Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics

Cycle: XXXIII

Disciplinary Field: SECS-S/01

# Models for the Natural History of Breast Cancer

Advisor: Marco Bonetti

PhD Thesis by

Laura Bondi

ID number: 3053727

**Academic Year: 2022**

# Acknowledgements

I would first like to express my sincere gratitude to my advisor Prof. Marco Bonetti. He devoted an incredible amount of time to constantly guiding me through this research work, always showing his patience, enthusiasm and great care for the details, in addition to continuously motivating me to do my best.

A big thank goes to the entire faculty of the PhD in Statistics at Bocconi University for introducing me to so many different topics in statistics, and in particular to Prof. Raffaella Piccarreta for contributing with her ideas to one of the chapters of this thesis. I need to thank also Denitsa Grigorova and Prof. Marcello Pagano for their remarkable contributions to parts of this thesis.

This PhD journey would certainly not have been the same without sharing it with the many wonderful colleagues and friends I had the luck to meet. I thank all of them for making these four years in Milan so memorable and for truly supporting and helping me whenever I needed.

Lastly and most importantly, I want to express my deep gratitude to all my loved ones for constantly encouraging me and for making me believe in my own abilities. Your unconditional trust was an indispensable source of energy and love.

# Contents

**Introduction** **1**

**1 Statistical models for the natural history of breast cancer, with application to data from a Milan cohort study** **5**

1.1 Introduction . . . . . 5

1.2 The motivating data . . . . . 10

1.3 A first model: observed data likelihood . . . . . 14

1.3.1 Model specification and results . . . . . 20

1.4 More flexible models . . . . . 22

1.4.1 Approximate Bayesian Computation . . . . . 22

1.4.2 Models . . . . . 26

1.4.3 Model selection and results . . . . . 30

1.4.4 Comparing alternative screening strategies . . . . . 37

1.5 Discussion . . . . . 41

1.6 Additional analyses and results . . . . . 45

1.6.1 Observed data likelihood for the model in Section 1.3.1 . . . . . 45

1.6.2 ABC vs MLE . . . . . 47

1.6.3 ABC estimation for the "Gamma + piecewise Exponential" model . . 48

1.6.4 Results under different data assumptions . . . . . 52

1.6.5 Sensitivity analysis for the ABC metric . . . . . 55
1.6.6 Distribution of the observed age at detection within a screening program: an analytical example . . . . . 55

**2 An exploration of ABC and dissimilarities 69**

2.1 Introduction . . . . . 69
2.2 Dissimilarity-based criteria . . . . . 74
2.3 A new ABC-inspired estimator . . . . . 78
2.4 The bivariate normal model . . . . . 80
2.5 An example with discrete data and different dissimilarities . . . . . 87
2.6 Discussion . . . . . 89

**3 Optimal estimation of the sparsity index in Poisson size-biased sampling 95**

3.1 Introduction . . . . . 95
3.2 Estimating $\lambda$ by maximum likelihood . . . . . 98
3.3 Estimation of the sparsity parameter $\mu$ . . . . . 100
3.3.1 Maximum likelihood estimation of $\mu$ . . . . . 100
3.4 Computation of the UMVUE of $\mu$ ($T_3$) . . . . . 103
3.4.1 A first exact algorithm to compute $T_3$ . . . . . 103
3.4.2 An improved exact algorithm to compute $T_3$ . . . . . 104
3.4.3 An efficient approximate algorithm to compute $T_3$ using the characteristic function . . . . . 106
3.4.4 Constructing confidence intervals for $\mu$ from $T_3$ . . . . . 108
3.5 Simulation study: MLE vs UMVUE for $\mu$ . . . . . 109
3.6 Discussion . . . . . 112
3.7 Tables of Rao-Blackwellized estimates for the sparsity index . . . . . 115

**Conclusions** **119**

**Bibliography** **121**

# List of Figures

1.1 Random effects meta-analysis of breast cancer mortality after 13 years of follow-up in breast cancer screening trials. . . . . 7

1.2 A graphical representation of the natural history from onset until detectability of the disease. . . . . 16

1.3 Prior and posterior probabilities of the nine models (global retention rate = 0.005). . . . . 31

1.4 Prior (red dashed line) and local linear regression adjusted approximate posterior (blue histogram and solid line) densities for each parameter of the "Rescaled Beta + piecewise Exponential" model. . . . . 34

1.5 Approximate posterior distribution of the mean age at asymptomatic detectability $\mu(x)$ and of the susceptible proportion $\mathrm{p}(x)$ across covariate groups. . . . . 36

1.6 Predictive distributions for $T_A$ in each covariate group and for $\Delta$ given the observed value of $T_A$. . . . . 38

1.7 Prior (red dashed line) and unadjusted posterior (blue histogram and solid line) densities for each parameter of the "Normal + Exponential" model without covariates. The green vertical lines represent the MLEs, together with their 95% confidence intervals (green areas). . . . . 48

1.8 Prior (red dashed line) and local-linear-regression adjusted posterior (blue histogram and solid line) densities for each parameter of the "Normal + Exponential" model without covariates. The green vertical lines represent the MLEs, together with their 95% confidence intervals (green areas). . . . . 48

1.9 Prior (red dashed line) and local linear regression adjusted approximate posterior (blue histogram and solid line) densities for each parameter of the "Gamma + piecewise Exponential" model. . . . . 49

