbf{x}^n$, written $T_{\mathbf{x}^n}$, is the probability distribution on $\mathcal{X}$ defined by*

$$T_{\mathbf{x}^n}(r) \stackrel{\triangle}{=} \frac{|\{i : x_i = r\}|}{n} \quad \forall r \in \mathcal{X}. \tag{70}$$

We let $\mathcal{T}^n$ denote the set of n-types, that is types with denominator n.

Note that the superscript n keeps track of the length of the sequence, which is also the denominator of the type. As is apparent, the type is a function summarizing the information included in the observed sequence $\mathbf{x}^n$ by mapping the n-dimensional observed sequence onto a $|\mathcal{X}|$-dimensional summary statistic.

The following quantities play a crucial role in the Method of Types. Below, we stipulate that $0 \cdot \log \frac{0}{r} \triangleq 0$ and that $r \cdot \log \frac{r}{0} \triangleq +\infty$ if $r > 0$. Given two probability distributions on $\mathcal{X}$, P and Q, we consider:

- the *entropy* of P, defined as

$$H(P) \triangleq -\sum_{r \in \mathcal{X}} P(r) \log P(r) ;$$

- the *Kullback–Leibler divergence* between P and Q, defined as

$$D(P\|Q) \triangleq \sum_{r \in \mathcal{X}} P(r) \quad \log \frac{P(r)}{Q(r)} .$$

With an abuse of notation, whenever the first argument of $D(\cdot\|Q)$ is a set of probability distributions, say E, $D(E\|Q)$ stands for $\inf_{P \in E} D(P\|Q)$. When $P^* = \operatorname{argmin}_{P \in E} D(P\|Q)$ exists, it is called the *information projection of* Q *onto* E.

Let $\mathbf{X}^n = \{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables, distributed according to $P_\theta \triangleq P(\cdot|\theta)$, for some $\theta \in \Theta$. In what follows, we let $\Pr(\cdot|\theta)$ be the probability measure on sequences induced by P_θ. The joint probability of n i.i.d. extractions x^n from P_θ, according to Proposition 5 can be written as :

$$\Pr(\mathbf{X}^n = x^n|\theta) = 2^{n\left(-D(T_{x^n}||P_\theta) - H(T_{x^n})\right)}. \tag{71}$$

From the Neyman–Fisher theorem [30, Ch. 2.2] follows that types are always sufficient statistics for θ, whatever P_θ.

Remark 3 (Types and ABC) *While the number of sequences of length n is exponential in n, it is easy to show that the cardinality of the set of types with denominator n, $\mathcal{T}^n$, is polynomial in n; in fact, $|\mathcal{T}^n| \leqslant (n+1)^{|\mathcal{X}|}$, see [27, Ch.11]. From an* ABC *perspective, it follows that using types as a summary statistics could mitigate the computational problems related to the comparison between the observed dataset and the pseudo-data, especially for large n. Furthermore, summarizing data through their empirical distributions is a way of overcoming the difficulties in finding sufficient statistics when P_θ is unknown (and $\Pr(\cdot|\theta)$ as well). Indeed, even when confined to discrete random variables, $P(\cdot|\,\theta)$ is an unknown model, not necessarily a Multinomial model, see Section 8.6 for examples. With no knowledge of the analytical form of the likelihood, finding sufficient summary statistics for θ, the vector of parameters given as an input to the simulator, is a central issue. In the literature there are several examples of models for conditionally independent discrete data in which the likelihood is analytically intractable and the required* ABC *method concerns empirical distributions. Examples are the* ABC *methods proposed in [66] and [17] to make inference on the mutation and selection parameters governing the Fisher–Wright model [49]. There, despite the conditional independence and the discreteness of the observations, the likelihood function is difficult to evaluate since the normalizing constant depends on the parameters: for small values of the selection parameter, numerical solutions have been found by Genz and Joice [57], in other cases, likelihood-free methods are required.*

Noting from (71) that the probability of the observed sequence decreases exponentially at a rate given by the Kullback–Leibler divergence between T_{x^n} and P_θ, we can say (informally) that a sequence x^n is *typical* if $D(T_{x^n}||P_\theta) < \delta$ for some small $\delta > 0$.

The Law of Large Numbers states that as the length of a typical sequence goes to infinity, its type converges in probability to P_θ. A proof is reported in Appendix 5.

Theorem 5 (Law of Large Numbers) *Let $X^n = \{X_i\}_{i=1}^n$ be a sequence of i.i.d random variables with $X_i \sim P_\theta$. Then for each $\delta > 0$*

$$\Pr(D(T_{X^n}||P_\theta) \leqslant \delta|\theta) \geqslant 1 - 2^{-n(\delta - |\mathcal{X}|\frac{\log(n+1)}{n})}.$$

Moreover,under $\Pr(\cdot|\theta)$, as $n \to \infty$, $D(T_{X^n}||P_\theta) \to 0$ with probability 1.

On the other hand, observing a sequence whose type is far from P_θ, called a *non-typical* sequence, is a rare event, and its probability obeys a fundamental result in LDT, Sanov's theorem; see [27, Th.11.4.1].

Theorem 6 (Sanov's Theorem) *Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables on $\mathcal{X}$ such that each $X_i \sim P_\theta$. Let $\Delta^{|\mathcal{X}|-1}$ be the simplex of probability distributions over $\mathcal{X}$ and let $E \subseteq \Delta^{|\mathcal{X}|-1}$. Then*

$$\Pr\left(T_{X^n} \in E \,|\, \theta\right) \leqslant (n+1)^{|\mathcal{X}|} 2^{-n D(P^*||P_\theta)}, \tag{72}$$

where $P^* = \underset{P \in E}{\operatorname{argmin}}\, D(P||P_\theta)$ is the information projection of P_θ onto E. Furthermore, if E is the closure of its interior,

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr(T_{X^n} \in E \,|\, \theta) = -D(E||P_\theta) = -D(P^*||P_\theta).$$

Suppose that E is composed of types of non-typical sequence. Then Sanov's theorem characterizes the exponential decrease rate of the probability of E. Taking into account this probability may provide a finer ABC approximation for the likelihood, as discussed in the next section.

5.2 LARGE DEVIATIONS THEORY IN ABC

In this section we provide a formal explanation of what is meant by "poor" parameter proposals and how they can contribute to the representation of the approximate posterior distribution by means of LDT. Suppose for simplicity that we are interested in obtaining an approximation of the posterior distribution, $\tilde{\pi}(\theta|x^n)$, via R-ABC or an equivalent IS-ABC (see Section 3.2.3) by assuming as given: a) the marginal importance density $q(\theta)$ to be the prior distribution on Θ; b) $\epsilon > 0$ as a threshold; c) types as summary statistics; d) the Kullback–Leibler divergence as distance function. For the sake of simplicity, from now on we will also assume T_{x^n} to be full support.

Given a budget of S iterations, both R-ABC and IS-ABC generate a sequence of pairs $(T_{y^m}^{(s)}, \theta^{(s)})$ with $s \in \{1, ..., S\}$. Each $T_{y^m}^{(s)}$ is an m-type resulting from a sequence of i.i.d. random variables, $Y^m = \{Y_j\}_{j=1}^m$, distributed according to $P\left(\cdot \,|\, \theta^{(s)}\right)$. We stress that the length of the simulated sequence, m, need not be equal to n, the length of the observed data sequence. Note also that, because of the independence assumption,

choosing $m = M \cdot n$ with $M \in \mathbb{N}$ means that the algorithm simulates M pseudo-datasets at each iteration like a marginal sampler.

Looking at Algorithm 8 and 10, being the whole pair $(\theta^{(s)}, T_{\mathbf{y}^m}^{(s)})$ accepted or rejected, one can define the *joint acceptance region* for these algorithms on the space $\Theta \times \mathcal{T}^m$. However, being the acceptance rule based only on the simulated type, regardless of the proposed parameter value, the acceptance region can be projected onto the probability simplex $\Delta^{|\mathcal{X}|-1} \supset \mathcal{T}^m$.

Definition 2 (Acceptance region) *Let $\Delta^{|\mathcal{X}|-1}$ be the simplex of probability distributions over $\mathcal{X}$ and let $T_{\mathbf{x}^n}$ be the type of the observed sequence $\mathbf{x}^n$. The* acceptance region $\mathcal{B}_\epsilon(T_{\mathbf{x}^n})$, *referred to as $\mathcal{B}_\epsilon$ for short, is defined for any $\epsilon \geqslant 0$, as*

$$\mathcal{B}_\epsilon \triangleq \left\{ P \in \Delta^{|\mathcal{X}|-1} : D(P||T_{\mathbf{x}^n}) \leqslant \epsilon \right\}$$

Now we can define a "poor" parameter proposal as a parameter $\theta^{(s)}$ such that $T_{\mathbf{x}^n}$ and the other types in the acceptance region are types of non-typical sequences w.r.t. $P(\cdot\,|\,\theta^{(s)})$. Accordingly, sampling a "poor" parameter means that there is a large divergence between $T_{\mathbf{x}^n}$ and $P(\cdot|\theta^{(s)})$. On the other hand, with m large enough, $T_{\mathbf{y}^m}^{(s)}$ is very likely to be close to $P(\cdot|\theta^{(s)})$, due to the Law of Large Numbers. Heuristically, this implies that the probability of simulating a sequence $\mathbf{y}^m$ whose type is in the acceptance region is very small. Recalling that in R-ABC and in IS-ABC a crude Monte Carlo estimate of the probability $\Pr(T_{\mathbf{Y}^m} \in \mathcal{B}_\epsilon|\,\theta^{(s)})$ is given by the indicator function $\mathbb{1}\{D(T_{\mathbf{y}^m}^{(s)}||T_{\mathbf{x}^n}) \leqslant \epsilon\}$, the vast majority of the "poor" parameter proposals are discarded altogether. Thus the posterior probability of "poor" parameters values is approximated as zero, even when the true posterior probability is strictly positive. We propose to mitigate this problem by assigning strictly positive weights to each proposal $\theta^{(s)}$, even if $T_{\mathbf{y}^m}^{(s)}$ is outside the acceptance region. To this end, we want to replace the indicator function with a finer estimate of the probability $\Pr(T_{\mathbf{Y}^m} \in \mathcal{B}_\epsilon|\,\theta^{(s)})$.

In principle, Sanov's theorem implies that, for m large enough, that probability can be approximated at each iteration by

$$\Pr\left(T_{\mathbf{Y}^m} \in \mathcal{B}_\epsilon \,|\, \theta^{(s)}\right) \approx 2^{-mD(\mathcal{B}_\epsilon||P_{\theta^{(s)}})}. \tag{73}$$

By substituting this probability to the indicator function in (58), the approximated posterior becomes

$$\tilde{\pi}(\theta, T_{\mathbf{y}^m}|T_{\mathbf{x}^n}) \propto \pi(\theta) P(T_{\mathbf{y}^m}|\theta) 2^{-mD(\mathcal{B}_\epsilon||P_\theta)}. \tag{74}$$

Unfortunately, the computation of the probability in (73) is still not feasible when the model $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ is unknown, as we do not know how to compute $D(\mathcal{B}_\epsilon||P_{\theta^{(s)}})$. The following theorem provides an asymptotic approximation to circumvent the problem. A proof is provided in Appendix D.1

Theorem 7 *Let $\mathbf{Y}^m = \{Y_j\}_{j=1}^m$ be a sequence of i.i.d. random variables taking values on the finite set $\mathcal{X} = \{r_1, ..., r_{|\mathcal{X}|}\}$, with each $Y_j \sim P_\theta$. Then under the measure $\Pr(\cdot|\theta)$*

$$\lim_{m\to\infty} D(\mathcal{B}_\epsilon||T_{\mathbf{Y}^m}) = D(\mathcal{B}_\epsilon||P_\theta) \quad \text{a.s.} \tag{75}$$

In essence, this result says that, as m increases and the type T_{y^m} converges to the distribution P_θ that has generated y^m, the information projection of T_{y^m} onto $\mathcal{B}_\epsilon$ converges to that of P_θ onto $\mathcal{B}_\epsilon$ (see Figure 2). From (73) and Theorem 7, for m large enough, $2^{-mD(\mathcal{B}_\epsilon||T_{y^m})}$ provides a feasible asymptotic estimate for the acceptance probability, $\Pr(T_{Y^m} \in \mathcal{B}_\epsilon \,|\, \theta)$. Replacing the indicator function in the ABC approximate posterior (58) with this estimate, we obtain the following new joint approximate posterior distribution:

$$\tilde{\pi}(\theta, T_{y^m}|T_{x^n}) \propto \pi(\theta) P(T_{y^m} \,|\, \theta) 2^{-mD(\mathcal{B}_\epsilon||T_{y^m})}. \tag{76}$$

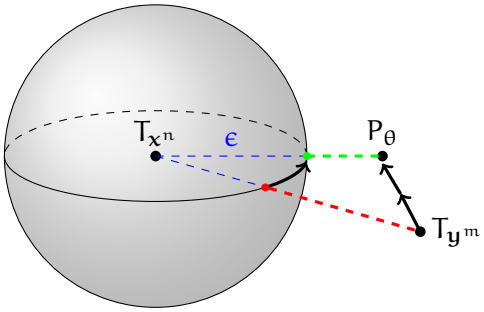


Figure 2: Acceptance region, $\mathcal{B}_\epsilon$, types, T_{x^n} and T_{y^m}, and the probability distribution P_θ that generated y^m. Asymptotically (as $m \to \infty$) T_{y^m} converges to P_θ and the distance $D(\mathcal{B}_\epsilon||T_{y^m})$ (red) converges to $D(\mathcal{B}_\epsilon||P_\theta)$ (green).

5.3 Large deviations approximate bayesian computation (LD-ABC)

The discussion in the previous section indicates that ABC methods can be improved by resorting to a better approximation for the likelihood. In particular, the (implicit) rejection step can be avoided by evaluating the positive probability of rare events via Sanov's theorem. Indeed, an easy way of sampling from (76) is a Large Deviations version of the ABC algorithms, which is what we will call *Large Deviations Approximate Bayesian Computation* (LD-ABC).

Here, we consider two different sampling schemes:

- an Importance Sampling LD-ABC;
- a Metropolis-Hastings LD-ABC.

In both the algorithms, the involved sufficient summary statistics are the types, the distance function is the Kullback–Leibler divergence and the kernel density function is defined by

$$K_{\epsilon,m}\left(T_{y^m}\right) = \begin{cases} 1 & \text{if } D(T_{y^m}||T_{x^n}) \leqslant \epsilon \\ 2^{-mD(\mathcal{B}_\epsilon||T_{y^m})} & \text{if } D(T_{y^m}||T_{x^n}) > \epsilon. \end{cases} \tag{77}$$

The output of the LD-ABC algorithms is a sample from the following approximate joint posterior distribution:

$$\tilde{\pi}(\theta, T_{y^m}|T_{x^n}) \propto \pi(\theta)\, K_{\epsilon,m}\left(T_{y^m}\right)\, P_\theta(T_{y^m}) \tag{78}$$

which, by marginalizing out simulated types, becomes

$$\tilde{\pi}(\theta|T_{x^n}) \propto \pi(\theta) \sum_{T_{y^m} \in \mathcal{T}^m} K_{\epsilon,m}\big(T_{y^m}\big)\, P_\theta(T_{y^m}) \tag{79}$$

where $\mathcal{T}^m$ denotes the set of the m-types, i.e., types with denominator m. Hence, the likelihood approximated by LD-ABC is

$$\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{x^n}) \stackrel{\triangle}{=} \sum_{T_{y^m} \in \mathcal{T}^m} K_{\epsilon,m}\big(T_{y^m}\big)\, P_\theta(T_{y^m})\,. \tag{80}$$

Note that the quality of the approximation depends both on the threshold ϵ and on the size of the pseudo-dataset, m. More precisely, the *adjustment* w.r.t. the likelihood approximate by R-ABC[1], here denoted $\tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{x^n}) \stackrel{\triangle}{=} \sum_{T_{y^m} \in \mathcal{B}_\epsilon} P_\theta(T_{y^m})$, depends on m and ϵ. In fact, from (77) and Definition 2, the approximate likelihood in (80) can be written as

$$\begin{aligned}\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{x^n}) &= \sum_{T_{y^m} \in \mathcal{B}_\epsilon} P_\theta(T_{y^m}) + \sum_{T_{y^m} \in \mathcal{B}^c_\epsilon} 2^{-mD(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) \\ &= \tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{x^n}) + \alpha_{\epsilon,m}(\theta)\end{aligned}$$

where the term $0 \leqslant \alpha_{\epsilon,m}(\theta) \leqslant 1$ is the adjustment. The following lemma gives an upper bound for that adjustment $\alpha_{\epsilon,m}(\theta)$, in two cases depending on P_θ.

Proposition 1 (The adjustment upper bound) *Let* $\alpha_{\epsilon,m}(\theta) = \tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{x^n}) - \tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; \mathcal{B}_\epsilon)$ *be the difference between the two likelihood functions approximated by* LD-ABC *and* R-ABC. *Let* $\mathcal{B}_\epsilon$ *be the* ABC *acceptance region and* $\mathring{\mathcal{B}}_\epsilon$ *its interior. We have the following upper bounds, depending on* θ, *which hold for all* $m \geqslant 1$.

(a) $P_\theta \in \mathring{\mathcal{B}}_\epsilon$. *Then* $D(\mathcal{B}^c_\epsilon || P_\theta) > 0$ *and* $\alpha_{\epsilon,m}(\theta) \leqslant (m+1)^{|\mathcal{X}|} 2^{-mD(\mathcal{B}^c_\epsilon || P_\theta)}$;

(b) $P_\theta \in \mathcal{B}^c_\epsilon$. *Let* $\gamma \stackrel{\triangle}{=} D(\mathcal{B}_\epsilon || P_\theta) > 0$. *Then there exists* $0 < \delta < \gamma$ *s.t.* $\alpha_{\epsilon,m}(\theta) \leqslant (m+1)^{|\mathcal{X}|} 2^{-m\delta}$.

Proof Let us consider the two cases separately, $P_\theta \in \mathring{\mathcal{B}}_\epsilon = \{P \in \Delta^{|\mathcal{X}|-1} : D(P||T_{x^n}) < \epsilon\}$ and $P_\theta \in \mathcal{B}^c_\epsilon = \{P \in \Delta^{|\mathcal{X}|-1} : D(P||T_{x^n}) > \epsilon\}$.

- $P_\theta \in \mathring{\mathcal{B}}_\epsilon$.

$$\alpha_{\epsilon,m} \leqslant \sum_{T_{y^m} \in \mathcal{B}^c_\epsilon} P_\theta(T_{y^m}) \leqslant (m+1)^{|\mathcal{X}|} 2^{-mD(\mathcal{B}^c_\epsilon || P_\theta)}$$

  where the last inequality follows from a direct application of Sanov's Theorem.

- $P_\theta \in \mathcal{B}^c_\epsilon$. Choose any $0 < \gamma' < \gamma \stackrel{\triangle}{=} D(\mathcal{B}^c_\epsilon || T_{x^n})$ (note that $\gamma > 0$) and apply Lemma D.1.1 with $E = \mathcal{B}_\epsilon$ and $Q = P_\theta$ to obtain $\delta > 0$ such that $D(\mathcal{B}_\epsilon || Q') \geqslant \gamma'$

1 Here we refer to an R-ABC involving types as summary statistics, the Kullback–Leibler divergence as distance function and the same tuning parameters, m and ϵ, as in the corresponding LD-ABC.

for each $Q' \in \mathcal{B}_\delta(P_\theta)$. We can assume without loss of generality that $\delta \leqslant \gamma'$. It follows that

$$\begin{aligned}
\alpha_{\epsilon,m} &= \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c} 2^{-mD(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) \\
&= \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} 2^{-mD(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-mD(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) \\
&\leqslant \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} P_\theta(T_{y^m}) + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-mD(\mathcal{B}_\epsilon || T_{y^m})} \\
&\leqslant \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} 2^{-mD(T_{y^m} || P_\theta)} + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-mD(\mathcal{B}_\epsilon || T_{y^m})} \qquad (81) \\
&\leqslant \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} 2^{-m\delta} + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-m\gamma'} \\
&\leqslant \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c} 2^{-m\delta} \\
&\leqslant (m+1)^{|\mathcal{X}|} 2^{-m\delta}
\end{aligned}$$

where (81) follows from (71) and the last step follows from an upper bound for the size of $\mathcal{T}^m$ (see [27, Ch. 11, Th. 11.1.1]).

□

5.3.1 *Importance Sampling* LD-ABC

Starting from the definition of an acceptance region satisfying the hypothesis of Sanov's theorem , as in Definition 2, a sample from the approximated posterior distribution $\tilde{\pi}(\theta, T_{y^m} | T_{x^n})$ can be obtained as illustrated in Algorithm 12.

Algorithm 12 LD-ABC Importance Sampling

for $s = 1, ..., S$ **do**
  Draw $\theta^{(s)} \sim q(\cdot)$
  Generate $\mathbf{Y}^m = \{Y_j\}_{j=1}^m$ with $Y_j \sim P(\cdot|\theta^{(s)})$ from the simulator
  if $D(T_{y^m}^{(s)} || T_{x^n}) \leqslant \epsilon$ **then**
    Set the IS weight for $(\theta^{(s)}, T_{y^m}^{(s)})$ to $\omega_s = \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$
  else
    Set the IS weights for $(\theta^{(s)}, T_{y^m}^{(s)})$ to

$$\omega_s = 2^{-mD(\mathcal{B}_\epsilon || T_{y^m}^{(s)})} \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$$

  end if
end for

Looking at Algorithm 12, it is apparent that this LD-ABC algorithm is a specialization of the more general IS-ABC. In fact, it relies on an instrumental distribution of the form

$$q(\theta, T_{\mathbf{y}^m}) = q(\theta)P(T_{\mathbf{y}^m}|\theta)$$

and returns a weighted sample with the following importance weights:

$$\frac{\pi(\theta^{(s)})P(T_{\mathbf{y}^m}^{(s)}|\theta^{(s)})K_{\epsilon,m}(T_{\mathbf{y}^m}^{(s)})}{q(\theta^{(s)})P(T_{\mathbf{y}^m}^{(s)}|\theta^{(s)})} = \frac{\pi(\theta^{(s)})K_{\epsilon,m}(T_{\mathbf{y}^m}^{(s)})}{q(\theta^{(s)})}. \tag{82}$$

However, in contrast with IS-ABC, at each iteration s, is assigned a positive weight to the proposed $\theta^{(s)}$ since the weight equals the value 0 only when $D(\mathcal{B}_\epsilon \| T_{\mathbf{y}^m}) = \infty$.

From Proposition 1 it follows that, as m goes to infinity, $\alpha_{\epsilon,m}(\theta) \to 0$ for almost all $\theta \in \Theta$. Therefore, the approximate likelihood from LD-ABC Importance Sampling achieves the approximate likelihood from R-ABC and preserves its asymptotic properties. Moreover, we speculate that LD-ABC Importance Sampling improves the efficiency of the standard IS-ABC by mitigating the sample degeneracy. An evaluation of the ESS in (25) might be a way of appreciating the improved induced by avoiding the implicit rejection.

Recalling that

$$\widehat{\mathrm{ESS}} = \frac{\left(\sum_{s=1}^{S} \omega_s\right)^2}{\sum_{s=1}^{S} \omega_s^2}$$

(with the proviso that $\widehat{\mathrm{ESS}} \triangleq 0$ if all ω_s's are zero). Let $\widehat{\mathrm{ESS}}_{\mathrm{IS}}$ and $\widehat{\mathrm{ESS}}_{\mathrm{LD}}$ be, respectively, the ESS achieved by S iterations of IS-ABC and LD-ABC by setting the same tuning parameters, distance function and importance density $q(\theta)$. Explicitly, let us assume that the kernel function for IS-ABC is 1 within the acceptance region $\mathcal{B}_\epsilon$ and 0 outside. Heuristically, adding positive weights increases the numerator more than the denominator in (25), suggesting that a non null weight assigned by LD-ABC to a parameter proposal rejected by IS-ABC is enough to have $\widehat{\mathrm{ESS}}_{\mathrm{LD}} > \widehat{\mathrm{ESS}}_{\mathrm{IS}}$. This is confirmed by the following simple result, whose proof is given in Appendix D.1.

Proposition 2 (Empirical ESS) *It holds that* $\widehat{\mathrm{ESS}}_{\mathrm{LD}} \geqslant \widehat{\mathrm{ESS}}_{\mathrm{IS}}$. *Moreover this inequality is strict, provided that in at least one iteration of the algorithm is generated a full support* $T_{\mathbf{y}^m}$ *falling outside* $\mathcal{B}_\epsilon$ *is generated.*

A tedious but straightforward analysis shows in fact that the event mentioned in the statement, upon which strict inequality holds, occurs with probability 1 as $S \longrightarrow +\infty$. The above result will be empirically validated in the experiments of Section 5.5, thus providing further evidence that LD-ABC achieves an improvement in terms of efficiency.

Below, we sum up the technical development so far with a discussion on the role of the parameters m and ϵ.

Remark 4 (On the role of the tuning parameters) *Concerning the role of* m *and* ϵ, *we can sum up the content of Propositions 1 and 2 as follows:*

1. *large* m *and small* ϵ *point to low* $\widehat{\text{ESS}}$ *and low* $\alpha_{\epsilon,m}$*;*

2. *small* m *and large* ϵ *point to high* $\widehat{\text{ESS}}$ *and high* $\alpha_{\epsilon,m}$*.*

If one regards $\widehat{\text{ESS}}$ *as a measure of efficiency, and* $\alpha_{\epsilon,m}$ *as a measure of (lack of) accuracy w.r.t. the* R-ABC *likelihood (but see also below), 1 and 2 above indicate how to trade off one for the other.*

In particular, as Theorem 7 requires a relatively large m *in order to get a good approximation for the posterior probability, 1 above says we can increase the tolerance* ϵ *to mitigate the resulting inefficiency. On the other hand, in cases where a small tolerance parameter* ϵ *is required, 2 above offers room to mitigate the resulting inefficiency, by decreasing* m*.*

Note, however, that when considering accuracy w.r.t. the target posterior density $\pi(\theta|T_{x^n})$*, the adjustment* $\alpha_{\epsilon,m}$ *cannot simply be regarded as a measure of imprecision: rather, it represents a compensation for those* θ*'s that would be assigned a too low probability by pure* R-ABC*. In this case, as sounder measure of precision can be obtained by directly comparing a kernel-estimated density (obtained with* LD-ABC *weights) and the target posterior density, e.g., in terms of the* mean integrated squared error (MISE)*. This measure is, however, impossible to evaluate analytically, since its calculation presupposes the knowledge of the target posterior density. From a more empirical point of view, further discussion of the consequences of different choices of* ϵ *and* m *on the performance of the posterior estimators is presented in Section 5.5, illustrated by a number of examples.*

5.3.2 *Metropolis-Hastings* LD-ABC

The Metropolis-Hastings sampling scheme represents an alternative for sampling from the approximate joint posterior distribution, $\tilde{\pi}(\theta, T_{y^m}|T_{x^n})$. Algorithm 13 is a LD version of the MCMC-ABC described in Section 3.2.4.

Algorithm 13 LD-ABC Metropolis-Hastings

Initialize $\theta^{(0)}$ and $y^{(0)}$
for $s = 1, ..., S$ **do**
  Draw $\theta^* \sim \tilde{q}\left(\theta^{(s-1)}, \cdot\right)$
  Generate Y^{m*} and compute $T^*_{y^m}$
  Compute $\alpha = \min\left\{1, \dfrac{\pi(\theta^*)K_{\epsilon,m}(T^*_{y^m})\tilde{q}\left(\theta^*, \theta^{(s-1)}\right)}{\pi(\theta^{(s-1)})K_{\epsilon,m}(T^{(s-1)}_{y^m})\tilde{q}\left(\theta^{(s-1)}, \theta^*\right)}\right\}$
  Draw $u \sim \text{Unif}[0, 1]$
  if $u < \alpha$ **then**
    Assign $(\theta^{(s)}, T^{(s)}_{y^m}) \leftarrow (\theta^*, T^*_{y^m})$ with
  else
    Assign $(\theta^{(s)}, T^{(s)}_{y^m}) \leftarrow (\theta^{(s-1)}, T^{(s-1)}_{y^m})$
  end if
end for

The LD-ABC Metropolis-Hastings builds a Markov chain on the joint space Θ $times\mathcal{T}^m$ by getting samples from the following proposal distribution:

$$\tilde{q}\left((\theta, T_{y^m}), (t, T))\right) = \tilde{q}(\theta, t) \cdot P_\theta(T),$$

where $\tilde{q}(\theta, \cdot)$ is the proposal distribution on the parameter space, (θ, T_{y^m}) represents the current state, the (t, T) the proposed state. The resulting Markov chain is characterized by the following transition probability

$$q((\theta, T_{y^m}), (t, T)) =$$
$$\begin{cases} \tilde{q}((\theta, T_{y^m}), (t, T)) r((\theta, T_{y^m}), (t, T)) & \text{if } \tilde{\pi}(t, T|T_{x^n}) > \tilde{\pi}(\theta, T_{y^m}|T_{x^n}) \\ \tilde{q}((\theta, T_{y^m}), (t, T)) & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} r((\theta, T_{y^m}), (t, T)) &= \frac{\tilde{\pi}(t, T|T_{x^n})\tilde{q}((t, T), (\theta, T_{y^m}))}{\tilde{\pi}(\theta, T_{y^m}|T_{x^n})\tilde{q}((\theta, T_{y^m}), (t, T))} \\ &= \frac{\tilde{\pi}(t)P(T|t)K_{\epsilon,m}(T)\tilde{q}(t, \theta)P(T_{y^m}|\theta)}{\tilde{\pi}(\theta)P(T_{y^m}|\theta)K_{\epsilon,m}(T_{y^m})\tilde{q}(\theta, t)P(T|t)} \\ &= \frac{\tilde{\pi}(t)K_{\epsilon,m}(T)\tilde{q}(t, \theta)}{\tilde{\pi}(\theta)K_{\epsilon,m}(T_{y^m})\tilde{q}(\theta, t)}. \end{aligned}$$

Note that the approximated posterior distribution satisfies the Detailed Balance condition:

$$\begin{aligned} &\tilde{\pi}(\theta, T_{y^m}|T_{x^n})q((\theta, T_{y^m}), (t, T)) \\ &= \tilde{\pi}(\theta, T_{y^m}|T_{x^n})\tilde{q}((\theta, T_{y^m}), (t, T))r((\theta, T_{y^m}), (t, T)) \qquad (83) \\ &= \tilde{\pi}(\theta, T_{y^m}|T_{x^n})\tilde{q}((\theta, T_{y^m}), (t, T))\frac{\tilde{\pi}(t, T|T_{x^n})\tilde{q}((t, T), (\theta, T_{y^m}))}{\tilde{\pi}(\theta, T_{y^m}|T_{x^n})\tilde{q}((\theta, T_{y^m}), (t, T))} \\ &= \tilde{\pi}(t, T|T_{x^n})q((t, T), (\theta, T_{y^m})). \qquad (84) \end{aligned}$$

It is worth noting that in the LD version of the MCMC-ABC the acceptance ratio $r((\theta, T_{y^m}), (t, T))$ assumes strictly positive values and does not lead to implicit rejections.

5.4 A COMPUTATIONAL ISSUE: THE MINIMIZATION OF THE KL-DIVERGENCE

In the proposed LD-ABC, the minimization of the KL divergence between the acceptance region and the simulated type poses a computational difficulty. This is a constrained minimization problem on a space of dimension $|\mathcal{X}|$. As $|\mathcal{X}|$ grows, this problem can rapidly become intractable.

A practical work-around to this problem can be found by considering a suitable path from T_{y^m} to T_{x^n}, passing through P^*. In *Information Geometry*, this path is represented by a linear interpolation on the logarithmic scale, the *exponential geodesic* [104].

Definition 3 (Exponential geodesic) *Let P_1 and P_2 be two probability distributions over $\mathcal{X}$ and let P_ξ be the probability distribution such that for each $r \in \mathcal{X}$*

$$\log P_\xi(r) = \xi \log P_1(r) + (1 - \xi) \log P_2(r) + \log c$$

where $\xi \in [0, 1]$ and c is a proper normalizing constant. The exponential geodesic *between P_1 and P_2 is the following set of distributions*

$$\gamma_e(P_1, P_2) \triangleq \{P_\xi : \xi \in [0, 1]\}. \qquad (85)$$

Min	Mean	Max	s.d.
$0.0052 \cdot 10^{-14}$	$1.0719 \cdot 10^{-6}$	0.0052	$5.3075 \cdot 10^{-7}$

Table 1: Summaries of the empirical distributions of the relative errors of the approximate distances.

Our approach when minimizing the KL divergence between $\mathcal{B}_\epsilon(T_{\mathbf{x}^n})$ and $T_{\mathbf{y}^m}$ is to focus on a path between the observed and the simulated type, that is the exponential geodesic $\gamma_e(T_{\mathbf{x}^n}, T_{\mathbf{y}^m})$. We search in this path the information projection P^*, or an approximation of it. This reduces the dimension of the minimization problem from $|\mathcal{X}|$ to 1, that of the parameter ξ. Specifically, let $P_{\xi^*} \in \mathcal{B}_\epsilon(T_{\mathbf{x}^n})$ be the element of $\gamma_e(T_{\mathbf{x}^n}, T_{\mathbf{y}})$ defined as

$$P^*_\xi(r) \stackrel{\triangle}{=} T_{\mathbf{x}^n}(r)^{\xi^*} \cdot T_{\mathbf{y}^m}(r)^{1-\xi^*} c^* \qquad (r \in \mathcal{X})$$

where $\xi^* \stackrel{\triangle}{=} \underset{\xi \in [0,1] : T_\xi \in \mathcal{B}_\epsilon}{\operatorname{argmin}} D(P_\xi \| T_{\mathbf{y}^m})$.

Hence, whatever $|\mathcal{X}|$, $D(\mathcal{B}_\epsilon \| T_{\mathbf{y}^m})$ is approximated by means of a minimization with respect to a single parameter, ξ.

In the experiments described in the next section the minimization procedure is implemented resorting to a Sequential Least Square Programming algorithm (SLSQP) optimizer available in the *scipy.optimize* package in Python. The optimizer uses a slightly modified version of Lawson and Hanson's nonlinear least-squares solver [77]. We have empirically verified that $D(P_{\xi^*} \| T_{\mathbf{y}^m})$ approximates with very good accuracy $D(\mathcal{B}_\epsilon \| T_{\mathbf{y}^m})$. Table 1 summarizes the distribution of the distances approximation relative errors w.r.t. the true distance, over the $S = 100,000$ simulations in the experiment in Section 5.5.2, with $m = 500$ and $\epsilon = 0.005$.

5.5 EXPERIMENTS

In order to evaluate the performance of the proposed method, we have put a proof-of-concept implementation of LD-ABC at work on two examples. We compare the results obtained from LD-ABC algorithms with those obtained from standard R-ABC and MCMC-ABC. To isolate the improvement introduced by LDT, we compare the standard algorithms with their LD counterpart by using the same tolerance threshold,the same summary statistic (the type), a kernel function assuming value 1 in the acceptance region (uniform kernel), the same proposal or importance distributions. Recall that by setting the importance distribution equal to the prior the Importance Sampling LD-ABC in Algorithm 12 represents the LD counterpart of the R-ABC in Algorithm 8. For both the examples, there is a closed form or a MCMC method for sampling from the exact posterior distribution, and the resulting posterior inference is taken as a reference for comparison. Recalling that our aim is to mitigate sample degeneracy and the resulting bias in the approximation of the posterior density in tail area, the comparison is mainly based on the evaluation of the ESS and of the MISE. We also consider the performance in terms of point estimators by computing the mean squared error (MSE).

Table 2: Squared errors, integrated squared errors and ESS averaged over 100 reruns, with $\epsilon = 0.01$, $m = 100$.

Algorithm	$\widehat{\text{MSE}}$	$\widehat{\text{MISE}}$	$\widehat{\text{ESS}}$
R-ABC	0.0002	0.6399	1,282
LD-IS-ABC	≈ 0	0.0153	3,051
MCMC-ABC	0.0002	0.6922	684
LD-MCMC-ABC	≈ 0	0.0203	1,750

5.5.1 *Example 1: Binomial distribution.*

Let $\mathbf{x}^{20}$ be a sample from i.i.d. Bernoulli random variables with parameter θ. Suppose that $\mathbf{x}^{20}$ has empirical distribution $T_{\mathbf{x}^n} = [0.3, 0.7]$. Assuming a uniform prior distribution for θ, the posterior distribution, $\pi(\theta|\mathbf{x}^{20})$, is a Beta distribution with parameters $\alpha = 15$ and $\beta = 7$.

We ran $S = 10,000$ iterations of IS-ABC with the proposal distribution equal to the prior and with both the uniform kernel and the proposed LD kernel. Note that in the first case the algorithm corresponds to the R-ABC and the only difference between R-ABC and its LD counterpart is the value assumed by the kernel function outside the acceptance region (inside both the kernels are uniform, see (77)). We also implemented the MCMC-ABC sampling scheme. For the sake of readability, we adopt the abbreviation LD, standing for Large Deviations, when is employed the kernel function is (77).

Since our speculation is that the evaluation of the probability of rare events provides a better approximation in the tail areas, we are interested in a comparison among the shapes of the approximate posterior distributions. Accordingly, besides the posterior mean of θ, we also approximate the posterior densities by means of a Gaussian kernel density estimation.

Figure 3 shows the posterior distributions and cumulative density functions (cdf) approximated by each algorithm. As it is apparent, the LD algorithms (blue lines) approximate better the true posterior (dashed grey line). Looking at the cdf's, we can see that using the standard algorithms (red lines) provide a worse approximation in the tail areas. Figure 4 shows also the 99% credible intervals for the estimated cdf's. Note that the estimates of the cdf achieved by standard R-ABC and MCMC-ABC are more variable then the estimate got via LD-ABC and LD-MCMC-ABC.

We can evaluate the point estimates of the posterior mean and the estimate of the posterior densities through MSE and the MISE. We also consider ESS as a measure of the degree of sample degeneracy. In Table 2 we report the estimated mean squared errors ($\widehat{\text{MSE}}$), the estimated mean integrated squared errors ($\widehat{\text{MISE}}$), and estimated effective sample size ($\widehat{\text{ESS}}$) computed by averaging the squared errors, the integrated squared errors and the ESS over 100 reruns of each algorithm. We can see that the proposed method outperforms the standard methods. Even though the quality of the point estimations is almost the same, the proposed kernel function leads to a clear improvement in terms of density estimation and ESS.

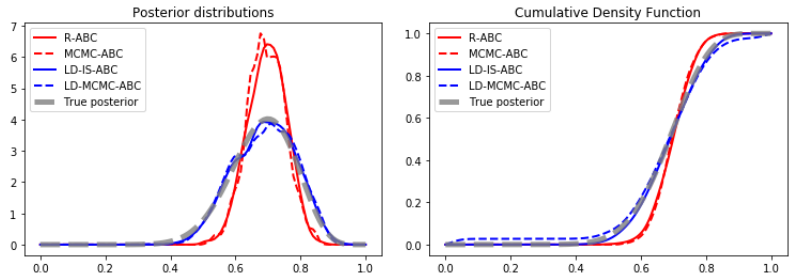


Figure 3: Posterior distributions (LHS) and posterior cumulative density functions (RHS) provided employing IS and MCMC schemes with the uniform kernel (red lines) and with the LD kernel (blue lines).