1.10 Approximate posterior distribution of the mean age at asymptomatic detectability $\mu(x)$ and of the susceptible proportion $\mathrm{p}(x)$, across the eight covariate groups, for the "Gamma + piecewise Exponential" model. . . . . 51

1.11 Predictive distributions for $T_A$ in each covariate group and for $\Delta$ given the observed value of $T_A$, for the "Gamma + piecewise Exponential" model. . . . 52

1.12 Prior (red dashed line) and local linear regression adjusted posterior (blue histogram and solid line) densities for each parameter of the "Gamma + piecewise Exponential" model. . . . . 54

1.13 Predictive distributions for $T_A$ in each covariate group and for $\Delta$ given the observed value of $T_A$. . . . . 54

1.14 True values (red vertical lines), prior (black dashed lines), and local linear regression adjusted posterior (blue histograms and blue solid lines) densities obtained using Metric 1 and the "Gamma + piecewise Exponential" model. . 56

1.15 Sample of observed values $T_S$ in a simulated sample from 10 million initial subjects. Parameter values were $p_A = 0.15$, $\mu_A = 65$, $\sigma_A^2 = 100$, $\lambda = 1/3$, $\mu_D = 80$, $\sigma_D^2 = 25$, $gap = 3$, and $Bmin = -50$. The red curve represents the density function in (1.5) 61

1.16 Sample of observed values $\widetilde{T}_A$ in a simulated sample from 10 million initial subjects. Parameter values were $p_A = 0.15$, $\mu_A = 65$, $\sigma_A^2 = 100$, $\lambda = 1/3$, $\mu_D = 80$, $\sigma_D^2 = 25$, $gap = 3$ and $Bmin = -50$. The red curve represents the density function in (1.13). 66

2.1 Distances between approximated posterior distributions obtained from ABC with a sequence of summary statistics having increasing fraction of information. 72

2.2 Samples from the prior distribution of the two mean components, and Wilcoxon-Mosler distance (in log-scale) of each generated dataset from the observed data. The blue curve is the estimated conditional quantile of order $\tau = 0.1$. . . . . 82

2.3 Posterior distributions for the bivariate normal model with known covariance matrix. . . . . 84

2.4 Case $n = 1000$. Prior samples for the two mean components and Wilcoxon-Mosler distance (in log-scale) of each generated dataset from the observed data. The blue curve is the estimated conditional quantile of order $\tau = 0.1$. . . . . 85

2.5 Results for the bivariate normal model with unknown variances. . . . . 86

2.6 Bivariate sample from the prior distributions of $(\mu_1, \mu_2)$ and $(\sigma_1, \sigma_2)$ colored according to the Wilcoxon-Mosler distance of each corresponding generated dataset from the observed data. The red dots indicate the true parameter value. 88

2.7 Results from the Euclidean distance. Plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5). . . . . 90

2.8 Results from the Manhattan distance. Plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5). . . . . 91

2.9 Results from the rough 0/1/2 distance. Plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5). . . . . 92

3.1 Probability mass function of S for $\lambda = 2$ and several values of $n$. The black points represent the distribution approximated by ifft and the red ones the empirical distribution computed from a sample of size $k = 30,000$. . . . . 107
3.2 Biases (a) and MSEs (b) of the two estimators for $\mu$ as functions of the sample size $n$, when $\lambda = 3$ ($\mu = 0.25$). . . . . 116

# List of Tables


1.1 Descriptive statistics of the data. Time is measured from birth (in years). *There were 2845 subjects with one or more of these covariates missing. . . . . 13

1.2 Observed outcomes in each covariate group and in the total sample. Ages are measured in years. $X_1$= at least one birth (0:No, 1:Yes); $X_2$=Education level (0:Low/Medium, 1:High); $X_3$=Family history of cancer (0:No, 1:Yes). . . . . 15

1.3 MLEs and 95% confidence intervals for the model parameters. Time is measured in years. . . . . 21

1.4 Posterior probabilities of the nine models (global retention rate = 0.005). . . 31

1.5 Counts of votes for the nine models out of a total of 1000 trees composing the random forest. . . . . 32

1.6 Posterior modes and the 95% highest posterior density (HPD) intervals. . . . 35

1.7 $X_1$= At least one birth (0:No, 1:Yes); $X_2$= Education level (0:Low/Medium, 1:High); $X_3$= Family history of cancer (0:No, 1:Yes). . . . . 36

1.8 Observed summary statistics on a sample of size 100,000 generated from the estimated "Rescaled Beta + piecewise Exponential" model under several different screening strategies. The screening strategies are defined by the screening age range, the gap between subsequent exams, and the overall adherence proportion. . . . . 40

1.9 Posterior modes and 95% highest posterior density intervals (HPDI). . . . . 57

3.1 Absolute and relative bias of $T_1$ in estimating $E(Y) = 1 + \lambda$. . . . . 99
3.2 Mean and standard deviation for the three estimators for $\mu$ computed from 1,000 simulations performed using $\mu = 0.25$ ($\lambda = 3$) and several different values of $n$. . . . . 110
3.3 Coverage and average width (within brackets) of confidence intervals computed from $1,000$ simulations where the theoretical coverage $1 - \alpha$ was set to 0.95. 111
3.4 Rao-Blackwellized estimates of $\mu$ given the sample size $n$ and $s = \sum_{i=1}^{n} x_i$ computed by the exact algorithm. . . . . 115
3.5 Rao-Blackwellized estimates of $\mu$ given the sample size n and $s = \sum_{i=1}^{n} x_i$ computed by the approximated algorithm based on the inverse fast Fourier transform. . . . . 117

# Introduction

This PhD thesis is composed of three projects concerning different, yet related, topics.

Chapter 1, which corresponds to the first project, is deeply applied in scope, but at the same time it presents methodological challenges. Both classical and Bayesian statistics techniques are exploited to deal with the complex missing data structure of the problem.

In that chapter we develop multi-state models for the natural history of breast cancer, where the main events of interest are the age at the start of asymptomatic detectability of the disease (through screening), denoted by $T_A$, and the age at symptomatic detection (through symptoms), denoted by $T_S$. The time interval between these two events represents the latent phase of the tumor, which is called *sojourn time* and which we denote by $\Delta$.

The goal is to draw conclusions about quantities that are mostly unobservable. Indeed, both $T_A$ and $T_S$ are never observed on any woman. Clearly, the distribution of the observed ages at detection, asymptomatic or symptomatic, would not be a good estimate of the underlying disease history, and the proportion of observed diagnoses is very different from the probability of developing breast cancer for a woman in her lifetime.

The aim of the work is to reconstruct the underlying latent process through the probabilistic description of the occurrence and subsequent evolution of the disease. We develop several parametric specifications, which have in common a cure rate structure to take into account that only a fraction of the women experience breast cancer in their lifetime.

We present the results of the analysis of data collected as part of a motivating study

from Milan. Participants in the study had a varying degree of compliance to a regional breast cancer screening program. The subjects' ten-year trajectories have been obtained from administrative data collection performed by the Italian national health care system. Additional covariates were collected by means of questionnaires.

We present a tractable model for which we develop the likelihood contributions of the observed trajectories, and perform maximum likelihood inference on the latent process. As likelihood-based inference is not feasible for more flexible models, we rely on Approximate Bayesian Computation (ABC) for inference for more complex models, among which we perform model selection. Issues that arise from the use of ABC for model choice and parameter estimation are also discussed, including the problem of the selection of appropriate summary statistics when implementing ABC.

The estimated parameters of the underlying disease process allow for the study of the effect of different examination schedules (age range and frequency of screening examinations) on a population of asymptomatic subjects, in terms of number and kind of diagnoses.

In Chapter 2, we report on an exploratory work that was motivated by our first project. We focus on the use of dissimilarities among observations to define a measure of the distance between two datasets. This problem is clearly relevant in the context of ABC, where observed and model-generated data need to be compared. We consider simple models, where we can investigate the ability of the dissimilarity approach to recover the true parameter values. As part of this study, we propose a new likelihood-free estimation procedure. The new estimator is based on calibration ideas, and makes more complete use of the datasets that are routinely generated when one performs ABC inference.

Chapter 3 is devoted to a purely methodological study that concerns results and algorithms developed for computing the optimal estimator in a size-biased sampling problem.

Size-bias can occur in a variety of contexts, whenever the sampling unit is the individual and the population consists of clusters of individuals. For example, in the study of the family

history of cancer, larger families have higher probability to have at least one case of cancer and to be, therefore, included into the Cancer Registry.

In this chapter, we obtain the uniformly minimum variance unbiased estimator (UMVUE) for the *sparsity* index in size-biased Poisson sampling. We first propose two exact algorithms, based on the enumeration of cases, where the second algorithm is the refined and speeded up version of the first. Despite their exact nature, these algorithms become not feasible already for quite small sample sizes. As an alternative, a third, approximate, algorithm, based on the inverse fast Fourier transform, is also developed to compute the distribution of the UMVUE. An exact confidence interval based on the UMVUE is also built by inversion of the associated test. The performance of the estimation procedure is compared to classical maximum likelihood inference, in terms of mean squared errors of the two estimators, as well as with respect to the average coverage and width of the confidence intervals.

# Chapter 1

# Statistical models for the natural history of breast cancer, with application to data from a Milan cohort study[1]

## 1.1 Introduction

Cancer screening is defined as the examination of asymptomatic subjects in order to classify them as likely or unlikely to be diseased [63]. The expected positive aspects of screening are the reduction in mortality and the avoidance of advanced morbidity. However, along with the benefits there may be negative effects of screening such as overdiagnosis, overtreatment, and false positive results that may lead to psychological distress.

Three recent meta-analyses reviews [42], [25], [46] examined data from past randomized clinical trials comparing breast cancer mortality in the treated (screened) and control group.

[1] Joint work with Marco Bonetti, Denitsa Grigorova (Sofia University) and Antonio Russo (ATS Milano)

They consider eleven RCTs, which were carried out in the US, in Sweden and in the UK. The number of trials varies (eight or eleven) because three of them consist of two parts. The earliest trial started in 1963 (New York Health Insurance Plan) and the most recent one in 1991 (UK Age trial). Even if they are collecting information and results from the same trials, the three meta-analyses reach very different conclusions. A first difference arises in the way they classify trials depending on the level of the randomization procedure, which can be adequate or sub-optimal

According to [42], the benefit can be quantified as a 20% relative risk (RR) reduction in breast cancer mortality in women invited to participate in a 20-year screening program In Figure 1.1 (Figure 1 of [42], p. 1780) they show this result, together with the marginal estimates of each RCT. They also provide an estimate of overdiagnosis, which occurs roughly in the 19% of cases, when expressed as a proportion of the cancers diagnosed during the active screening period. From these results, they conclude that the UK breast screening programs, which invite women aged $50 - 70$ every three years, confer significant benefit and should continue.