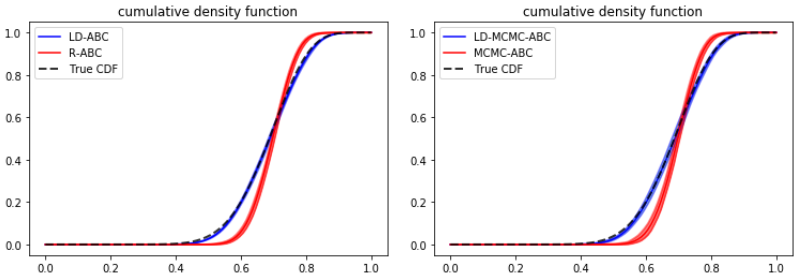


Figure 4: Posterior cumulative density functions with 99% credible intervals derived from 100 reruns of IS (RHS) and MCMC algorithms (LHS) with the uniform kernel (red) and the LD kernel (blue).

To evaluate the improvement induced by the LD kernel in the mixing of the chain induced by MCMC-ABC, we consider the distributions of the *sojourn time*. The sojourn time is defined by the number of consecutive iterations at which the sampled parameter remains above (or under) a given threshold [139]. When a MCMC sampler gets stuck in regions of low posterior density, the chain exhibits longer sojourn times above (or under) the threshold. In Figure 5, we can see that the LD-MCMC-ABC provides distributions of the sojourn time more concentrated around small values then MCMC-ABC, for each of the four considered threshold (0.55, 0.6,0.75 and 0.8). We can conclude that our method leads also to a better mixing of the chain in regions of the parameter space characterized by low probability posterior densities.

A possible drawback of the method is represented by the scalability af the algorithms with respect to the cardinality of $\mathcal{X}$. To investigate this feature, we tested both R-ABC and LD-ABC at work on five sequences of $n = 250$ random variables distributed according to a $\mathtt{Binomial}(\theta^{true}, N)$ distribution with parameters $\theta^{true} = 0.3$ and N equal to $3, 4, 5, 6$ and 7, respectively. We adapt the threshold ϵ to each case by choosing an α-quantile (with $\alpha = 0.0005$) of the distribution of the distances between the observed and the simulated types, according to Beaumont et al. [6]. The resulting values are displayed in Table 3. Here, we exploited the fact that both the algorithms are embarrassingly parallel, thus running them in parallel on a machine with 16 vCPU setting $S = 100,000$ and $m = 500$.

Figure 6 shows the posterior distributions for θ computed from the five datasets. The black dashed lines represent the true $\mathtt{Beta}(\alpha, \beta)$ posteriors. The quality of the approximation seems to not be affected by the cardinality of $\mathcal{X}$ and the LD algorithm

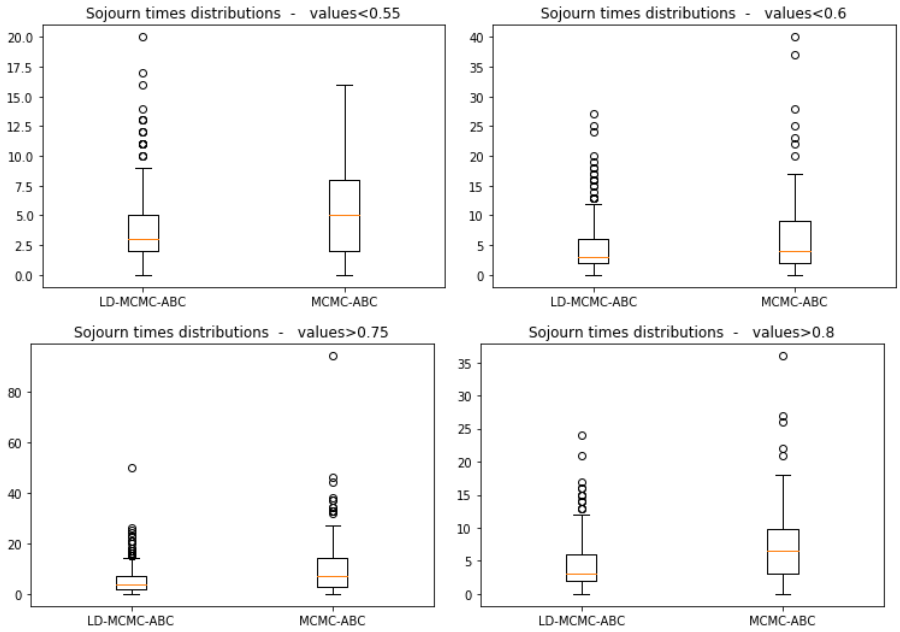


Figure 5: Boxplots of the distributions of the sojourn time in the LD-MCMC-ABC and MCMC-ABC with four different thresholds: 0.55 (top-left), 0.6 (top-right), 0.75 (bottom-left), 0.8 (bottom-right).

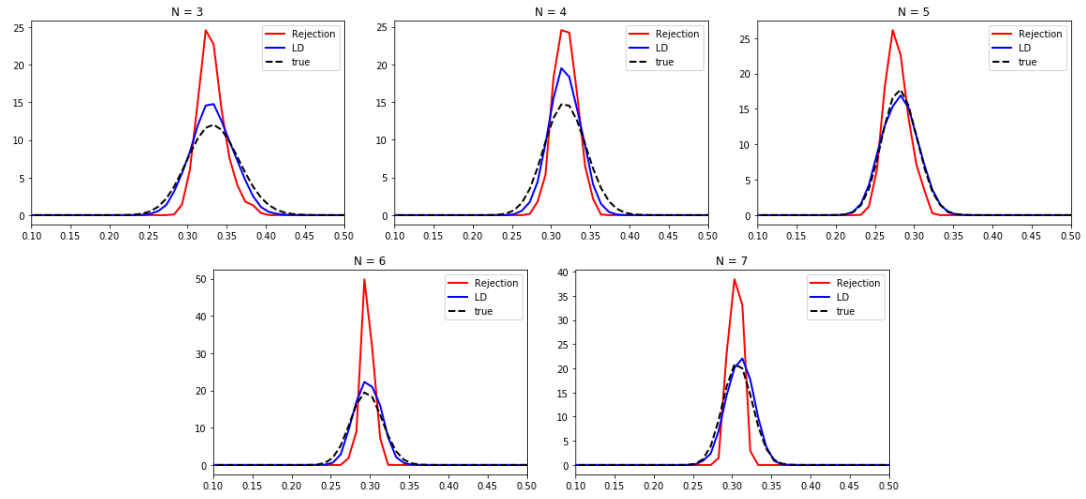


Figure 6: Posterior distributions of θ approximated by LD-ABC (blue lines) and R-ABC (red lines). The black dashed lines represent the true Beta posterior distributions. Each panel corresponds to a different a value of N (3,4,5,6 and 6).

Table 3: $\widehat{\text{ESS}}$'s, running times and tolerance thresholds for each algorithm and dataset.

	Algorithm	$\widehat{\text{ESS}}$	Time	ϵ
$N = 3$	LD	4,807	25.06 s	0.0007
	R	61	1.38 s	
$N = 4$	LD	4,448	24.96	0.0002
	R	31	1.42 s	
$N = 5$	LD	4,382	16.67 s	0.0095
	R	137	1.40 s	
$N = 6$	LD	2,763	16.07 s	0.0012
	R	48	1.42 s	
$N = 7$	LD	1,346	12.13 s	0.0062
	R	45	1.42 s	

always outperforms the standard R-ABC. Regarding the efficiency of the algorithms, Table 3 displays the $\widehat{\text{ESS}}$ and the computational times of each experiment. The $\widehat{\text{ESS}}$ achieved by LD-ABC decreases as N increases, however it is always much greater than the one achieved by R-ABC. Note that, in this example the generative model is very simple and requires only to get samples from a Binomial distribution. Accordingly, the running time of R-ABC appears short and very stable. Running the LD-ABC requires a greater computational time to evaluate the importance weights employing the minimization procedure described in Section 5.4. However, being the $\widehat{\text{ESS}}$ about 40 times greater, one can choose a smaller number of iterations when is implemented the LD version (see Section 5.5.3). Finally, we want to emphasize that as N increases the running time decreases. This is due to the fact that the procedure proposed in Section 5.4 makes the dimension of the minimization problem independent from $|\mathcal{X}|$. Accordingly, the computational time depends only on the number of iterations required by the least-squares solver.

5.5.2 *Example 2: Mixture of binomial distributions .*

Let $\mathbf{X}^n = \{X_i\}_{i=1}^n$ be a sequence of i.i.d, discrete random variables distributed according to the following parametric finite mixture model:

$$\lambda \text{Bin}(\theta_1, N = 4) + (1 - \lambda)\text{Bin}(\theta_2, N = 4). \tag{86}$$

Here we assume a uniform prior distribution on the mixture weight λ and that (θ_1, θ_2) are uniformly distributed on the set $\{(\theta_1, \theta_2) : 0 \leqslant \theta_2 \leqslant \theta_1 \leqslant 1\}$ by imposing the following *identifiability constraint*:

$$\theta_1 \geqslant \theta_2.$$

An analytical computation of the posterior distribution requires the evaluation of the likelihood

$$P(\mathbf{x}^n|\lambda,\theta_1,\theta_2) = \prod_{i=1}^{n}\binom{N}{x_i}\left[\lambda\theta_1^{x_i}(1-\theta_1)^{N-x_i}+(1-\lambda)\theta_2^{x_i}(1-\theta_2)^{N-x_i}\right]. \tag{87}$$

The direct computation of (87) is infeasible, even with few hundred observations, as it involves the expansion of the likelihood into 2^n terms. In the literature there are several methods to deal with this problem, which allow sampling from the parameters' posterior distributions, see [91]. A widespread method is a Gibbs Sampling handling the finite mixtures issue as a missing data problem, see [39]. Samples from the joint posterior distribution are obtained by means of a hierarchical model involving a vector of latent random variables, $\mathbf{Z}^n = \{Z_i\}_{i=1}^n$, where each $Z_i \sim \texttt{Bernoulli}(1-\lambda)$ indicates to which component the i-th observation belongs:

$$\begin{cases} X_i \sim \texttt{Bin}(\theta_1, N) & \text{if } z_i = 0 \\ X_i \sim \texttt{Bin}(\theta_2, N) & \text{if } z_i = 1. \end{cases} \tag{88}$$

Table 4: Details for the simulation of the data-set and for the Gibbs implementation.

θ_1^{true}	θ_2^{true}	λ^{true}	N	n
0.9	0.2	0.8	4	100

S	Burn-in	Thinning
100,000	50,000	10

Table 5: Posterior estimates derived via Gibbs Sampling

	MCMC posterior estimates		
	θ_1	θ_2	λ
Mean	0.8998	0.1556	0.8281
Variance	0.0004	0.0036	0.0018

Here the generative model consists of simulating each of the n values from one of the two binomials according to the result of a $\texttt{Bernoulli}(1-\lambda)$ experiment. The same generative model has been run by a plug-in of the true values of the parameters displayed in Table 4 (LHS) to obtain the observed data. The implemented Gibbs sampling is illustrated in the Direct Acyclic Graph (DAG) in Figure 7 and displayed in Algorithm 14. We ran Algorithm 14 as detailed in Table 4 (RHS) and after burn-in and thinning we got 5,000 values for each parameter regarded as drawn independently from the true posterior distributions. The Posterior means and variances are displayed in Table 5.

In order to compare LD-ABC performance with that of a R-ABC, the marginal importance distributions are set to be the prior distributions and types are used as summary statistics also in R-ABC. We ran R-ABC and LD-ABC with $S = 100,000$ and with four different pairs (m, ϵ). The $\widehat{\text{MSE}}$ and $\widehat{\text{MISE}}$, are computed by averaging

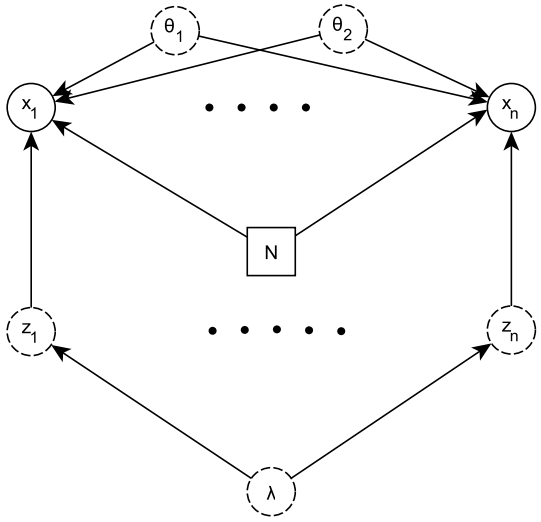


Figure 7: DAG representation of the finite mixture of binomials distributions.

Algorithm 14 Gibbs sampling

Require: x^n

Initialize $\mathbf{p}^{(0)} = p_1^{(0)}, ..., p_n^{(0)}$

for $s = 1, ..., S$ **do**

Draw $Z_i^{(s)} \sim \text{Ber}(p_i^{(s-1)}) \quad \forall i \in \{1, ..., n\}$

Draw $\theta_1^{(s)} \sim \text{TruncatedBeta}\big(1 + \sum_{i=1}^n x_i \mathbb{1}\{z_i = 0\}, 1 + \sum_{i=1}^n (N - x_i)\mathbb{1}\{z_i = 0\}, \theta_2^{(s-1)}, 1\big)$

Draw $\theta_2^{(s)} \sim \text{TruncatedBeta}\big(1 + \sum_{i=1}^n x_i \mathbb{1}\{z_i = 1\}, (1 + \sum_{i=1}^n N - x_i)\mathbb{1}\{z_i = 1\}, 0, \theta_1^{(s)}\big)$

Draw $\lambda^{(s)} \sim \text{Beta}\big(1 + n - \sum_{i=1}^n z_i^{(s)}, 1 + \sum_{i=1}^n z_i^{(s)}\big)$

Compute $p_i^{(s)} = \dfrac{(1-\lambda^{(s)})(\theta_2^{(s)})^{x_i}(1-\theta_2^{(s)})^{N-x_i}}{(1-\lambda^{(s)})(\theta_2^{(s)})^{x_i}(1-\theta_1^{(s)})^{N-x_i} + \lambda^{(s)}(\theta_1^{(s)})^{x_i}(1-\theta_2^{(s)})^{N-x_i}}$

end for

over 100 reruns of each ABC algorithm (with $S = 100,000$) the squared errors and the integrated squared errors (w.r.t. the output given by the Gibbs sampler), respectively. The results are summarized in Table 6.

First, we note that both the $\widehat{\text{MSE}}$ and the $\widehat{\text{MISE}}$ achieved by LD-ABC are always lower for LD-ABC than for R-ABC. Hence, in our example, taking into account the probability of large deviation events has improved both the point estimates and the approximation of the posterior distributions. Moreover, as already pointed out in Section 5.3.1, LD-ABC mitigates the sample degeneracy by achieving an $\widehat{\text{ESS}}$ up to more than five times that achieved by R-ABC (see Table 7).

In order to evaluate the sample degeneracy, Table 7 (RHS) also displays the *normalized perplexity*, which equals $2^{H(\tilde{\omega})}/S$, where $H(\tilde{\omega})$ denotes the entropy of the normalized weights. Cappé et al. [19] show that the normalized perplexity repre-

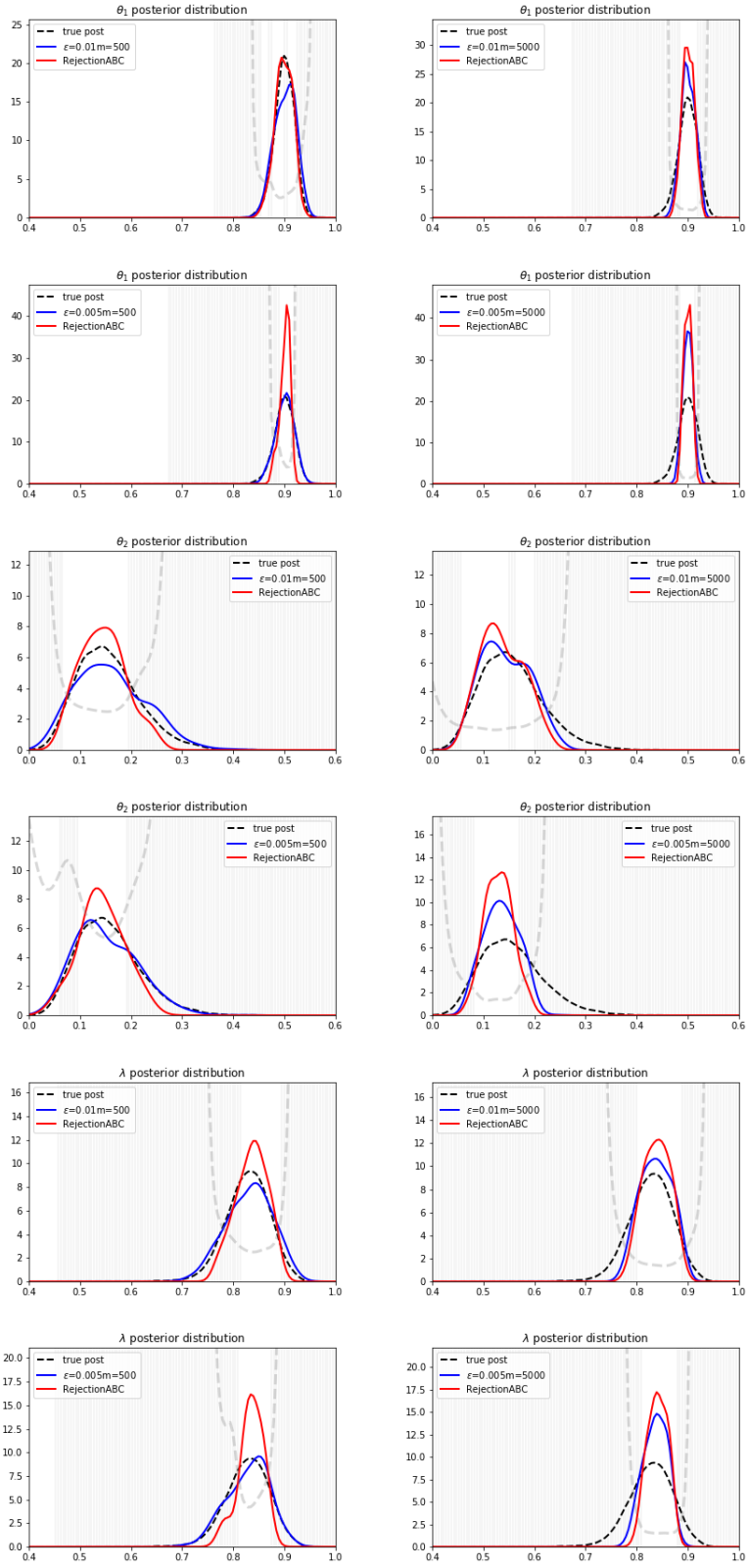


Figure 8: Posterior distributions corresponding to four different pairs of tuning parameters (m, ϵ). Each panel refers to one of the three model parameters. Red lines represent the posterior density estimates provided via R-ABC. The blue lines represent the estimates provided via LD-ABC. The dashed black lines are the output of the Gibbs sampler. The gray dashed lines are the ratios $\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{\mathbf{x}^n})/\tilde{\mathcal{L}}^{R}_{\epsilon,m}(\theta; T_{\mathbf{x}^n})$ providing a representation of the adjustment $\alpha_{\epsilon,m}$.

Table 6: Squared errors and integrated squared errors averaged over 100 reruns. Each column contains results for one of the model parameters both for LD-ABC and R-ABC.

		$m = 500$, $\epsilon = 0.005$		
		θ_1	θ_2	λ
$\widehat{\text{MSE}}_{mean}$	LD	$0.0121 \cdot 10^{-4}$	$0.1516 \cdot 10^{-4}$	$0.1444 \cdot 10^{-4}$
	R	$0.0679 \cdot 10^{-4}$	$2.4437 \cdot 10^{-4}$	$0.9420 \cdot 10^{-4}$
$\widehat{\text{MSE}}_{var}$	LD	$0.0000 \cdot 10^{-4}$	$0.0012 \cdot 10^{-4}$	$0.0004 \cdot 10^{-4}$
	R	$0.0006 \cdot 10^{-4}$	$0.055 \cdot 10^{-4}$	$0.0135 \cdot 10^{-4}$
$\widehat{\text{MISE}}$	LD	0.1445	0.0479	0.0799
	R	2.9162	0.8656	1.6831
		$m = 500$, $\epsilon = 0.01$		
		θ_1	θ_2	λ
$\widehat{\text{MSE}}_{mean}$	LD	$0.015 \cdot 10^{-4}$	$0.0581 \cdot 10^{-4}$	$0.0251 \cdot 10^{-4}$
	R	$0.0288 \cdot 10^{-4}$	$1.6023 \cdot 10^{-4}$	$0.681 \cdot 10^{-4}$
$\widehat{\text{MSE}}_{var}$	LD	$0.0000 \cdot 10^{-4}$	$0.0079 \cdot 10^{-4}$	$0.0013 \cdot 10^{-4}$
	R	$0.0003 \cdot 10^{-4}$	$0.0277 \cdot 10^{-4}$	$0.0065 \cdot 10^{-4}$
$\widehat{\text{MISE}}$	LD	0.4662	0.2019	0.2344
	R	0.8509	0.2744	0.4634
		$m = 5\,000$, $\epsilon = 0.005$		
		θ_1	θ_2	λ
$\widehat{\text{MSE}}_{mean}$	LD	$0.0189 \cdot 10^{-4}$	$2.7694 \cdot 10^{-4}$	$0.9609 \cdot 10^{-4}$
	R	$0.0281 \cdot 10^{-4}$	$3.8095 \cdot 10^{-4}$	$1.1787 \cdot 10^{-4}$
$\widehat{\text{MSE}}_{var}$	LD	$0.0006 \cdot 10^{-4}$	$0.0639 \cdot 10^{-4}$	$0.0148 \cdot 10^{-4}$
	R	$0.0009 \cdot 10^{-4}$	$0.0854 \cdot 10^{-4}$	$0.0196 \cdot 10^{-4}$
$\widehat{\text{MISE}}$	LD	3.2921	1.0344	1.6270
	R	7.3482	2.3676	3.4212
		$m = 5\,000$, $\epsilon = 0.01$		
		θ_1	θ_2	λ
$\widehat{\text{MSE}}_{mean}$	LD	$0.0184 \cdot 10^{-4}$	$1.489 \cdot 10^{-4}$	$0.6666 \cdot 10^{-4}$
	R	$0.024 \cdot 10^{-4}$	$2.3092 \cdot 10^{-4}$	$0.9049 \cdot 10^{-4}$
$\widehat{\text{MSE}}_{var}$	LD	$0.0003 \cdot 10^{-4}$	$0.029 \cdot 10^{-4}$	$0.0068 \cdot 10^{-4}$
	R	$0.0005 \cdot 10^{-4}$	$0.0495 \cdot 10^{-4}$	$0.0115 \cdot 10^{-4}$
$\widehat{\text{MISE}}$	LD	0.7410	0.2175	0.3856
	R	1.8753	0.5775	0.9733

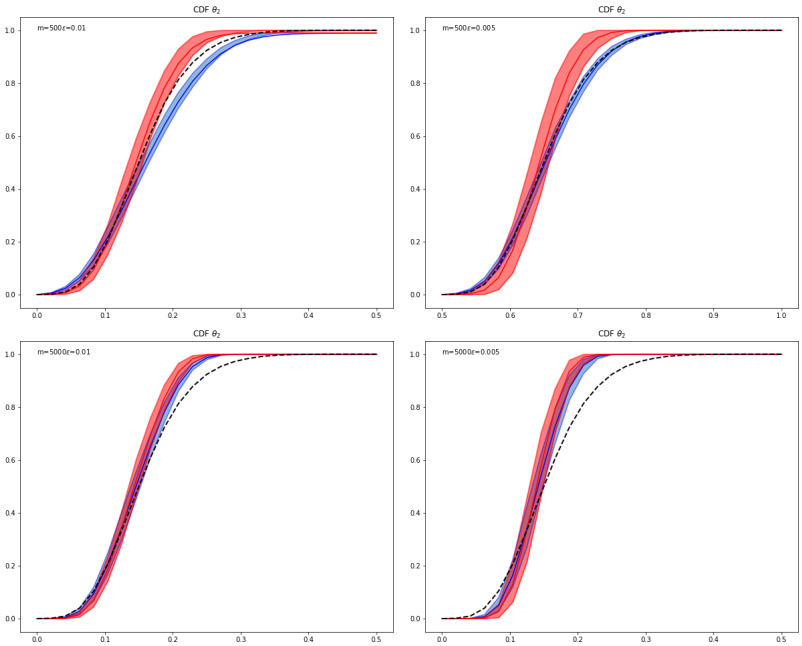


Figure 9: Posterior cumulative density functions for θ_2. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cumulative density function for a pair (m, ϵ). For $\theta_2 < 0.5$ both the cumulative density functions equal 0. 90% intervals over 100 reruns of each algorithm are also shown.

sents an estimate of $2^{-D\left(\tilde{\pi}(\theta, T_{y^m} | T_{x^n}) || q(\theta, T_{y^m})\right)}$, meaning that when the perplexity is larger, the sample degeneracy is smaller.

The following comments are consistent with Remark 4. Concerning the role of the tuning parameters, m and ϵ, we note that by fixing a large m (e.g., $5,000$), as ϵ increases both $\widehat{\mathrm{ESS}}$ and the perplexity increase as well. Moreover, both $\widehat{\mathrm{MSE}}$ and $\widehat{\mathrm{MISE}}$ decrease. The same happens by reducing m with ϵ fixed to a small value (e.g., 0.005).

In Figure 8 three matrices of plots, one for each parameter, show the posterior densities: the size of the pseudo-dataset, m, equals 500 in the plots on the LHS of each panel, and $5,000$ on the RHS. The topmost plots show the approximate distributions with $\epsilon = 0.01$, the others the distributions corresponding to $\epsilon = 0.005$. We note that as m increases the blue lines (LD-ABC) overlap the red ones (R-ABC). In principle, we would expect that both the algorithms achieve a better approximation of the posterior shapes with $\epsilon = 0.005$ than $\epsilon = 0.01$. However, in the case of R-ABC, we see a deviation from the true posterior distributions (dotted lines), when moving from the first to the second row of each matrix. The same deviation occurs for LD-ABC, but only in the second column, when $m = 5,000$. This suggests that the quality of the R-ABC approximation is affected by a low value of the ESS which is in turn determined by a too ambitious value of ϵ and m. On the other hand, when $m = 500$, LD-ABC manages to mitigate the effect of a small ϵ, but it fails when a large value of m causes a too small ESS for the LD-ABC as well. In the figures we also superimposed the ratio $\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{x^n}) / \tilde{\mathcal{L}}^{R}_{\epsilon,m}(\theta; T_{x^n}) = 1 + \alpha_{\epsilon,m}(\theta; T_{x^n}) / \tilde{\mathcal{L}}^{R}_{\epsilon,m}(\theta; T_{x^n})$ evaluated pointwise and shown by the gray dashed lines. This quantity depends on the contribution of the adjustment w.r.t. the R-ABC likelihood and shows how the adjustment

Table 7: $\widehat{\text{ESS}}$ and normalized perplexity averaged over 100 reruns for each pair of tuning parameters.

		$\widehat{\text{ESS}}$	
		$\epsilon = 0.005$	$\epsilon = 0.01$
$m = 500$	LD	261	445
	R	25	81
$m = 5\,000$	LD	71	168
	R	31	94

		Normalized Perplexity	
		$\epsilon = 0.005$	$\epsilon = 0.01$
$m = 500$	LD	0.0034	0.0055
	R	0.0002	0.0008
$m = 5\,000$	LD	0.0008	0.0018
	R	0.0003	0.0009

acts in modifying this latter when the R-ABC posterior density is underestimated (gray areas). Figure 9, shows the posterior cdf for θ_2. The Posterior cdf for the other two parameters are given in Appendix D.3, Figures 21 and 22. We also show the 90% credible intervals for the estimated cumulative density functions. The red areas are always larger than the blue areas, meaning that the estimates provided by R-ABC exhibit greater variability. This is more significant when $\epsilon = 0.005$, due to the small acceptance probability.

To wrap up, as suggested by the $\widehat{\text{MISE}}$'s, the posterior distributions approximated by LD-ABC appear more faithful to the true shapes. Moreover, the ESS and the variability of the estimates are less sensitive to small values of ϵ.

5.5.3 *Concluding remarks on the choice of the tolerance threshold*

The discussion of the results in the previous subsection, together with Remark 4, gives some useful suggestions for tuning the parameters m and ϵ. Further research may provide a more formal way of choosing the optimal pair (m, ϵ). A possible strategy is to automaticly tune m and ϵ by minimizing an objective function. For instance, one can follow the method suggested by Soubeyrand et al. [145], which employs the minimization of the Bayesian mean squared error for selecting the optimal weighting function and tolerance threshold. However, it is worth to note that, regarding the LD method, the role of the threshold ϵ is quite different w.r.t. to other ABC strategies. In order to discuss such differences, let us consider a standard procedure for the choice of ϵ. As already mentioned, it can be set equal to the α-th quantile of the empirical distribution of the distances between the observed and the simulated summary statistics. Note that the number of the accepted parameter proposals (and the ESS) is strongly related to α: small values of α correspond to a small acceptance rate. On

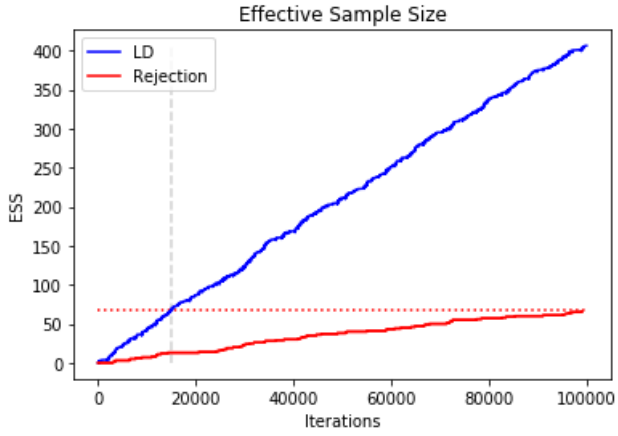


Figure 10: ESS of the two algorithms vs. the number of iterations. The dotted red line represents the $\widehat{\text{ESS}}$ achieved by R-ABC after $S = 100,000$ iterations. The gray dashed lines indicates the number of iterations $(14,985)$ needed by the LD-ABC to get the same ESS as the R-ABC.

the other hand, large values of α correspond to large values of ϵ and to a consequent bias in the approximation of the posterior distribution. To wrap up:

- large α $(\epsilon) \longrightarrow$ large ESS (bias);
- small α $(\epsilon)\longrightarrow$ small ESS (bias).

Most of the more sophisticated ABC strategies (such as sequential methods) improve the exploration of the parameter space and provide a more targeted sample from the parameter space. This leads to observing smaller distances between the observed and the simulated summary statistics. Accordingly, by comparing two different sampling schemes one note that the same value of α corresponds to different ϵ's: sampling schemes providing an efficient explorations of the parameters space are associated with small values of ϵ.

The LD method, as already discussed, can be applied to different sampling schemes (in this thesis we give an IS and a MCMC sampling scheme). It improves the ABC performances both in terms of bias reduction and of efficiency (ESS). For a given value of α, the LD version and the standard version of an ABC sampling scheme lead to the same empirical distribution of the distances and the same value of ϵ.

Let us consider the R-ABC and its LD version described in the previous subsection. The two algorithms produce the same samples $(\theta^{(s)}, T_{y^{(s)}_m})$ but they assign different weights. By resorting to a pilot run, with $\alpha = 0.0005$, we get $\epsilon = 0.0089$. Figure 11 shows that the two algorithms achieve very different results, even though they employ the same ϵ and number of iterations $(100,000)$. Looking at Figure 10 we can see that R-ABC achieves an $\widehat{\text{ESS}}$ of about 50, meaning that the pairs $(\theta^{(s)}, T_{y^{(s)}_m})$ in the acceptance region are about the $\alpha \cdot S$. The LD-ABC, borrowing information from the pairs outside the acceptance region, achieves an $\widehat{\text{ESS}}$ greater then 400. Finally, it is noteworthy that Figures 10 and 11 show that LD-ABC needs only $14,985$ iterations to achieve the same $\widehat{\text{ESS}}$ as the R-ABC and a better approximation of the posterior distributions.

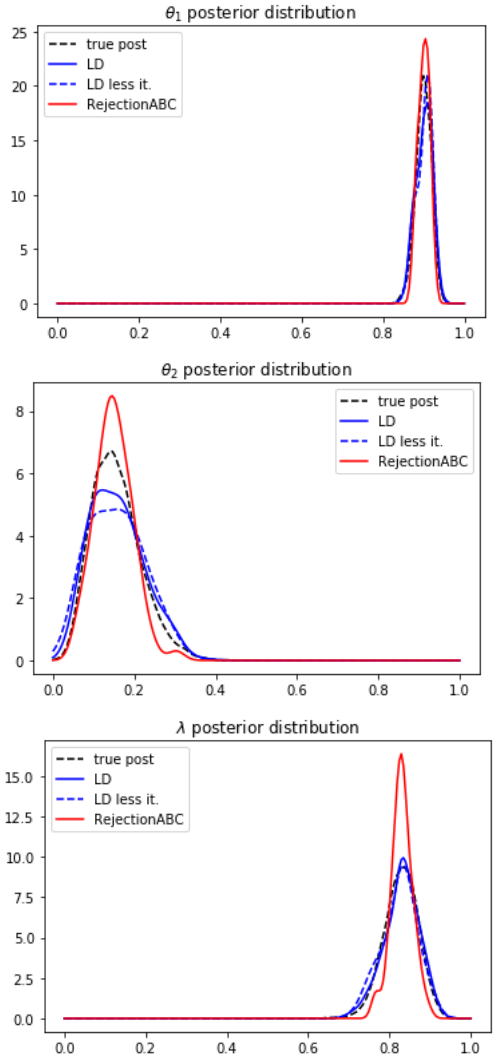


Figure 11: Posterior density functions of the three parameters of the Binomial mixture model in Section 5.5.2. The true posterior distributions are represented by the black dashed lines. The continuous red lines represent the posterior density estimates provided via R-ABC and the blue ones via LD-ABC. The dashed blue lines represent the density estimates achieved by the LD-ABC after 14,985 iterations.

6

LD-ABC FOR FINITE STATE MARKOV CHAINS

In the previous chapter we introduced two LD-ABC algorithms restricting ourselves to discrete i.i.d. random variables. Here, we extend the proposed method to sequences of dependent random variables, in particular to homogeneous finite state Markov chains. This will requires to consider more sophisticated versions of LDT, where the i.i.d. assumption is relaxed, as well as extensions of the Method of types.

In the following section we define the *doublet probability distribution*, which plays a key role in this framework. We also introduce some additional notation and some preliminary concepts.

6.1 FINITE STATE MARKOV CHAINS AND DOUBLET PROBABILITY DISTRIBUTIONS: SET UP AND NOTATION

Let $\mathcal{X}$ be a finite set of cardinality k

$$\mathcal{X} \triangleq \{r_1, ..., r_k\}.$$

For the sake of simplicity, hereafter we will denote the elements of the set $\mathcal{X}$ by their labels $\{1, ..., k\}$. The set of all possible probability measures over $\mathcal{X}$ can be identified as the $(k-1)$-dimensional probability simplex:

$$\Delta^{k-1} \triangleq \{\mathbf{p} \in [0,1]^k : \sum_{i=1}^{k} p_i = 1\}.$$

A *stationary* Markov Process $\{X_t\}$ on $\mathcal{X}$ can be equivalently characterized: 1) by the *transition matrix* or 2) by the *doublet probability distribution*. Here, we resort to the second approach since the doublet probability distribution will play a key role in the ABC framework. See Appendix C for the alternative formulation.

Define the *doublet probability distribution*, P, as a non negative matrix of order $k \times k$ summing to 1 and inducing a probability measure, $\Pr(\cdot,\cdot)$, over $\mathcal{X}^2 \triangleq \mathcal{X} \times \mathcal{X}$. Thus, denoted by p_{ij}, the entries of P are

$$p_{ij} = \Pr(X_t = i, X_{t+1} = j) \qquad \forall (i,j) \in \mathcal{X}^2. \tag{89}$$

Let us denote by $\mathcal{M}(\mathcal{X}^2) \subseteq \Delta^{k^2-1}$ the set of the stationary doublet probability distributions over $\mathcal{X}^2$ defined as follows.

Definition 4 (Stationary doublet probability distribution) *A distribution* $P \in \mathcal{M}(\mathcal{X}^2)$ *is said to be* stationary *if*

$$\sum_{j \in \mathcal{X}} p_{ij} = \sum_{j \in \mathcal{X}} p_{ji} \quad \forall i \in \mathcal{X}. \tag{90}$$

Suppose that $\{X_t\}$ is a stationary Markov process assuming values in $\mathcal{X}$. Then the doublet probability distribution, P, captures all the relevant information about the process. In particular, the distribution of the process, $\mathbf{p} = (p_1, ..., p_k)$, can be retrieved from P as follows

$$p_i \stackrel{\triangle}{=} \sum_{j \in \mathcal{X}} p_{ij} = \sum_{j \in \mathcal{X}} p_{ji} \quad \forall i \in \mathcal{X}. \tag{91}$$

The corresponding *state transition matrix* of the Markov chain, Q, is the stochastic matrix of order $k \times k$ composed by entries

$$q_{ij} \stackrel{\triangle}{=} \Pr(X_{t+1} = j | X_t = i) \quad \forall (i,j) \in \mathcal{X}^2. \tag{92}$$

Note that the doublet probability distribution does not represent a stochastic matrix, hence cannot correspond to the transition matrix Q. However, each entry of Q can be derived from P as follows:

$$q_{ij} = \frac{p_{ij}}{p_i}. \tag{93}$$

Hence, given a stationary doublet probability distribution P, from Definition 4 and (93), follows that the marginal distribution, $\mathbf{p}$, is the distribution satisfying *stationarity* (as defined in C.3) being a (normalized) row eigenvector of Q corresponding to the eigenvalue 1:

$$(\mathbf{p}Q)_j = \sum_{i \in \mathcal{X}} p_i q_{ij} = \sum_{i \in \mathcal{X}} p_{ij} = p_j \qquad \forall j \in \mathcal{X}. \tag{94}$$

This implies that the probability of each state does not change during the process.