In [46] the authors conclude that breast cancer mortality is generally reduced with mammography screening, although estimates are not statistically significant at all ages, and the magnitudes of the effect are small. Advanced cancer is reduced with screening for women aged 50 years or older.

The Cochrane review [25], instead, is very skeptical about the benefit of breast screening programs, stating that the only three RCTs with an adequate randomization do not show a significant relative risk reduction. This estimate becomes statistically significant only when trials conducted under a sub-optimal randomization procedure are included in the analysis.

While these three meta-analyses did not reach consistent conclusions, one could also be concerned about the current validity of results provided by very old trials (28-56 years ago). As pointed out in [25], "it is likely that the absolute effect of screening today is smaller

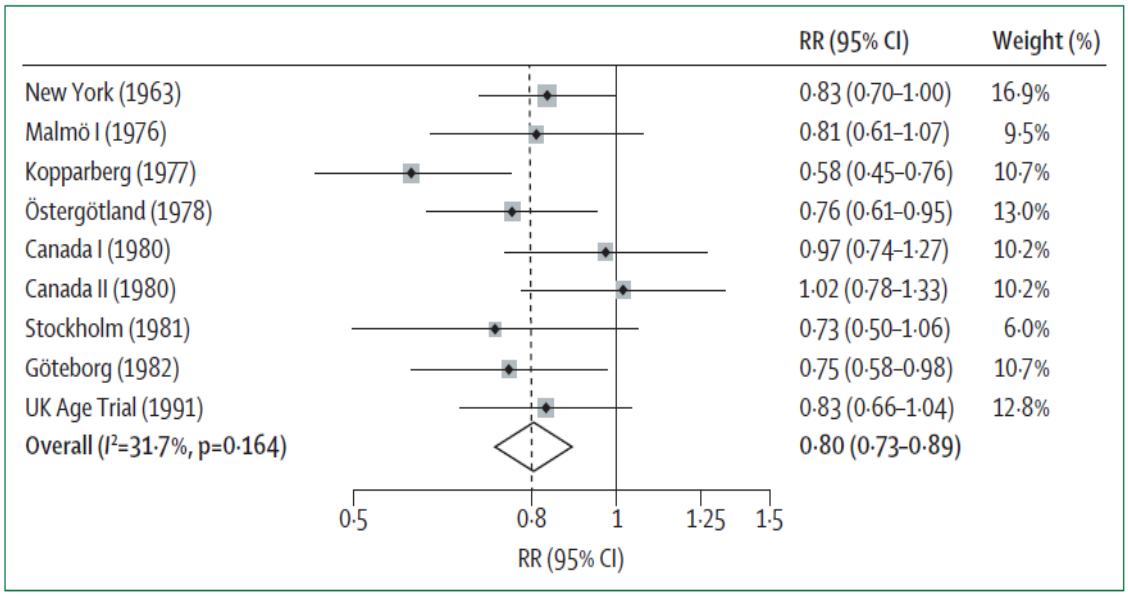



Figure 1.1: Random effects meta-analysis of breast cancer mortality after 13 years of follow-up in breast cancer screening trials.

than in the trials, because of substantial advances in treatment and greater breast cancer awareness."

Once a screening program is established in a country, it is difficult to conduct randomized trials to assess the effectiveness of screening. Sound, updated, and country-specific evidence is needed to decide whether to establish breast cancer screening programs and to identify the optimal screening policy in terms of the age range of the women invited, and the lag between successive examinations. As a consequence, there is a strong interest in learning about the natural history of the disease.

Hu and Zelen [30] discuss a theoretical model for planning screening trials in order to compare mortality rates between a control group and a screened group. The authors model the natural history of the disease and how the disease could be detected by regular screening examinations. The work is used for planning the National Lung Screening Trial.

More recent evidence about screening effectiveness has been obtained from observational

data, even though a higher risk of bias and confounding must be dealt with.

Six groups were commissioned by the U.S. Preventive Services Task Force to evaluate the benefits and harms of many possible breast cancer screening strategies and they were carried out by the Cancer Intervention and Surveillance Modeling Network (CISNET) of the National Cancer Institute [62]. Six universities collaborated to this project, proposing each one a model. They believe that, since all models make assumptions about unobservable events, it is appropriate to consider several models to provide a more comprehensive picture of the problem and to illustrate the effects of differences in model assumptions.

Between 2007 and 2015, a very accurate evaluation of the Norwegian Breast Cancer Screening Program (NBCSP) has been conducted by a scientific committee appointed by the Research Council of Norway [61], in order to investigate whether the program was fulfilling its intentions and purpose. They quantify the reduction in breast cancer mortality attributable to the implementation of the NBCSP, compared to a situation with no screening program, to be in the range 20-30% for women aged 50-79 years. They also focus on the estimation of overdiagnosis which, together with lead time effect, makes the incidence rates of breast cancer increase compared to a situation without screening and they propose several ways to estimate those quantities.

A commentary by Aalen [1] introduces a different class of models whose aim is to understand disease processes beyond the simple survival setting and integrating into the analysis all the information collected at each clinical examination.[59, 21] In Sweeting et al.[59], the authors implement multi-state Markov models to analyze the longitudinal disease progression when transition times between disease states are interval censored, and taking into account different assumptions on the possibly non-ignorable missing data process occurring during follow-up. This setting reflects the screening context in which, even though examination times are scheduled, subjects can decide not to attend them and the decision to adhere to the scheduled examinations is possibly not independent of the underlying disease status or of the (perceived

or real) risk of the subject. Similarly, Chen et al.[21] is concerned with the analysis of incomplete longitudinal data, where the observation process may contain information about the life history of the disease. They consider progressive multi-state Markov response models where the parameter estimation is performed by maximizing the likelihood function.

An alternative to multi-state Markov models consists in modeling the underlying biological tumour growth as a continuous process. Recent work [3, 31] proposes a continuous tumour growth model and derives theoretical results for jointly estimating tumour growth, time to symptomatic detection and mammography screening sensitivity as a function of mammographic density. These models evaluate mammography screening in terms of mortality, to estimate overdiagnosis, and to estimate the impact of lead-time bias when comparing survival times between screen-detected cancers and cancers found outside of the screening program. The models are fitted using likelihood-based estimation, but a recent work explores ways to move to a likelihood-free approach, consisting of calibrating the parameters via summary statistics at the population level [8].

The aim of this Chapter is to present a new class of statistical models to describe the natural history of breast cancer, focusing on the insurgence of the disease, and on the detection of cases as it progresses from asymptomatic to symptomatic.

In Section 1.2 we describe the motivating observational study conducted in Milan. While observational studies do not typically provide trusted evidence to answer the same questions as randomized trials do, here we reconstruct the underlying latent process through the probabilistic description of the occurrence and development of breast cancer.

All the models that we discuss can be seen as multi-state semi-Markov models, where the future evolution depends not only on the current state, but also on the entry time into that state. The estimation procedure that we employ depends on the complexity of the model. In principle, it is possible to compute the observed data likelihood[38] of each model, in order to find the maximum likelihood estimates for the parameters. However, the likelihood

calculation and maximization can be numerically complicated or infeasible, unless the model has a simple structure.

Section 1.3 describes the modeling approach, and describes one such simple model for which likelihood inference is feasible. In Section 1.4 we move to the Bayesian inferential framework and develop a likelihood-free estimation procedure based on Approximate Bayesian Computation (ABC)[57] that allows us to implement a variety of models and to perform both model selection and parameter estimation on the motivating data. In Section 1.5 we discuss the use of ABC in this modeling setting, and we close with some final remarks. Section 1.6 contains supplementary analyses and results about the models and the estimation procedures.

## 1.2 The motivating data

The data that motivated this study concern a cohort of $n = 78051$ women, aged between 41 and 76 years, resident in the municipality of Milan, who were invited to participate (with a varying degree of adherence) to the mammographic screening program and in particular to a study with the acronym of FRiCaM (Risk Factors for Breast Cancer: Fattori di Rischio per il Carcinoma della Mammella), supported by a specific grant of the Italian League of Cancer Prevention. Italy does not have a universal screening program for all regions in the country. However, currently all Italian regions have implemented screening programs. [64] Screening examinations in Milan are normally offered to women 50-74 years old every two years (recently extended from the previous 50-69 policy), but under specific circumstances high-risk women can also be included in the program earlier. All women had to be disease-free when they entered the study.

To collect data for the motivating study, a questionnaire was sent by mail or handed out to a total of 151246 eligible women who had received no diagnosis of breast cancer at the time of entry, and about 60% of them completed it and returned it at their upcoming screening

examination, or through postal delivery. The date when a woman filled out the questionnaire (which included the informed consent form) marked her date of entry into the study. Study entry dates range from January 1st, 2003 to December 31st, 2007.

The subjects' health trajectories were obtained from administrative data collected by the Italian National Health Service and from the Cancer Registry database. Follow-up ended when an invasive cancer diagnosis occurred or, for women without an observed diagnosis, when censoring occurred. The censoring date coincides with the earliest among date of administrative censoring (December 31st, 2016), date of cancellation from the study, date of emigration, and date of death. The median follow-up was 12.29 years.

The available data also include the date of birth, the timing of the screening examinations (either mammograms or ecographies, which we treat equally) that were performed, and the dates of the outside-screening examinations and of the diagnoses (invasive tumors only). Due to lack of permission to obtain such information, the data did not include the examination results, and we had to infer whether each examination likely gave a positive or negative result based on some assumptions. Different assumptions could lead to different conclusions, and our analysis were therefore repeated under several scenarios. Even when changing the assumptions on the examination outcomes and on the dates and kinds of detections, the results did not show considerable change.

Below we present the results obtained under what seemed to be the most plausible set of assumptions, also after discussion with an investigator who is familiar with the data. In section 1.6 we include results produced under one different set of assumptions.

For those women without an observed diagnosis of breast cancer, we assumed that all the examinations had given a negative result. For women having a breast cancer diagnosis recorded in the Cancer Registry, we had to determine whether the detection was symptomatic or asymptomatic, and to establish the date of the last negative examination before detection. A key piece of information was available from the variable which differentiated between

screening and non-screening examinations. Indeed, non-screening examinations may be due to suspicious symptoms. We first checked if there were any screening examinations within the six months prior to the diagnosis. If yes, then the last one before the diagnosis was assumed to have yielded a positive result, and to have led to an asymptomatic detection. In this case the date of detection was defined as the date of that positive exam.