NOTATION We stress that throughout this chapter we still denote by $\mathcal{X} = \{r_1, ..., r_{|\mathcal{X}|}\}$ a finite, nonempty set. We denote with the upper-case letter P a matrix representing a pmf over $\mathcal{X}^2$ and with Q a stochastic transition matrix. A generic entry of P is denoted by p_{ij}, which is the element in the i-th row and j-th column. The entries of the transition matrix Q are denoted by q_{ij}, which represents the probability of going from the state i to the state j. The bold lower-case letters $\mathbf{p}$ and $\mathbf{q}_i$ are vectors of probabilities representing a pmf over $\mathcal{X}$, marginal probabilities and conditional probabilities, respectively. We will let $\mathbf{X}^n = \{X_i\}_{i=1}^n$, $\mathbf{Y}^m = \{Y_i\}_{i=1}^m$ and so on denote sample paths from finite state Markov chains taking values in $\mathcal{X}$.

6.1.1 *Entropy, Relative Entropy and Conditional Entropy*

Here we recall the definitions of *entropy* and *relative entropy* and introduce the *joint entropy* and the c*onditional (relative) entropy*.

Let X be a discrete random variable taking values in the finite set $\mathcal{X} = \{1, .., k\}$ and $\mathbf{p} \in \Delta^{k-1}$ a probability mass function over $\mathcal{X}$. We define the *entropy* of $\mathbf{p}$ as follows:

$$H(\mathbf{p}) \stackrel{\triangle}{=} -\sum_{i=1}^{k} p_i \log(p_i). \tag{95}$$

Suppose that $\mathbf{p}, \mathbf{p}' \in \Delta^{k-1}$ are two probability mass functions such that $\mathbf{p}$ is dominated by $\mathbf{p}'$, i.e., $\text{supp}(\mathbf{p}') \stackrel{\triangle}{=} \{i \in \mathcal{X} : p'_i > 0\} \supseteq \text{supp}(\mathbf{p}) \stackrel{\triangle}{=} \{i \in \mathcal{X} : p_i > 0\}$, we define:

- The Kullback–Leibler divergence or *relative entropy* between $\mathbf{p}$ and $\mathbf{p}'$:

$$D(\mathbf{p}\|\mathbf{p}') \triangleq \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{p_i'}\right);$$

- The *cross entropy* between $\mathbf{p}$ and $\mathbf{p}'$:

$$J(\mathbf{p}, \mathbf{p}') \triangleq -\sum_{i=i}^{k} p_i \log p_i'.$$

Let (X_t, X_{t+1}) be a pair of discrete random variables defined on the finite set $\mathcal{X}^2$. Let $P = \{p_{ij} : (i,j) \in \mathcal{X}^2\} \in \mathcal{M}(\mathcal{X}^2)$ be their joint probability mass function over $\mathcal{X}^2$ and $\mathbf{p} = \{p_i : i \in \mathcal{X}\}$ the marginal distribution over $\mathcal{X}$.

We define:

- the *joint entropy*

$$H(P) \triangleq -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} \log(p_{ij}); \tag{96}$$

- the *conditional entropy*

$$\begin{aligned} H_c(P) &\triangleq \sum_{i\in\mathcal{X}} \Pr(X=i) H(\mathbf{q}_i) \\ &= -\sum_{i\in\mathcal{X}} p_i \sum_{j\in\mathcal{X}} q_{ij} \log q_{ij} \\ &= -\sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{X}} p_{ij} \log q_{ij}, \end{aligned} \tag{97}$$

  where the vector $\mathbf{q}_i = (q_{i1}, ..., q_{ik}) \in \Delta^{k-1}$ is the conditional probability distribution defined by

$$\Pr(X_{t+1} = j | X_t = i) = q_{ij} \triangleq \frac{p_{ij}}{p_i}.$$

Let $P, P' \in \mathcal{M}(\mathcal{X}^2)$ be two probability mass functions such that $\mathrm{supp}(P') \supseteq \mathrm{supp}(P)$. We define the *conditional relative entropy* between P and P'.

$$\begin{aligned} D_c(P\|P') &\triangleq \sum_{i\in\mathcal{X}} \Pr(X=i) D(\mathbf{q}_i\|\mathbf{q}_i') \\ &= \sum_{i\in\mathcal{X}} p_i \sum_{j\in\mathcal{X}} q_{ij} \log \frac{q_{ij}}{q_{ij}'} \\ &= \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{X}} p_{ij} \log \frac{p_{ij}}{q_{ij}' p_i} \end{aligned} \tag{98}$$

where $q_{ij}' = \frac{p_{ij}'}{p_i'}$. Note that in what follows we impose the conditions $\mathrm{supp}(P') \supseteq \mathrm{supp}(P)$ and $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$ by considering only full-support distributions.

The following two *chain rules* prove that the (relative) entropy of a pair of random variables is the sum of the (relative) entropy of one of the two variables plus the conditional (relative) entropy.

Theorem 8 (Chain rule for entropy)

$$H(P) = H_c(P) + H(\mathbf{p})$$

Proof

$$\begin{aligned}
H(P) &= -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log(p_{ij}) \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log(q_{ij}p_i) \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log q_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log p_i \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log q_{ij} - \sum_{i=1}^{k} p_i \log p_i \\
&= H_c(P) + H(\mathbf{p})
\end{aligned}$$

□

Theorem 9 (Chain rule for relative entropy)

$$D(P\|P') = D_c(P\|P') + D(\mathbf{p}\|\mathbf{p}') \tag{99}$$

Proof

$$\begin{aligned}
D(P\|P') &= \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log\frac{p_{ij}}{p'_{ij}} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log\frac{p_i}{p'_i}\frac{q_{ij}}{q'_{ij}} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log\frac{p_i}{p'_i} + \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log\frac{q_{ij}}{q'_{ij}} \\
&= \sum_{i=1}^{k} p_i \log\frac{p_i}{p'_i} + \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log\frac{q_{ij}}{q'_{ij}} \\
&= D(\mathbf{p}\|\mathbf{p}') + D_c(P\|P')
\end{aligned}$$

□

From Theorems 8 and 9 we derive the following alternative definitions for the conditional entropy and the conditional relative entropy, respectively:

$$H_c(P) \triangleq H(P) - H(\mathbf{p}) \tag{100}$$

$$D_c(P\|P') \triangleq D(P\|P') - D(\mathbf{p}\|\mathbf{p}'). \tag{101}$$

A type is commonly defined by the empirical distribution of a sequence of random variables, as in Chapter 5. More precisely, given a sequence of n samples from a finite set, the *first order type* of that sequence is its empirical distribution. In principle ,the Method of Types is suitable for i.i.d. random variables but extensions to other forms of dependence are possible by considering the more general definition of l-th order types.

In the rest of the chapter, we let $n \geqslant 1$ be an arbitrarily fixed integer, denoting the length of the sequences.

Definition 5 *The l-th order type of a sequence $\mathbf{x}^n = x_1, ..., x_n \in \mathcal{X}^n$ is defined as the probability distribution $T^{(l)}_{\mathbf{x}^n} \in \Delta^{k^l - 1}$ with*

$$T^{(l)}_{\mathbf{x}^n}(\mathbf{z}) = \frac{1}{n-l+1} \sum_{i=1}^{n-l+1} \mathbb{1}\{(x_i, ..., x_{i+l-1}) = \mathbf{z}\} \quad \forall \mathbf{z} \in \mathcal{X}^l. \tag{102}$$

Given a Markov process $\{X_t\}$ and an observed sample path $\mathbf{x} = x_1, ..., x_n$, the appropriate type concept is the *second order type* [31]:

$$T^{(2)}_{\mathbf{x}^n}(i,j) = \frac{1}{n-1} \sum_{t=1}^{n-1} \mathbb{1}\{x_t = i, x_{t+1} = j\} \quad \forall (i,j) \in \mathcal{X}^2. \tag{103}$$

$T^{(2)}_{\mathbf{x}^n}$ can be thought as a matrix of order $k \times k$ representing an empirical estimate of the doublet probability distribution.

The second order type defined by (103) does not ensure the stationarity in the sense of Definition 4. However, by resorting to the *cyclic convention* that the $(n+1)$-th element of the path is always equal to x_1, the following stationary second order type is given:

$$\dot{T}^{(2)}_{\mathbf{x}^n} \triangleq \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{x_t = i, x_{t+1} = j\} \quad \forall (i,j) \in \mathcal{X}^2, \tag{104}$$

where $x_{n+1} = x_1$.

We can see that $\dot{T}^{(2)}_{\mathbf{x}^n}$ can be obtained by applying to $T^{(2)}_{\mathbf{x}^n}$ the one-to-one function $h : \Delta^{k^2-1} \to \Delta^{k^2-1}$.

In fact,

$$\dot{T}^{(2)}_{\mathbf{x}^n}(ij) = h(T^{(2)}_{\mathbf{x}^n}(ij)) = \begin{cases} \left(T^{(2)}_{\mathbf{x}^n}(ij)(n-1)+1\right)\frac{1}{n} & \text{if } i = x_n,\ j = x_1 \\ T^{(2)}_{\mathbf{x}^n}(ij)\frac{n-1}{n} & \text{otherwise.} \end{cases} \tag{105}$$

In what follows we deal with the type under cyclic convention, $\dot{T}^{(2)}_{\mathbf{x}^n}$, since it always satisfies stationarity.

6.2.1 *Second order types as summary statistics*

Let us assume $\mathbf{X}^n = X_1, ..., X_n \in \mathcal{X}^n$ to be a sample path from a *parametric* Markov process which is characterized by a doublet probability distribution whose entries depend on an unknown parameter (or a vector) $\theta \in \Theta$. In order to make an inference on θ, one can assume a model specifying a family of doublet probability distributions $\mathcal{F} \triangleq \{P_\theta : \theta \in \Theta\}$).

The second order type maps the n dimensional sequence onto a matrix of size $k \times k$, meaning that, depending on the length of the sequence and on the cardinality of the finite set $\mathcal{X}$, it can be considered a summary statistic. In particular, as already shown in Section 5.1 for the i.i.d. case, the type represents a *sufficient statistic* for θ, providing data reduction and capturing all the relevant information about θ contained in the sample $\mathbf{x}^n$: given two sample paths $\mathbf{x}^n$ and $\mathbf{y}^n$ such that $T^{(2)}_{\mathbf{x}^n} = T^{(2)}_{\mathbf{y}^n}$, the inference on θ should be the same whether $\mathbf{X}^n = \mathbf{x}^n$ or $\mathbf{X}^n = \mathbf{y}^n$.

Proposition 3 (Sufficiency) *Let $\{X_t\}$ be a Markov process taking values in the finite set $\mathcal{X}$ with stationary doublet probability distribution $P_\theta \in \Delta^{k^2-1}$. Let the initial value of the chain be $X_1 = x_1$. Then the second order type $T^{(2)}_{\mathbf{x}^n}$ is a sufficient summary statistic for θ.*

Proof Let $\mathbf{X}^n$ denotes the path $\{X_t\}_{t=1}^n$ and $\mathbf{x}^n$ denotes an observed sequence $x_1, ..., x_n$. Let us consider the probability of observing $\mathbf{x}^n$ given the initial value x_1

$$\begin{aligned}\Pr(\mathbf{X}^n = \mathbf{x}^n | X_1 = x_1, \theta) &= \prod_{t=1}^{n-1} \Pr(X_{t+1} = x_{t+1} | X_t = x_t, \theta) \\ &= \prod_{t=1}^{n-1} \frac{P_\theta(x_{t+1}, x_t)}{\mathbf{p}_\theta(x_t)}\end{aligned}$$

where $\mathbf{p}_\theta$ is the marginal distribution retrieved from P_θ, $P_\theta(x_{t+1}, x_t)$ denotes the element of P_θ corresponding to the pair (x_{t+1}, x_t) and $\mathbf{p}_\theta(x_t)$ denotes the element of the vector $\mathbf{p}_\theta$ corresponding to x_t. By taking the logarithms we can write

$$\log \Pr(\mathbf{X}^n = \mathbf{x}^n | X_1 = x_1) = \sum_{t=1}^{n-1} \big(\log P_\theta(x_{t+1}, x_t) - \log \mathbf{p}_\theta(x_t) \big). \tag{106}$$

Now consider the first order type for the sequence $x_1, ..., x_{n-1}$:

$$T^{(1)}_{\mathbf{x}}(i) = \frac{1}{n-1} \sum_{t=1}^{n-1} \mathbb{1}\{x_t = i\}.$$

The event $\{X_t = i\}$ occurs $(n-1) \cdot T^{(1)}_{\mathbf{x}^n}(i)$ times and the event $\{X_t = i, X_{t+1} = j\}$ occurs $(n-1) \cdot T^{(2)}_{\mathbf{x}^n}(ij)$ times. Thus, the (106) can be rewritten as follows

$$\begin{aligned}&\log \Pr(\mathbf{X}^n = \mathbf{x}^n | X_1 = x_1, \theta) = \\ &= \sum_{i \in \mathcal{X}} \sum_{j \in \mathcal{X}} (n-1) T^{(2)}_{\mathbf{x}^n}(ij) \log P_\theta(ij) - \sum_{i \in \mathcal{X}} (n-1) T^{(1)}_{\mathbf{x}^n} \log \mathbf{p}_\theta(i) \\ &= (n-1)\big(J(\mathbf{p}_\theta, T^{(1)}_{\mathbf{x}^n}) - J(P_\theta, T^{(2)}_{\mathbf{x}^n}) \big) \\ &= -(n-1)\big(D_c(T^{(2)}_{\mathbf{x}^n} || P_\theta) + H_c(T^{(2)}_{\mathbf{x}^n}) \big),\end{aligned} \tag{107}$$

where (107) is obtained by adding and subtracting the conditional entropy. It follows that, according to the Neyman–Fisher factorization theorem, $T^{(2)}_{\mathbf{x}^n}$ is a sufficient summary statistic. □

Note that, given $X_1 = x_1$ and $X_n = x_n$, the second order type under the cyclic convention, being a one-to-one function of the type without cyclic convention, is a sufficient summary statistic for θ as well.

6.3 ABC and large deviations for markov chains

Let $\mathbf{x}^n = x_1, ..., x_n \in \mathcal{X}^n$ be an observed sample path from a parametric stationary Markov process, whose doublet probability distribution is assumed to be in the family $\mathcal{F} \stackrel{\triangle}{=} \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{M}(\mathcal{X}^2)$. Consider the case in which P_θ is unavailable, or the following joint probability is computationally demanding:

$$\Pr(\mathbf{X}^n = \mathbf{x}^n|\theta) = \mathbf{p}_\theta(x_1) \prod_{t=1}^{n-1} \frac{P_\theta(x_{t+1}, x_t)}{\mathbf{p}_\theta(x_t)}.$$

When a simulator is available, one can resort to ABC methods to derive the following approximate posterior distributions:

$$\tilde{\pi}(\theta, \mathbf{y}^n|\mathbf{x}^n) \propto \pi(\theta)\ \Pr(\mathbf{Y}^n = \mathbf{y}^n|\theta)\ \mathbb{1}\{d(\mathbf{y}^n, \mathbf{x}^n) \leqslant \epsilon\},$$
$$\tilde{\pi}(\theta|\mathbf{x}^n) \propto \pi(\theta) \Pr\left(d(\mathbf{y}^n, \mathbf{x}^n) \leqslant \epsilon|\theta\right).$$

The basic idea behind the method proposed in the previous chapter was to resort to LDT in order to provide a better estimate for the probability $\Pr\left(d(\mathbf{y}^n, \mathbf{x}^n) \leqslant \epsilon|\theta\right)$. As discussed in the previous chapter, one of the major results in LDT is Sanov's Theorem, which establishes the *rate function*, i.e., a function quantifying the decline of the probability of rare events at least asymptotically, for sequences of i.i.d. random variables. The analog of Sanov's theorem for Markov chains is the Donsker and Varadhan theorem [42], or it can be estabilished by means of an easier counting approach that can be traced back to Boza [14] and Natarajan [102]. It was also presented as an application of the method of types by Csiszár [31].

Theorem 10 *Let $\{X_t\}$ be a Markov process taking values in the finite set $\mathcal{X}$, with stationary doublet probability distribution $P_\theta \in \mathcal{M}(\mathcal{X}^2)$ and let $\mathbf{X}^n = X_1, ..., X_n$. If $E \subseteq \mathcal{M}(\mathcal{X}^2)$, then for each $\theta \in \Theta$*

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr(\dot{T}^{(2)}_{\mathbf{X}^n} \in E|\theta) = -\inf_{P\in E} D_c(P||P_\theta). \tag{108}$$

Proof See [102, Th. 1]. □

Exploiting the same arguments as in Chapter 5, one can rely on the second order type as summary statistic and resort to the conditional relative entropy as measure of discrepancy between the simulated and the observed data. Accordingly, given an observed path $\mathbf{x}^n$, we can define the following ABC acceptance region

$$\Gamma_\epsilon(\dot{T}^{(2)}_{\mathbf{x}^n}) \stackrel{\triangle}{=} \{P \in \Delta^{k^2-1} : D_c(P||\dot{T}^{(2)}_{\mathbf{x}^n}) \leqslant \epsilon\}, \tag{109}$$

where $P_\theta \in \Delta^{k^2-1}$ is the true doublet stationary probability distribution. For the sake of notational simplicity, $\Gamma_\epsilon(\dot{T}^{(2)}_{x^n})$ is referred to as Γ_ϵ for short.

From Theorem 10 follows that by drawing from the simulator a sample path of length m, $\mathbf{Y}^m$, at each iteration s the acceptance probability can be approximated by

$$\Pr\left(\dot{T}^{(2)}_{Y^m} \in \Gamma_\epsilon | \theta^{(s)}\right) \approx 2^{-m D_c(\Gamma_\epsilon || P_\theta^{(s)})}. \tag{110}$$

Note that, since we are assuming that P_θ is not available, in (110) we cannot directly compute $D_c(\Gamma_\epsilon||P_\theta)$. However, the following theorem establishes the approximation $D_c(\Gamma_\epsilon||P_\theta) \approx D_c(\Gamma_\epsilon||\dot{T}^{(2)}_{y^m})$. A proof is presented in Appendix D.2

Theorem 11 *Let $\{Y_t\}$ be a Markov process taking values in the finite set $\mathcal{X}$, such that the stationary doublet probability distribution is $P_\theta \in \mathcal{M}(\mathcal{X}^2)$ and let $\mathbf{Y}^m = Y_1, ..., Y_m$. Then, under the measure induced by P_θ*

$$\lim_{m\to\infty} D_c(\Gamma_\epsilon||T^{(2)}_{Y^m}) = D_c(\Gamma_\epsilon||P_\theta) \quad \text{a.s.} \tag{111}$$

Accordingly, we define the Markov analogue of the i.i.d. kernel function as follows:

$$K_\epsilon(\dot{T}^{(2)}_{y^m}) \triangleq \begin{cases} 1 & \text{if } D_c(\dot{T}^{(2)}_{y^m}||\dot{T}^{(2)}_{x^n}) \leqslant \epsilon \\ 2^{-m D_c(P^*||\dot{T}^{(2)}_{y^m})} & \text{if } D_c(\dot{T}^{(2)}_{y^m}||\dot{T}^{(2)}_{x^n}) > \epsilon \end{cases}. \tag{112}$$

It follows that the joint ABC approximate posterior becomes

$$\tilde{\pi}(\theta, \dot{T}^{(2)}_{y^m}|\dot{T}^{(2)}_{x^n}) \propto \pi(\theta) P_\theta(\dot{T}^{(2)}_{y^m}) K_\epsilon(\dot{T}^{(2)}_{y^m}) \tag{113}$$

and the approximate marginal posterior becomes

$$\tilde{\pi}(\theta|\dot{T}^{(2)}_{x^n}) \propto \pi(\theta) \sum_{\dot{T}^{(2)}_{y^m} \in \mathcal{T}(m,2)} P_\theta(\dot{T}^{(2)}_{y^m}) K_\epsilon(\dot{T}^{(2)}_{y^m}), \tag{114}$$

where $\mathcal{T}(m, 2)$, referred to as $\mathcal{T}$ for short, is the set of second order types generated from a Markov chain of length m. This implies that the ABC likelihood is

$$\tilde{\mathcal{L}}_\epsilon(\theta; \dot{T}^{(2)}_{x^n}) = \sum_{\dot{T}^{(2)}_{y^m} \in \mathcal{T}} P_\theta(\dot{T}^{(2)}_{y^m}) K_\epsilon(\dot{T}^{(2)}_{y^m}). \tag{115}$$

In order to sample from (113), we introduce both an IS and a MCMC scheme displayed in Algorithm 15 and Algorithm 16, respectively.

Again LDT allows defining a kernel on a non-compact support, thus avoiding the implicit rejection step in both the sampling schemes.

6.4 EXPERIMENTS

Let us consider a categorical time series $\{X_t\}$ taking values in the finite set $\mathcal{X}$. Specifically, we consider the Pegram's operator-based autoregressive AR(1) process dealt with in [1]. The Pegram's operator [110], denoted by "$*$", is a mixing operator which

Algorithm 15 Importance Sampling LD-ABC for Markov Chains

for $s = 1, ..., S$ **do**
  Draw $\theta^{(s)} \sim q$
  Generate $\mathbf{y}^{(s)} = y_1^{(s)}, ..., y_m^{(s)}$ from $P(\cdot|\theta^{(s)})$ and compute $\dot{T}^{(2)}_{\mathbf{y}^{(s)}}$
  if $D_c(\dot{T}^{(2)}_{\mathbf{y}^{(s)}} || \dot{T}^{(2)}_{\mathbf{x}^n}) \leqslant \epsilon$ **then**
    Set the IS weight for $(\theta^{(s)}, \dot{T}^{(2)}_{\mathbf{y}^{(s)}})$ to

$$\omega_s = \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$$

  else
    Set the IS weights for $(\theta^{(s)}, \dot{T}^{(2)}_{\mathbf{y}^{(s)}})$ to $\omega_s = 2^{-mD(\Gamma_\epsilon || \dot{T}^{(2)}_{\mathbf{y}^{(s)}})} \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$
  end if
end for

Algorithm 16 Metropolis-Hastings LD-ABC for Markov Chains

Initialize $\theta^{(0)}$ and $\mathbf{y}^{(0)}$
for $s = 1, ..., S$ **do**
  Draw $\theta^* \sim \tilde{q}\left(\theta^{(s-1)}, \cdot\right)$
  Draw $\mathbf{y}^* = y_1^*, ..., y_m^*$ from $P(\cdot|\theta^*)$ and compute $\dot{T}^{(2)}_{\mathbf{y}^*}$
  Compute $\alpha = \min\left\{1, \frac{\pi(\theta^*) K_\epsilon(\dot{T}^{(2)}_{\mathbf{y}^*}) \tilde{q}\left(\theta^*, \theta^{(s-1)}\right)}{\pi(\theta^{(s-1)}) K_\epsilon(\dot{T}^{(2)}_{\mathbf{y}^{(s-1)}}) \tilde{q}\left(\theta^{(s-1)}, \theta^*\right)}\right\}$
  Draw $u \sim \text{Unif}[0, 1]$
  if $u < \alpha$ **then**
    Assign $(\theta^{(s)}, \dot{T}^{(2)}_{\mathbf{y}^{(s)}}) \leftarrow (\theta^*, \dot{T}^{(2)}_{\mathbf{y}^*})$ with
  else
    Assign $(\theta^{(s)}, \dot{T}^{(2)}_{\mathbf{y}^{(s)}}) \leftarrow (\theta^{(s-1)}, \dot{T}^{(2)}_{\mathbf{y}^{(s-1)}})$
  end if
end for

mixes two or more random variables. Here we assume that, at each time t, the random variable X_t is a mixture of two discrete random variables, the random variable X_{t-1} and the so called *innovation term* ϵ_t:

$$X_t = (X_{t-1}, \lambda) * (\varepsilon_t, 1 - \lambda). \tag{116}$$

Stated otherwise

$$X_t = \begin{cases} X_{t-1} & \text{with probability } \lambda \\ \varepsilon_t & \text{with probability } 1 - \lambda. \end{cases}$$

The mixing weights are $\lambda \in [0, 1]$ and $1 - \lambda$. The innovation term, ε_t, is a discrete random variable distributed over $\mathcal{X}$ according to the probability distribution

$\theta \triangleq (\theta_1, ..., \theta_k) \in \Delta^{k-1}$. Accordingly, the Markov chain $\{X_t\}$ is characterized by the following transitions probabilities:

$$Q_{\theta,\lambda}(i,j) \triangleq \Pr(X_{t+1} = i | X_t = j, \lambda, \theta) = \lambda \mathbb{1}\{i = j\} + (1-\lambda)\theta_i.$$

Let us assume to observe a sample path $\mathbf{X}^n = X_1, ...X_n$ from the autoregressive process described above. For the sake of simplicity, we assume that the starting point of the chain, x_1, is fixed. Accordingly, the related likelihood function can be retrieved as follows:

$$\begin{aligned}
&\Pr(X_2 = x_2, ..., X_{n+1} = x_{n+1} | x_1, \lambda, \theta) \\
&= \prod_{t=1}^{n} \Pr(X_{t+1} = x_{t+1} | X_t = x_t, \lambda, \theta) \\
&= \prod_{t=1}^{n} \left(\lambda \mathbb{1}\{x_{t+1} = x_t\} + (1-\lambda)\theta(x_{t+1})\right) \\
&= \prod_{t=1}^{n} \left(\lambda + (1-\lambda)\theta(x_{t+1})\right)^{\mathbb{1}\{x_{t+1}=x_t\}} \left((1-\lambda)\theta(x_{t+1})\right)^{1-\mathbb{1}\{x_{t+1}=x_t\}} \\
&= \prod_{j=1}^{k} \left(\lambda + (1-\lambda)\theta_j\right)^{n T_x^{(2)}(j,j)} \left((1-\lambda)\theta_j\right)^{n \sum_{i \neq j} T_x^{(2)}(i,j)} \\
&= \lambda^{n \sum_{j=1}^{k} T_x^{(2)}(j,j)} (1-\lambda)^{n \sum_{j=1}^{k} \sum_{i \neq j} T_x^{(2)}(i,j)} \prod_{j=1}^{k} \left(1 + \frac{1-\lambda}{\lambda}\theta_j\right)^{n T_x^{(2)}(j,j)} \theta_j^{n \sum_{i \neq j} T_x^{(2)}(i,j)}
\end{aligned}$$

where $\theta(x_{t+1})$ denotes the the probability of x_{t+1} according to θ.

In [1], they showed that there is no real gain in implementing a MCMC method and proposed a standard Importance Sampling for sampling from the posterior distribution obtained by assuming the following prior distributions:

$$\begin{aligned}
&\theta \sim \mathrm{Dirichlet}(\alpha_1, ..., \alpha_k) \\
&\lambda \sim \mathrm{Beta}(a, b).
\end{aligned}$$

In such a case, the joint posterior distribution becomes

$$\begin{aligned}
\pi(\lambda, \theta | \mathbf{x}^n) &\propto \pi(\lambda)\pi(\theta) \Pr(X_2 = x_2, ..., X_n = x_n | x_1, \lambda, \theta) \\
&\approx \pi(\lambda)\pi(\theta) \Pr(X_2 = x_2, ..., X_{n+1} = x_{n+1} | x_1, \lambda, \theta) \\
&\propto \left[\prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \lambda^{a-1} (1-\lambda)^{b-1}\right] \\
&\times \left[\lambda^{n \sum_{j=1}^{k} T_{x^n}^{(2)}(j,j)} (1-\lambda)^{n \sum_{j=1}^{k} \sum_{i \neq j} T_{x^n}^{(2)}(i,j)} \prod_{j=1}^{k} \left(1 + \frac{1-\lambda}{\lambda}\theta_j\right)^{n T_{x^n}^{(2)}(j,j)} \theta_j^{n \sum_{i \neq j} T_{x^n}^{(2)}(i,j)}\right] \\
&= \left[\prod_{j=1}^{k} \theta_j^{\alpha_j + n \sum_{i \neq j} T_{x^n}^{(2)}(i,j) - 1}\right] \left[\lambda^{a + n \sum_{j=1}^{k} T_{x^n}^{(2)}(j,j) - 1} (1-\lambda)^{b + n \sum_{j=1}^{k} \sum_{i \neq j} T_{x^n}^{(2)}(i,j) - 1}\right] \\
&\times \left[\prod_{j=1}^{k} \left(1 + \frac{1-\lambda}{\lambda}\theta_j\right)^{n T_{x^n}^{(2)}(j,j)}\right].
\end{aligned}$$

It follows that samples from the posterior distribution can be got via a standard Importance Sampling with importance distributions for θ and λ equal respectively to

$$\mathrm{Dirichlet}\big(\alpha_1 + n\sum_{i\neq 1} \dot{T}_x^{(2)}(i,1), ..., \alpha_k + n\sum_{i\neq k} \dot{T}_x^{(2)}(i,k)\big)$$

$$\mathrm{Beta}\big(a + n\sum_{j=1}^{k} \dot{T}_x^{(2)}(j,j), b + n\sum_{j=1}^{k}\sum_{i\neq j} \dot{T}_x^{(2)}(i,j)\big).$$

Accordingly, the importance weights equal

$$\prod_{j=1}^{k}\left(1 + \frac{1-\lambda^{(s)}}{\lambda^{(s)}}\theta_j^{(s)}\right)^{n\dot{T}_x^{(2)}(j,j)} \qquad \forall s \in \{1, ..., S\}$$

and are easy to compute.

However we note that this simplification arises only with proper prior distributions. In all the other cases MCMC methods require the computation of the likelihood at each iteration.

In Section 6.4.1, for the sake of comparing the performances of standard ABC methods and LD-ABC, we consider the example described above. We test the performance of our method at work on simulated data by assuming the results provided by the Importance Sampling proposed in [1] as a benchmark. Noting that the size of the second order type scales quadratically with the cardinality of $|\mathcal{X}|$, we are also interested in evaluating how this fact affects the quality of the approximation. To this end, we also compare the performance of LD-ABC with a standard R-ABC using a non-sufficient lower-dimensional statistic.

In Section 6.4.2 we test both ABC and LD-ABC at work on real data by assuming non-conjugate prior distributions.

6.4.1 *Example 1*

In this toy example we consider a simulated time series $\mathbf{X}^{60} = X_1, ..., X_{60}$ taking values in $\mathcal{X} = \{1, 2, 3\}$.

We assume the following prior distributions:

$$(\theta_1, \theta_2, \theta_3) \sim \mathrm{Dirichlet}(1, 1, 1)$$
$$\lambda \sim \mathrm{Beta}(1, 1).$$

Note that this choice allows implementing the Importance Sampling proposed in [1].

The ABC threshold is choosen as the 0.0001-quantile of the distribution of the distances between the observed and simulated types, thus setting $\epsilon = 0.005$. We ran both the R-ABC and the LD-ABC (Algorithm 15) with $m = 120$ and $S = 100,000$. We also ran the Importance Sampling described in [1] with $S = 100,000$. Table 8 shows the results of the simulations both in terms of $\widehat{\mathrm{MSE}}$ and $\widehat{\mathrm{MISE}}$, computed by averaging over 100 reruns. We can see that LD-ABC outperforms the standard ABC in terms of point estimates for each of the four parameters. Moreover, our method leads to $\widehat{\mathrm{MISE}}$'s about ten times smaller then the R-ABC, as also shown in Figure 12. The figure shows also the posterior distributions approximated by the R-ABC using

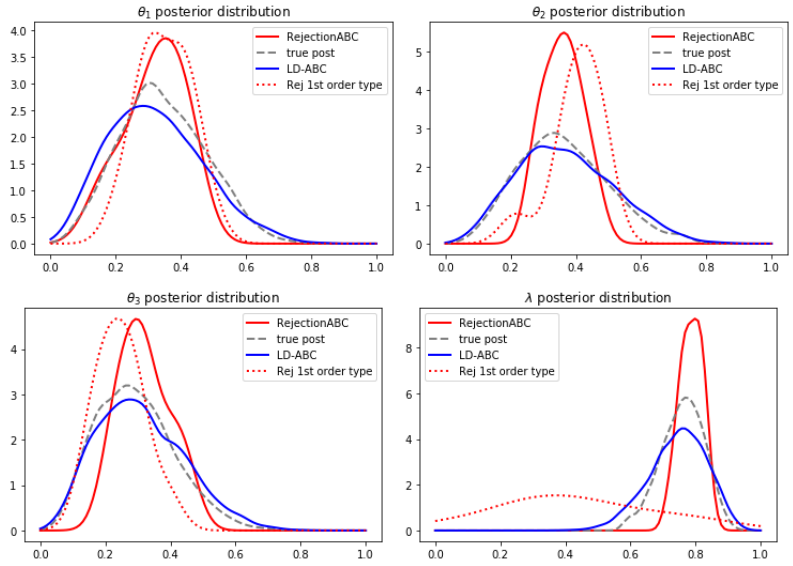


Figure 12: Posterior distributions corresponding to $m = 120$ and $\epsilon = 0.005$. Each plot refers to one of the four parameters of the model. Red lines represent the posterior density estimates provided via R-ABC. The blue lines represent the estimates provided via LD-ABC. The dotted red lines are the estimates provided by the R-ABC using the first order type. The dashed gray lines are the true posterior distributions.

a lower-dimensional summary statistic (dotted red lines): the first order type. As is apparent, such non-sufficient summary statistic leads to results even worse then the R-ABC using the second order type.

Another interesting feature is the variability of the estimates. Looking at Figure 13 we can see that the approximation of the cdf's is highly variable when relying on the standard ABC, while the LD-ABC exhibits narrow intervals. This result is caused by the serious sample degeneracy leading the R-ABC to an average $\widehat{\text{ESS}}$ equal to just 11. Hence, in this example is apparent that LD-ABC is able to mitigate sample degeneracy leading to $\widehat{\text{ESS}}$ equal to $4,619$ and to less variable estimates.

Table 8: Squared errors and integrated squared errors averaged over 100 reruns. Each column contains results for one of the parameters of the model both for LD-ABC and R-ABC.

		$m = 120$, $\epsilon = 0.005$			
		θ_1	θ_2	θ_3	λ
$\widehat{\text{MSE}}_{mean}$	LD	$4.56 \cdot 10^{-4}$	$0.76 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$	$1.54 \cdot 10^{-4}$
	R	$13.59 \cdot 10^{-4}$	$16.35 \cdot 10^{-4}$	$8.99 \cdot 10^{-4}$	$6.63 \cdot 10^{-4}$
$\widehat{\text{MSE}}_{var}$	LD	$0.21 \cdot 10^{-4}$	$0.24 \cdot 10^{-4}$	$0.12 \cdot 10^{-4}$	$0.12 \cdot 10^{-4}$
	R	$0.64 \cdot 10^{-4}$	$0.59 \cdot 10^{-4}$	$0.54 \cdot 10^{-4}$	$0.06 \cdot 10^{-4}$
$\widehat{\text{MISE}}$	LD	0.0780	0.028	0.0274	0.1922
	R	0.2575	0.3162	0.3679	1.0681

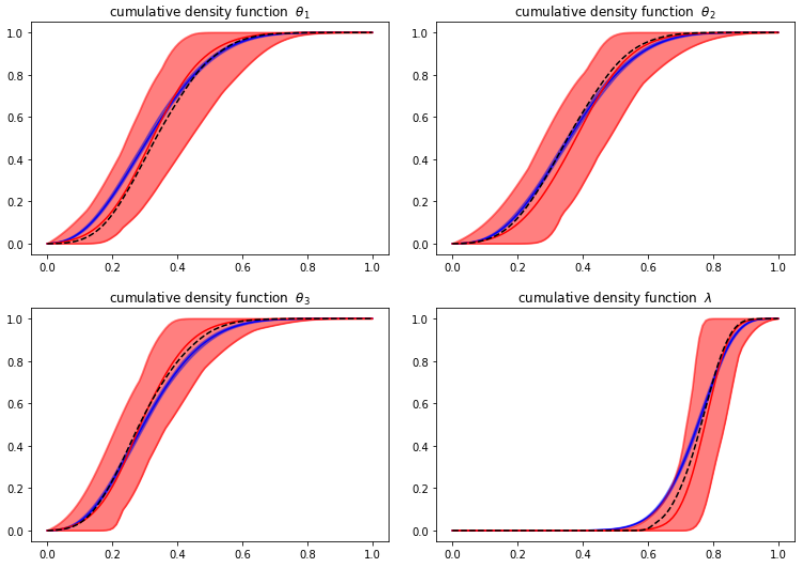


Figure 13: Posterior cumulative density functions for $\theta_1, \theta_2, \theta_3$ and λ. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cdf. 99% intervals over 100 rerun of each algorithm are also represented.

6.4.2 *Example 2*

In this example we consider the same model in Example 1 but assuming different prior distributions. In such a case, the evaluation of the likelihood function is needed at each iteration to implement a MCMC. To avoid this computational effort and can resort to ABC methods.

We test both LD-ABC and R-ABC on real-world data reported in [146]. The dataset consists of a collection of categorical time series of infant sleep status in an EEG study. Infant sleep states are categorized into six possible states: 1. quiet sleep-trace alternant; 2. quiet sleep-high voltage; 3. indeterminate sleep; 4. active sleep-low voltage; 5. active sleep-mixed; 6. awake. Here we consider as observed data $x_1, ..., x_{120}$ one of the 24 reported time series, represented in Figure 14. We drop the status *awake* as never observed in the time series and merge the other status in 1. quite sleep; 2. indeterminate sleep; 3. active sleep.

We assume two logistic normal distributions [2] as prior distributions. Specifically, we assume $\theta \sim \mathtt{LogisticNormal}(\mu_\theta, \Sigma)$ and $\lambda \sim \mathtt{LogisticNormal}(\mu_\lambda, \sigma^2)$ with

$$\mu_\theta = (0,0) \quad \Sigma = \begin{bmatrix} 1.45 & 0 \\ 0 & 1.45 \end{bmatrix}$$
$$\mu_\lambda = 0 \quad \sigma^2 = 1.$$

Note that assuming a logistic normal distribution over the simplex corresponds to assume a multivariate normal distribution for a random variable $\mathbf{Z} \in \mathbb{R}^d$ whose additive logistic transformation maps the d-dimensional random variable to a $(d+1)$-dimensional vector in the simplex Δ^d.

We run both the algorithms for $S = 1,000,000$ iterations both with $\epsilon = 0.05$ and $\epsilon = 0.01$. Table 9 displays the $\widehat{\text{ESS}}$'s. We can see that with a low threshold (0.01)

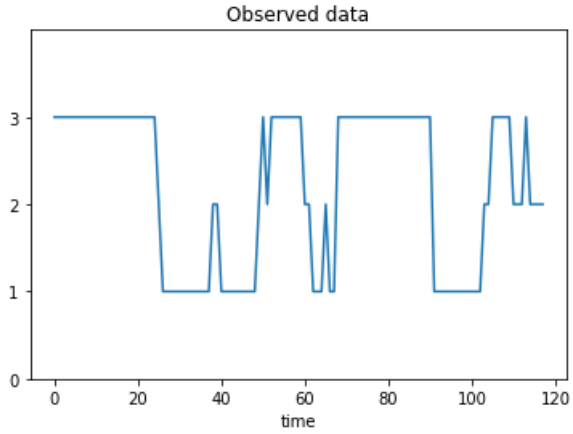


Figure 14: Observed time series $\mathbf{x}^{120} = x_1, ..., x_{120}$.