If, instead, there were no screening exams within six months of the date of diagnosis, we classified that detection as symptomatic, and we set the date of detection equal to the date of the most recent non-screening exam, if there were any within the six months prior to diagnosis. If no exams at all were recorded in the six months prior to diagnosis, then we set the date of the symptomatic detection back by a number of days equal to the average shift applied to the symptomatic detections which had that information (42.6 days).

Once the dates of detection were defined, we picked the last negative exam as the most recent exam performed at least six months before the detection. We decided to impose a distance of at least six months between the last negative exam and the detection because most diagnoses are preceded by a few examinations very close to each other, and those were likely performed to confirm the presence of the tumor.

These limitations of the available data are such that the results of our analyses should be taken with some caution (for example, no sensitivity/specificity of the examinations can be taken into account). However, also given the large sample size, we feel that they still provide useful information in particular on the effect of covariates, and most importantly these analyses let one explore the issues that one must address when developing and estimating disease history models.

Out of the 78051 women in the sample, 3034 (3.89%) were diagnosed with invasive breast cancer during the observation period and 75017 (96.11%) were without diagnosis at the end of their follow-up. We do not consider DCIS (ductal carcinoma in situ) cases, which were not included in the Cancer Registry database. Under the assumptions described above, the

asymptomatic detections were 572 and the symptomatic ones 2462. The total number of exams was 396183, performed on 74345 women. The remaining 3706 women did not undergo any examination during the observation period. For additional descriptive statistics we refer to Table 1.1.

| | Min | Median | Mean | Max |
|---|---|---|---|---|
| Age at questionnaire | 41.30 | 60.91 | 60.82 | 76.85 |
| Age at first exam after entry | 41.37 | 61.02 | 60.80 | 84.64 |
| Age at asymptomatic detection | 45.05 | 64.93 | 64.18 | 76.23 |
| Age at symptomatic detection | 46.40 | 67.74 | 67.34 | 86.35 |
| # screening examinations | | Mammographies | Ecographies | Total |
| | 0 | 12215 (0.16) | 75694 (0.970) | 12213 (0.16) |
| | 1 | 21452 (0.28) | 2219 (0.028) | 21412 (0.27) |
| | 2 | 17679 (0.23) | 131 (0.002) | 17459 (0.22) |
| | 3 | 12612 (0.16) | 7 (0.000) | 12303 (0.16) |
| | $\geq 4$ | 14003 (0.18) | 0 | 14664 (0.19) |
| # outside-screening examinations | | Mammographies | Ecographies | Total |
| | 0 | 8812 (0.11) | 62609 (0.80) | 8256 (0.11) |
| | 1 | 10903 (0.14) | 7866 (0.10) | 10283 (0.13) |
| | 2 | 26900 (0.34) | 2699 (0.03) | 25063 (0.32) |
| | 3 | 18451 (0.24) | 1466 (0.02) | 17239 (0.22) |
| | $\geq 4$ | 12985 (0.17) | 3411 (0.04) | 17210 (0.22) |
| Breast cancer diagnoses | Yes | No | | |
| | 3034 (0.04) | 75017 (0.96) | | |
| Observed Follow-up | Median | Mean | Min | Max |
| | 12.29 | 11.66 | 0 | 13.93 |
| Status at end of follow-up | Alive | Cancelled | Dead | Emigrated |
| (only for non-diseased subjects) | 65494 (0.873) | 232 (0.003) | 7410 (0.099) | 1881 (0.025) |
| | No | Yes | Missing* | |
| At least one birth ($X_1$) | 11933 | 63935 | 2183 | |
| High level of education ($X_2$) | 47315 | 29994 | 742 | |
| Family history of cancer ($X_3$) | 47419 | 30562 | 70 | |

Table 1.1: Descriptive statistics of the data. Time is measured from birth (in years). *There were 2845 subjects with one or more of these covariates missing.

Additional variables, including level of education, comorbidities, family structure and family history of cancer, were collected by means of a questionnaire filled in by the participants. We focused on three dichotomous covariates, which divide the women in eight groups as shown

in Table 1.2: having had at least one child birth $X_1$; level of education $X_2$; and family history of cancer $X_3$. These covariates are defined as follows:

$$X_1 = \begin{cases} 0 & \text{no child birth,} \\ 1 & \text{at least one child birth;} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{education} < \text{high school diploma,} \\ 1 & \text{education} \geq \text{high school diploma;} \end{cases}$$

$$X_3 = \begin{cases} 0 & \text{no family history of cancer,} \\ 1 & \text{family history of cancer.} \end{cases}$$

Table 1.2 also shows the number of asymptomatic and symptomatic detections and the median age at detection, both in the total sample and within the eight covariate groups. Single imputation of missing values was performed on the three covariates by replacing them with draws from independent Bernoulli variables with parameters equal to the proportion of ones among the non-missing values for each variable.

Note that the three covariates were assessed at the time of entry into the motivating study. However, given the rather advanced age at entry, we may consider the first two as being definitively measured at that time. On the other hand, family history is still potentially evolving, and we plan to study that specific issue elsewhere. As an approximation, in our models we treat these covariates as being baseline covariates that summarize the life-long effect of parity, education and family history on breast cancer development and evolution.

## 1.3 A first model: observed data likelihood

We assume that after the onset of the disease (which may or may not occur) there is a time interval in which not even a screening examination is able to detect the presence of the disease

| Group | $(x_1, x_2, x_3)$ | Size | Dx (%) | #Asymp Dx | #Symp Dx | % Symp Dx among all Dx | Median age Asymp Dx | Median age Symp Dx |
|---|---|---|---|---|---|---|---|---|
| 1 | (0,0,0) | 3377 | 142 (4.2%) | 27 | 115 | 81% | 65.49 | 69.10 |
| 2 | (0,0,1) | 2107 | 115 (5.5%) | 31 | 84 | 73% | 66.24 | 66.80 |
| 3 | (0,1,0) | 3430 | 154 (4.5%) | 24 | 130 | 84% | 59.36 | 64.99 |
| 4 | (0,1,1) | 3354 | 153 (4.6%) | 23 | 130 | 85% | 63.70 | 64.54 |
| 5 | (1,0,0) | 27964 | 939 (3.4%) | 183 | 756 | 81% | 65.67 | 69.69 |
| 6 | (1,0,1) | 14338 | 599 (4.2%) | 109 | 490 | 82% | 65.83 | 68.54 |
| 7 | (1,1,0) | 12694 | 479 (3.8%) | 98 | 381 | 80% | 62.95 | 66.60 |
| 8 | (1,1,1) | 10787 | 453 (4.2%) | 77 | 376 | 83% | 62.78 | 64.45 |
| Total | | 78051 | 3034 (3.9%) | 572 | 2462 | 81% | 64.93 | 67.74 |

Table 1.2: Observed outcomes in each covariate group and in the total sample. Ages are measured in years. $X_1$= at least one birth (0:No, 1:Yes); $X_2$=Education level (0:Low/Medium, 1:High); $X_3$=Family history of cancer (0:No, 1:Yes).

(see Figure 1.2). All times are measured from birth of the woman. The two main quantities of interest are the time (from birth) to the start of asymptomatic detectability of the disease (which we denote by $T_A$) and the time to the symptomatic detection of the disease (denoted by $T_S$). At time $T_A$ the disease becomes detectable through screening. Between time $T_A$ and $T_S$ the tumor can only be detected through screening (the "sojourn time", denoted by $\Delta$), while starting from time $T_S$ the disease becomes evident because of symptoms. In other words we have $T_S = T_A + \Delta$. Further, we assume that symptomatic detection occurs exactly when the first symptoms appear, i.e. we assume that we observe $T_S$ precisely for subjects with a symptomatic detection.

While studying the latent evolution of the disease, we are also interested in studying the probability of insurgence of the disease in a woman's lifetime. For this reason, to allow for the direct estimation of such probability we introduce a cure rate structure, i.e. that there exists a proportion of women, which we call the "cured proportion", denoted by $(1-p)$ with $p \in (0,1)$, who will never experience the event of interest of developing breast cancer no matter how long they live. This is equivalent to assigning positive probability $(1-p)$ to the event $\{T_A = +\infty, T_S = +\infty\}$. The probability $p$ is one of the parameters of the model. The

standard terminology "cure" is not easy to interpret in this context, so we will instead refer to the fraction $p$ of women who will develop the disease as the "susceptible" proportion.

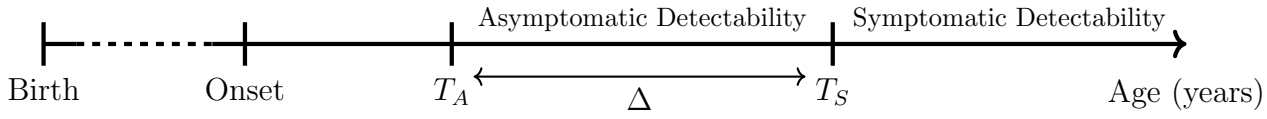



Figure 1.2: A graphical representation of the natural history from onset until detectability of the disease.

Importantly, we work under the stable disease population assumption, in which the rate of births and the distribution of ages at tumor onset are constant across calendar time.[31] We also assume stationarity of the joint distribution of $(T_A, \Delta)$ across birth cohorts.

Note that the goal is to draw conclusions about quantities that are mostly unobservable. Indeed, both $T_A$ and $T_S$ are never observed on any woman, and clearly the observed data would not be a good representation of the latent variables of interest.

First of all, the time to the start of the asymptomatic detectability $T_A$ is always interval censored. That is, even when we observe an asymptomatic detection, we never observe $T_A$ precisely but we can only conclude that it happened before the observed age at detection.

Second, there is a selection of women who enter the study (and the sample), since women who have already had a breast cancer diagnosis before entering the screening program are excluded from the sample.

Third, once a woman has entered the study, she is not followed until her death, but follow-up lasts around 12 years. Therefore, we do not have any information about tumors with onset, or that will be detected later on.

Another relevant consideration concerns the relationship of these latent quantities to the observed data: the mean of the ages at symptomatic detection in the sample should be smaller than the expected value of $T_S$ in the population, due to selection into the set of the observed $T_S$ ages; indeed, subjects with larger sojourn time $\Delta$ (e.g., $T_S$) are less likely to

have their $T_S$ value observed (since asymptomatic detection is more likely).

Hence the distribution of the observed ages at detection, asymptomatic or symptomatic, would not represent a good estimate of the underlying disease history, and why the proportion of observed diagnoses out of the total should not be confused with the probability of developing breast cancer for a woman in her lifetime.