Table 9: $\widehat{\text{ESS}}$ achieved by R-ABC and LD-ABC after $S = 1,000,000$ iterations with $\epsilon = 0.01$ and $\epsilon = 0.05$

	$\widehat{\text{ESS}}$	
	$\epsilon = 0.01$	$\epsilon = 0.05$
LD	22 785	115 140
R	3	2 290

the R-ABC accepts only three parameter proposals. It follows that the resulting sample is inadequate to approximate the posterior distributions. On the other hand, the LD-ABC provides an adequate sample size in both cases ($\epsilon = 0.05$ and $\epsilon = 0.01$). This is also apparent looking at Figure 15. The posterior distributions approximated by LD-ABC with the two different thresholds (solid and dashed blue lines) are quite similar while R-ABC is strongly affected by the choice of the threshold. Thus, we can conclude that also in this example our method mitigates the sample degeneracy problem by increasing the ESS and possibly improves the approximation of the posterior distributions relaying on a larger sample.

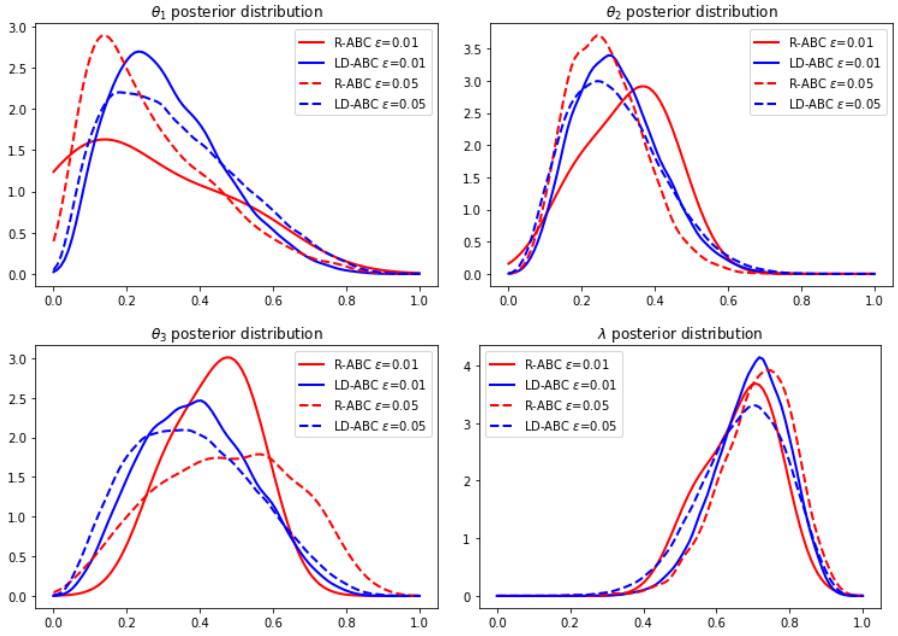


Figure 15: Approximate posterior distributions of each parameter with $S = 1,000,000$, $\epsilon = 0.01$ (solid lines) and $\epsilon = 0.05$ (dashed lines).

Part III

SIMULATED INFERENCE FOR LEARNING FROM ANONYMIZED DATA

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

— Marie Curie

7

INTRODUCTION

In this part of the thesis we consider the problem of learning from data anonymized through group-based anonymization scheme, which is a popular approach to data publishing. This method aims at protecting the privacy of the individuals involved in a dataset, by releasing an obfuscated version of the original data, where the exact correspondence between individuals and attribute values is hidden. When publishing data about individuals, one must balance the *learner*'s utility against the risk posed by an *attacker*, potentially targeting individuals in the dataset. We consider two class of group-based anonymization schemes, the *horizontal* and *vertical* schemes, and propose a unified Bayesian model of group-based schemes and a related MCMC method to learn the population parameters from an anonymized table. This allows one to analyse the risk for any individual in the dataset to be linked to a specific sensitive value, when the attacker knows the individual's nonsensitive attributes, beyond what is implied for the general population. We call this *relative threat* analysis. We illustrate the results obtained with the proposed methodology on a real-world dataset.

Furthermore, we consider ABC methods as an alternative strategy for inferring the population parameters from obfuscated data. We define a *generative model* reproducing the stochastic process leading to the observation of an anonymized dataset, thus enabling inference via ABC. Finally, we test the method proposed in Chapter 5 at work on an example of anonymized table.

STRUCTURE OF PART III In Chapter 8 we introduce the problem of learning from anonymized tables and the relative threats approach. In particular, in Section 8.2 we propose a unified formal definition of vertical and horizontal schemes. Based on that, measures of (relative) privacy threats and utility are introduced in Section 8.4. In Section 8.5, we study a MCMC algorithm to learn the population parameters posterior distributions and the attacker's probability distribution learned from the anonymized data. In Section 8.6, we illustrate the results of an experiment conducted on a real-world dataset. A few concluding remarks and perspectives for future research are reported in Section 8.7. In Chapter 9 we introduce ABC methods as an alternative for learning population parameters from anonymized data. In Section 9.2 we test LD-ABC at work on an obfuscated table. Some technical material has been confined to Appendix E.

8

RELATIVE PRIVACY THREATS AND LEARNING FROM ANONYMIZED DATA

We consider a scenario where datasets containing personal microdata are released in anonymized form. The goal here is to enable the computation of general population characteristics with reasonable accuracy, at the same time preventing leakage of sensitive information about individuals in the dataset. The Database of Genotype and Phenotype [89], the U.K. Biobank [106] and the UCI Machine Learning repository [75] are well-known examples of repositories providing this type of datasets.

Anonymized datasets always have "personal identifiable information", such as names, SSNs and phone numbers, removed. At the same time, they include information derived from nonsensitive (say, gender, ZIP code, age, nationality) as well as sensitive (say, disease, income) attributes. Certain combinations of nonsensitive attributes, like ⟨gender, date of birth, ZIP code⟩, may be used to uniquely identify a significant fraction of the individuals in a population, thus forming so-called *quasi-identifiers*. For a given target individual, the *victim*, an attacker might easily obtain this piece of information (e.g. from personal web pages, social networks etc.), use it to identify him/her within a dataset and learn the corresponding sensitive attributes. This attack was famously demonstrated by L. Sweeney, who identified Massachusetts' Governor Weld medical record within the Group Insurance Commission (GIC) dataset [147]. Note that *identity disclosure*, that is the precise identification of an individual's record in a dataset, is not necessary to arrive at a privacy breach: depending on the dataset, an attacker might infer the victim's sensitive information, or even a few highly probable candidate values for it, without identity disclosure involved. This more general type of threat, *sensitive attribute disclosure*, is the one we focus on here[1].

In an attempt to mitigate such threats for privacy, regulatory bodies mandate complex, often baroque syntactic constraints on the published data. As an example, here is an excerpt from the HIPAA *safe harbour* deidentification standard [153], which prescribes a list of 18 identifiers that should be removed or obfuscated, such as

> *all geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) the geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) the initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.*

There exists a large body of research, mainly in Computer Science, on syntactic methods. In particular, group-based anonymization techniques have been systematically investigated, starting with L. Sweeney's proposal of k-*anonymity* [147], followed

1 Depending on the nature of the dataset, the mere *membership disclosure*, i.e. revealing that an individual is present in a dataset, may also be considered as a privacy breach: think of data about individuals who in the past have been involved in some form of felony. We will not discuss membership disclosure privacy breaches in this thesis.

Table 10: A table (top) anonymized according to 2-anonymity via local recoding (bottom-left) and Anatomy (bottom-right).

ID	Nat.	ZIP	Dis.
1	Malaysia	45501	Heart
2	Japan	45502	Flu
3	Japan	55503	Flu
4	Japan	55504	Stomach
5	China	66601	HIV
6	Japan	66601	Diabetes
7	India	77701	Flu
8	Malaysia	77701	Heart

a) Original table

ID	Nat.	ZIP	Dis.
1	{M, J}	4550*	Heart
2	{M, J}	4550*	Flu
3	Japan	5550*	Flu
4	Japan	5550*	Stomach
5	{C, J}	66601	HIV
6	{C, J}	66601	Diabetes
7	{I, M}	77701	Flu
8	{I, M}	77701	Heart

b) 2-anonymity via local recoding

GID	Nat.	ZIP	Dis.
1	Japan	45502	Heart
1	Malaysia	45501	Flu
2	Japan	55504	Flu
2	Japan	55503	Stomach
3	Japan	66601	HIV
3	China	66601	Diabetes
4	Malaysia	77701	Flu
4	India	77701	Heart

c) Anatomy

by its variants, like *ℓ-diversity* [87] and *Anatomy* [158]. In group-based methods, the anonymized – or obfuscated – version of a table is obtained by partitioning the set of records into groups, which are then processed to enforce certain properties. The rationale is that, even knowing that an individual belongs to a group of the anonymized table, it should not be possible for an attacker to link that individual to a specific sensitive value in the group. Two examples of group based anonymization are in Table 10, adapted from [157]. The topmost, original table collects medical data from eight individuals; here *Disease* is considered as the only sensitive attribute. The central table is a 2-anonymous, 2-diverse table: within each group the nonsensitive attribute values have been generalized following group-specific rules (*local recoding*) so as to make them indistinguishable; moreover, each group features 2 distinct sensitive values. In general, each group in a k-anonymous table consists of at least k records, which are indistinguishable when projected on the nonsensitive attributes; ℓ-diversity additionally requires the presence in each group of at least ℓ distinct sensitive values, with approximately the same frequency. This is an example of *horizontal* scheme. Table 10 (c) is an example of application of the *Anatomy* scheme: within each group, the nonsensitive part of the rows are *vertically* and *randomly* permuted, thus breaking the link between sensitive and nonsensitive values. Again, the table is 2-diverse.

In recent years, the effectiveness of syntactic anonymization methods has been questioned, as offering weak guarantees against attackers with strong background knowledge – very precise contextual information about their victims. *Differential pri-*

vacy [45], which promises protection in the face of *arbitrary* background knowledge, while valuable in the release of summary statistics, still appears not of much use when it comes to data publishing (see the following section). As a matter of fact, release of syntactically anonymized tables appears to be the most widespread data publishing practice, with quite effective tool support (see e.g. [113]).

Here, discounting the risk posed by attackers with strong background knowledge, we pose the problem in relative terms: given that whatever is *learned about the general population* from an anonymized dataset represents legitimate and useful information ("smoke is associated with cancer"), one should prevent an attacker from drawing conclusions about specific individuals in the table ("almost certainly the target individual has cancer"): in other words, learning sensitive information for an individual in the dataset, *beyond* what is implied for the general population. To see what is at stake here, consider dataset (b) in Table 10. Suppose that the attacker's victim is a Malaysian living at ZIP code 45501, and known to belong to the original table. The victim's record must therefore be in the first group of the anonymized table. The attacker may reason that, with the exception of the first group, a Japanese is never connected to Heart Disease; this hint can become a strong evidence in a larger, real-world table. Then the attacker can link with high probability the Malaysian victim in the first group to Heart Disease. In this attack, the attacker combines knowledge of the nonsensitive attributes of the victim (Malaysian, ZIP code 45501) with the group structure and the knowledge learned from the anonymized table.

We propose a unified probabilistic model to reason about such forms of leakage. In doing so, we clearly distinguish the position of the *learner* from that of the *attacker*: the resulting notion is called *relative privacy threat*. In our proposal, both the learner and the attacker activities are modeled as forms of Bayesian inference: the acquired knowledge is represented as a joint posterior probability distribution over the sensitive and nonsensitive values, given the anonymized table *and*, in the case of the attacker, knowledge of the victim's presence in the table. A comparison between these two distributions determines what we call relative privacy threat. Since posterior distributions are in general impossible to express analytically, we also put forward a MCMC method to practically estimate such posteriors. We also illustrate the results of applying our method to the Adult dataset from the UCI Machine Learning repository [75], a common benchmark in anonymization research.

8.1 RELATED WORKS

Sweeney's k-anonymity [147] is among the most popular proposals aiming at a systematic treatment of syntactic anonymization of microdata. The underlying idea is that every individual in the released dataset should be hidden in a "crowds of k". Over the years, k-anonymity has proven to provide weak guarantees against attackers who know much about their victims, that is have a strong background knowledge. For example, an attacker may know from sources other than the released data that his victim does *not* suffer from certain diseases, thus ruling out all possibilities but one in the victims's group. Additional constraints may be enforced in order to mitigate those attacks, like ℓ-diversity [87] and *t-closeness* [80]. Differential Privacy [45] promises protection in the face of arbitrary background knowledge. In its basic, interactive version, this means that, when querying a database via a differentially private

mechanism, one will get approximately the same answers, whether the data of any specific individual is included or not in the database. This is typically achieved by injecting controlled levels of noise in the reported answer, e.g. Laplacian noise. Differential Privacy is very effective when applied to certain summary statistics, such as histograms. However, it raises a number of difficulties when applied to table publishing: in concrete cases, the level of noise necessary to guarantee an acceptable degree of privacy would destroy utility [34, 35, 135]. Moreover, due to correlation phenomena, it appears that Differential Privacy cannot in general be used to control evidence about the participation of individuals in a database [12, 74]. In fact, the no-free-lunch theorem of Kifer and Machanavajjhala [74] implies that it is impossible to guarantee both privacy and utility, without making assumptions about how the data have been generated (e.g., independence assumptions). Clifton and Tassa [25] critically review issues and criticisms involved in both syntactic methods and Differential Privacy, concluding that both have their place, in Privacy Preserving– Data Publishing and Data Mining, respectively. Both approaches have issues that call for further research. A few proposals involve blending the two approaches, with the goal to achieve both strong privacy guarantees and utility, see e.g. [81].

A major source of inspiration for our work has been Kifer's [73]. The main point of [73] is to demonstrate a pitfall of the *random worlds* model, where the attacker is assumed to assign equal probability to all cleartext tables compatible with the given anonymized one. Kifer shows that a Bayesian attacker willing to learn from the released table can draw sharper inferences than those possible in the random worlds model. In particular, Kifer shows that it is possible to extract from (anatomized) ℓ-diverse tables belief probabilities greater than $1/\ell$, by means of the so-called deFinetti attack. While pinpointing a deficiency of the random worlds model, it is questionable if this should be considered an attack, or just a legitimate learning strategy. Quoting [25] on the deFinetti attack:

> The question is whether the inference of a general behavior of the population in order to draw belief probabilities on individuals in that population constitutes a breach of privacy (...). To answer this question positively for an attack on privacy, the success of the attack when launched against records that *are* part of the table should be significantly higher than its success against records that *are not* part of the table. We are not aware of such a comparison for the deFinetti attack.

It is this very issue that we tackle in the present chapter. Specifically, our main contribution here is to put forward a concept of relative privacy threat, as a means to assess the risks implied by publishing tables anonymized via group-based methods. To this end, we introduce: (a) a unified probabilistic model for group-based schemes; (b) rigorous characterizations of the learner and the attacker's inference, based on Bayesian reasoning; and, (c) a related MCMC method, which generalizes and systematizes that proposed in [73].

Very recently, partly inspired by differential privacy, a few authors have considered what might be called a relative or *differential* approach to assessing privacy threats, in conjunction with some notion of learning or inference from the anonymized data. Especially relevant to our work is *differential inference*, introduced in a recent paper by Kassem et al. [72]. These authors make a clear distinction between two different

types of information that can be inferred from anonymized data: learning of "public" information, concerning the population, should be considered as legitimate; on the contrary, leakage of "private" information about individuals should be prevented. To make this distinction formal, given a dataset, they compare two probability distributions that can be machine-learned from two distinct training sets: one including and one excluding a target individual. An attack exists if there is a significant difference between the two distributions, measured e.g. in terms of Earth Moving Distance. While similar in spirit to ours, this approach is conceptually and technically different from what we do here. Indeed, in our case the attacker explicitly takes advantage of the extra piece of information concerning the presence of the victim in the dataset to attack the target individual, which leads to a more direct notion of privacy breach. Moreover, in [72] a Bayesian approach to inference is not clearly posed, so the obtained results lack a semantic foundation, and strongly depend on the adopted learning algorithm. Pyrgelis et al. [117] use Machine Learning for membership inference on aggregated location data, building a binary classifier that can be used to predict if a target user is part of the aggregate data or not. A similar goal is pursued in [101]. Again, a clear semantic foundation of these methods is lacking, and the obtained results can be validated only empirically. In a similar vein, Bichsel et al. [10] and Ding et al. [41] have proposed statistical techniques to detect privacy violations, but they only apply to differential privacy. Other works, such as [64] and [90], have just considered the problem of how to effectively learn from anonymized datasets, but not of how to characterize legitimate, as opposed to non-legitimate, inference.

On the side of the random worlds model, Chi-Wing Wong et al. [157] show how information on the population extracted from the anonymized table – in the authors' words, the *foreground* knowledge – can be leveraged by the attacker to violate the privacy of target individuals. The underlying reasoning, though, is based on the random worlds model, hence is conceptually and computationally very different from the Bayesian model adopted in the present chapter. Bewong et al. [9] assess relative privacy threat for transactional data by a suitable extension of the notion of t-*closeness*, which is based on comparing the relative frequency of the victim's sensitive attribute in the whole table with that in the victim's group. Here the underlying assumption is that the attacker's prior knowledge about sensitive attributes matches the public knowledge, and that the observed sensitive attributes frequencies provide good estimates both for the public knowledge and the attacker's belief. Our proposal yields more sophisticated estimates via a Bayesian inferential procedure. Moreover, in our scenario the assumption on the attacker's knowledge is relaxed requiring only the knowledge of the victim's presence in whatever group of the table.

A concept very different from the previously discussed proposals is Rubin's *multiple imputation* approach [130], by which only tables of *synthetic* data, generated sampling from a predictive distribution learned from the original table, are released. This avoids syntactic masking/obfuscation, whose analysis requires customized algorithms on the part of the learner, and leaves to the data producer the burden of synthesis. Note that this task can be nontrivial and raises a number of difficulties concerning the availability of auxiliary variables for non-sampled units, see [119]. In Rubin's view, synthetic data overcome all privacy concerns, in that no real individual's data is actually released. However, this position has been questioned, on the grounds that information about participants may leak through the chain: original ta-

ble → posterior parameters → synthetic tables. In particular, Machanavajjhala et al. [88] study Differential Privacy of synthetic categorical data. They show that the release of such data can be made differentially private, at the cost of introducing very powerful priors. However, such priors can lead to a serious distortion in whatever is learned from the data, thus compromising utility. In fact, [138] argues that, in concrete cases, the required pseudo sample size hyperparameter could be larger than the size of the table. Experimental studies [21, 22] appear to confirm that such distorting priors are indeed necessary for released synthetic data to provide acceptable guarantees, in the sense of Differential Privacy. See [138] for a recent survey of results about synthetic data release and privacy.

In [107], Park and Jitkrittum propose a strategy for allowing posterior inference obaying to Differential Privacy requirments by means of ABC. Specifically, their proposal, the ABCDP mechanism, produces samples from an approximated posterior distribution obeying to the notion of Differential Privacy. The key idea is that the ABC threshold, ϵ, is related to the privacy guarantees in the posterior samples. In order to obtain the smallest privacy loss, despite the repeated use of the data (in the typical comparison step in ABC), they rely on Renyi Differential Privacy [99]. Further considerations about the connection between ABC and data anonymization are in Chapter 9.

8.2 GROUP BASED ANONYMIZATION SCHEMES

A dataset consists of a collection of rows, where each row corresponds to an individual. Formally, let $\mathcal{R}$ and $\mathcal{S}$, ranged over by r and s respectively, be finite non-empty sets of *nonsensitive* and *sensitive* values, respectively. A *row* is a pair $(s, r) \in \mathcal{S} \times \mathcal{R}$. There might be more than one sensitive and nonsensitive characteristic, so s and r can be thought of as vectors.

A *group-based anonymization algorithm* $\mathcal{A}$ is an algorithm that takes a multiset of rows as input and yields an obfuscated table as output, according to the scheme

$$\text{multiset of rows} \longrightarrow \text{cleartext table} \longrightarrow \text{obfuscated table.}$$

Formally, fix $N \geqslant 1$. Given a multiset of N rows, $d = \{\!\{(s_1, r_1), ..., (s_N, r_N)\}\!\}$, $\mathcal{A}$ will first arrange d into a sequence of *groups*, $\mathbf{x} = g_1, ..., g_k$, the *cleartext table*. Each group in turn is a sequence of n_i rows, $g_i = (s_{i,1}, r_{i,1}), ..., (s_{i,n_i}, r_{i,n_i})$, where n_i can vary from group to group. Note that both the number of groups, $k \geqslant 1$, and the number of rows in each group, n_i, depend in general on the original multiset d as well as on properties of the considered algorithm – such as ensuring k-anonymity and ℓ-diversity (see below). The obfuscated table is then obtained as a sequence $\mathbf{x}^* = g_1^*, ..., g_k^*$, where the obfuscation of each group g_i is a pair $g_i^* = (m_i, l_i)$. Here, each $m_i = s_{i,1}, ..., s_{i,n_i}$ is the sequence of *sensitive* values occurring in g_i; each l_i, called *generalized nonsensitive value*, is one of the following:

- for *horizontal* schemes, a *superset* of g_i's nonsensitive values: $l_i \supseteq \{r_{i,1}, ..., r_{i,n_i}\}$;
- for *vertical* schemes, the *multiset* of g_i's nonsensitive values: $l_i = \{\!\{r_{i,1}, ..., r_{i,n_i}\}\!\}$.

Note that the generalized nonsensitive values in vertical schemes include all and only the values, with multiplicities, found in the corresponding original group. On

the other hand, generalized nonsensitive values in horizontal schemes may include additional values, thus generating a superset. What values enter the superset depends on the adopted technique, e.g., micro-aggregation, generalization or suppression; in any case this makes the rows in each group indistinguishable when projected onto the nonsensitive attributes. For example, each of 45501, 45502 is generalized to the superset $4550* = \{45500, 45501, ..., 45509\}$ in the first group of Table 10 (b).

Sometimes it will be notationally convenient to ignore the group structure of $\mathbf{x}$ altogether, and regard the cleartext table $\mathbf{x}$ simply as a sequence of rows, $(s_1, r_1), (s_2, r_2), ..., (s_1, s_N)$. Each row (s_j, r_j) is then uniquely identified within the table $\mathbf{x}$ by its index $1 \leqslant j \leqslant N$.

An instance of horizontal schemes is k-*anonymity* [147]: in a k-anonymous table, each group consists of at least $k \geqslant 1$ rows, where the different nonsensitive values appearing within each group have been generalized so as to make them indistinguishable. In the most general case, different occurrences of the same nonsensitive value might be generalized in different ways, depending on their position (index) within the table $\mathbf{x}$: this is the case of *local recoding*. Alternatively, each occurrence of a nonsensitive value is generalized in the same way, independently of its position: this is the case of *global recoding*. Further conditions may be imposed on the resulting anonymized table, such as ℓ-*diversity*, requiring that at least $\ell \geqslant 1$ distinct values of the sensitive attribute appear in each group. Table 10 (bottom-left) shows an example of $k = 2$-anonymous and $\ell = 2$-diverse table: in each group the nonsensitive values are indistinguishable and two different sensitive values (diseases) appear in each group.

An instance of vertical schemes is *Anatomy* [158]: within each group, the link between the sensitive and nonsensitive values is hidden by randomly permuting one of the two parts, for example the nonsensitive one. As a consequence, an anatomized table may be seen as consisting of *two* sub-tables: a sensitive and a nonsensitive one. Table 10 (c) shows an example of anatomized table: in the nonsensitive sub-table, the reference to the corresponding sensitive values is lost; only the multiset of nonsensitive values appears for each group.

Remark 5 (disjointness) Some anonymization schemes enforce the following disjointness property on the obfuscated table $\mathbf{x}^*$:

> Any two generalized nonsensitive values in $\mathbf{x}^*$ are disjoint: $i \neq j$ implies $l_i \cap l_j = \emptyset$.

We need not assume this property in our treatment – although assuming it may be computationally useful in practice (see Section 8.3).

For ease of reference, we provide a summary of the notation that will be used throughout the chapter in Table 11.

8.3 A UNIFIED PROBABILISTIC MODEL

We provide a unified probabilistic model for reasoning on group-based schemes. We first introduce the random variables of the model together with their joint density function. On top of these variables, we then define the probability distributions on $\mathcal{S} \times \mathcal{R}$ that formalize the *learner* and the *attacker* knowledge, given the obfuscated table.

Table 11: Summary of notation.

Symbol	Description	Symbol	Description
A	attacker	β	$\theta_{R\|S}$ hyperparameters
α	θ_S hyperparameters	δ	nonsensitive freq.
γ	sensitive freq.	g_i^*	obfuscated group i
g_i	group i	$\mathbf{GT}_A$	global threat level
ETV	emp. total variation	k	number of groups
I	evaluator (ideal)	k	min size of groups s
l_i	group i nonsens. values	L	learner
ℓ	min n. of sens. val.	m_i	group i sens. values
N	n. of rows in the table	θ	parameters of R, S
$\theta_{R\|s}$	parameters of $R\|s$	θ_S	parameters of S
R	nonsensitive r.v.	S	sensitive r.v.
$\mathbf{x}$	clear text table	$\mathbf{x}^*$	obfuscated table
Ti	rel. threat level	**TV**	total variation
RF	rel. faithfulness level	v	victim

8.3.1 *Random variables*

The model consists of the following random variables.

- θ, taking values in the set of full support probability distributions Θ over $\mathcal{S} \times \mathcal{R}$, is the joint probability distribution of the sensitive and nonsensitive attributes in the population.
- $\mathbf{X} = G_1, ..., G_k$, taking values in the set of cleartext tables $\mathcal{X}$. Each group G_i is in turn a sequence of $n_i \geqslant 1$ consecutive rows in $\mathbf{X}$, $G_i = (S_{i,1}, R_{i,1}), ..., (S_{i,n_i}, R_{i,n_i})$. The number of groups k is not fixed, but depends on the anonymization scheme and the specific tuples composing $\mathbf{X}$.
- $\mathbf{X}^* = G_1^*, ..., G_k^*$, taking values in the set of obfuscated tables $\mathcal{X}^*$.

We assume that the above three random variables form a Markov chain:

$$\theta \longrightarrow \mathbf{X} \longrightarrow \mathbf{X}^* . \tag{117}$$

In other words, uncertainty on $\mathbf{X}$ is driven by θ, and $\mathbf{X}^*$ solely depends on the table $\mathbf{X}$ and the underlying obfuscation algorithm. As a result, $\mathbf{X}^* \perp\!\!\!\perp \theta \mid \mathbf{X}$. Equivalently, the joint probability density function $f(\cdot,\cdot,\cdot)$ of these variables can be factorized as follows, where $\theta, \mathbf{x}, \mathbf{x}^*$ range over Θ, $\mathcal{X}$ and $\mathcal{X}^*$, respectively:

$$f(\theta, \mathbf{x}, \mathbf{x}^*) = f(\theta) f(\mathbf{x}|\theta) f(\mathbf{x}^*|\mathbf{x}) . \tag{118}$$

Additionally, we shall assume the following:

- $\theta \in \Theta$ is encoded as a pair $\theta = (\theta_S, \theta_{R|S})$ where $\theta_{R|S} = \{\theta_{R|s} : s \in \mathcal{S}\}$. Here, θ_S are the parameters of a full support categorical distribution over $\mathcal{S}$, and, for each $s \in \mathcal{S}$, $\theta_{R|s}$ are the parameters of a full support categorical distribution over $\mathcal{R}$. For each $(s, r) \in \mathcal{S} \times \mathcal{R}$

$$f(s, r|\theta) \quad = \quad f(s|\theta) \cdot f(r|\theta_{R|s}).$$

We also posit that the θ_S and the $\theta_{R|s}$'s are chosen independently, according to Dirichlet distributions of hyperparameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{|\mathcal{S}|})$ and $\boldsymbol{\beta}^s = (\beta^s_1, ..., \beta^s_{|\mathcal{R}|})$, respectively. In other words

$$f(\theta) \quad = \quad \mathrm{Dir}(\theta_S \,|\, \boldsymbol{\alpha}) \cdot \prod_{s \in \mathcal{S}} \mathrm{Dir}(\theta_{R|s} \,|\, \boldsymbol{\beta}^s)\,. \tag{119}$$

The hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may incorporate prior (background) knowledge on the population, if this is available. Otherwise, a uniformative prior can be chosen setting $\alpha_i = \beta^s_j = 1$ for each i, s, j. When $r \in \mathcal{R}$ is a tuple of attributes, we shall assume conditional independence of those attributes given s, so that the joint probability of $r|s$ can be determined by factorization [2].

- The N individual rows composing the table $\boldsymbol{x}$, say $(s_1, r_1), ..., (s_N, r_N)$, are assumed to be drawn i.i.d. according to $f(\cdot|\theta)$. Equivalently

$$f(\boldsymbol{x}|\theta) \quad = \quad f(s_1, r_1|\theta) \cdots f(s_N, r_N|\theta)\,. \tag{120}$$

Instances of the above model can be obtained by specifying an anonymization mechanism $\mathcal{A}$. In particular, the distribution $f(\boldsymbol{x}^*|\boldsymbol{x})$ only depends on the obfuscation algorithm that is adopted, say $\mathrm{obf}(\boldsymbol{x})$. In the important special case $\mathrm{obf}(\boldsymbol{x})$ acts as a deterministic function on tables, $f(\boldsymbol{x}^*|\boldsymbol{x}) = 1$ if and only if $\mathrm{obf}(\boldsymbol{x}) = \boldsymbol{x}^*$, otherwise $f(\boldsymbol{x}^*|\boldsymbol{x}) = 0$.

8.3.2 *Learner and attacker knowledge*

We shall denote by p_L the probability distribution over $\mathcal{S} \times \mathcal{R}$ that can be learned given the anonymized table $\boldsymbol{x}^*$. This distribution we take to be the average of $f(s, r|\theta)$ with respect to the density $f(\theta|\boldsymbol{X}^* = \boldsymbol{x}^*)$. Formally, for each $(s, r) \in \mathcal{S} \times \mathcal{R}$:

$$p_L(s, r|\boldsymbol{x}^*) \stackrel{\triangle}{=} \mathbb{E}_{\theta \sim f(\theta|\boldsymbol{x}^*)}[f(s, r|\theta)] \;= \int_\Theta f(s, r|\theta) f(\theta|\boldsymbol{x}^*)\, d\theta. \tag{121}$$

Of course, we can condition p_L on any given r and obtain the conditional probability $p_L(s|r, \boldsymbol{x}^*)$. Equivalently, we can compute

$$p_L(s|r, \boldsymbol{x}^*) \stackrel{\triangle}{=} \mathbb{E}_{\theta \sim f(\theta|\boldsymbol{x}^*)}[f(s|r, \theta)] \;= \int_\Theta f(s|r, \theta) f(\theta|\boldsymbol{x}^*)\, d\theta. \tag{122}$$

In particular, one can read off this distribution on a victim's nonsensitive attribute, say r_v, and obtain the corresponding distribution on $\mathcal{S}$.

2 This assumption in some context may be strong. In the next chapter we introduce an ABC method which allows overcoming it.

We shall assume the attacker knows the values of $\mathbf{X}^* = \mathbf{x}^*$ and the nonsensitive value r_v of a target individual, the victim; *moreover the attacker knows the victim is an individual in the table*. Accordingly, in what follows we fix once and for all $\mathbf{x}^*$ and r_v: these are the values observed by the attacker. Given knowledge of a victim's nonsensitive attribute r_v *and* knowledge that the victim is actually in the table $\mathbf{X}$, we can define the attacker's distribution on $\mathcal{S}$ as follows.

Let us introduce in the above model a new random variable V, identifying the index of the victim within the cleartext table $\mathbf{X}$. We posit that V is uniformly distributed on $\{1, \ldots, N\}$, and independent from $\theta, \mathbf{X}, \mathbf{X}^*$. Recalling that each row (S_j, R_j) is identified within $\mathbf{X}$ by a unique index j, we can define the attacker's probability distribution on $\mathcal{S}$, after seeing $\mathbf{x}^*$ and r_v, as follows, where it is assumed that $f(R_V = r_v, \mathbf{x}^*) > 0$, that is the observed victim's r_v is compatible with $\mathbf{x}^*$:

$$p_A(s|r_v, \mathbf{x}^*) \stackrel{\triangle}{=} f(S_V = s \mid R_V = r_v, \mathbf{x}^*) . \tag{123}$$

The following crucial lemma provides us with a characterization of the above probability distribution that is only based on a selection of the marginals R_j given $\mathbf{x}^*$. This will be the basis for actually computing $p_A(s|r_v, \mathbf{x}^*)$. Note that, on the right-hand side, only those rows whose sensitive value - known from $\mathbf{x}^*$ - is s contribute to the summation. A proof of the lemma is given in Appendix E.1.

Lemma 8.3.1 *Let* $\mathbf{X} = (S_j, R_j)_{j \in 1 \ldots N}$. *Let* s_j *be the sensitive value in the* j*-th entry of* $\mathbf{x}^*$. *Let* r_v *and* $\mathbf{x}^*$ *such that* $f(R_V = r_v, \mathbf{x}^*) > 0$. *Then*

$$p_A(s|r_v, \mathbf{x}^*) \propto \sum_{j \,:\, s_j = s} f(R_j = r_v \,|\, \mathbf{x}^*) . \tag{124}$$

Note that the disjointness of generalized nonsensitive values of the groups can make the computation of (124) more efficient, restricting the summation on the right-hand side to a unique group.

Example 1. In order to illustrate the difference between the learner's and the attacker's inference, we reconsider the toy example at the beginning of this chapter. Let $\mathbf{x}^*$ be the 2-anonymous, 2-diverse Table 10(b). Assume the attacker's victim is the first individual of the original dataset, who is from Malaysia(=M) and lives in the ZIP code 45501 area, hence $r_v = (M, 45501)$. Table 12 shows the belief probabilities of the learner, $p_L(s|r_v, \mathbf{x}^*)$, and of the attacker, $p_A(s|r_v, \mathbf{x}^*)$, for the victim's disease s. We also include the random worlds model probabilities, $p_{RW}(s|r_v, \mathbf{x}^*)$, which are just proportional to the frequency of each sensitive value within the victim's group. Note that the learner and the attacker distributions have the same mode, but the attacker is more confident about his prediction of the victim's disease. The random worlds model produces a multi-modal solution.

As to the computation of the probabilities in Table 12, a routine application of the equations (118) – (124) shows that p_L and p_A reduce to the expressions (125) and (126) below, given in terms of the model's density (118). The crucial point here is that the adversary knows the group his victim is in, i.e., the first two lines of $\mathbf{x}^*$ in the example. Below, $s \in \mathcal{S}$; for $j = 1, 2$, s_j denotes the sensitive value of the j-th row, while

Table 12: Posterior distributions of diseases for a victim with $r_v = (M, 45501)$, for the anonymized $\mathbf{x}^*$ in Table 10(b). NB: figures affected by rounding errors.

	Heart	Flu	Stomach	HIV	Diabetes
$p_L(s|r_v, \mathbf{x}^*)$	0.343	0.317	0.113	0.114	0.113
$p_A(s|r_v, \mathbf{x}^*)$	0.580	0.420	0	0	0
$p_{RW}(s|r_v, \mathbf{x}^*)$	0.500	0.500	0	0	0

.

$\mathbf{x}$ is a cleartext table, from which $\mathbf{x}_{-j}$ is obtained by removing (s_j, r_v). It is assumed that the obfuscation algorithm $\mathcal{A}$ is deterministic, so that $f(\mathbf{x}^*|\mathbf{x}) \in \{0, 1\}$.

$$p_L(s|r_v, \mathbf{x}^*) \propto \int_\Theta f(\theta) f(s, r_v|\theta) \sum_{\mathbf{x}:\mathcal{A}(\mathbf{x})=\mathbf{x}^*} f(\mathbf{x}|\theta)\, d\theta \tag{125}$$

$$p_A(s_j|r_v, \mathbf{x}^*) \propto \int_\Theta f(\theta) f(s_j, r_v|\theta) \sum_{\mathbf{x}_{-j}:\mathcal{A}(\mathbf{x})=\mathbf{x}^*} f(\mathbf{x}|\theta)\, d\theta\,. \tag{126}$$

Unfortunately, the analytic computation of the above integrals, even for the considered toy example, is a daunting task. For instance, the summation in (125) has as many terms as $\mathbf{x}^*$-compatible tables $\mathbf{x}$, that is 6.4×10^5 for Example 1 – although the resulting expression can be somewhat simplified using the independence assumption (120). Accordingly, the figures in Table 12 have been computed resorting to simulation techniques, see Section 8.5.

An alternative, more intuitive description of the inference process is as follows. The learner and the attacker first learn the parameters θ given $\mathbf{x}^*$, that is they evaluate $f(\theta_{Dis}|\mathbf{x}^*)$, $f(\theta_{ZIP|s}|\mathbf{x}^*)$ and $f(\theta_{Nat|s}|\mathbf{x}^*)$, for all $s \in \mathcal{S}$. Due to the uncertainty on the ZIP code and/or Nationality, learning θ takes the form of a mixture (this is akin to learning with soft evidence, see Corradi et al. [26]). After that, the learner, ignoring the victim is in the table, predicts the probability of r_v, $p_L(r_v|s, \mathbf{x}^*)$, for all s, by using a mixture of Multinomial-Dirichlet. The attacker, on the other hand, while still basing his prediction $p_A(r_v|s, \mathbf{x}^*)$ on the parameter learning outlined above, restricts his attention to the first two lines of $\mathbf{x}^*$, thus realizing that $s \in \{\text{Heart, Flu}\}$. Then, by Bayes theorem, and adopting the relative frequencies of the diseases in $\mathbf{x}^*$ as an approximation of $f(s|\mathbf{x}^*)$, the posterior probability of the diseases for the victim can be computed.

Remark 6 (attacker's inference and forensic identification) The attacker's inference is strongly reminiscent of two famous settings in forensic science: the *Island Problem* (IP) and the *The Data Base Search Problem* (DBS), see e.g. [3, 36] and more recently [142]. In an island with N inhabitants a crime is committed; a characteristic of the criminal (e.g. a DNA trait) is found on the crime scene. It is known that the island's inhabitants posses this characteristic independently with probability p. It is assumed the existence of exactly one culprit C in the island. In IP, one island's inhabitant I, the suspect, is found to have the given characteristic, while the others are not tested. An investigator is interested in the probability that $I = C$.

When we cast this scenario in our framework, the individuals in the table play the role of the inhabitants (including the culprit), while r_v plays the role of the characteristic found on the crime scene, matching that of the suspect. In other words - perhaps

ironically - our framework's victim plays here the role of the suspect S, while our attacker is essentially the investigator. Letting $\mathcal{S} = \{0, 1\}$ (innocent/guilty) and $\mathcal{R} = \{0, 1\}$ (characteristic absent/present), the investigator's information is then summarized by an obfuscated horizontal table $\boldsymbol{x}^*$ of N rows with as many groups, where exactly one row, say the j-th, has $S_j = 1$ and $R_j^* = R_j = 1$ (the culprit), while for $i \neq j$, $S_i = 0$ and $R_i^* = *$ ($N - 1$ innocent inhabitants). Recalling that the variable V in our framework represents the suspect's index within the table, the probability that $I = C$ is

$$\begin{aligned} \Pr(V = j | R_V = 1, \boldsymbol{x}^*) &= \Pr(S_V = 1 | R_V = 1, \boldsymbol{x}^*) \\ &= p_A(s = 1 | r_v = 1, \boldsymbol{x}^*) . \end{aligned}$$

Then applying (124), we find

$$\begin{aligned} p_A(s = 1 | r_v = 1, \boldsymbol{x}^*) &= \frac{f(R_j = 1 | \boldsymbol{x}^*)}{f(R_j = 1 | \boldsymbol{x}^*) + (N-1) f(R_{i \neq j} = 1 | \boldsymbol{x}^*)} \\ &= \frac{1}{1 + (N-1) f(R_{i \neq j} = 1 | \boldsymbol{x}^*)} . \end{aligned} \tag{127}$$

For ease of comparison with the classical IP and DBS settings, rather than relying on a learning procedure, we just assume here $f(R_i = 1 | \boldsymbol{x}^*) = p$ for $i \neq j$, so that (127) simplifies to

$$p_A(s = 1 | r_v = 1, \boldsymbol{x}^*) = \frac{1}{1 + (N-1)p} \tag{128}$$

which is the classical result known from the literature.
In DBS, the indicted exhibiting r_v is found after testing $1 \leqslant k < N$ individuals that do not exhibit r_v. This means the table $\boldsymbol{x}^*$ consists now of k rows $(s, r) = (0, 0)$ (the k innocent, tested inhabitants not exhibiting r_v), one row $(s, r) = (1, 1)$ (the culprit) and $N - 1 - k$ rows $(s, r^*) = (0, *)$ (the $N - 1 - k$ innocent, non-tested inhabitants). Accordingly, (127) becomes (letting $j = k + 1$, and possibly after rearranging indices):

$$\begin{aligned} &p_A(s = 1 | r_v = 1, \boldsymbol{x}^*) = \\ &\qquad \frac{f(R_{k+1} = 1 | \boldsymbol{x}^*)}{f(R_{k+1} = 1 | \boldsymbol{x}^*) + k f(R_{i \in \{1,k\}} = 1 | \boldsymbol{x}^*) + (N - 1 - k) f(R_{i > k+1} = 1 | \boldsymbol{x}^*)} . \end{aligned} \tag{129}$$

Letting $f(R_i = 1 | \boldsymbol{x}^*) = p$ for $i > k + 1$, equation (129) becomes

$$p_A(s = 1 | r_v = 1, \boldsymbol{x}^*) = \frac{1}{1 + (N - 1 - k)p}$$

which again is the classical result known from the literature. Finally note that our methodology also covers the possibility to learn about the probability of the characteristic, $f(R_i = 1 | \boldsymbol{x}^*)$, but here we have only stressed how the attacker strategy solves the IP and DBS forensic problems. Uncertainty about population parameters and identification has been considered in [20].

We now briefly discuss an extension of our framework to the more general case where the attacker has only partial information about his victim's nonsensitive attributes. For a typical application, think of a dataset where $\mathcal{R}$ and $\mathcal{S}$ are individuals' genetic profiles and diseases, respectively, with an adversary knowing only a partial DNA

profile of his victim; e.g., only the alleles at a few loci. Formally, fix a nonempty set $\mathcal{Y}$ and let $g : \mathcal{R} \to \mathcal{Y}$ be a (typically non-injective) function, modeling the attacker's observation of the victim's nonsensitive attribute. With the above introduced notation, consider the random variable $Y \stackrel{\triangle}{=} g(R_V)$. It is natural to extend definition (123) as follows, where $g(r_v) = y_v \in \mathcal{Y}$ and $f(Y = y_v, \boldsymbol{x}^*) > 0$:

$$p_A(s|y_v, \boldsymbol{x}^*) \stackrel{\triangle}{=} f(S_V = s \mid Y = y_v, \boldsymbol{x}^*) . \tag{130}$$

It is a simple matter to check that (124) becomes the following, where $g^{-1}(y) \subseteq \mathcal{R}$ denotes the counter-image of y according to g:

$$p_A(s|r_v, \boldsymbol{x}^*) \propto \sum_{j\,:\,s_j = s} f(R_j \in g^{-1}(y_v) \,|\, \boldsymbol{x}^*) . \tag{131}$$

Also note that one has $f(R_j \in g^{-1}(y_v) \,|\, \boldsymbol{x}^*) = \sum_{r \in g^{-1}(y_v)} f(R_j = r \,|\, \boldsymbol{x}^*)$. An extension to the case of partial and *noisy* observations can be modeled similarly, by letting $Y = g(R_V, E)$, where E is a random variable representing an independent source of noise. We leave the details of this extension for future work.

8.4 MEASURES OF PRIVACY THREAT AND UTILITY

We are now set to define the measures of *privacy threat* and *utility* we are after. We will do so from the point of view of a person or entity, the *evaluator*, who:

(a) has got a copy of the cleartext table $\boldsymbol{x}$, and can build an obfuscated version $\boldsymbol{x}^*$ of it;

(b) must decide whether to release $\boldsymbol{x}^*$ or not, weighing the privacy threats and the utility implied by this act.

The evaluator clearly distinguishes the position of the *learner* from that of the *attacker*. The learner is interested in learning from $\boldsymbol{x}^*$ the characteristics of the general population, via p_L. The attacker is interested in learning from $\boldsymbol{x}^*$ the sensitive value of a target individual, the *victim*, via p_A. The last probability distribution is derived by exploiting the additional piece of information that the victim is an individual known to be in the original table, of whom the attacker gets to know the nonsensitive values. As pointed out in [100], information about the victim's nonsensitive attributes can be easily gathered from other sources such as personal blogs and social networks. These assumptions about the attacker's knowledge allow a comparison between the risks of a sensitive attribute disclosure for an individual *who is part of the table* and for individuals who are not. The evaluator adopts the following *relative*, or differential, point of view:

> a situation where, for some individual, p_A conveys much more information than that conveyed by p_L (learner's legitimate inference on general population), must be deemed as a privacy threat.

Generally speaking, the evaluator should refrain from publishing $\boldsymbol{x}^*$ if, for some individual, the *level* of relative privacy threat exceeds a predefined threshold. Concerning the definition of the level of threat, the evaluator adopts the following Bayesian

decision-theoretic point of view. Whatever distribution p is adopted to guess the victim's sensitive value, the attacker is faced with some utility function. Here, we consider a simple 0-1 utility function for the attacker, yielding 1 if the sensitive attribute is guessed correctly and 0 otherwise. The resulting attacker's expected utility is maximized by the Bayes act, i.e., by choosing $s = \operatorname{argmax}_{s' \in \mathcal{S}} p(s')$, and equals $p(s)$. The above discussion leads to the following definitions. Note that we consider threat measures both for individual rows and for the overall table. For each threatened row, the relative threat index **Ti** says how many times the probability of correctly guessing the secret is increased by the attacker's activity, i.e., by exploiting the knowledge of the victim's presence in the table. At a global, table-wise level, the evaluator also considers the fraction $\mathbf{GT}_A$ of rows threatened by the attacker.

Definition 6 (privacy threat) *We define the following privacy threat measures.*

- *Let* q *be a full support distribution on* $\mathcal{S}$ *and* (s, r) *be a row in* $\boldsymbol{x}$*. We say* (s, r) *is* threatened under q *if* $q(s) = \max_{s'} q(s')$*, and that its* threat level under q *is* $q(s)$*.*

- *For a row* (s, r) *in* $\boldsymbol{x}$ *that is threatened by* $p_A(\cdot|r, \boldsymbol{x}^*)$*, its* relative threat level *is*

$$\mathbf{Ti}(s, r, \boldsymbol{x}, \boldsymbol{x}^*) \triangleq \frac{p_A(s|r, \boldsymbol{x}^*)}{p_L(s|r, \boldsymbol{x}^*)}. \tag{132}$$

- *Let* $N_A(\boldsymbol{x}, \boldsymbol{x}^*)$ *be the number of rows* (s, r) *in* $\boldsymbol{x}$ *threatened by* $p_A(\cdot|r, \boldsymbol{x}^*)$*. The* global threat level $\mathbf{GT}_A(t, \boldsymbol{x}^*)$ *is the fraction of rows that are threatened, that is*

$$\mathbf{GT}_A(\boldsymbol{x}, \boldsymbol{x}^*) \triangleq \frac{N_A(\boldsymbol{x}, \boldsymbol{x}^*)}{N}. \tag{133}$$

  Similarly, we denote by $\mathbf{GT}_L(\boldsymbol{x}, \boldsymbol{x}^*)$ *the fraction of rows* (s, r) *in* $\boldsymbol{x}$ *that are threatened under* $p_L(\cdot|r, \boldsymbol{x}^*)$*.*

- *As a measure of how better the attacker performs than learner at a global level, we introduce* relative global threat*:*

$$\mathbf{RGT}_{\mathbf{A}}(\boldsymbol{x}, \boldsymbol{x}^*) \triangleq \max\{0, \mathbf{GT}_A(\boldsymbol{x}, \boldsymbol{x}^*) - \mathbf{GT}_L(\boldsymbol{x}, \boldsymbol{x}^*)\}. \tag{134}$$

Remark 7 (setting a threshold for Ti) A difficult issue is how to set an acceptable threshold for the relative threat level **Ti**. This is conceptually very similar to the question of how to set the level of ϵ in differential privacy: its proponents have always maintained that the setting of ϵ is a policy question, not a technical one. Much depends on the application at hand. For instance, when the US Census Bureau adopted differential privacy, this task was delegated to a committee (the Data Stewardship Executive Policy committee, DSEP); details on the operations of this committee can be found in [51, Sect.3.1]. We think that similar considerations apply when setting the threshold of **Ti**. For instance, an evaluator might consider the distribution of the **Ti** values in the dataset (see Figure 17a–17h in Section 8.6) and then choose a percentile as a cutoff. This is reminiscent of the strategy for tuning the tolerance threshold in ABC methods. See also the introductory part of Chapter 9 for a discussion of the connection between these two thresholds.

The evaluator is also interested in the potential *utility* conveyed by an anonymized table for a learner. Note that the learner's utility is distinct from the attacker's one. Indeed, the learner's interest is to make inferences that are as close as possible to the ones that could be done using the cleartext table. Accordingly, obfuscated tables that are *faithful* to the original table are the most useful. This leads us to compare two distributions on the population: the distribution learned from the anonymized table, p_L, and the *ideal* (I) distribution, p_I, one can learn from the cleartext table $\mathbf{x}$. The latter is formally defined as the expectation[3] of $f(s, r|\theta)$ under the posterior density $f(\theta|\mathbf{x})$. Explicitly, for each (s, r)

$$p_I(s, r|\mathbf{x}) \triangleq \int_\Theta f(s, r|\theta) f(\theta|\mathbf{x}) \, d\theta \,. \tag{135}$$

Note that the posterior density $f(\theta|\mathbf{x})$ is in turn a Dirichlet density (see next section) and therefore a simple closed form of the above expression exists, based on the frequencies of the pairs (s, r) in $\mathbf{x}$. In particular, recalling the α_s, β_r^s notation for the prior hyperparameters introduced in Section 8.3, let $\alpha_0 = \sum_s \alpha_s$ and $\beta_0^s = \sum_r \beta_r^s$, and $\gamma_s(\mathbf{x})$ and $\delta_r^s(\mathbf{x})$ denote the frequency counts of s and (s, r), respectively, in $\mathbf{x}$. Then we have

$$p_I(s, r|\mathbf{x}) = \frac{\alpha_s + \gamma_s(\mathbf{x})}{\alpha_0 + N} \cdot \frac{\beta_r^s + \delta_r^s(\mathbf{x})}{\beta_0^s + \gamma_s(\mathbf{x})}. \tag{136}$$

The comparison between p_L and p_I can be based on some form of *distance* between distributions. One possibility is to rely on *total variation* (a.k.a. statistical) distance. Recall that, for discrete distributions q, q' defined on the same space $\mathcal{X}$, the total variation distance is defined as

$$\mathbf{TV}(q, q') \triangleq \sup_{A \subseteq \mathcal{X}} |q(A) - q'(A)| = \frac{1}{2} \sum_x |q(x) - q'(x)| \,.$$

Note that $\mathbf{TV}(q, q') \in [0, 1]$. The total variation distance is a quite conservative notion of diversity since it based on the event that shows the largest difference between distributions. This allows evaluating the greatest error made by a learner when using p_L in place of p_I.

Definition 7 (faithfulness) *The* relative faithfulness level of $\mathbf{x}^*$ w.r.t. $\mathbf{x}$ *is defined as*

$$\mathbf{RF}(\mathbf{x}, \mathbf{x}^*) \triangleq 1 - \mathbf{TV}\big(p_I(\cdot|\mathbf{x}) \,,\, p_L(\cdot|\mathbf{x}^*) \big) \,.$$

Remark 8 In practice, the total variation of two high-dimensional distributions might be very hard to compute. Pragmatically, we note that for M large enough, $\mathbf{TV}(q, q') = \frac{1}{2}\mathbb{E}_{x\sim q(x)}[|1 - \frac{q'(x)}{q(x)}|] \approx \frac{1}{2M} \sum_{i=1}^{M} |1 - \frac{q'(x_i)}{q(x_i)}|$, where the x_i are drawn i.i.d. according to $q(x)$. Then a Monte Carlo estimate of the total variation is the *empirical* total variation defined below, where (s_i, r_i), for $i = 1, ..., M$, are generated i.i.d. according to $p_I(\cdot, \cdot|\mathbf{x})$:

$$\mathbf{ETV}(\mathbf{x}, \mathbf{x}^*) \triangleq \frac{1}{2M} \sum_{i=1}^{M} \left| 1 - \frac{p_L(s_i, r_i \,|\mathbf{x}^*)}{p_I(s_i, r_i \,|\mathbf{x})} \right| . \tag{137}$$

3 Another sensible choice would be taking $p_I(s, r|\mathbf{x}) = f(s, r|\theta_{MAP})$, where $\theta_{MAP} = \text{argmax}_\theta f(\theta|\mathbf{x})$ is the maximum a posteriori distribution given $\mathbf{x}$. This choice would lead to essentially the same results.

Remark 9 (ideal knowledge vs. attacker's knowledge) The following scenario is meant to further clarify the extra power afforded to the attacker, by the mere knowledge that his victim is in the table. Consider a trivial anonymization mechanism that simply releases the cleartext table, that is $\mathbf{x}^* = \mathbf{x}$. As $p_L = p_I$ in this case, it would be tempting to conclude that the attacker cannot do better than the learner, hence there is no *relative* risk involved. However, this conclusion is wrong: for instance, $p_I(\cdot|r_v, \mathbf{x})$ can fail to predict the vicitim's correct sensitive value if this value is rare, as we show below.

For the sake of simplicity, consider the case where the observed victim's nonsensitive attribute r_v occurs just once in $\mathbf{x}$ in a row (s_0, r_v). Also assume a noninformative Dirichlet prior, that is, in the notation of Section 8.3, set the hyperparameters to $\alpha_s = \beta_r^s = 1$ for each $s \in \mathcal{S}, r \in \mathcal{R}$. Then, simple calculations based on (136) and the attacker's distribution characterization (124), show the following. Here for each $s \in \mathcal{S}$, $\gamma_s = \gamma_s(\mathbf{x})$ denotes the frequency count of s in $\mathbf{x}$, and c a suitable normalizing constant:

$$p_I(s|r_v, \mathbf{x}) = \begin{cases} \frac{1+\gamma_s}{|\mathcal{R}|+\gamma_s} c & \text{if } s \neq s_0 \\ \frac{2(1+\gamma_{s_0})}{|\mathcal{R}|+\gamma_{s_0}} c & \text{if } s = s_0 \end{cases} \tag{138}$$

$$p_A(s|r_v, \mathbf{x}^*) = \begin{cases} 0 & \text{if } s \neq s_0 \\ 1 & \text{if } s = s_0 \,. \end{cases}$$

As far as the target individual $(s_0, r_v) \in \mathbf{x}$ is concerned, we see that while p_A predicts s_0 with certainty, predictions based on $p_L = p_I$ will be blatantly wrong, if there are values $s \neq s_0$ that occur very frequently in $\mathbf{x}$, while s_0 is rare, and N is large compared to $|\mathcal{R}|$. To make an extreme numeric case, consider $|\mathcal{S}| = 2$, $|\mathcal{R}| = 1{,}000$ and $\gamma_{s_0} = 1$ in a table $\mathbf{x}$ of $N = 10^6$ rows: plugging these values in (138) yields $p_L(s_0|r_v, \mathbf{x}^*) = p_I(s_0|r_v, t) \approx 0.004$, hence a relative threat for (s_0, r_v) of $1/p_L(s_0|r_v, \mathbf{x}^*) \approx 250$.

8.5 LEARNING FROM THE OBFUSCATED TABLE BY MCMC

Estimating the privacy threat and faithfulness measures defined in the previous section, for specific tables $\mathbf{x}$ and $\mathbf{x}^*$, implies being able to compute the distributions (121), (122) and (124). Unfortunately, these distributions, unlike (135), are not available in closed form, since $f(\theta|\mathbf{X}^* = \mathbf{x}^*) = f(\theta|\mathbf{x}^*)$ cannot be derived analytically. Indeed, in order to do so, one should integrate $f(\theta, \mathbf{x}|\mathbf{x}^*)$ with respect to the density $f(\mathbf{x}|\mathbf{x}^*)$, which appears not to be feasible.

To circumvent this difficulty, we will introduce a *Gibbs sampler*, defining a Markov chain $\{Z_i\}_{i \geqslant 0}$, with $Z_i = (\theta_i, \mathbf{X}_i)$, converging to the density

$$f(\theta = \theta, \mathbf{X} = \mathbf{x}|\mathbf{x}^*) = f\big(\theta = \theta,\, S_1 = s_1, R_1 = r_1, ..., S_N = s_N, R_N = r_N \,|\, \mathbf{x}^*\big)$$

(note that the sensitive values s_j in $\mathbf{X}$ are in fact fixed and known, given $\mathbf{x}^*$). General results (see Chapter 2) ensure that, if $\theta_0, \theta_1, ...$ are the samples drawn from the θ-marginal of such a chain, then for each $(s, r) \in \mathcal{S} \times \mathcal{R}$

$$\frac{1}{M}\sum_{\ell=0}^{M} f(s, r|\theta_\ell) \to \int_\Theta f(s, r|\theta) f(\theta|\mathbf{x}^*) d\theta \;=\; p_L(s, r|\mathbf{x}^*) \tag{139}$$

$$\frac{1}{M}\sum_{\ell=0}^{M} f(s|r, \theta_\ell) \to \int_\Theta f(s|r, \theta) f(\theta|\mathbf{x}^*) d\theta \;=\; p_L(s|r, \mathbf{x}^*) \tag{140}$$

almost surely as $M \longrightarrow +\infty$. Therefore, by selecting an appropriately large M, one can build approximations of $p_L(s, r|\mathbf{x}^*)$ and $p_L(s|r, \mathbf{x}^*)$ using the arithmetical means on the left-hand side of (139) and (140), respectively. Moreover, for each index $1 \leqslant j \leqslant N$, using samples drawn from the R_j-marginals of the same chain, one can build an estimate of $f(R_j = r_j|\mathbf{x}^*)$. Consequently, using (124) (resp. (131), in the case of partial observation) one can estimate $p_A(s|r_v, \mathbf{x}^*)$ (resp. $p_A(s|y_v, \mathbf{x}^*)$) for any given r_v (resp. y_v).

In the rest of the section, we will first introduce MCMC for this problem and then show its convergence. We will then discuss details of the sampling procedures for each of the two possible schemes, horizontal and vertical.

8.5.1 *Definition and convergence of the Gibbs sampler*

Simply stated, our problem is sampling from the marginals of the following target density function, where $\mathbf{x}^* = g_1^*, ..., g_k^*$ and $\mathbf{x} = g_1, ..., g_k$ (note that the number of groups k is known and fixed, given $\mathbf{x}^*$),

$$f(\theta, \mathbf{x}|\mathbf{x}^*)\,. \tag{141}$$

Note that the r_j's of interest, for $1 \leqslant j \leqslant N$, are the elements of the groups g_i's, for $1 \leqslant i \leqslant k$. The Gibbs scheme allows for some freedom as to the blocking of variables. Here we consider $k+1$ blocks, coinciding with θ and $g_1, ..., g_k$. This is natural as, in the considered schemes, $(R_i, S_i) \perp\!\!\!\perp (R_j, S_j)|\theta, \mathbf{x}^*$ for (R_i, S_i) and (R_j, S_j) occurring in distinct groups. Formally, let $z^0 = \theta^0, \mathbf{x}^0$ (with $\mathbf{x}^0 = g_1^0, ..., g_k^0$) denote any initial state satisfying $f(\theta^0, \mathbf{x}^0|\mathbf{x}^*) > 0$. Given a state at step h, $z^h = \theta^h, \mathbf{x}^h$ ($\mathbf{x}^h = g_1^h, ..., g_k^h$), one lets $z^{h+1} \stackrel{\triangle}{=} \theta^{h+1}, \mathbf{x}^{h+1}$, where $\mathbf{x}^{h+1} = g_1^{h+1}, ..., g_k^{h+1}$ and

$$\theta^{h+1} \quad \text{is drawn from} \quad f(\theta|\mathbf{x}^h, \mathbf{x}^*) \tag{142}$$

$$g_i^{h+1} \quad \text{is drawn from} \quad f(g|\theta^{h+1}, g_1^{h+1}, ..., g_{i-1}^{h+1}, g_{i+1}^h, ..., g_k^h, \mathbf{x}^*) \quad (1 \leqslant i \leqslant k). \tag{143}$$

Running this chain presupposes we know how to sample from the *full conditional* distributions on the right-hand side of (142) and (143). In particular, there are several possible approaches to sample from g. In this subsection we provide a general discussion about convergence, postponing the details of sampling from the full conditionals to the next subsection.

Let us denote by $\mathbf{x}_{-i} \stackrel{\triangle}{=} g_1, ..., g_{i-1}, g_{i+1}, ..., g_k$ the table obtained by removing the i-th group g_i from $\mathbf{x}$. The following relations for the full conditionals of interest can

be readily checked, relying on the conditional independencies of the model (118) and (120) (we presuppose that in each case the conditioning event has nonzero probability)

$$f(\theta|\mathbf{x},\mathbf{x}^*) = f(\theta|\mathbf{x}) \tag{144}$$

$$f(g|\theta,\mathbf{x}_{-i},\mathbf{x}^*) \propto f(g|\theta)f(\mathbf{x}^*|g,\mathbf{x}_{-i}) \quad (1 \leqslant i \leqslant k). \tag{145}$$

As we shall see, each of the above two relations enables sampling from the densities on the left-hand side. Indeed, (144) is a posterior Dirichlet distribution, from which effective sampling can be easily performed (see next subsection). A straightforward implementation of (145) in a Rejection Sampling (RS) perspective is as follows: draw g according to $f(g|\theta)$ and accept it with probability $f(\mathbf{x}^*|g,\mathbf{x}_{-i}) = f(\mathbf{x}^*|\mathbf{x})$. Here, $f(\mathbf{x}^*|\mathbf{x})$ is just the probability that the obfuscation algorithm returns $\mathbf{x}^*$ as output when given $\mathbf{x} = g,\mathbf{x}_{-i}$ as input [4]. Actually, to make sampling from the RHS of (145) effective, further assumptions will be introduced (see next subsection). Note that, since the sensitive values are fixed in $\mathbf{x}$ and known from the given $\mathbf{x}^*$, sampling g in (145) is actually equivalent to sampling the *nonsensitive* values of the group.

In addition to (145), to simplify our discussion about convergence, we shall henceforth assume that, for each group index $1 \leqslant i \leqslant k$, the set of instances of the i-th group that are compatible with $\mathbf{x}^*$ does *not* depend on the rest of the table, $\mathbf{x}_{-i}$. That is, we assume that for each i ($1 \leqslant i \leqslant k$):

$$\begin{aligned}\{g : f(\mathbf{x}^*|g,\mathbf{x}_{-i}) > 0\} &= \{g : f(\mathbf{x}^*|g,\mathbf{x}'_{-i}) > 0\}\ \forall \mathbf{x}_{-i} \text{ and } \mathbf{x}'_{-i} \\ &\triangleq \mathcal{G}_i\,. \end{aligned} \tag{146}$$

For instance, (146) holds true if the anonymization algorithm ensures $\mathbf{x}^*$ is independent from $\mathbf{x}_{i-1}$ given a i-th group g: $\mathbf{x}^* \perp\!\!\!\perp \mathbf{x}_{-i} \,|\, g$.

Let $z = (\theta, g_1, ..., g_k)$ denote a generic state of this Markov chain. Under the assumption (146), the *support* of the target density $f(z|\mathbf{x}^*)$ is the product space

$$\mathcal{Z} \triangleq \Theta \times \mathcal{G}_1 \times \cdots \times \mathcal{G}_k\,. \tag{147}$$

By this, we mean that $\{z : f(z|\mathbf{x}^*) > 0\} = \mathcal{Z}$. This is a consequence of: (a) the fact that Dirichlet only considers full support distributions; and (b) equation (145), taking into account the assumption (146). Let $Z_0, Z_1, ...$ denote the Markov chain defined by the sampler over $\mathcal{Z}$ and denote by $\kappa(\cdot|\cdot)$ its conditional kernel density over $\mathcal{Z}$. Slightly abusing notation, let us still indicate by $f(\cdot|\mathbf{x}^*)$ the probability distribution over $\mathcal{Z}$ induced by the density $f(z|\mathbf{x}^*)$. Convergence in distribution follows from the following proposition, which is an instance of general results dealt with in Section 2.3.3.1.

Proposition 4 (convergence) *Assume* (146). *For each (measurable) set* $A \subseteq \mathcal{Z}$ *such that* $f(A|\mathbf{x}^*) > 0$ *and each* $z^0 \in \mathcal{Z}$, *we have* $\kappa(Z^1 \in A|Z^0 = z^0) > 0$. *As a consequence, the Markov chain* $\{Z_i\}_{i\geqslant 0}$ *is irreducible and aperiodic, and its stationary density is* $f(z|\mathbf{x}^*)$ *in* (141).

4 A similar reasoning applies in the likelihood-free method proposed in Chapter 9.

Proof Let us assume that (146) holds and that $\mathcal{Z}$ is the product space defined in (147). It follows that both (144) and (145) are well-defined for each $z \in \mathcal{Z}$. Thus, being $\kappa(\cdot|\cdot)$ the product among the full conditional distributions, from the Fubini's theorem follows that $\int_A \kappa(z|Z^0 = z^0)dz > 0$. □

8.5.2 *Sampling from the full conditionals*

Let us consider (144) first. It is a standard fact that the posterior of the Dirichlet distribution $f(\theta|x)$, given the N i.i.d. observations x drawn from the categorical distribution $f(\cdot|\theta)$, is still a Dirichlet, where the hyperparameters have been updated as follows. Denote by $\gamma(x) = (\gamma_1, ..., \gamma_{|\mathcal{S}|})$ the vector of the frequency counts γ_i of each s_i in x. Similarly, given s, denote by $\delta^s(x) = (\delta^s_1, ..., \delta^s_{|\mathcal{R}|})$ the vector of the frequency counts δ_i of the pairs (r_i, s), for each r_i, in x. Then, for each $\theta = (\theta_S, \theta_{R|S})$, we have

$$f(\theta|x) \quad = \quad \mathrm{Dir}(\theta_S \mid \alpha + \gamma(x)) \cdot \prod_{s \in \mathcal{S}} \mathrm{Dir}(\theta_{R|s} \mid \beta^s + \delta^s(x)) \,. \tag{148}$$

Let us now discuss (145). In what follows, for the sake of notation we shall write a generic i-th group as $g_i = (s_1, r_1), ..., (s_n, r_n)$ (thus avoiding double subscripts), and let $g^*_i = (m_i, l_i)$ denote the corresponding obfuscated group in x^*. As already observed, given an obfuscated i-th group $g^*_i = (l_i, m_i)$, when sampling a i-th group g from (145), one actually needs to generate only the nonsensitive values of g, which are constrained by l_i, as the sensitive ones are already fixed by the sequence m_i. In what follows, to make sampling from (145) effective, will shall work under the following assumptions, which are stronger than (146).

(a) Deterministic obfuscation function: for each x and x^*, $f(x^*|x)$ is either 0 or 1.

(b) For each $1 \leqslant i \leqslant k$, letting $g^*_i = (l_i, m_i)$, with $m_i = s_1, ..., s_n$, the i-th obfuscated group in x^*, the following holds true:

Horizontal schemes

$$\mathcal{G}_i = \{g = (s_1, r_1), ..., (s_n, r_n) \,:\, r_\ell \in l_i \;\text{ for } 1 \leqslant \ell \leqslant n\} \tag{149}$$

Vertical schemes

$$\mathcal{G}_i = \{g = (s_1, r_{i_1}), ..., (s_n, r_{i_n}) \,:\, \text{for } r_{i_1}, ..., r_{i_n} \text{ a permutation of } l_i\}. \tag{150}$$

Assumption (a) is realistic in practice. In horizontal schemes, assumption (b) makes the considered sets $\mathcal{G}_i$'s possibly larger than the real ones, that is $l_i \supset \{r_1, ..., r_n\}$. This happens, for instance, if in certain groups the ZIP code is constrained to just, say, two values, while the generalized code "5013*" allows for all values in the set $\{50130, \ldots, 50139\}$. We will not attempt here a formal analysis of this assumption. In some cases, such as in schemes based on global recoding, this assumption is realistic. Otherwise, we only note that the support $\mathcal{Z}$ of the resulting Markov chain may be (slightly) larger than the one that would be obtained not assuming (149) or (150). Heuristically, this leads one to sampling from a more dispersed density than the target one. At least, the resulting distributions can be taken to represent a lower bound of what the attacker can actually learn.

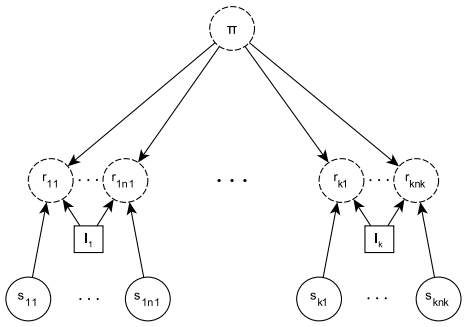


Figure 16: Sampling from $f(g|\theta, \mathbf{x}_{-i}, \mathbf{x}^*)$ ($g \in \mathcal{G}_i$) for horizontal schemes, across all the groups.

Under assumptions (a) and (b) above, for each $1 \leqslant i \leqslant k$, it holds that $g \in \mathcal{G}_i$ if and only if $f(\mathbf{x}^*|g, \mathbf{x}_{-i}) = 1$. Therefore sampling according to the right-hand side of (145) reduces to the following:

$$\text{draw } g \in \mathcal{G}_i \text{ with probability } \propto f(g|\theta) \quad (1 \leqslant i \leqslant k). \tag{151}$$

We discuss now how to implement (151) effectively. This will achieve sampling from the full conditionals (145) without resorting to a presumably inefficient RS method. We deal with the two cases, horizontal and vertical, separately.

HORIZONTAL SCHEMES In order to generate $g = (r_1, s_1), ..., (r_n, s_n) \in \mathcal{G}_i$, for each $\ell = 1, .., n$, we draw $r_\ell \in l_i$ with probability $\propto f(r_\ell|s_\ell, \theta)$. Explicitly, (145) now becomes

$$f(g|\theta, \mathbf{x}_{-i}, \mathbf{x}^*) \;=\; \begin{cases} 0 & \text{if } g \notin \mathcal{G}_i \\ \prod_{\ell=1}^{n} \dfrac{f(r_\ell|s_\ell, \theta)}{\sum_{r \in l_i} f(r|s_\ell, \theta)} & \text{if } g \in \mathcal{G}_i \end{cases} \tag{152}$$

thus satisfying (151). Note that this is equivalent to sampling each row independently. The sampling process of $f(g|\theta, \mathbf{x}_{-i}, \mathbf{x}^*)$ for horizontal schemes across all the groups of the table is illustrated graphically in Fig. 16.

VERTICAL SCHEMES Let $l_i = \{\!\{ r_1, ..., r_n \}\!\}$. We have that $g \in \mathcal{G}_i$ if and only if $g = (s_1, r_{i_1}), ..., (s_n, r_{i_n})$, for some permutation $(r_{i_\ell})_{1 \leqslant \ell \leqslant n}$ of $r_1, ..., r_n$. Here, sampling the nonsensitive values of g row by row would involve to gradually reduce the sample space. A sampling procedure along these lines is possible, but nontrivial, see Appendix E.2. Another possibility is to resort to ABC methods as shown in Chapter 9.

Here we discuss a more straightforward sampling procedure, based on generating $g_i \in \mathcal{G}_i$ in a single shot. We adopt a *single-iteration Metropolis within Gibbs* scheme. Essentially, this consists in running a Metropolis method that targets the distribution $\propto f(g|\theta)$ with support $\mathcal{G}_i$, for one iteration [5]. Specifically, let us write the current value of the i-th group in the Gibbs Markov chain as g_i^h. The Metropolis step consists in

5 Note that here we only assume that the obfuscation function is such that $f(\mathbf{x}^*|\mathbf{x})$ represents a multiplicative constant in (145). Thus, in Anatomy the assumption that the obfuscation function is deterministic, i.e., Assumption (a), is slightly relaxed.

drawing $g \in \mathcal{G}_i$ according to a proposal distribution $J(g|g_i^h)$ and accepting it, that is letting $g_i^{h+1} = g$, with probability

$$\varepsilon \triangleq \min\left\{1, \frac{f(g|\theta)J(g_i^h|g)}{f(g_i^h|\theta)J(g|g_i^h)}\right\} \tag{153}$$

while keeping $g_i^{h+1} = g_i^h$ with probability $1 - \varepsilon$. The resulting MCMC method is an example of the MwG introduced in Section 2.3.4 As to the proposal distribution $J(g|g_i^h)$, a possibility is generating $g \in \mathcal{G}_i$ via a pure random permutation of the n nonsensitive values in l_i; or just to swap the nonsensitive values of two randomly chosen positions in g_i^h. In both cases, the proposal is symmetric, and (153) simplifies accordingly as follows, where $r_1, ..., r_n$ is the sequence of sensitive values in the poposed g:

$$\varepsilon = \min\left\{1, \frac{\prod_{\ell=1}^{n} f(r_\ell|s_\ell, \theta)}{\prod_{\ell=1}^{n} f(r_\ell^h|s_\ell, \theta)}\right\}.$$

8.6 EXPERIMENTS

We have put a proof-of-concept implementation of our method at work on a subset of the Adult dataset extracted by Barry Becker from the 1994 US Census database and available from the UCI machine learning repository [75]. This is a common benchmark for experiments on anonymization [114]. In particular, we have focused on the subset of 5692 rows also considered by the authors of [114], with the following categorical attributes: *sex*, *age*, *race*, *marital status*, *education*, *native country*, *workclass*, *salary class*, *occupation*, with *occupation* (14 values) considered as the only sensitive attribute. We consider the case in which both the learner and the attacker have no prior information about the phenomena in the general population and assume non-informative prior distributions (with all the hyperparameters equal to 1). We will discuss implementation and results details separately for vertical and horizontal schemes. We will then briefly discuss convergence issues of the employed MCMC method.

8.6.1 *Horizontal schemes: k-anonymity*

Using the ARX anonymization tool [113] we obtained two different k-anonymous versions of the considered dataset, enjoying respectively k-anonymity and ℓ-diversity[6] for $k = \ell = 4$ and $k = \ell = 6$. The average size of the groups was respectively of 38 rows ($k = \ell = 4$) and of 355 rows ($k = \ell = 6$).

The results we have obtained are summarized in Table 13. For reference, we include the following information in the last two lines: *baseline accuracy*, the fraction of rows correctly classified using the empirical distribution obtained from the frequencies of the sensitive values in the anonymized table, i.e., the fraction of the most frequent sensitive value; and *ideal accuracy*, the fraction of tuples threatened under p_I. As a further element of comparison, we also consider an attacker whose reasoning is based on the random worlds models, and include in the table $\mathbf{GT}_{RW}$, the fraction of

6 Recall that ℓ-diversity requires at least ℓ distinct values of the sensitive attribute in each group.

Table 13: Summary of threat and faithfulness measures for anonymization according to k-anonymity and ℓ- diversity.

		Group size and diversity	
		$k = \ell = 4$	$k = \ell = 6$
Global threat level under p_A	$\mathbf{GT}_A$	0.2930	0.2994
Global threat level under p_L	$\mathbf{GT}_L$	0.2681	0.2756
Global threat level under p_{RW}	$\mathbf{GT}_{RW}$	0.2131	0.2890
Relative global threat	$\mathbf{RGT}_A$	0.0249	0.0232
Empirical relative faithfulness level	**RF**	0.3106	0.3011
Absolute error under p_A	$\mathbf{ABS}_A$	9795.58	9699.09
Absolute error under p_{RW}	$\mathbf{ABS}_{RW}$	9980.35	9451.53
Baseline accuracy		0.1656	
Ideal accuracy		0.3534	

rows correctly classified assuming all tables compatible with $\boldsymbol{x}^*$ equally likely. Like in [73], we compute $\boldsymbol{ABS}_A$ and $\boldsymbol{ABS}_{RW}$, the *absolute error* under the distribution derived under p_A and under the random worlds distribution p_{RW}, respectively. $\boldsymbol{ABS}$ is defined as $\sum_{i=1}^{N} \sum_{s \in \mathcal{S}} |\mathbb{1}\{s_i = s\} - p(s|r_i, \boldsymbol{x}^*)|$, where $p(\cdot)$ might be either of $p_A(\cdot)$ or $p_{RW}(\cdot)$. Note that, since the considered anonymized tables do not enjoy disjointness between groups (see Remark 5), also in the random worlds perspective the probability of each sensitive attribute may well be $\geqslant 1/\ell$. In our experiments, when $\ell = 4$ the attacker outperforms random worlds classification, while when a more powerful obfuscation is adopted the two results are quite similar.

The remaining rows in Table 13 consider the privacy threats and faithfulness measures introduced in Section 8.4. As a general comment, small variations of ℓ and/or k do not produce dramatic changes. The faithfulness level is stable, but does not reach a satisfactory level. The attacker is anyway in a position to correctly classify the sensitive attribute of individuals in the table $\approx 2.3 - 2.5\%$ more often than the learner. We found the maximum value of $\mathbf{Ti}_A$ for the threatened rows is about 13.8, meaning the attacker can be up to $\approx$14 times more confident than the learner about the guessed value.

A more informative summary of our analysis is provided by the scatter plots and histograms of Figure 17. The scatter plots are obtained from the threat levels under p_L and under p_A. The number of rows (s, r) in which $p_A(s|r, \boldsymbol{x}^*) \geqslant p_L(s|r, \boldsymbol{x}^*)$ roughly equals those in which $p_A(s|r, \boldsymbol{x}^*) \leqslant p_L(s|r, \boldsymbol{x}^*)$, although globally the attacker has a slight advantage in terms of number of threatened rows. In Figure 17 we also report the empirical distribution $\log_2 \mathbf{Ti}_A$ for tuples threatened under p_A and under p_L. We also have evidence of positive skewness, as shown by the value of γ (the third standardized moments of the empirical distributions). Recalling that $\log_2 \mathbf{Ti}_A = 1$ means $p_A(s|r, \boldsymbol{x}^*) = 2p_L(s|r, \boldsymbol{x}^*)$, the histograms show that $p_A(s|r, \boldsymbol{x}^*)$ is often more than twice $p_L(s|r, \boldsymbol{x}^*)$ leading to a $\log_2 \mathbf{Ti}_A \geqslant 1$. In particular, when $k = \ell = 4$, $\log_2 \mathbf{Ti}_A$ is at least 1 for $\approx 6\%$ of the individuals threatened under p_A, meaning $\approx 0.6\%$ of

Table 14: Summary of threat and faithfulness measures for anonymization according to Anatomy.

		Group size and diversity	
		$\ell = 4$	$\ell = 6$
Global threat level under p_A	$\mathbf{GT}_A$	0.3273	0.2396
Global threat level under p_L	$\mathbf{GT}_L$	0.2653	0.2136
Global threat level under p_{RW}	$\mathbf{GT}_{RW}$	0.1669	0.1689
Relative global threat	$\mathbf{RGT}_\mathbf{A}$	0.0620	0.0260
Empirical relative faithfulness level	**RF**	0.6493	0.5341
Absolute error under p_A	$\mathbf{ABS}_A$	8391.66	9276.25
Absolute error under p_{RW}	$\mathbf{ABS}_{RW}$	9471.94	9889.07
Baseline accuracy		0.1656	
Ideal accuracy		0.3534	

the whole table. Conversely, $\log_2 \mathbf{Ti}_A$ is close to 0 for most of the rows in which $p_A(s|r, \boldsymbol{x}^*) \leqslant p_L(s|r, \boldsymbol{x}^*)$.

8.6.2 *Vertical schemes: Anatomy*

Using a freely available anonymization tool [118], we have obtained two anatomized versions of the considered dataset, with groups of size $\ell = 4$ and $\ell = 6$, respectively. The resulting tables also enjoy ℓ-diversity. The results we have obtained are summarized in Table 14. Concerning the random worlds approach, we note the following. Anatomy partitions the tables in groups all of size ℓ. Therefore, although disjointness is not satisfied, just as in the horizontal case, the sensitive attribute frequencies equal $1/\ell$ in each group. This implies that the probability of a sensitive value depends on how many groups contain the victim's nonsensitive attributes and on their frequencies in each group, leading often to multimodal distributions. We assume that a guess may be obtained randomly choosing between the equally likely sensitive attributes. Accordingly, the fractions of threatened rows, $\mathbf{GT}_{RW}$, are averaged over 500 different sampling. Here, it is apparent that the our attacker is able to classify better than the random worlds scenario. We note that, as ℓ increases from 4 to 6, the fraction of rows threatened under the distributions derived by the learner ($\mathbf{GT}_L$) and by the attacker ($\mathbf{GT}_A$) decreases significantly. Moreover, as ℓ grows both the relative threat $\mathbf{RGT}_\mathbf{A}$ and the faithfulness level **RF** decrease, which implies a trade-off between privacy and the utility conveyed by the table.

Again, for a more informative summary of our analysis, we look at scatter plots and histograms, displayed in Figure 18, where we compare p_A and p_L on threatened rows. It is apparent here that the attacker is more confident than the learner in the majority of the cases, even when focusing on the rows threatened under p_L. This is in contrast with the horizontal case, where the attacker exhibits smaller threat levels on the rows threatened under p_L (Figure 17, (d) and (h)). As far as the histograms are concerned, an even greater skewness than the horizontal case is evident here. In

particular, the attacker can be up to $\approx$ 287 times more confident then the learner, being the maximum $\mathbf{Ti}_A$ about 286.19. Moreover, when $\ell = 4$, the individuals with $\log_2 \mathbf{Ti}_A \geqslant 1$ are $\approx$ 26% of the rows threatened under p_A ($\approx$ 8% of the whole table). This means that there are 483 individuals in the dataset for which the threat level under p_A is at least twice as much the threat level under p_L.

8.6.3 *Discussion*

Comparing the horizontal and the vertical cases for the considered dataset, the following considerations are in order.

- In the horizontal case, we have a situation of low faithfulness and low privacy threat, irrespective of the value of k and ℓ. Indeed, in both cases the average group size is well above k, and this has a negative effect on the inference capabilities of both the learner and the attacker. The slight numerical differences observed between the cases $k = \ell = 4$ and $k = \ell = 6$ are basically an artifact of the anonymization tool. Yet, in relative terms, one can observe a significant increase in the number of tuples threatened by the attacker, over the learner.
- In the vertical case, one obtains a greater faithfulness at the price of a greater privacy threat. This difference from the horizontal case is partly explained by the smaller group size, which now coincides with ℓ. Now moving from $\ell = 4$ to $\ell = 6$ has a tangible negative impact on the inference capabilities of both the learner and the attacker. In relative terms, one can observe an even more marked increase of the number of tuples threatened by the attacker, over the learner.

The above considerations partly depend on both the original dataset and the details of the employed anonymization tool.

8.6.4 *Assessing* MCMC *convergence*

For each of the considered anonymized datasets, we ran a MCMC as introduced in Section 8.5 for $M = 100,000$ runs. The convergence of each chain to the stationary distribution was assessed via a method based on comparing sub-sequences of the sample sequences with one another. More precisely, as for the population parameters distribution (148), we used the method proposed by Geweke [59] described in Section 2.3.5

After a burn-in of 50,000 iterations, we compared the last 25,000 samples against 5 blocks of of 5,000 consecutive samples each, taken starting from the 50,000-th iteration. We found that all the distributions $\theta_{R|S}$ produced a test statistic within two standard deviations from zero, thus providing evidence of convergence.

As for the distribution of the cleartext table, $f(\mathbf{x}|\theta, \mathbf{x}^*)$, we used the procedure designed for categorical distributions by Deonovich and Smith (see Section 2.3.5). After a burn-in of 50,000 observations, we compared 5 sub-sequences of 10,000 consecutive samples each. For the vertical scheme, we assessed the convergence for each row of the table, thereby demonstrating the stationary of $f(\mathbf{x}|\theta, \mathbf{x}^*)$. For the horizontal scheme, some of the rows did not exhibit evidence of convergence. However, we

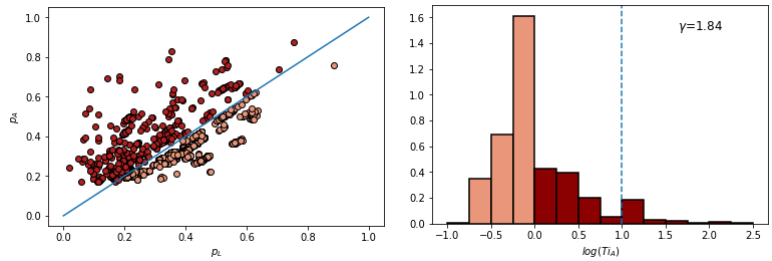
found that, starting with several independent chains, very similar results in terms of the proposed assessment measures were obtained.

In the vertical case, within the Metropolis step both the pure random permutation and the swap group generation strategies (Section 8.5.2) were experimented. The obtained results are consistent; however, the pure random permutation strategy shows a much higher rate of rejection, suggesting that the swap strategy should be preferred.

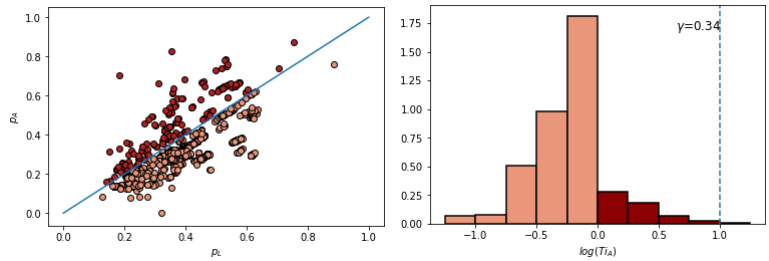
8.7 CONCLUSIONS

We have put forward a notion of relative privacy threat that applies to group-based anonymization schemes. Our proposal is based on a rigorous characterization of the learner's and of the attacker's inference, in a unified Bayesian model of group-based schemes. A related MCMC algorithm for posterior parameters estimation has also been introduced. Experiments conducted on the well-known Adult dataset [75] have been illustrated.

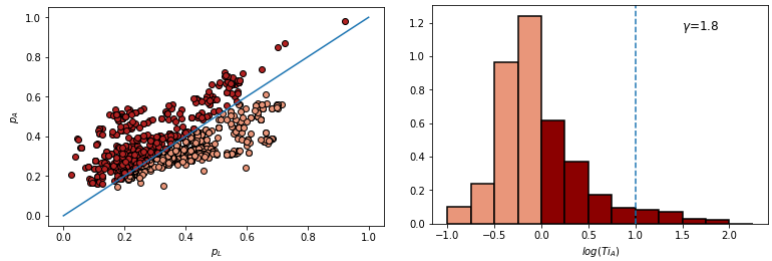
Our analysis emphasizes the risks posed by the mere fact that an attacker can look up a released anonymized table. This prompts an obvious alternative: release the parameters of the posterior distribution learned from the cleartext table (p_I, in our notation). This may not always be possible, or be a good idea, for several reasons. First, certain organizations must release datasets as part of their mission, e.g. census bureaus. Second, especially in the case of high-dimensional data, the computation of the posterior is feasible only assuming suitable conditional independencies, whereby potentially important correlations are lost; see [25] and references therein. Third, parameters release itself is not exempt from risks for privacy. In particular, although differentially private release of the parameters is possible [40], it seems that quite strong priors are necessary to obtain acceptable guarantees; see [138, Ch.6] and references therein. An attempt of releasing differentially private posterior distributions by resorting to ABC will be discussed in the next chapter. However, further research is called for an understanding of the circumstances under which data and/or parameters release can be done safely.



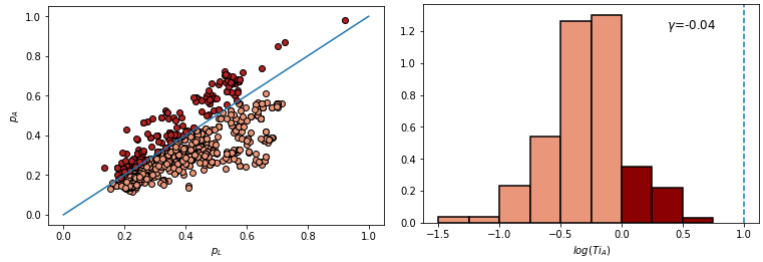
(a) (b)



(c) (d)

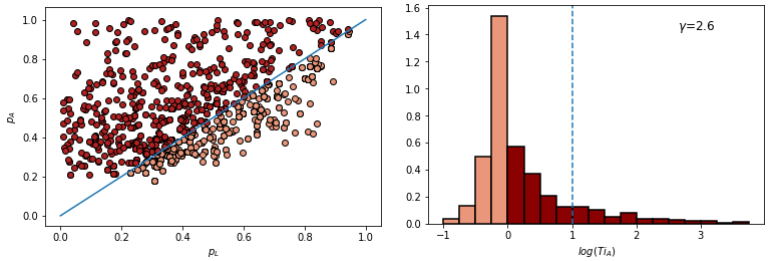


(e) (f)

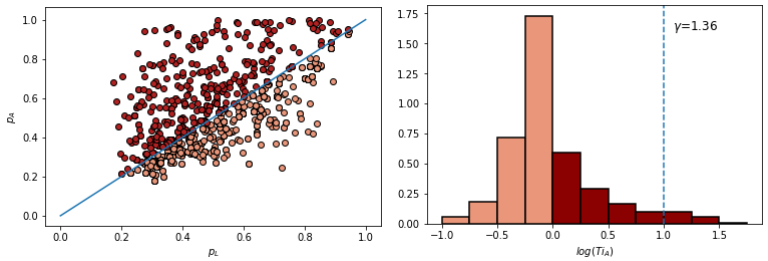


(g) (h)

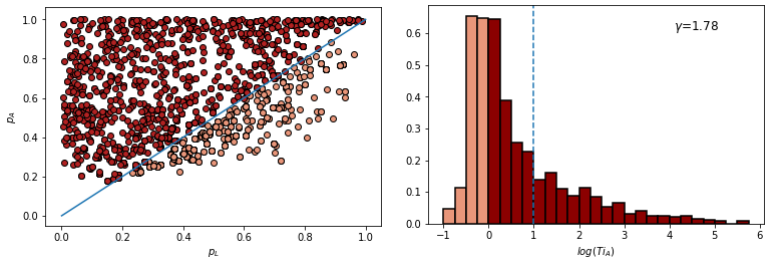
Figure 17: Results for k-anonymity. Top ($\ell = k = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = k = 4$. The skewness value (γ) represents the third standardized moment of the empirical distribution. Dark red areas show where the attacker performs better than the learner.



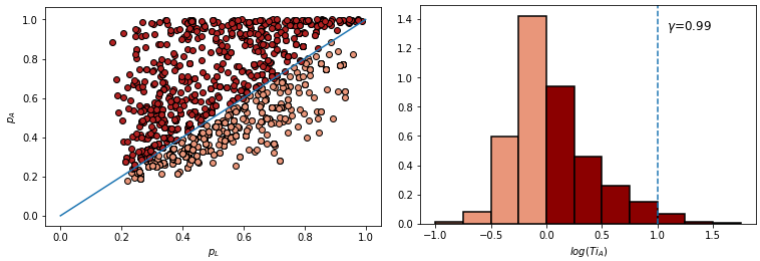
(a) (b)



(c) (d)



(e) (f)



(g) (h)

Figure 18: Results for Anatomy. Top ($\ell = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = 4$. The skewness value (γ) represents the third standardized moment of the empirical distribution. Dark red areas show where the attacker performs better than the learner.

9

APPROXIMATE BAYESIAN INFERENCE FROM ANONYMIZED DATA: ABC VS LD-ABC

In the previous chapter we provided a MCMC method for sampling from the posterior distribution of the population parameters, θ, both in the horizontal and vertical scheme case. We noted that Systematic scan Gibbs sampler does not apply when the table is anonymized resorting to Anatomy. Accordingly, we proposed both a MwG algorithm and a Random scan Gibbs sampler (in Appendix E.2). As already mentioned, ABC represents an alternative solution.

Very recently a new research line investigating the relations between data anonymization and ABC has been developing. The directions investigated are mainly two: 1) setting up strategies for providing posterior distributions obeying to Differential Privacy requirements by means of ABC algorithms; 2) defining ABC methods for learning population characteristics from anonymized data.

As regards the first point, a proposal comes from Park and Jitkrittum in [107]. Here the key idea is that the ABC tolerance level is related to the Differential privacy budget. More specifically, they provide a way of getting samples from the posterior distribution by defining a framework in which two entities exist: a *data owner* and a *modeler*. In this framework their ABCDP is performed in two steps: a *non private* step and a *private* step. In the non private step the modeler gets samples from the parameters prior distribution and simulates pseudo-data giving those parameters as input to a simulator. In the second step the data owner takes the pairs of parameters and simulated data and returns a set of binary indicators determining whether pseudo-data are resembling the original data. Taking the binary indicators from the data owner, the modeler is able to convert samples from the prior into samples from the posterior distribution.

In [61], Gong investigates the property of a posterior distribution derived via ABC starting from Differential Private data. He proves that a proper ABC algorithm, conditional on obfuscated data, provides samples from the exact posterior distribution of the parameters given the non-obfuscated data, i.e., no approximations occur thanks to the tolerance parameter ϵ. A likelihood-based strategy for inference is also discussed. This latter, in the same vein of the MCMC method proposed in the previous chapter, is based on a data augmentation strategy relaying on the introduction of the complete data as latent variable.

Here, we discuss how to properly define a generative model producing tables anonymized according to Anatomy and enabling an ABC implementation also in the group-based setting. Furthermore, we test LD-ABC at work on a subset of the dataset analysed in the previous chapter. The MwG introduced in Section 8.5.2 provides us a benchmark for the comparison between LD-ABC and standard ABC methods.

9.1 LEARNING FROM OBFUSCATED DATA VIA ABC

Let us assume that our aim is learning population parameters θ from a table obfuscated resorting to Anatomy, thus assuming an honest learner point of view. The computation of the distribution in (121) implies the ability of sampling from the posterior distribution[1]

$$\pi(\theta|\mathbf{x}^*) \propto \pi(\theta) f(\mathbf{x}^*|\theta) = \pi(\theta)\mathcal{L}(\theta;\mathbf{x}^*).$$

However, there is no tractable analytical expression for the likelihood $\mathcal{L}(\theta;\mathbf{x}^*)$.

In Chapter 8 this problem is circumvented by defining a MCMC scheme for sampling from the joint posterior $\pi(\theta,\mathbf{x}|\mathbf{x}^*)$ on an augmented space. Here, we present an alternative solution based on ABC. Specifically, we define the following generative model providing tables anonymized according to Anatomy:

1. Generate a table of n i.i.d. rows $(s,r) \in \mathcal{S}\times\mathcal{R}$ distributed according to the vector θ given as input;
2. Partition the table into k groups of dimension $\{n_i\}_{i=1}^k$;
3. Randomly permute the nonsensitive attribute values $r_1,...,r_{n_i}$ within each group g_i.

Here n is the number of rows and k the number of groups in the observed anonymized table $\mathbf{x}^*$, while n_i is the number of rows in g_i^*.

By resorting to the above described generative model one is able to get samples from $f(\cdot|\theta)$ despite the unavailability of its analytical form. This allows considering complex relations among the sensitive and nonsensitive characteristics and relaxing the assumption of conditional independence among the nonsensitive attributes in the previous chapter.

The output of ABC algorithms will be a sample from the approximate joint posterior distribution $\tilde{\pi}(\theta,\mathbf{y}^*|\mathbf{x}^*)$. Note that, as discussed in Chapter 3, also ABC methods are based on a data augmentation strategy relying on the introduction of the simulated obfuscated dataset $\mathbf{y}^*$. Accordingly, the main differences between MCMC and ABC strategies are that 1) in the MCMC method the instrumental variable is the cleartext table $\mathbf{x}$, while in the ABC method is the simulated obfuscated table $\mathbf{y}^*$; 2) the MCMC method provides samples from the true posterior distribution while the ABC method gets samples from an approximate posterior distribution.

9.2 COMPARING ABC AND LD-ABC

For the sake of evaluating the performances of the LD-ABC method, we tested both R-ABC and LD-ABC at work on an obfuscated table $\mathbf{x}^*$ obtained by anonymizing a subset of the Adult dataset (5692 rows) where we take *race* (four possible values) as a sensitive attribute and *workclass* (four possible values) as the nonsensitive attribute. The resulting anonymized table is composed of $k = 1,423$ groups, with $n_i = 4$ for each group g_i. Note that in order to apply the method proposed in Chapter 5, the m simulated rows in each dataset $\mathbf{y}^*$ must satisfy the assumptions of Sanov's

1 Note that in the previous chapter we let $f(\cdot)$ denotes all the probability distributions. Here we adopt the same notation as in Chapter 5 denoting by $\pi(\cdot)$ the pdf over the parameter space Θ.

	MEAN INTEGRATED SQUARED ERRORS				$\widehat{\text{ESS}}$	
	Gov.	Self-empl.	Priv.	Without-pay		
R	0.6372	0.9185	16.3660	0.1477	R	16 704
LD	0.4147	0.5814	7.3125	0.106	LD	35 231

Table 15: The leftmost table shows the Squared errors integrated over the 3-simplex and averaged over 100 reruns of ABC. Each column corresponds to an element of $\{\theta_{R|s} : s \in \{\text{Governament, Self-employed, Private, Without-pay}\}\}$. The rightmost table shows the Effective Sample Sizes acheved by R-ABC and LD-ABC averaged over 100 reruns.

theorem, i.e., they must be independent and identically distributed. Recalling that the m pairs (s, r)'s are generated independently and that the permutation is completely at random, we conjecture that the i.i.d. assumption is satisfied. We have positively verified this assumption empirically via the *permutation test* based on the periodicity test statistic described in the National Institute of Standards and Technology Special Publication 800-90B [152].

Here we assume, as in Chapter 8, that θ_S an the $\theta_{R|s}$'s are independently distributed according to non-informative Dirichlet prior distributions. Note that, despite the ABC strategy allows relaxing most of the assumptions needed in the previous chapter, here we still assume all of them in order to take the output of the MwG algorithm as benchmark. The output of ABC is a sample from the approximate joint posterior distribution $\tilde{\pi}(\theta_{R|S}, T_{y^*}|T_{x^*})$, since the sensitive part is not changed by the anonymization algorithm and the posterior distribution for θ_S exists in closed form [2].

We consider the output of 100,000 MCMC runs as a reference, in order to compute the $\widehat{\text{MSE}}$'s and $\widehat{\text{MISE}}$'s and thus comparing the accuracy of LD-ABC and R-ABC. The posterior means derived via MCMC are displayed in Appendix E.3 (Table 18). Results with $m = 100$ and $\epsilon = 1$ are displayed in Tables 16 and 15.

In terms of point estimations the performance of LD-ABC and R-ABC are quite similar. Nevertheless, the $\widehat{\text{MSE}}$'s achieved by LD-ABC are almost always smaller than the ones achieved by R-ABC. Concerning the approximations of the multivariate posterior distributions, looking at the $\widehat{\text{MISE}}$ we can conclude that LD-ABC outperforms R-ABC (see Table 15). Moreover, by focusing on the improvement in efficiency we note that the value of $\widehat{\text{ESS}}$ for LD-ABC is more than twice that for R-ABC.

2 The posterior distribution of θ_s is simply a Dirichlet distribution where the parameters are updated by the frequency counts of each $s \in \mathcal{S}$, as shown in Section 8.5.2.

Table 16: Squared errors averaged over 100 reruns of ABC. Each column corresponds to an element of $\{\theta_{R|s} : s \in \{\text{Governament, Self-employed, Private, Without-pay}\}\}$.

MEAN SQUARED ERRORS					
		Government	Self-emp	Private	Without-pay
White	LD	$1.997 \cdot 10^{-3}$	$3.121 \cdot 10^{-3}$	$3.958 \cdot 10^{-2}$	$1.088 \cdot 10^{-6}$
	R	$3.141 \cdot 10^{-3}$	$5.056 \cdot 10^{-3}$	$7.893 \cdot 10^{-2}$	$2.343 \cdot 10^{-6}$
Asian-Pac-Islander	LD	$2.976 \cdot 10^{-4}$	$3.908 \cdot 10^{-4}$	$6.109 \cdot 10^{-3}$	$8.338 \cdot 10^{-7}$
	R	$4.576 \cdot 10^{-4}$	$6.416 \cdot 10^{-4}$	$1.147 \cdot 10^{-2}$	$1.902 \cdot 10^{-6}$
Black	LD	$3.489 \cdot 10^{-6}$	$1.194 \cdot 10^{-5}$	$3.555 \cdot 10^{-4}$	$1.212 \cdot 10^{-6}$
	R	$2.410 \cdot 10^{-6}$	$2.628 \cdot 10^{-6}$	$1.782 \cdot 10^{-3}$	$2.651 \cdot 10^{-6}$
Other	LD	$6.701 \cdot 10^{-4}$	$1.078 \cdot 10^{-3}$	$1.039 \cdot 10^{-2}$	$1.02 \cdot 10^{-6}$
	R	$1.241 \cdot 10^{-3}$	$2.038 \cdot 10^{-3}$	$1.733 \cdot 10^{-2}$	$2.353 \cdot 10^{-6}$

Part IV

CONCLUSIONS, DISCUSSION AND FUTURE RESEARCH

"Doubt is the origin of wisdom."

— René Descartes

DISCUSSION AND CONCLUSIONS

Models involving several latent variables and nuisance parameters allows a comprehensive representation of complex phenomena, however, in such a case, the inference on the parameters of interest is characterized by a high computational complexity. In fact, the marginalization w.r.t. latent variables and nuisance parameters require the computation of demanding integrals (or summations) on high-dimensional spaces.

Monte Carlo (MC) methods represent a well-known approach to avoid such complex computations and conduct Bayesian inference via simulations. We have reviewed the most important families of MC methods providing, either formally or informally, comparisons among them. Furthermore, we have highlighted that, apart from the standard MC integration, all the MC methods provide samples from the posterior distribution resorting to easy-to-sample proposal distributions. Obviously, sampling directly from the target distribution, when it is possible, provides more efficient MC estimators. However, in most cases, drawing samples from the target is infeasible and the choice of the proposal distribution strongly affects the efficiency of the resulting estimators. We have shown that the evaluation of the efficiency of an algorithm can be based on the effective sample size (ESS), which compares the variability of the estimator based on the actual sampling procedure with that based on direct sampling from the target. A low value of ESS is indicative of sample degeneracy. The problem of sample degeneracy is the focus of the thesis. We have emphasized that it arises when the involved proposal distribution is far from the target and that it becomes more serious in the likelihood-free framework, from Random Weights Importance Sampling to all approximate Bayesian computation (ABC) sampling schemes.

In Part II we have proposed a way of addressing sample degeneracy in ABC methods. Our proposal consists in the definition of a convenient kernel function which allows taking into account the probability of rare events via large deviations theory (LDT). By relying on the Method of Types formulation of LDT we have also overcome the difficulty of selecting the summary statistics summarizing data via their empirical distributions. The proposed kernel function has been involved both in an IS and MCMC sampling scheme. Being defined on a non-compact support, it avoids any implicit or explicit rejection step thus increasing the ESS and improving the mixing of the Markov chain built by the MCMC-ABC algorithm. Moreover, we have shown through several examples that the resulting approximate likelihood, assuming positive values also for "poor" parameter proposals, leads to a better approximation of the posterior density in the tail areas. We have also provided formal guarantees of the improvement induced by our method as well as of the convergence of our ABC approximate likelihood to the true likelihood.

In Part III we have dealt with an application to a real-world problem in the framework of data anonymization. In particular, we have considered data anonymized via group-based anonymization schemes. We have adopted the point of view of an evaluator interested in publishing data in an obfuscated form preserving their utility and, at the same time, protecting the privacy of the involved individuals. We have put forward a notion of relative privacy threat that applies to group-based anonymization

schemes. To rigorously characterize the learner's and attacker's inference, we have defined a unified Bayesian probabilistic model. To perform Bayesian simulated inference, we have proposed a MCMC method relying on the introduction in the model of a high-dimensional auxiliary random variable: the cleartext table. Finally, we have shown that ABC represents a valid alternative for conducting inference from data anonymized resorting to Anatomy. In particular we have emphasized that the ABC approach relies on the introduction of an auxiliary random variable as well. Specifically, it consider as auxiliary variables the simulated obfuscated data, i.e., pseudo data, rather then the cleartext tables. Furthermore, we tested the LD-ABC methodology at work on anonymized data.

LIMITATIONS AND FUTURE RESEARCH In the literature, a variety of methods for performing Bayesian inference via simulations are available. Here, we have focused on the most important sampling schemes, both in a standard framework and in a likelihood-free framework. However, we have not covered more sophisticated algorithms such as Sequential Monte Carlo methods (see [143] among others), Adaptive Markov Chain Monte Carlo and Gradient-Based Markov Chain Monte Carlo techniques (see e.g. [16, Ch. 4-5]). In the literature ABC sequential and adaptive methods have been proposed as well. We speculate that the LD-ABC method can be combined with them by adopting sampling schemes such as Population Monte Carlo [7] and Sequential Monte Carlo [37], rather than the involved IS-ABC and MCMC-ABC. Furthermore, we have only considered the case of a uniform kernel for the pairs in the acceptance region but other kernels can be introduced, e.g., a Gaussian kernel. This would imply a discrimination among the pairs in the acceptance region leading to different importance weights or to an improved mixing of the chain built by the MCMC.

Even though the developments proposed in this thesis deal with a relevant problem and offer a novel perspective on ABC methods, its applicability at the moment is restricted to i.i.d. discrete random variables or finite state Markov Chains. Further research is needed to deepening LDT in order to extend the proposed method to other forms of dependence and possibly to the continuous setting. Moreover, further developments are called in to provide an automatic way of selecting the two tuning parameters, m and ϵ. However, in the last part of the thesis, we have shown the utility of LD-ABC to address real world problems. The application of ABC to data anonymized employing Anatomy allows relaxing several assumptions needed to implement MCMC methods (e.g., conditional independence of the nonsensitive attribute, the assumption that the obfuscation function is deterministic, etc.). Finally, it proceeds on a modern research line offering scope for further developments.

APPENDIX

A

INEQUALITIES AND CONVERGENCE

A.1 LAWS OF LARGE NUMBERS

Theorem 12 (Strong law of large numbers) *Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables, each having the same finite mean μ. Then*

$$\Pr\left(\lim_{n\to\infty} \frac{Y_1 + \ldots + Y_n}{n} = \mu\right) = 1$$

Proof For a proof see [127, Ch 5 Th. 5.4.4] among others. In Appendix D is given a proof of a Method of Types formulation of the Law of Large Numbers, both for sequences of i.i.d random variables and finite state Markov chains. □

A.2 CENTRAL LIMIT THEOREM

Theorem 13 (Central limit theorem) *Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables with mean, μ, and finite variance σ^2. Denoted by $\bar{Y}$ the sample mean $\frac{Y_1+\ldots+Y_n}{n}$, then as $n \to \infty$*

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

meaning that $\frac{(\bar{Y}-\mu)}{\frac{\sigma}{\sqrt{n}}}$ converges in distribution to a standard normal distribution.

Proof See e.g. [127, Ch 11 Th. 11.2.2] □

A.3 INEQUALITIES

Theorem 14 (Jensen's Inequality) *Let Y be a real valued random variable on $\mathcal{Y} \subseteq \mathbb{R}$. If $h(\cdot)$ is convex function on $\mathcal{Y}$, then*

$$\mathbb{E}[h(Y)] \geqslant h\big(\mathbb{E}[Y]\big)$$

provided both expectations exist.

For a strictly convex function $h(\cdot)$, equality holds iff $\mathbb{E}[Y] = Y$ almost surely.

Proof See e.g. [78, Ch 3 Prop. 3.5.1]. □

B

APPROXIMATE METHODS

Let us consider the random variable X distributed according to $p(\cdot)$ and assume that we are able to derive both $\mathbb{E}_p[X] = \mu_X$ and $Var_p[X] = \sigma_X^2$. Suppose that we are interest in deriving the expected value and the variance of the random variable

$$Y = g(X).$$

When the function $g(\cdot)$ is non-linear one can resort to the so called *Delta Method* (see [124]). Such method proceeds by a linearisation carried out through a Taylor series expansion. Accordingly,

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)\frac{dg(\mu_X)}{dX} \tag{154}$$

where $\frac{dg(\mu_X)}{dX}$ denotes the first derivative evaluated at μ_X. Thus, being Y approximated as a linear function of X one can simply take the expectation and the variance in (154).

Now consider two random variables $X \sim p$ and $Y \sim f$ and assume that we are able to compute

$$\begin{aligned} \mathbb{E}_p[X] &= \mu_X \quad Var_p[X] = \sigma_X^2 \\ \mathbb{E}_f[Y] &= \mu_Y \quad Var_f[Y] = \sigma_Y^2. \end{aligned}$$

The expected value and the variance of

$$Z = g(X, Y)$$

can be computed by resorting to the Taylor expansion:

$$Z \approx g(\mu_X, \mu_Y) + (X - \mu_X)\frac{dg(\mu_X, \mu_Y)}{dX} + (Y - \mu_Y)\frac{dg(\mu_X, \mu_Y)}{dY}.$$

Thus,

$$Var[Z] \approx \sigma_X^2\left(\frac{dg(\mu_X, \mu_Y)}{dX}\right)^2 + \sigma_Y^2\left(\frac{dg(\mu_X, \mu_Y)}{dY}\right)^2 + 2\sigma_{XY}\left(\frac{dg(\mu_X, \mu_Y)}{dX}\frac{dg(\mu_X, \mu_Y)}{dY}\right). \tag{155}$$

In order to derive a good approximation for $\mathbb{E}[Z]$ one can resort to the second order Taylor expansion

$$\begin{aligned} Z \approx & g(\mu_X, \mu_Y) + (X - \mu_X)\frac{dg(\mu_X, \mu_Y)}{dX} + (Y - \mu_Y)\frac{dg(\mu_X, \mu_Y)}{dY} \\ & + \frac{1}{2}(X - \mu_X)^2\frac{d^2g(\mu_X, \mu_Y)}{dX^2} + \frac{1}{2}(Y - \mu_Y)^2\frac{d^2g(\mu_X, \mu_Y)}{dY^2} \\ & + (X - \mu_X)(Y - \mu_Y)\frac{d^2g(\mu_X, \mu_Y)}{dXY} \end{aligned}$$

and compute

$$\mathbb{E}[Z] \approx g(\mu_X, \mu_Y) + \frac{1}{2}\sigma_X^2\frac{d^2g(\mu_X, \mu_Y)}{dX^2} + \frac{1}{2}\sigma_Y^2\frac{d^2g(\mu_X, \mu_Y)}{dY^2} + \sigma_{XY}\frac{d^2g(\mu_X, \mu_Y)}{dXY}. \tag{156}$$

B.0.1 *Expectation and Variance of a Ratio*

Let us consider

$$Z = g(X, Y) = \frac{Y}{X}.$$

By computing the following derivatives

$$\begin{aligned}
&\frac{dg(\mu_X, \mu_Y)}{dX} = -\frac{\mu_Y}{\mu_X^2} \qquad \frac{dg(\mu_X, \mu_Y)}{dY} = \frac{1}{\mu_X} \\
&\frac{d^2g(\mu_X, \mu_Y)}{dX^2} = \frac{2\mu_Y}{\mu_X^3} \qquad \frac{d^2g(\mu_X, \mu_Y)}{dY^2} = 0 \\
&\frac{d^2g(\mu_X, \mu_Y)}{dXY} = -\frac{1}{\mu_X^2}
\end{aligned}$$

one can approximate $\mathbb{E}[Z]$ and $Var[Z]$ respectively from (156) and (155):

$$\begin{aligned}
\mathbb{E}[Z] &\approx \frac{\mu_Y}{\mu_X} + \sigma_X^2 \frac{\mu_Y}{\mu_X^3} - \frac{\sigma_{XY}}{\mu_X}, \\
Var[Z] &\approx \sigma_X^2 \frac{\mu_Y^2}{\mu_X^4} + \frac{\sigma_Y^2}{\mu_X^2} - 2\sigma_{XY} \frac{\mu_Y}{\mu_X^3}.
\end{aligned} \tag{157}$$

C

MARKOV CHAINS

A *Markov chain* is a sequence of random variables, $\{X_t\}$, such that the probability distribution of each $X_t \in \mathcal{X}$ depends only on the state attained by X_{t-1}, for each t. Depending on the space on which is defined the parameter t, a Markov process can be classified as a *discrete* or *continuous time* stochastic process. Moreover, the Markov chain is defined as a *finite state* Markov chain when $\mathcal{X}$ is a finite set, otherwise is defined as *infinite state* Markov chain. According to the nature of the Markov Chains described in Section 2.3 and in Chapter 6, throughout this chapter we restrict our attention to discrete time Markov chains, firstly focusing on the finite case. The infinite case is also mentioned. For further details we refer the reader to [50, Ch. 4], [127, Ch. 8].

C.1 DEFINITIONS AND MAIN PROPERTIES

Definition 8 *A finite state and discrete time Markov chain is a stochastic process* $\{X_t\}_{t\in\mathbb{N}}$ *with each* X_t *assuming values in the finite set* $\mathcal{X}$ *and such that*

$$\begin{aligned}\Pr(X_{t+1} = j | X_t = i, X_{t-1} = x_{t-1} ..., X_0 = x_0) &= \Pr(X_{t+1} = j | X_t = i) \\ &= p_{ij}^{(t)} \qquad \forall h \in \mathbb{N}\end{aligned}$$

A typical example of Markov chain is represented by the *random walk*. A *random walk* is a stochastic process generated by a sequence of random variables $\{X_t\}_{\in\mathbb{N}}$ satisfying the following equality:

$$X_{t+1} = X_t + \epsilon_t \tag{158}$$

where $\epsilon_t \perp\!\!\!\perp X_t, X_{t-1}, ..., X_0$.

Accordingly,

$$X_{t+1} \perp\!\!\!\perp X_{t-1} | X_t.$$

A Markov Chain is characterized by a) a *state space*, say $\mathcal{X}$; b) an initial probability distribution over $\mathcal{X}$, $\mathbf{p}^{(0)}$; c) a *transition kernel*[1] – i.e. the conditional probabilities $X_{t+1}|X_t$ for each pair $(X_t, X_{t+1}) \in \mathcal{X}^2$.

Definition 9 *Let* $\{X_t\}_{t\in\mathbb{N}}$ *be a Markov chain assuming values in the finite set* $\mathcal{X}$ *with cardinality* $|\mathcal{X}| = k$. *The* transition matrix *at time* t, $Q^{(t)}$, *is the* $k \times k$ *stochastic matrix composed by entries*

$$q_{ij}^{(t)} \triangleq \Pr(X_{t+1} = j | X_t = i) \quad \forall (i,j) \in \mathcal{X}^2.$$

Each entry of the transition matrix, $q_{ij}^{(t)}$, represents a transition probability –i.e. the probability of going from the state i to the state j. It follows that the transition matrix is a *stochastic matrix*, meaning that

1 An alternative characterization is based on the doublet probability distribution as in Section 6.1.

1. $q_{ij}^{(t)} \in [0,1] \qquad \forall (i,j) \in \mathcal{X}^2$

2. $\sum_{j=1}^{k} q_{ij}^{(t)} = 1.$

The Markov chain is said to be *stationary* or *homogeneous* when the transition probabilities do not depend on the parameter t.

Definition 10 *Let $\{X_t\}_{t\in\mathbb{N}}$ be an homogeneous finite state Markov chain. Then,*

$$Q^{(t_1)} = Q^{(t_2)} \qquad \forall (t_1, t_2) \in \mathbb{N} \times \mathbb{N}$$

For the sake of an easier notation, in the homogeneous case we shortly denote the transition matrix as Q and its entries as q_{ij}.

Let us denote as $\mathbf{p}^{(t)} = (p_1^{(t)}, ..., p_k^{(t)})$ the probability distribution over the set of possible states of the chain at time t. In particular, $\mathbf{p}^{(0)}$ denotes the initial probability distribution. From the *low of total probability* follows that the probability distribution of X_{t+1} is retrieved as

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} Q^{(t)}. \tag{159}$$

Note that $\mathbf{p}^{(t)}$ can be in turn retrieved as $\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)} Q^{(t-1)}$, thus (159) becomes

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t-1)} Q^{(t-1)} Q^{(t)}.$$

Going backwards, the probability distribution at the $(t+r)$-th step is

$$\mathbf{p}^{(t+r)} = \mathbf{p}^{(t)} Q^{(t)} ... Q^{(t+r-1)}. \tag{160}$$

Equation (160) is known as Chapman - Kolmogorov equation. Since in the homogeneous case $Q^{(t)} = ... = Q^{(t+r-1)} = Q$, the (160) becomes

$$\mathbf{p}^{(t+r)} = \mathbf{p}^{(t)} Q^r$$

where Q^r is the r-th power of the transition matrix composed by entries $q_{ij}^r = \Pr(X_{t+r} = j | X_t = i)$.

C.2 CLASSIFICATION OF STATES

Let us consider an homogeneous discrete time Markov chain defined on $\mathcal{X}$. The state space $\mathcal{X}$ (finite or infinite) can be classified according to two possible partitions of $\mathcal{X}$:

$$\mathcal{X} = \mathcal{A} \cup \mathcal{P} \tag{161}$$

$$= \mathcal{R}_0 \cup \mathcal{R}_+ \cup \mathcal{T} \tag{162}$$

In (161) we denote as $\mathcal{A}$ the set of *aperiodic states* and as $\mathcal{P}$ the set of *periodic states*. Each possible state in $\mathcal{X}$ is said to be periodic or aperiodic depending on the value assumed by its period. The period is defined as follows.

Definition 11 (Period) *For each state* $i \in \mathcal{X}$*, the period* $d(i)$ *is defined as the greatest common divisor (gcd) of the set*

$$\{t \geqslant 1 : q_{ii}^t > 0\}. \tag{163}$$

The state i *is* periodic *if* $d(i) > 1$ *and* aperiodic *if* $d(i) = 1$.

Thus, a strictly positive probability of remaining in a state $i \in \mathcal{X}$ – i.e. $q_{ii}^1 > 0$ – implies that the above-defined gcd equals 1 and represents a sufficient condition for the aperiodicity. However, such condition is not necessary since the gcd can be equal to 1 also when $q_{ii}^1 = 0$ (e.g. when the set in (163) contains only prime numbers).

Another possible classification of states follows from the partition in (162), where $\mathcal{T}$, $\mathcal{R}_+$, $\mathcal{R}_0$ represent the set of *transient*, *positive recurrent* and *null recurrent* states, respectively. Let us denote as f_i^t the probability that the first return to a given state, say i, occurs after t steps. The probability that a return to i does occur, at whatever time, is defined as

$$f_i = \sum_{t=1}^{\infty} f_i^t \qquad \forall i \in \mathcal{X}.$$

Accordingly, each state $i \in \mathcal{X}$ is defined as

transient iff $f_i < 1$;

recurrent iff $f_i = 1$.

Let T_i be the random variable representing the time at which the first return to i occurs. Its expected value $\mathbb{E}(T_i|X_0 = i)$ allows for discriminating between positive and null recurrent states. In fact, the state i is classified as

positive recurrent iff $\mathbb{E}(T_i|X_0 = i) < +\infty$;

null recurrent iff $\mathbb{E}(T_i|X_0 = i) = +\infty$.

Note that recurrent states can be periodic or aperiodic since the two classifications derive from different partitions of $\mathcal{X}$.

Given a pair of states $(i, j) \in \mathcal{X}^2$ we say that j is *accessible* from i, written $i \to j$, if there is a direct path from i to j meaning that

$$\exists t \in \mathbb{N} : \Pr(X_t = j|X_0 = i) = q_{ij}^t > 0.$$

States i and j *communicate*, written $i \leftrightarrow j$, if j is accessible from i and viceversa.

Definition 12 (Classes of states) *A* class *of states,* $\mathcal{C}$*, is a non-empty subset of* $\mathcal{X}$*.*

A class $\mathcal{C}$ *is said to be a* closed class *if*

$$q_{ij} = 0 \quad \forall i \in \mathcal{C} \text{ and } \forall j \in \mathcal{X} \setminus \mathcal{C}$$

or to be irreducible class *if*

$$i \leftrightarrow j \quad \forall (i, j) \in \mathcal{C}^2.$$

Theorem 15 *For any Markov chain* $\{X_t\}$ *all the states in the same* irreducible class *have the same period.*

Proof Let $(i,j) \in \mathcal{C}^2$ be a pair of states in the irreducible class $\mathcal{C}$. Then $i \leftrightarrow j$ meaning that there exist some r and s such that $Pr(X_{t+r} = j|X_t = i)$ and $Pr(X_{t+s} = i|X_t = j)$. It follows that there exists a path of length $r+s$ going from i to j and back to i. Thus, according to Definition 11, $r+s$ must be divisible by $d(i)$. Let h be any integer such that $q_{jj}^h > 0$. Since there is a path of length $r+h+s$ going from i to j, then again to j and then back to i (see Figure 19), $r+h+s$ must be divisible by $d(i)$ and thus h is divisible by $d(i)$. Since this holds true for any h such that $q_{jj}^h > 0$, $d(j)$ is divisible by $d(i)$. Reversing the roles of i and j, $d(i)$ is also divisible by $d(j)$ so $d(i) = d(j)$. □

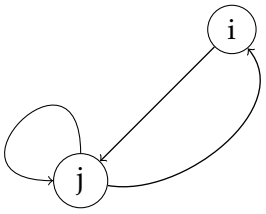


Figure 19: Representation of the $r+h+s$ steps path between i and j.

Theorem 16 *Let $\mathcal{C}$ be an irreducible class. The states of the Markov chain in $\mathcal{C}$ are all transient or all recurrent.*

Proof Let us consider $(i,j) \in \mathcal{C}^2$. Suppose that i is a transient state. Thus, there exists a state $k : i \to k, k \nrightarrow i$. Being $j \to i$ and $i \to k$, there exists a path such that $j \to k$. If $k \to j$, it would exists a path going from k to j and then from j to i, thus contradicting the hypothesis. It follows that must be $k \nrightarrow j$, meaning that j is a transient state (see Figure 20). □

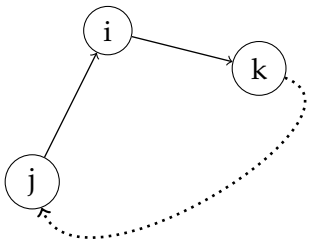


Figure 20: Representation of two transient states $(i,j) \in \mathcal{C}$.

A Markov chain $\{X_t\}_{t\in\mathbb{N}}$ assuming values in $\mathcal{X}$ is *irreducible* when $\mathcal{X}$ corresponds to an unique irreducible class, meaning that each state in $\mathcal{X}$ communicates with each other. Hence, given an irreducible Markov chain its states are all periodic or aperiodic and all transient or recurrent. In particular, considering a finite state Markov chain, all its states are positive recurrent (see [127, Th. 8.4.9]).

Definition 13 (Ergodic Markov chain) *A Markov chain is said to be* ergodic *when it is aperiodic, irreducible and positive recurrent.*

From Definition 13 follows that for an irreducible finite state Markov chain, the aperiodicity is a necessary and sufficient condition under which the chain is ergodic.

	Finite		Infinite	
	Reducible	**Irreducible**	**Reducible**	**Irreducible**
Positive	*one or infinite*	*one*	*infinite or none*	*one ore more*
Not pos.	*one or infinite*	\\	*infinite or none*	*none*

Table 17: Conditions for the existence and uniqueness of the stationary distribution.

C.3 THE STATIONARY DISTRIBUTION

Definition 14 *Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov chain taking values in $\mathcal{X}$ and let Q be its transition matrix. A probability distribution over $\mathcal{X}$, $\boldsymbol{\pi} = (\pi_1, ...\pi_k)$, is said to be the* stationary distribution *when*

$$\boldsymbol{\pi} = \boldsymbol{\pi}Q.$$

Generally speaking, a Markov chain may or may not admit a stationary distribution. Furthermore, when a stationary distribution exists it may or may not be unique. A finite states Markov chain always admits at least one stationary distribution and the irreducibility of the chain ensures the uniqueness. As regards infinite states Markov chains, the stationary distribution may not exists.

Table 17 summarizes some important results about the conditions for the existence and uniqueness of the stationary distribution.

Definition 15 (Asymptotic distribution) *Let $\{X_t\}_{t\in\mathbb{N}}$ to be a Markov chain taking values in $\mathcal{X}$ with transition kernel Q. The* asymptotic distribution *for that chain is the unique stationary distribution $\boldsymbol{\pi}$ satisfying*

$$\lim_{n\to+\infty} \mathbf{p}^{(0)}\mathbf{Q}^{(n)} = \boldsymbol{\pi} \tag{164}$$

for each distribution $\mathbf{p}^{(0)}$ over $\mathcal{X}$.

Theorem 17 (Ergodic theorem) *Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov chain with transition kernel Q. If $\{X_t\}_{t\in\mathbb{N}}$ is ergodic, the following asymptotic distribution exists*

$$\boldsymbol{\pi} = (\pi_i = 1/\mathbb{E}[T_i|X_0 = i])_{i\in\mathcal{X}}.$$

Furthermore, as n goes to $+\infty$ the transition matrix Q^n tends to a matrix in which each rows equals $\boldsymbol{\pi}$.

On one side the *irreducibility* ensures the existence of the stationary distribution for a *recurrent positive* Markov chain. On the other side the *aperiodicity* implies that the probability of each state does not depends on the initial state.

In any case, the *reversibility* of the chain is a sufficient condition for the stationarity of the chain.

Definition 16 (Reversibility) *Let $\boldsymbol{\pi}$ be a probability distribution over $\mathcal{X}$. The distribution $\boldsymbol{\pi}$ is said to be* reversible *if the following* detailed balance *equation is satisfied:*

$$\pi_i q_{ij} = \pi_j q_{ji} \qquad \forall (i,j) \in \mathcal{X}^2 \tag{165}$$

The detailed balance condition in (165) implies the stationarity being

$$\sum_{i \in \mathcal{X}} \pi_i q_{ij} = \sum_{i \in \mathcal{X}} \pi_j q_{ji} = \pi_j \sum_{i \in \mathcal{X}} q_{ji}.$$

If the transition matrix is a doubly stochastic matrix – i.e. a squared matrix in which each row and column sum to one – we obtain

$$\sum_{i \in \mathcal{S}} \pi_j q_{ji} = \pi_j.$$

C.4 SOME USEFUL CONCEPTS ABOUT CONTINUOUS STATE MARKOV CHAINS

In the previous section we dealt with Markov chains taking values in a finite set. Let us consider the case in which each random variable X_t is defined on a continuous space $\mathcal{X} \subseteq \mathbb{R}$. In such cases the Markov property can be reformulated as follows

$$f_{X_{t+1}|X_t,\dots,X_0}(y|x,\dots,x_0) = f_{X_{t+1}|X_t}(y|x)$$

where $f(\cdot)$ denotes a probability density function. Moreover the transition matrix is replaced by a *transition density function* $q(\cdot,\cdot)$ such that

$$\int_{\mathcal{X}} q(x_t, x_{t+1}) dx_{t+1} = 1.$$

Thus, the probability density function for the random variable X_{t+1} is retrieved as

$$f_{X_{t+1}}(y) = \int_{\mathcal{X}} f_{X_t}(x) q(x,y) dx.$$

The two partitions of the state space in (161) and (162) are also valid in the continuous-state case. However, in the continuous setting another important property is the *Harris recurrence.*

Definition 17 (Harris recurrence) *Let us denote as η_A the number of times the chain visits the measurable set $A \subset \mathcal{X}$. The set $A \subset \mathcal{X}$ is said* Harris recurrent *if $\Pr(\eta_A = \infty) = 1$ for all $x \in A$. The chain $\{X_t\}_{t \in \mathbb{N}}$ is said* Harris recurrent *if it is irreducible and every measurable set $A \subset \mathcal{X}$ is Harris recurrent.*

D

APPENDIX TO PART II

D.1 THEOREMS AND PROOFS IN CHAPTER 5

Proposition 5 *Let* $\mathbf{X}^n = X_1, ..., X_n$ *be a sequence of i.i.d. random variables distributed over* $\mathcal{X}$ *according to* P *and let* T_{x^n} *denotes its type. Then*

$$\Pr(\mathbf{X}^n = x^n) = 2^{n\left(-D(T_{x^n}||P)-H(T_{x^n})\right)}. \tag{166}$$

Proof

$$\begin{aligned}
\Pr(\mathbf{X}^n = x^n) &= \prod_{i=1}^{n} \Pr(X_i = x_i) \\
&= \prod_{r\in\mathcal{X}} P(r)^{nT_{x^n}(r)} \\
&= \prod_{r\in\mathcal{X}} 2^{nT_{x^n}(r)\log P(r)} \\
&= \prod_{r\in\mathcal{X}} 2^{n(T_{x^n}(r)\log P(r) - T_{x^n}(r)\log T_{x^n}(r) + T_{x^n}(r)\log T_{x^n}(r))} \\
&= 2^{n\sum_{r\in\mathcal{X}}(-T_{x^n}(r)\log\frac{T_{x^n}(r)}{P(r)} + T_{x^n}(r)\log T_{x^n}(r))} \\
&= 2^{n(-D(T_{x^n}||P)-H(T_{x^n}))}.
\end{aligned}$$

□

In what follows, we will make use of a few basic notions and facts about the Method of Types and information projections, for which we refer the reader to [33, Ch.1].

The simplex of the distributions over $\mathcal{X}$, seen a subset $\Delta^{|\mathcal{X}|-1} \subseteq \mathbb{R}^{|\mathcal{X}|}$, inherits the standard topology from $\mathbb{R}^{|\mathcal{X}|}$. W.r.t. this topology, the function $D(P||Q)$ is lower semi-continuous in the pair of arguments (P, Q), and continuous at (P, Q) whenever Q has full support, that is whenever $\mathrm{supp}(Q) \stackrel{\triangle}{=} \{r \in \mathcal{X} : Q(r) > 0\} = \mathcal{X}$. Convergence to Q in KL divergence, $D(Q_n||Q) \to 0$, implies convergence in the standard topology, $Q_n \to Q$. As a function of P, $D(P||Q)$ is strictly convex, and continuous whenever Q is full support. Hence for any convex and closed set $E \subseteq \Delta^{|\mathcal{X}|-1}$ the information projection of Q onto E, $P^* = \mathrm{argmin}_{P\in E} D(P||Q)$, exists and is unique. The following is a fundamental result about information projections. The support of E is defined as $\mathrm{supp}(E) \stackrel{\triangle}{=} \bigcup_{P\in E} \mathrm{supp}(P)$.

Theorem 18 (Pythagorean inequality, [33] Th.3.1) *Let* E *be a closed and convex set and* Q *be full support. Let* $P^* = \mathrm{argmin}_{P\in E} D(P||Q)$. *Then* $\mathrm{supp}(P^*) = \mathrm{supp}(E)$. *Moreover, for each* $P \in E$, $D(P||Q) \geqslant D(P||P^*) + D(P^*||Q)$.

Proof of Theorem 5 (see [27, Ch 11.2.1]:) Let us consider a δ-*typical set* of probability distributions defined as follows:

$$\mathcal{B}_\delta(P_\theta) \stackrel{\triangle}{=} \{P \in \Delta^{|\mathcal{X}|-1} : D(P||P_\theta) \leqslant \delta\}.$$

The probability that a sequence $\mathbf{X}^n$ leads to a *non-typical* type can be bounded as follows:

$$\begin{aligned}
\Pr\left(T_{\mathbf{X}^n} \notin \mathcal{B}_\delta(P_\theta)\right) &= \sum_{P \in \mathcal{T}^m \cap \mathcal{B}_\delta^C} \Pr\left(T_{\mathbf{X}^n} = P|\theta\right) \\
&\leqslant \sum_{P \in \mathcal{T}^m \cap \mathcal{B}_\delta^C} 2^{-nD(P||P_\theta)} \qquad (167) \\
&\leqslant \sum_{P \in \mathcal{T}^m \cap \mathcal{B}_\delta^C} 2^{-n\delta} \\
&\leqslant (n+1)^{|\mathcal{X}|} 2^{-n\delta} \qquad (168) \\
&= 2^{-n(\delta - |\mathcal{X}|\frac{\log(n+1)}{n})}
\end{aligned}$$

where (167) and (168) follow from the bounds for the probability of the *type class* and the size of $\mathcal{T}^m$, respectively (see [27, Th. 11.1.4] and [27, Th. 11.1.1]). Accordingly,

$$\begin{aligned}
\Pr\left(D(T_{\mathbf{X}^n}||P_\theta) \leqslant \delta|\theta\right) &= 1 - \Pr\left(T_{\mathbf{X}^n} \notin \mathcal{B}_\delta(P_\theta)\right) \\
&\geqslant 1 - 2^{-n(\delta - |\mathcal{X}|\frac{\log(n+1)}{n})}.
\end{aligned}$$

Moreover, summing over n

$$\sum_{n=1}^{\infty} \Pr\left(D(T_{\mathbf{X}^n}||P_\theta) > \delta|\theta\right) > \infty.$$

Thus, applying the Borel-Cantelli lemma [127, Th. 3.4.2] to the event $\{D(T_{\mathbf{X}^n}||P_\theta) > \delta\}$ follows that

$$\Pr\left(D(T_{\mathbf{X}^n}||P_\theta) > \delta \text{ i.o. } |\theta\right) = 0$$

where "i.o." is read as " infinitely often". Since this holds for every $\delta > 0$ under $\Pr(\cdot|\theta)$, as $n \to +\infty$, $D(T_{\mathbf{X}^n}||P_\theta) \to 0$ with probability 1. □

Proof of Theorem 7: Fix an infinite sequence $\tau \in \mathcal{Y}^\infty$, $\tau = (y_1, y_2, y_3, ...)$. For each $m \geqslant 1$, let $T_{\mathbf{y}^m}(\tau)$, $T_{\mathbf{y}^m}$ for short, denote the type of the first m symbols of τ, the sequence $(y_1, ..., y_m)$. Assume τ is such that $T_{\mathbf{y}^m} \to P_\theta$ as $m \to +\infty$. Note that, since P_θ is full support, this implies that for all sufficiently large m, $T_{\mathbf{y}^m}$ is full support as well. Define P^* and, for any such sufficiently large m, P^*_m as follows:

$$P^* \stackrel{\triangle}{=} \operatorname*{argmin}_{P \in \mathcal{B}_\epsilon} D(P||P_\theta) \quad \text{and} \quad P^*_m \stackrel{\triangle}{=} \operatorname*{argmin}_{P \in \mathcal{B}_\epsilon} D(P||T_{\mathbf{y}^m}).$$

Note that as $T_{\mathbf{y}^m}$ is full support and $\mathcal{B}_\epsilon = \{P : D(P||T_{\mathbf{x}^n}) \leqslant \epsilon\}$ is convex and closed, the projection P^*_m exists and is unique. Moreover, as $T_{\mathbf{x}^n}$ is by assumption full support, it is easily seen that $\text{supp}(\mathcal{B}_\epsilon) = \mathcal{X}$: hence, by the first part of Theorem 18, the projection P^*_m is full support as well.

As $\mathcal{B}_\epsilon$ is closed and $D(\cdot||P_\theta)$ is continuous, $P^* \in \mathcal{B}_\epsilon$. We can now apply the Pythagorean Inequality, considering P^*_m as a projection and $P^* \in E = \mathcal{B}_\epsilon$, and obtain

$$D(P^*||T_{\mathbf{y}^m}) \geqslant D(P^*||P^*_m) + D(P^*_m||T_{\mathbf{y}^m}). \qquad (169)$$

As P_θ is assumed to be full support, $D(\cdot\|\cdot)$ as a function of its second argument is continuous at P_θ, hence

$$\lim_{m\to\infty} D(P^*\|T_{y^m}) = D(P^*\|P_\theta)\,. \tag{170}$$

Assuming $\{P^*_m\}$ converges, let $P^{**} \stackrel{\triangle}{=} \lim_{m\to\infty} P^*_m$, where clearly $P^{**} \in \mathcal{B}_\epsilon$; if $\{P^*_m\}$ does not converge, we can equivalently take any convergent subsequence of it. Taking lim inf on both sides of (169), and exploiting (170) on the left-hand side, and lower semi-continuity on the right-hand side, we can write

$$\begin{aligned} D(P^*\|P_\theta) &= \lim_{m\to\infty} D(P^*\|T_{y^m}) \\ &\geqslant \liminf_{m\to\infty} \left(D(P^*\|P^*_m) + D(P^*_m\|T_{y^m}) \right) \\ &\geqslant \liminf_{m\to\infty} D(P^*\|P^*_m) + \liminf_{m\to\infty} D(P^*_m\|T_{y^m}) \\ &\geqslant D(P^*\|P^{**}) + D(P^{**}\|P_\theta)\,. \end{aligned}$$

Summing up

$$D(P^*\|P_\theta) \geqslant D(P^*\|P^{**}) + D(P^{**}\|P_\theta)\,. \tag{171}$$

Recalling that P^* is the information projection of P_θ onto $\mathcal{B}_\epsilon$, that $P^{**} \in \mathcal{B}_\epsilon$ and that $D(\cdot\|\cdot)$ is nonnegative, the only possibility for (171) to hold is that $D(P^*\|P^{**}) = 0$, which implies $P^* = P^{**}$. In other words

$$\lim_{m\to\infty} P^*_m = P^*\,. \tag{172}$$

This way, we have shown that $(P^*_m, T_{y^m}) \to (P^*, P_\theta)$. Under $D(\cdot\|\cdot)$ this limit becomes, by continuity at (P^*, P_θ):

$$\lim_{m\to\infty} D(P^*_m\|T_{y^m}) = D(P^*\|P_\theta)\,. \tag{173}$$

We have shown that (173) holds true for any sequence $\tau \in \mathcal{X}^\infty$ such that $T_{y^m} = T_{y^m}(\tau) \to P_\theta$. Now let $\Pr(\cdot|\theta)$ be the probability measure on $\mathcal{Y}^\infty$ induced by P_θ. The LLN (Theorem 5) says that, under $\Pr(\cdot|\theta)$, the set of such τ's has probability 1. Hence (173) under $\Pr(\cdot|\theta)$ holds with probability 1, that is, almost surely. □

Recall that, for each Q and $\delta \geqslant 0$, $\mathcal{B}_\delta(Q) \subseteq \Delta^{|\mathcal{X}|-1}$ denotes the ball of radius δ centered at Q:

$$\mathcal{B}_\delta(Q) \stackrel{\triangle}{=} \{P : D(P\|Q) \leqslant \delta\}\,.$$

Lemma D.1.1 *Let $E \subseteq \Delta^{|\mathcal{X}|-1}$ be a convex and closed set. Let $Q \in \Delta^{|\mathcal{X}|-1}$ be such that $\gamma \stackrel{\triangle}{=} D(E\|Q) > 0$. Then for each $0 < \gamma' < \gamma$ there is $\delta > 0$ such that for each $Q' \in \mathcal{B}_\delta(Q)$ one has $D(E\|Q') \geqslant \gamma'$.*

Proof The fact that E is closed and convex ensures that the projection $D(E\|Q)$ exists and is finite. Consider the strictly descending chain of balls of radius $\delta_n = 1/n$ centered at Q: $\mathcal{B}_{\delta_1}(Q) \supseteq \mathcal{B}_{\delta_2}(Q) \supseteq \cdots \supseteq \mathcal{B}_{\delta_n}(Q) \supseteq \cdots$. By contradiction, assume

that there exists $0 < \gamma' < \gamma$ such that for each $\delta > 0$, there is $Q' \in \mathcal{B}_\delta(Q)$ such that $D(E\|Q') < \gamma'$. In particular, we then have that

$$\text{for each } n \geqslant 1 \text{ there is } Q'_n \in \mathcal{B}_{\delta_n}(Q) \text{ s.t. } D(E\|Q'_n) < \gamma'. \tag{174}$$

We can therefore assume without loss of generality that

$$\lim_{n\to\infty} D(E\|Q'_n) < \gamma'. \tag{175}$$

(if not, we can anyway extract from $\{D(E\|Q'_n)\}$ a subsequence with the desired property). On the other hand, being $\lim_{n\to\infty} D(Q'_n\|Q) = 0$, we have $\lim_{n\to\infty} Q'_n = Q$. Being $D(\cdot\|\cdot)$ lower semi-continuous, we obtain

$$\liminf_{n\to\infty} D(E\|Q'_n) \geqslant D(E\|Q) = \gamma > \gamma'. \tag{176}$$

But this contradicts (174). □

Proof of Proposition 2: Let us consider $\widehat{ESS} : \mathbb{R}^S \to \mathbb{R}$ as a function of S variables, $\widehat{ESS}(x_1, ..., x_S)$, defined for nonnegative reals x_i's, not all zero, representing the weights. The partial derivative of $\widehat{ESS}$ w.r.t. x_i has the form

$$\frac{\partial}{\partial x_i}\widehat{ESS}(x_1, ..., x_S) = C \cdot \sum_{j\neq i}(x_j^2 - x_i x_j)$$

for a function C that is > 0 in the domain of definition of $\widehat{ESS}$. Therefore, $\frac{\partial}{\partial x_i}\widehat{ESS}$ is nonnegative when evaluated at any point $(x_1, ..., x_S)$ in the domain of $\widehat{ESS}$ with the following property: for each $j \neq i$ s.t. $x_j > 0$, one has $0 \leqslant x_i \leqslant x_j$. If additionally at least one $j \neq i$ exists s.t. $x_j > x_i$, then $\frac{\partial}{\partial x_i}\widehat{ESS}$ is strictly positive.

An execution of the IS algorithm consists of $S \geqslant 1$ independent iterations of the main loop: let us denote by ω_s and ρ_s the unnormalized weights (82) generated using the LD-ABC and IS-ABC kernel functions, respectively, at iteration $s = 1, ..., S$, and by $\omega = (\omega_1, ..., \omega_S)$ and $\rho = (\rho_1, ..., \rho_S)$ the resulting sequences. By definition, the set of indices $s = 1, ..., S$ can be partitioned into three subsets: the subset A where $\rho_s = \omega_s > 0$, the subset B where $\rho_s = 0$ and $\omega_s > 0$, and the subset C where $\rho_s = \omega_s = 0$. Moreover, for each $s \in A$ and $s' \in B$, $\omega_s > \omega_{s'}$. For notational simplicity, assume $A = \{1, ..., h\}$, $B = \{h+1, ..., S'\}$ and $C = \{S'+1, ..., S\}$, for some $0 \leqslant h \leqslant S' \leqslant S$. Also assume, again only for notational simplicity, that $\omega_{h+1} \geqslant \omega_{h+2} \geqslant \cdots$.

If $S' = 0$, then $h = 0$ and by definition $\widehat{ESS}_{LD} = \widehat{ESS}_{IS} = 0$, hence assume $S' > 0$. If $h = S$, then $S' = S$ and $\omega = \rho$, hence the inequality in the statement again holds trivially as equality. Consider now a case where $0 < h < S$, that is $\omega \neq \rho$. For each $i = h+1, ..., S'$, consider a point $\rho_i(x) \stackrel{\triangle}{=} (\omega_1, ..., \omega_{i-1}, x, 0, ...0)$, with $0 \leqslant x \leqslant \omega_i$. The fact that $0 \leqslant x \leqslant \omega_j$ for each $j < i$, and moreover that $\omega_1 > x_i$, by the above considerations entails the strict positivity of $\frac{\partial}{\partial x_i}\widehat{ESS}$ when evaluated at $\rho_i(x)$, for $0 \leqslant x \leqslant \omega_i$. Therefore, considering $i = h+1, ..., S'$ in turn, we have:

$$\begin{aligned}
\widehat{ESS}_{IS} &= \widehat{ESS}(\omega_1, ..., \omega_h, 0, .., 0) \\
&< \widehat{ESS}(\omega_1, ..., \omega_h, \omega_{h+1}, 0, .., 0) \\
&< \cdots \\
&< \widehat{ESS}(\omega_1, ..., \omega_h, \omega_{h+1}, ..., \omega_{S'}, 0, ..., 0) \\
&= \widehat{ESS}_{LD}\,.
\end{aligned}$$

□

Let $\mathcal{X} = \{1, ..., |\mathcal{X}|\}$ be a non-empty finite set. We recall that capital letters P and P' denotes doublet probability distributions in the probability simplex $\Delta^{|\mathcal{X}|^2-1}$. The corresponding marginal distributions are denoted by $\mathbf{p} = (p_1, ..., p_{|\mathcal{X}|})$ and $\mathbf{p'} = (p'_1, ..., p'_{|\mathcal{X}|})$ with $\mathbf{p}, \mathbf{p'} \in \Delta^{|\mathcal{X}|-1}$. Q and Q' denote the stochastic matrices whose elements are retrieved as $Q(ij) = P(ij)/p_i$ and $Q'(ij) = P'(ij)/p'_i$. P_θ and $\mathbf{p}_\theta$ denote a doublet probability distribution over $\mathcal{X}^2$ parametrized according to θ and the corresponding marginal distribution over $\mathcal{X}$.

The simplex of the distributions over $\mathcal{X}^2$, seen a subset $\Delta^{|\mathcal{X}|^2-1} \subseteq \mathbb{R}^{|\mathcal{X}|^2}$, inherits the standard topology from $\mathbb{R}^{|\mathcal{X}|^2}$. W.r.t. this topology, the function $D_c(P||P')$ is continuous at (P, P') whenever P' has full support, that is whenever $\mathrm{supp}(P') \stackrel{\triangle}{=} \{(i,j) \in \mathcal{X}^2 : P'(ij) > 0\} = \mathcal{X}^2$. Convergence to P' in conditional relative entropy, $D_c(P'_n||P') \to 0$, implies the convergence of the conditional distribution in standard topology, $Q'_n \to Q'$. As a function of P, $D_c(P||P')$ is strictly convex, and continuous whenever P' is full support. Hence for any convex and closed set $E \subseteq \Delta^{|\mathcal{X}|^2-1}$, $P^* = \mathrm{argmin}_{P\in E} D_c(P||P')$ exists and is unique. The support of E is defined as $\mathrm{supp}(E) \stackrel{\triangle}{=} \bigcup_{P\in E} \mathrm{supp}(P)$.

In what follows we extend to the second order types some important results about the first order type reported in [27]. Throughout this section we assume the *cyclic convention*.

Theorem 19 (Pythagorean inequality in conditional relative entropy) *Let E be a closed and convex set and P' be full support. Let $P^* = \mathrm{argmin}_{P\in E} D_c(P||P')$. Then $\mathrm{supp}(P^*) = \mathrm{supp}(E)$. Moreover, for each $P \in E$, $D_c(P||P') \geqslant D_c(P||P^*) + D_c(P^*||P')$.*

Proof Consider $P \in E$ and the following convex combination

$$P_\lambda(ij) = \lambda P(ij) + (1-\lambda)P^*(ij) \quad \forall (ij) \in \mathcal{X}^2. \tag{177}$$

Note that $P_\lambda \to P^*$ as $\lambda \to 0$ and , since E is convex, $P_\lambda \in E$ for all $\lambda \in [0, 1]$. Since $D_c(P*||P')$ is the minimum along the path from P^* to P, the derivative of $D_c(P_\lambda||P')$ must be non-negative at $\lambda = 0$.

Let us denote the conditional relative

$$\frac{dD_c^\lambda}{d\lambda} = \sum_{i\in\mathcal{X}} \sum_{j\in\mathcal{X}} \big[P(ij) - P^*(ij)\big] \log \frac{P_\lambda(ij)}{p_\lambda(i)Q'(ij)} \tag{178}$$

$$+ \sum_{i\in\mathcal{X}} \sum_{j\in\mathcal{X}} \big[\big(P(ij) - P^*(ij)\big) - Q_\lambda(ij)\big(p_i - p_i^*\big)\big] \tag{179}$$

where $p_\lambda(i) = \sum_{j\in\mathcal{X}} P_\lambda(ij)$ and $Q_\lambda(ij) = P_\lambda(ij)/p_\lambda(i)$.

By noting that the second term in (179) equals 0, the derivative evaluated at $\lambda = 0$ equals

$$\begin{aligned} 0 \leqslant \frac{dD_c^{\lambda}}{d\lambda}|_{\lambda=0} &= \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{X}} \left[P(ij) - P^*(ij)\right] \log \frac{P^*(ij)}{p^*(i)Q'(ij)} \\ &= \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{X}} P(ij) \log \frac{P^*(ij)}{p^*(i)Q'(ij)} - D_c(P^*||P') \\ &= \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{X}} P(ij) \log \frac{P(ij)}{p(i)Q'(ij)} \frac{P^*(ij)p(i)}{p^*(i)P(ij)} - D_c(P^*||P') \\ &= D_c(P||P') - D_c(P||P^*) - D_c(P^*||P'). \end{aligned} \tag{180}$$

It follows that

$$D_c(P||P') \geqslant D_c(P||P^*) + D_c(P^*||P'). \tag{181}$$

Moreover, the inequality (180) rules out the contingency that $\text{supp}(P^*) \subset \text{supp}(P)$ since it would imply that as $\lambda \to 0$ the derivative $\frac{dD_c^{\lambda}}{d\lambda} \to -\infty$. Since this holds for each $P \in E$, $\text{supp}(P^*) = \text{supp}(E)$. □

Theorem 20 *Let $\mathcal{T}(n,2) \subseteq \Delta^{|\mathcal{X}|^2-1}$ denote the set of the second order n-types.*

$$|\mathcal{T}(n,2)| \leqslant (n+1)^{|\mathcal{X}|^2}.$$

Proof This theorem is the analogue of [27, Th. 11.1.1] for second order types. As in [27, Th. 11.1.1], this result follows from the fact that the type has $|\mathcal{X}|^2$ components and the numerator of each component can assume $n+1$ values. □

Definition 18 (Type class) *Let $T \in \mathcal{T}(n,2)$ denotes the second order type of a markovian sequence of length n. Its* type class, *denoted by $\mathcal{C}(T)$, is the following set of sequences of length n*

$$\mathcal{C}(T) \stackrel{\triangle}{=} \{x^n \in \mathcal{X}^n : T^{(2)}_{x^n} = T\}$$

Lemma D.2.1 (Lemma 3 from [102]) *Let us consider a sample path $\mathbf{X}^n = X_1, ..., X_n$ from a Markov process with stationary doublet probability distribution P_θ. Let us denote by $x^n = x_1, ..., x_n$ a realization from that sample path with second order type $T^{(2)}_{x^n}$. Assume that $\alpha = \min_{i\in\mathcal{X}} p_\theta(i)$ and $\beta = \min_{(ij)\in\mathcal{X}^2} P_\theta(ij)$. Then*

1. $\Pr(\mathbf{X}^n = x^n|\theta) = \dfrac{p_\theta(x_1)}{P_\theta(x_n, x_1)} 2^{-n[D_c(T^{(2)}_{x^n}||P_\theta) + H_c(T^{(2)}_{x^n})]}$
2. $n^{-|\mathcal{X}|}(n+1)^{-|\mathcal{X}|^2}\, 2^{nH_c(T^{(2)}_{x^n})} \leqslant |\mathcal{C}(T^{(2)}_{x^n})| \leqslant |\mathcal{X}|\, 2^{nH_c(T^{(2)}_{x^n})}$
3. $n^{-|\mathcal{X}|}(n+1)^{-|\mathcal{X}|^2}\, \alpha 2^{-nD_c(T^{(2)}_{x^n}||P_\theta)} \leqslant \Pr(T^{(2)}_{\mathbf{X}^n} = T^{(2)}_{x^n}|\theta) \leqslant \frac{|\mathcal{X}|}{\beta} 2^{-nD_c(T^{(2)}_{x^n}||P_\theta)}.$

Theorem 21 (Law of Large Numbers for Finite state Markov Chains) *Let $\mathbf{X}^n = X_1, ..., X_n$ be a sample path from a Markov process with doublet probability distribution P_θ. Then $\forall \delta > 0$:*

$$\Pr\left(D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) \leqslant \delta|\theta\right) \geqslant 1 - 2^{-n(\delta - |\mathcal{X}|^2 \log \frac{n+1}{n})} \tag{182}$$

Moreover, under $\Pr(\cdot|\theta)$, as $n \to \infty$, $D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) \to 0$ with probability 1.

Proof Let us consider a δ-*typical set* of doublet probability distributions defined as follows:

$$\Gamma_\delta(P_\theta) \stackrel{\triangle}{=} \{P \in \Delta^{|\mathcal{X}|^2-1} : D_c(P||P_\theta) \leqslant \delta\}.$$

The probability that a sequence $\mathbf{X}^n$ leads to a *non-typical* second order type can be bounded as follows:

$$\begin{aligned}
\Pr(T^{(2)}_{\mathbf{X}^n} \notin \Gamma_\delta(P_\theta)) &= \sum_{P \in \mathcal{T}(n,2) \cap \Gamma_\delta^C} \Pr(T^{(2)}_{\mathbf{X}^n} = P|\theta) \\
&\leqslant \sum_{P \in \mathcal{T}(n,2) \cap \Gamma_\delta^C} \frac{|\mathcal{X}|}{\beta} 2^{-n D_c(P||P_\theta)} \qquad (183) \\
&\leqslant (n+1)^{|\mathcal{X}|^2} 2^{-n\delta} \qquad (184) \\
&= 2^{-n(\delta - |\mathcal{X}|^2 \log \frac{n+1}{n})}.
\end{aligned}$$

where (183) and (184) follows from Lemma D.2.1 3) and Theorem 20. Accordingly

$$\begin{aligned}
\Pr\left(D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) \leqslant \delta|\theta\right) &= 1 - \Pr(T^{(2)}_{\mathbf{X}^n} \notin \Gamma_\delta(P_\theta)) \\
&\geqslant 1 - 2^{-n(\delta - |\mathcal{X}|^2 \log \frac{n+1}{n})}.
\end{aligned}$$

Moreover, summing over n

$$\sum_{n=1}^{\infty} \Pr\left(D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) > \delta|\theta\right) < \infty.$$

Thus, applying the Borel-Cantelli lemma [127, Th. 3.4.2] to the event $\{D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) > \delta\}$ follows that

$$\Pr\left(D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) > \delta \text{ i.o.}|\theta\right) = 0.$$

Since this holds for every $\delta > 0$ under $\Pr(\cdot|\theta)$, as $n \to +\infty$, $D_c(T^{(2)}_{\mathbf{X}^n}||P_\theta) \to 0$ with probability 1.

□

Proof of Theorem 11: The proof follows the same arguments as the proof of Theorem 7, *mutatis mutandis*. □

D.3.1 *Example 2: mixture of binomial distributions*

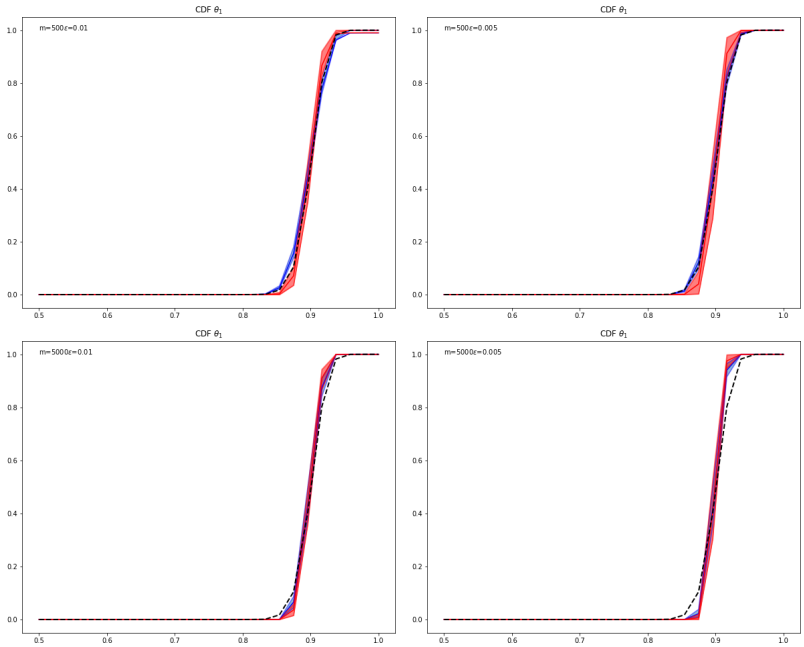


Figure 21: Posterior cumulative density functions for θ_2. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cumulative density function for a pair (m, ϵ). For $\theta_1 < 0.5$ both the cumulative density functions equal to 0. The 90% intervals over 100 rerun of each algorithm are also shown.

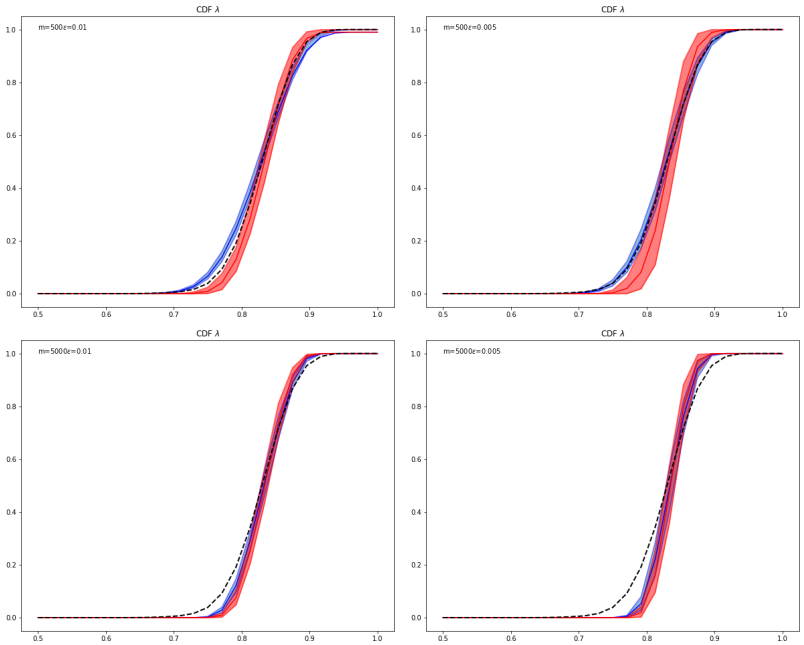


Figure 22: Posterior cumulative density functions for λ. Each plot shows in blue the output of LD-ABC, in red the output of R-ABC and in black the true cumulative density function for a pair (m, ϵ). For $\lambda > 0.5$ both the cumulative density functions equal 1. The 90% intervals over 100 rerun of each algorithm are also shown.

E

APPENDIX TO PART III

E.1 PROOFS

Proof of Lemma 8.3.1: We first characterize the probability $f(V = j|R_V = r_v, \boldsymbol{x}^*)$, for an arbitrary $j \in \{1, ..., N\}$. Bayes theorem yields

$$\begin{aligned} f(V = j|R_V = r_v, \boldsymbol{x}^*) &\propto f(R_V = r_v|V = j, \boldsymbol{x}^*) f(V = j|\boldsymbol{x}^*) \\ &= f(R_j = r_v|V = j, \boldsymbol{x}^*) f(V = j|\boldsymbol{x}^*) \\ &\propto f(R_j = r_v|V = j, \boldsymbol{x}^*) \quad (185) \\ &= f(R_j = r_v|\boldsymbol{x}^*) \quad (186) \end{aligned}$$

where (185) follows from $f(V = j|\boldsymbol{x}^*) = f(V = j) = 1/N$ (independence of V), and (186) follows because, as easily checked, for any fixed j, independence of R_j and V is preserved by conditioning on $\boldsymbol{x}^*$. Now we have, for every $s \in \mathcal{S}$

$$\begin{aligned} &p_A(s|r_v, \boldsymbol{x}^*) = \quad (187) \\ &= f(S_V = s \mid R_V = r_v, \boldsymbol{x}^*) \\ &= \sum_j f(S_V = s, V = j|R_V = r_v, \boldsymbol{x}^*) \\ &= \sum_j f(S_V = s|V = j, R_V = r_v, \boldsymbol{x}^*) f(V = j|R_V = r_v, \boldsymbol{x}^*) \\ &= \sum_j f(S_j = s|V = j, R_j = r_v, \boldsymbol{x}^*) f(V = j|R_V = r_v, \boldsymbol{x}^*) \\ &= \sum_{j\,:\,s_j = s} f(S_j = s|V = j, R_j = r_v, \boldsymbol{x}^*) f(V = j|R_V = r_v, \boldsymbol{x}^*) \quad (188) \\ &= \sum_{j\,:\,s_j = s} f(V = j|R_V = r_v, \boldsymbol{x}^*) \quad (189) \\ &\propto \sum_{j\,:\,s_j = s} f(R_j = r_v|\boldsymbol{x}^*)\,. \quad (190) \end{aligned}$$

where (188) and (189) follow from the fact that, for $s_j \neq s$, $f(S_j = s, \boldsymbol{x}^*) = 0$, while for $s_j = s$ obviously $f(S_j = s|V = j, R_j = r_v, \boldsymbol{x}^*) = 1$. Finally, (190) follows from (186).

Note that in (190) each term on the RHS actually is the joint probability $f(R_j = r_v, S_j = s|\boldsymbol{x}^*)$, being $s_j = s$ embedded in the range of the summation. □

E.2 AN ALTERNATIVE GROUP SAMPLING METHOD FOR VERTICAL SCHEMES

We consider the following method for sampling $g \in \mathcal{G}_i$. Draw n values r_{i_ℓ}, $\ell = 1, ..., n$, as follows:

1. draw r_{i_1} from l_i according to a distribution $\propto f(r|s_1, \theta)$;
2. draw r_{i_2} from $l_i \setminus \{\!|\, r_{i_1} \,|\!\}$ according to a distribution $\propto f(r|s_2, \theta)$;

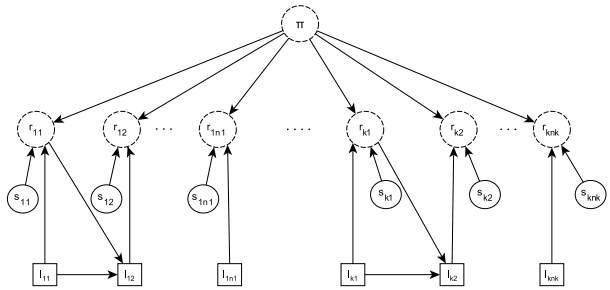


Figure 23: Sampling from $\psi(g|\theta, x^*)$ for vertical schemes.

...

n. draw r_{i_n} from $l_i \setminus \{\!\{ r_{i_1}, ..., r_{i_{n-1}} \}\!\}$ according to a distribution $\propto f(r|s_n, \theta)$.

For a multiset l', let $\sigma(l'|s_\ell, \theta) \stackrel{\triangle}{=} \sum_{r \text{ in } l'} f(r|s_\ell, \theta)$ denote the probability of extracting some element appearing in l' (disregarding multiplicities) according to $f(\cdot|s_\ell, \theta)$. Using this notation, the probability of returning exactly the sequence $r_{i_1}, ..., r_{i_n}$, hence $g = (s_1, r_{i_1}), ..., (s_n, r_{i_n}) \in \mathcal{G}_i$, as a result of the above n drawings, can be written as

$$\begin{aligned} \psi(g|\theta, x^*) &\stackrel{\triangle}{=} \frac{f(r_{i_1}|s_1, \theta)}{\sigma(l_i|s_1, \theta)} \cdot \frac{f(r_{i_2}|s_2, \theta)}{\sigma(l_i \setminus \{\!\{ r_{i_1} \}\!\}|s_2, \theta)} \cdots \frac{f(r_{i_n}|s_n, \theta)}{f(r_{i_n}|s_n, \theta)} \\ &= \frac{\prod_{\ell=1}^{n} f(r_{i_\ell}|s_\ell, \theta)}{\nu(g|\theta)} \end{aligned}$$

where we denote by $\nu(g|\theta)$ the denominator of the expression on the RHS of $\stackrel{\triangle}{=}$ above. The sampling process of $\psi(g|\theta, x^*)$ for vertical schemes across all the groups of the table is illustrated in Figure 23. We note that $\psi(g|\theta, x^*)$ is dependent on the chosen ordering of the sensitive values $s_1, ..., s_n$, which may invalidate condition (151). A possible solution could be to sweep the order of sampling according to the Random scan Gibbs sampler scheme described in Section 2.3.3.

E.3 ADDITIONAL RESULTS FROM EXPERIMENTS

Table 18: Posterior means via MCMC. Each column corresponds to the vector of posterior means for an element of $\{\theta_{R|s} : s \in \{\text{Governament, Self-employed, Private, Without-pay}\}\}$.

		Posterior Means			
		Gov.	Self-emp	Private	Without-pay
White	R	0.3991	0.3854	0.3859	0.2507
	MCMC	0.249	0.2494	0.2505	0.2501
	LD	0.3909	0.3774	0.3805	0.2389
Asian-Pac-Islander	R	0.1968	0.2015	0.1918	0.2507
	MCMC	0.2512	0.2495	0.2501	0.2501
	LD	0.1999	0.2041	0.1938	0.2530
Black	R	0.2428	0.2375	0.2527	0.2486
	MCMC	0.2502	0.2519	0.2492	0.2496
	LD	0.2438	0.2389	0.2505	0.2492
Other	R	0.1613	0.1756	0.1696	0.2500
	MCMC	0.2496	0.2492	0.2501	0.2502
	LD	0.1654	0.1796	0.1727	0.2438

ACKNOWLEDGEMENTS

I cannot thank enough Prof. Corradi for his valuable guidance. I will always be grateful to him for the interest, dedication, and generosity with which he supervised me. I feel very lucky to have met him.
I am also grateful to Prof. Boreale, for whom I have sincere esteem and admiration and without whom this thesis would have taken on a completely different form.
I thank Prof. Mira for inviting me to Università della Svizzera Italiana (USI), for giving me several opportunities for growth, but above all for the valuable suggestions and stimulating conversations.
I also thank Prof. Nardi for introducing me to large deviations theory and for being an example of dedication to teaching.
I would like to thank the two reviewers, Prof. Ventura and Prof. Tancredi, whose comments and remarks improved the presentation of the thesis as well as gave me new ideas for future research.
I express my gratitude to Prof. Mealli and Prof. Marchetti for their guidance beyond the support in technical stuff. I am grateful also to Prof. Baccini, Giulia Cereda and all the professors, researchers and students of the Department "G. Parenti" with whom I have interacted in these years, first as MS student and then as PhD student. Here, with them and thanks to them, was born my love for Statistics, for research and for sharing knowledge .
Thanks to Fiammetta and Giulia for all the moments, the anxieties and the fears experienced together. Thanks also (maybe above all) for the excellent dinners, the good wine and the trips to Chianti region.
I would also like to thank Giammarco, Giuseppe and all the other PhD students I met during these years for making me feel always in good company. In particular, I thank Matteo and Andrea with whom, in some sense, my adventure started. I hope to continue seeing the whole world as a Bayesian hierarchical model together with you.

I also need to thank all those who have been part of my life during these years: putting up with a PhD student in crisis is not an easy task!
Obviously thanks to Roberta, my little big supportive and irreplaceable friend.
Thanks to Monica, who cannot imagine how much of her there is in this thesis: by teaching me dance she taught me the value of research.
Thanks to my lifelong friends and to all those who have discreetly entered my life, lending me a hand or giving me a smile when needed.
I thank Pietro, because in everything I have done there is and there will always be a part of us.
Thank you to my entire family for encouraging me and making me feel appreciated. To my parents for the freedom (yes, but supervised!) and trust they have always given me. To my sister Anna for saying the right words to me at the right time.
Finally, thanks to my Grandpa Lino for teaching me the importance of self-sacrifice and the sense of duty: thanks to that I am here. This thesis is dedicated to him. All the sacrifices I made and those I will make are dedicated to him.

Non ci sono parole per ringraziare il Prof. Corradi per la sua preziosa guida. Gli sarò per sempre profondamente grata per l'interesse, la dedizione e la generosità con cui mi ha seguita. Mi ritengo davvero fortunata ad averlo incontrato.
Esprimo la mia gratitudine anche al Prof. Boreale, per cui nutro sincera stima e profonda ammirazione e senza il quale questa tesi avrebbe assunto tutt'altra forma.
Ringrazio la Prof.ssa Mira per avermi ospitato presso l'Università della Svizzera Italiana, per avermi offerto diverse occasioni di crescita ma soprattutto per gli utili suggerimenti e le stimolanti conversazioni.
Ringrazio anche la Prof.ssa Nardi per avermi introdotto nel mondo della teoria delle grandi deviazioni e per essere stata un esempio di dedizione all'insegnamento.
Inoltre, ringrazio i due revisori, la Prof.ssa Ventura ed il Prof. Tancredi, i cui commenti e le cui osservazioni hanno migliorato la presentazione della tesi, oltre ad avermi dato spunti per la ricerca futura.
Esprimo la mia gratitudine alla Prof.ssa Mealli e al Prof. Marchetti per la loro guida, oltre che per l'aiuto nella gestione delle cose tecniche e organizzative.
Sono grata anche alla Prof.ssa Baccini, a Giulia Cereda e a tutti i professori, ricercatori e studenti del Dipartimento "G. Parenti" con cui ho interagito in questi anni, prima da studentessa e poi da dottoranda. E' qui, con loro e grazie a loro, che è nato l'amore per la statistica, per la ricerca e per la condivisione.
Grazie a Fiammetta e Giulia per tutti i momenti, le ansie, le paure e le insicurezze vissute insieme. Grazie anche (forse soprattutto) per le ottime cene, il buon vino e le gite nel Chianti.
Inoltre, ringrazio Giammarco, Giuseppe e tutti gli altri dottorandi conosciuti in questi anni per avermi fatto sentire sempre in buona compagnia. In particolare, ringrazio Matteo ed Andrea con cui, in un certo senso, la mia avventura ha avuto inizio. Mi auguro di poter continuare a ricondurre tutto il mondo ad un modello gerarchico Bayesiano insieme a voi!
Sento di dover ringraziare anche tutti coloro che hanno fatto parte della mia vita in questi anni: sopportare una dottoranda in crisi non è un'impresa facile!
Ovviamente grazie a Roberta, piccolo grande sostegno, presenza costante e amica insostituibile.
Ringrazio Monica, che non immagina quanto di lei ci sia in questa tesi: insegnandomi la danza mi ha insegnato il valore della ricerca.
Grazie agli amici di sempre e a tutti quelli che sono entrati con discrezione nella mia vita, tendendomi una mano o strappandomi un sorriso quando ce ne è stato bisogno.
Ringrazio Pietro, perché in tutto quello che ho fatto c'è e ci sarà sempre una parte di noi.
Grazie a tutta la mia famiglia per avermi incoraggiata e fatto sentire apprezzata. Ai miei genitori per la libertà (si, ma vigilata!) e la fiducia che mi hanno sempre dato. A mia sorella Anna per avermi detto le parole giuste al momento giusto.
Infine, grazie a mio Nonno Lino per avermi trasmesso lo spirito di sacrificio ed il senso del dovere: è soprattutto grazie a quello che mi trovo qui. A lui è dedicata questa tesi. A lui sono dedicati tutti i sacrifici che ho fatto e quelli che farò.

Firenze, 15 Febbraio 2021

Cecilia Viscardi

BIBLIOGRAPHY

[1] Jean-François Angers, Atanu Biswas, and Raju Maiti. "Bayesian forecasting for time series of categorical data." In: *Journal of Forecasting* 36.3 (2017), pp. 217–229.

[2] Jhon Atchison and Sheng M. Shen. "Logistic-normal distributions: Some properties and uses." In: *Biometrika* 67.2 (1980), pp. 261–272.

[3] David J. Balding and Peter Donnelly. "Inference in forensic identification." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 158.1 (1995), pp. 21–40.

[4] Chris P. Barnes et al. "Considerate approaches to constructing summary statistics for ABC model selection." In: *Statistics and Computing* 22.6 (2012), pp. 1181–1197.

[5] Mark A. Beaumont. "Approximate Bayesian computation in evolution and ecology." In: *Annual review of ecology, evolution, and systematics* 41 (2010), pp. 379–406.

[6] Mark A. Beaumont, Wenyang Zhang, and David J. Balding. "Approximate Bayesian computation in population genetics." In: *Genetics* 162.4 (2002), pp. 2025–2035.

[7] Mark A. Beaumont et al. "Adaptive approximate Bayesian computation." In: *Biometrika* 96.4 (2009), pp. 983–990.

[8] Espen Bernton et al. "Approximate Bayesian computation with the Wasserstein distance." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.2 (2019), pp. 235–269.

[9] Michael Bewong et al. "A relative privacy model for effective privacy preservation in transactional data." In: *Concurrency and Computation: Practice and Experience* 31.23 (2019), e4923.

[10] Benjamin Bichsel et al. "Dp-finder: Finding differential privacy violations by sampling and optimization." In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 508–524.

[11] Michael GB Blum et al. "A comparative review of dimension reduction methods in approximate Bayesian computation." In: *Statistical Science* 28.2 (2013), pp. 189–208.

[12] Michele Boreale and Michela Paolini. "Worst-and average-case privacy breaches in randomization mechanisms." In: *Theoretical Computer Science* 597 (2015), pp. 40–61.

[13] Luke Bornn et al. "The use of a single pseudo-sample in approximate Bayesian computation." In: *Statistics and Computing* 27.3 (2017), pp. 583–590.

[14] Luis B. Boza. "Asymptotically optimal tests for finite Markov chains." In: *The Annals of Mathematical Statistics* (1971), pp. 1992–2007.

[15] Alessandra R. Brazzale, Anthony C. Davison, Nancy Reid, et al. *Applied asymptotics: case studies in small-sample statistics*. Vol. 23. Cambridge University Press, 2007.

[16] Steve Brooks et al. *Handbook of markov chain monte carlo*. CRC press, 2011.

[17] Erkan O. Buzbas and Noah A. Rosenberg. "AABC: approximate approximate Bayesian computation for inference in population-genetic models." In: *Theoretical population biology* 99 (2015), pp. 31–42.

[18] Stefano Cabras, Maria Eugenia Castellanos, and Erlis Ruli. "A Quasi likelihood approximation of posterior distributions for likelihood-intractable complex models." In: *Metron* 72.2 (2014), pp. 153–167.

[19] Olivier Cappé et al. "Adaptive importance sampling in general mixture classes." In: *Statistics and Computing* 18.4 (2008), pp. 447–459.

[20] David Cavallini and Fabio Corradi. "Forensic identification of relatives of individuals included in a database of DNA profiles." In: *Biometrika* 93.3 (2006), pp. 525–536.

[21] Anne-Sophie Charest. "How can we analyze differentially-private synthetic datasets?" In: *Journal of Privacy and Confidentiality* 2.2 (2011).

[22] Anne-Sophie Charest. "Empirical evaluation of statistical inference from differentially-private contingency tables." In: *International Conference on Privacy in Statistical Databases*. Springer. 2012, pp. 257–272.

[23] Yuguo Chen. "Another look at rejection sampling through importance sampling." In: *Statistics & probability letters* 72.4 (2005), pp. 277–283.

[24] Manuel Chiachio et al. "Approximate Bayesian computation by subset simulation." In: *SIAM Journal on Scientific Computing* 36.3 (2014), A1339–A1358.

[25] Chris Clifton and Tamir Tassa. "On syntactic anonymity and differential privacy." In: *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*. IEEE. 2013, pp. 88–93.

[26] Fabio Corradi et al. "Probabilistic classification of age by third molar development: the use of soft evidence." In: *Journal of Forensic Sciences* 58.1 (2013), pp. 51–59.

[27] Thomas M. Cover. *Elements of information theory*. John Wiley & Sons, 2006.

[28] Mary Kathryn Cowles and Bradley P. Carlin. "Markov chain Monte Carlo convergence diagnostics: a comparative review." In: *Journal of the American Statistical Association* 91.434 (1996), pp. 883–904.

[29] David R. Cox and Ole E. Barndorff-Nielsen. *Inference and asymptotics*. Vol. 52. CRC Press, 1994.

[30] David R. Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

[31] Imre Csiszár. "The method of types [information theory]." In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2505–2523.

[32] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Academic Press, 1981.

[33] Imre Csiszár and Paul C. Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[34] Fida K. Dankar and Khaled El Emam. "The application of differential privacy to health data." In: *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. 2012, pp. 158–166.

[35] Fida K. Dankar and Khaled El Emam. "Practicing differential privacy in health care: A review." In: *Transaction on Data Privacy* 6.1 (2013), pp. 35–67.

[36] Philip A. Dawid. "The island problem: coherent use of identification evidence." In: *Aspects of Uncertainty: A Tribute to DV Lindley,(ed. PR Freeman and AFM Smith)* (1994), pp. 159–70.

[37] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. "An adaptive sequential Monte Carlo method for approximate Bayesian computation." In: *Statistics and Computing* 22.5 (2012), pp. 1009–1020.

[38] Benjamin E. Deonovic and Brian J. Smith. "Convergence diagnostics for MCMC draws of a categorical variable." In: *arXiv preprint arXiv:1706.04919* (2017).

[39] Jean Diebolt and Christian P. Robert. "Estimation of finite mixture distributions through Bayesian sampling." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.2 (1994), pp. 363–375.

[40] Christos Dimitrakakis et al. "Robust, secure and private bayesian inference." In: *Arxiv, abs/1306.1066* (2013).

[41] Zeyu Ding et al. "Detecting violations of differential privacy." In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 475–489.

[42] Monroe D. Donsker and S.R. Srinivasa Varadhan. "Asymptotic evaluation of certain Markov process expectations for large time, I." In: *Communications on Pure and Applied Mathematics* 28.1 (1975), pp. 1–47.

[43] Christopher C. Drovandi, Anthony N. Pettitt, and Malcolm J. Faddy. "Approximate Bayesian computation using indirect inference." In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60.3 (2011), pp. 317–337.

[44] Christopher C. Drovandi, Anthony N. Pettitt, and Anthony Lee. "Bayesian indirect inference using a parametric auxiliary model." In: *Statistical Science* (2015), pp. 72–95.

[45] Cynthia Dwork. "Differential privacy: A survey of results." In: *International conference on theory and applications of models of computation*. Springer. 2008, pp. 1–19.

[46] Roger Eckhardt, Stan Ulam, and John Von Neumann. "The Monte Carlo method." In: *Los Alamos Science* 15 (1987), p. 131.

[47] Víctor Elvira, Luca Martino, and Christian P. Robert. "Rethinking the effective sample size." In: *arXiv preprint arXiv:1809.04129* (2018).

[48] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 419–474.

[49] Ronald Aylmer Fisher. *The genetical theory of natural selection*. The Clarendon Press, 1930.

[50] Robert G. Gallager. *Discrete stochastic processes*. Vol. 321. Springer Science & Business Media, 2012.

[51] Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. "Issues encountered deploying differential privacy." In: *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*. 2018, pp. 133–137.

[52] Alan E. Gelfand and Adrian F.M. Smith. "Sampling-based approaches to calculating marginal densities." In: *Journal of the American statistical association* 85.410 (1990), pp. 398–409.

[53] Andrew Gelman, Gareth O. Roberts, and Walter R. Gilks. "Efficient metropolis jumping rules." In: *Bayesian statistics* (1996).

[54] Andrew Gelman and Donald B. Rubin. "Inference from iterative simulation using multiple sequences." In: *Statistical science* 7.4 (1992), pp. 457–472.

[55] Andrew Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[56] Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[57] Alan Genz and Paul Joyce. "Computation of the normalization constant for exponentially weighted Dirichlet distribution integrals." In: *Computing Science and Statistics* 35 (2003), pp. 557–563.

[58] John Geweke. "Bayesian inference in econometric models using Monte Carlo integration." In: *Econometrica: Journal of the Econometric Society* (1989), pp. 1317–1339.

[59] John Geweke. "Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments." In: *Bayesian statistics* 4 (1992), pp. 641–649.

[60] Alexander Gleim and Christian Pigorsch. "Approximate Bayesian computation with indirect summary statistics." In: *Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers* (2013).

[61] Ruobin Gong. "Exact inference with approximate computation for differentially private data via perturbations." In: *arXiv preprint arXiv:1909.12237* (2019).

[62] Christian Gourieroux, Alain Monfort, and Eric Renault. "Indirect inference." In: *Journal of applied econometrics* 8.S1 (1993), S85–S118.

[63] Keith W. Hastings. "Monte Carlo sampling methods using Markov chains and their applications." In: *Biometrika* 57.1 (1970), pp. 97–109.

[64] Ali Inan, Murat Kantarcioglu, and Elisa Bertino. "Using anonymized data for classification." In: *2009 IEEE 25th International Conference on Data Engineering*. IEEE. 2009, pp. 429–440.

[65] Bai Jiang. "Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy." In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 1711–1721.

[66] Paul Joyce, Alan Genz, and Erkan O. Buzbas. "Efficient simulation and likelihood methods for non-neutral multi-allele models." In: *Journal of Computational Biology* 19.6 (2012), pp. 650–661.

[67] Paul Joyce and Paul Marjoram. "Approximately sufficient statistics and Bayesian computation." In: *Statistical applications in genetics and molecular biology* 7.1 (2008).

[68] George Karabatsos and Fabrizio Leisen. "An approximate likelihood perspective on ABC methods." In: *Statistics Surveys* 12 (2018), pp. 66–104.

[69] Robbert E. Kass, Luke Tierney, and Joseph B. Kadane. "The validity of posterior expansions based on Laplace's method." In: *Bayesian and likelihood methods in statistics and econometrics* 7 (1990), pp. 473–488.

[70] Robert E. Kass, Luke Tierney, and Joseph B. Kadane. "Approximate methods for assessing influence and sensitivity in Bayesian analysis." In: *Biometrika* 76.4 (1989), pp. 663–674.

[71] Robert E. Kass and Larry Wasserman. "The selection of prior distributions by formal rules." In: *Journal of the American statistical Association* 91.435 (1996), pp. 1343–1370.

[72] Ali Kassem et al. "Differential inference testing a practical approach to evaluate anonymized data." In: *[Research Report] INRIA.* (2018).

[73] Daniel Kifer. "Attacks on privacy and deFinetti's theorem." In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 2009, pp. 127–138.

[74] Daniel Kifer and Ashwin Machanavajjhala. "No free lunch in data privacy." In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 2011, pp. 193–204.

[75] Ronny Kohavi and Barry Becker. *UCI Machine Learning Repository: Adult Data Set*. URL: `http://archive.ics.uci.edu/ml`.

[76] Augustine Kong. "A note on importance sampling using standardized weights." In: *University of Chicago, Dept. of Statistics, Tech. Rep* 348 (1992).

[77] Dieter Kraft. *A software package for sequential quadratic programming*. Wiss. Berichtswesen d. DFVLR, 1988.

[78] Kenneth Lange. *Applied probability*. Springer Science & Business Media, 2010.

[79] Erich L. Lehmann and Howard J. D'Abrera. *Nonparametrics: statistical methods based on ranks.* Holden-day, 1975.

[80] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE. 2007, pp. 106–115.

[81] Ninghui Li, Wahbeh Qardaji, and Dong Su. "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy." In: *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*. 2012, pp. 32–33.

[82] Dennis V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, 1965.

[83] Jarno Lintusaari et al. "Fundamentals and recent developments in approximate Bayesian computation." In: *Systematic biology* 66.1 (2017), e66–e82.

[84] Jun S. Liu. *Metropolized Gibbs sampler: an improvement*. Tech. rep. Stanford University,Department of Statistics, 1996.

[85] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[86] David Luengo et al. "A survey of Monte Carlo methods for parameter estimation." In: *EURASIP Journal on Advances in Signal Processing* 2020.1 (2020), pp. 1–62.

[87] Ashwin Machanavajjhala et al. "l-diversity: Privacy beyond k-anonymity." In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 3–es.

[88] Ashwin Machanavajjhala et al. "Privacy: Theory meets practice on the map." In: *2008 IEEE 24th international conference on data engineering*. IEEE. 2008, pp. 277–286.

[89] Matthew D. Mailman et al. "The NCBI dbGaP database of genotypes and phenotypes." In: *Nature genetics* 39.10 (2007), pp. 1181–1186.

[90] Koray Mancuhan and Chris Clifton. "Statistical Learning Theory Approach for Data Classification with ℓ-diversity." In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM. 2017, pp. 651–659.

[91] Jean-Michel Marin, Kerrie Mengersen, and Christian P. Robert. "Bayesian modelling and inference on mixtures of distributions." In: *Handbook of statistics* 25 (2005), pp. 459–507.

[92] Jean-Michel Marin et al. "Approximate Bayesian computational methods." In: *Statistics and Computing* 22.6 (2012), pp. 1167–1180.

[93] Paul Marjoram et al. "Markov chain Monte Carlo without likelihoods." In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15324–15328.

[94] Andy W. Marshall. *The Use of Multi-stage Sampling Schemes in Monte Carlo Computations*. P (Rand Corporation). Rand Corporation, 1954. URL: `https://books.google.it/books?id=60kLYAAACAAJ`.

[95] Gael M. Martin et al. "Approximate Bayesian computation in state space models." In: *arXiv preprint arXiv:1409.8363* (2014).

[96] Kerrie L. Mengersen, Pierre Pudlo, and Christian P. Robert. "Bayesian computation via empirical likelihood." In: *Proceedings of the National Academy of Sciences* 110.4 (2013), pp. 1321–1326.

[97] Nicholas Metropolis and Stanislaw Ulam. "The monte carlo method." In: *Journal of the American statistical association* 44.247 (1949), pp. 335–341.

[98] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines." In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[99] Ilya Mironov. "Rényi differential privacy." In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE. 2017, pp. 263–275.

[100] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset). The University of Texas at Austin." In: *arXiv preprint cs* 610105 (2008).

[101] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Machine learning with membership privacy using adversarial regularization." In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 634–646.

[102] S. Natarajan. "Large deviations, hypotheses testing, and source coding for finite Markov chains." In: *IEEE Transactions on Information Theory* 31.3 (1985), pp. 360–365.

[103] John von Neumann. "Various Techniques Used in Connection with Random Digits." In: *Monte Carlo Method*. Ed. by A. S. Householder, G. E. Forsythe, and H. H. Germond. Vol. 12. National Bureau of Standards Applied Mathematics Series. Washington, DC: US Government Printing Office, 1951. Chap. 13, pp. 36–38.

[104] Frank Nielsen. "What is... an information projection." In: *Notices of the AMS* 65.3 (2018), pp. 321–324.

[105] Matthew A. Nunes and David J. Balding. "On optimal selection of summary statistics for approximate Bayesian computation." In: *Statistical applications in genetics and molecular biology* 9.1 (2010).

[106] William Ollier, Tim Sprosen, and Tim Peakman. "UK Biobank: from concept to reality." In: *Pharmacogenomics* 6.6 (2005), pp. 639–646.

[107] Mijung Park and Wittawat Jitkrittum. "ABCDP: Approximate Bayesian Computation Meets Differential Privacy." In: *arXiv preprint arXiv:1910.05103* (2019).

[108] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. "K2-ABC: Approximate Bayesian computation with kernel embeddings." In: *Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, 2016, pp. 398–407.

[109] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[110] Geoffrey G. S. Pegram. "An autoregressive model for multilag Markov chains." In: *Journal of Applied Probability* 17.2 (1980), pp. 350–362.

[111] Dennis Prangle. "Lazy abc." In: *Statistics and Computing* 26.1-2 (2016), pp. 171–185.

[112] Dennis Prangle, Richard G. Everitt, and Theodore Kypraios. "A rare event approach to high-dimensional approximate Bayesian computation." In: *Statistics and Computing* 28.4 (2018), pp. 819–834.

[113] Fabian Prasser and Florian Kohlmayer. "Putting statistical disclosure control into practice: The ARX data anonymization tool." In: *Medical Data Privacy Handbook*. Springer, 2015, pp. 111–148.

[114] Fabian Prasser, Florian Kohlmayer, and Klaus A. Kuhn. "A benchmark of globally-optimal anonymization methods for biomedical data." In: *2014 IEEE 27th international symposium on computer-based medical systems*. IEEE. 2014, pp. 66–71.

[115] Leah F. Price et al. "Bayesian synthetic likelihood." In: *Journal of Computational and Graphical Statistics* 27.1 (2018), pp. 1–11.

[116] Jonathan K. Pritchard et al. "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." In: *Molecular biology and evolution* 16.12 (1999), pp. 1791–1798.

[117] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. "Knock knock, who's there? Membership inference on aggregate location data." In: *arXiv preprint arXiv:1708.06145* (2017).

[118] Gong Qiyuan. *Anatomize*. URL: `https://github.com/qiyuangong/Anatomize`.

[119] Trivellore E. Raghunathan, Jerome P. Reiter, and Donald B. Rubin. "Multiple imputation for statistical disclosure limitation." In: *Journal of official statistics* 19.1 (2003), p. 1.

[120] Howard Raiffa and Robert Schlaifer. *Applied statistical decision theory*. Division of Research, Graduate School of Business Adminitration, Harvar, 1961.

[121] Louis Raynal et al. "ABC random forests for Bayesian parameter inference." In: *Bioinformatics* 35.10 (2019), pp. 1720–1728.

[122] Nancy Reid. "Asymptotics and the theory of inference." In: *The Annals of Statistics* 31.6 (2003), pp. 1695–2095.

[123] Nancy Reid and Y. Sun. "Assessing sensitivity to priors using higher order approximations." In: *Communications in Statistics—Theory and Methods* 39.8-9 (2010), pp. 1373–1386.

[124] John A. Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[125] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[126] Gareth O. Roberts and Adrian F.M. Smith. "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms." In: *Stochastic processes and their applications* 49.2 (1994), pp. 207–216.

[127] Jeffrey S. Rosenthal. *First Look At Rigorous Probability Theory, A*. World Scientific Publishing Company, 2006.

[128] Donald B. Rubin. "Inference and missing data." In: *Biometrika* 63.3 (1976), pp. 581–592.

[129] Donald B. Rubin. "Bayesianly justifiable and relevant frequency calculations for the applies statistician." In: *The Annals of Statistics* (1984), pp. 1151–1172.

[130] Donald B. Rubin. "Statistical disclosure limitation." In: *Journal of official Statistics* 9.2 (1993), pp. 461–468.

[131] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*. Vol. 10. John Wiley & Sons, 2016.

[132] Erlis Ruli, Nicola Sartori, and Laura Ventura. "Marginal posterior simulation via higher-order tail area approximations." In: *Bayesian Analysis* 9.1 (2014), pp. 129–146.

[133] Erlis Ruli, Nicola Sartori, and Laura Ventura. "Approximate Bayesian computation with composite score functions." In: *Statistics and Computing* 26.3 (2016), pp. 679–692.

[134] Erlis Ruli, Nicola Sartori, and Laura Ventura. "Robust approximate Bayesian inference." In: *Journal of Statistical Planning and Inference* 205 (2020), pp. 10–22.

[135] Rathindra Sarathy and Krishnamurty Muralidhar. "Evaluating Laplace noise addition to satisfy differential privacy for numeric data." In: *Transaction on Data Privacy* 4.1 (2011), pp. 1–17.

[136] Mohamed Sedki and Pierre Pudlo. "Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation." In: *Journal of the Royal Statistical Society: Series B* 74 (2012), pp. 466–467.

[137] Thomas A. Severini. *Likelihood methods in statistics*. Oxford University Press, 2000.

[138] Zhang Shijie. "The differential privacy of Bayesian inference." Bachelor's thesis. Harvard College, 2015. URL: `http://nrs.harvard.edu/urn-3:HUL.InstRepos:14398533`.

[139] Scott A. Sisson and Yanan Fan. *Likelihood-free MCMC*. Chapman & Hall/CRC, New York.[839], 2011.

[140] Scott A. Sisson, Yanan Fan, and Mark A. Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[141] Scott A. Sisson, Yanan Fan, and Mark M. Tanaka. "Sequential monte carlo without likelihoods." In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1760–1765.

[142] Klaas Slooten and Ronald Meester. "Forensic identification: the island problem and its generalisations." In: *Statistica Neerlandica* 65.2 (2011), 202–237.

[143] Adrian Smith. *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[144] Samuel Soubeyrand and Emilie Haon-Lasportes. "Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC." In: *Statistics & Probability Letters* 107 (2015), pp. 84–92.

[145] Samuel Soubeyrand et al. "Approximate Bayesian computation with functional statistics." In: *Statistical Applications in Genetics and Molecular Biology* 12.1 (2013), pp. 17–37.

[146] David S. Stoffer et al. "A Walsh—Fourier Analysis of the Effects of Moderate Maternal Alcohol Consumption on Neonatal Sleep-State Cycling." In: *Journal of the American Statistical Association* 83.404 (1988), pp. 954–963.

[147] Latanya Sweeney. "k-anonymity: A model for protecting privacy." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[148] Simon Tavaré et al. "Inferring coalescence times from DNA sequence data." In: *Genetics* 145.2 (1997), pp. 505–518.

[149] Luke Tierney. "Markov chains for exploring posterior distributions." In: *the Annals of Statistics* (1994), pp. 1701–1728.

[150] Luke Tierney and Joseph B. Kadane. "Accurate approximations for posterior moments and marginal densities." In: *Journal of the american statistical association* 81.393 (1986), pp. 82–86.

[151] Minh-Ngoc Tran et al. "Importance sampling squared for Bayesian inference in latent variable models." In: *arXiv preprint arXiv:1309.3339* (2013).

[152] Meltem Sönmez Turan et al. "Recommendation for the entropy sources used for random bit generation." In: *NIST Special Publication* 800.90B (2018).

[153] U.S. Office for Civil Rights. 2012. URL: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

[154] Cristiano Varin, Nancy Reid, and David Firth. "An overview of composite likelihood methods." In: *Statistica Sinica* (2011), pp. 5–42.

[155] Laura Ventura and Nancy Reid. "Approximate Bayesian computation with modified log-likelihood ratios." In: *Metron* 72.2 (2014), pp. 231–245.

[156] Mark Graham Moody Webster. "Convergence Properties of Approximate Bayesian Computation." PhD thesis. University of Leeds, 2016.

[157] Raymond Chi-Wing Wong et al. "Can the utility of anonymized data be used for privacy breaches?" In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5.3 (2011), pp. 1–24.

[158] Xiaokui Xiao and Yufei Tao. "Anatomy: Simple and effective privacy preservation." In: *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment. 2006, pp. 139–150.