When defining a model, there are basically three decisions to make that characterize its structure. The first one is the choice of the marginal distributions for $T_A$ and $\Delta$ for the diseased subjects. Any distribution having support on the non-negative real line may work, but even distributions on the real line could be appropriate under some specific parameter combinations that make the negative tail negligible.

The second assumption concerns the dependence structure between $T_A$ and $\Delta$. While modeling them as independent random variables may facilitate the form of the likelihood function and the estimation of the model parameters, such assumption may be too simplistic and not reflect the link between these two quantities that has been documented in the literature [67], [34].

Lastly, one should decide on how to include covariates (and which ones) in the model, both in the joint distribution of $(T_A, \Delta)$ and in the probability $p$.

The estimation procedure that we will follow will depend on the complexity of the model. In principle, it is possible to compute the observed data likelihood, and obtain the maximum likelihood estimates for the parameters. However, the calculation and maximization of the observed data likelihood can be complicated or not feasible, especially when the number of parameters grows. Indeed, such observed data likelihood involves many (bivariate) integrals which may not be solvable in closed form, but may need to be approximated numerically - thus introducing numerical difficulties in the estimation process.

Indeed, as we have seen above, each performed screening examination provides some information about the value of $T_A$, which is necessarily interval censored. On the other hand,

$T_S$ is observed precisely in the case of symptomatic (outside-screening) detections, or we only have partial information on it. Integrating the joint probability density of $(T_A, T_S)$, denoted by $f_{(T_A,T_S)}(t_a, t_s)$, on an appropriate subset of the domain as determined by the observed events, provides the observed data likelihood contribution for a generic $i$-th subject, which we denote by $L_i$.

Importantly, we condition on the observed mammography/ecography exams. Depending on the presence of a positive or negative exam, diagnosis and/or right censoring, one can observe different types of data configurations: cases with a symptomatic detection, cases with an asymptomatic detection and cases without an observed diagnosis. These three kinds of configurations contribute to the observed data likelihood in different ways. Recall that we are assuming perfect sensitivity and specificity of the examinations.

For a subject with an observed symptomatic detection, $T_S$ is fixed at the observed value $t_s$ and one should integrate the joint density function over all possible values of $T_A$. The lower bound of the integral ($l$) is the last negative examination if there is one, or the lower bound of the support otherwise. Thus, the contribution of such configuration to the observed data likelihood is:

$$L_i = p \cdot \int_l^\infty f_{T_A,T_S|\text{susceptible}}(u, t_s)du.$$

Note that, clearly, since $T_A < T_S$ with probability one (i.e. $\Delta > 0$), the integrand function is 0 for $u > t_s$.

For a subject with an observed asymptomatic detection, $T_S$ is greater than the last observed exam (denoted by $d$ since it coincides with the date of detection) and $T_A$ is necessarily between $l$, the last negative exam if there is one, and the detection time $d$. This

defines the integration region for this kind of trajectories:

$$L_i = p \cdot \int_l^d \int_d^\infty f_{T_A,T_S|\text{susceptible}}(u,v)dvdu.$$

Lastly, a subject who has not developed the disease (yet) may experience breast cancer after the last negative exam or the end of follow-up (with probability $p$), or never experience it (with probability $(1-p)$). In the first case, the likelihood contribution $L_i$ is obtained by integrating the joint density of $(T_A, T_S)|$susceptible over all values of $T_A$ greater than the last negative exam $l$ and over values of $T_S$ greater than the age at the end of follow-up. In the second case, the result of the analogous integration is 1 since the conditional distribution of $(T_A, T_S)|$non-susceptible is concentrated on $\{T_A = +\infty, T_S = +\infty\}$. Since these two events are disjoint, the total contribution to the likelihood is the sum of their probabilities:

$$L_i = (1-p) + p \cdot \int_l^\infty \int_c^\infty f_{T_A,T_S|\text{susceptible}}(u,v)dvdu.$$

Lastly, all likelihood contributions should take into account the fact that only asymptomatic women can enter the study, i.e.the distributions of the quantities of interest should all be conditional on the event $\{T_S >$ Age at entry$\}$: each likelihood contribution $L_i$ should be divided by the probability of the conditioning event

$$\text{c}_i = P(T_S > \text{Age at entry}|\text{Age at entry}) = (1-p) + p\ P(T_S > \text{Age at entry}|\text{diseased}, \text{Age at entry}).$$

Note how this expression is also based on the assumptions that once $T_S$ is reached, a symptomatic detection (and diagnosis) is immediately observed. While this is exactly not the case, we believe that it is most consistent with the study entry requirement. Notably, the condition does not require that $T_A >$ Age at entry.

The observed data likelihood is then given by the product of all the (independent)

subjects' contributions: $L = \prod_{i=1}^{n} \frac{L_i}{\mathrm{c}_i}$.

Even if not explicitly indicated in the notation above, $L$ clearly is a function of all the model parameters. For numerical maximization (and for estimation of the variance-covariance matrix of the MLEs), it is more convenient to work with the log-likelihood, which takes the form $l = log(L) = \sum_{i=1}^{n} log(L_i) - \sum_{i=1}^{n} log(\mathrm{c}_i)$.

The degree of difficulty of calculating the two-dimensional integrals which form the observed data likelihood function varies greatly according to the specific distributional assumptions. Only specific model formulations lead to analytical or partially analytical solutions. In the following Section we describe one such simple model.

### 1.3.1 Model specification and results

We consider a simple model that assumes independence between $T_A$ and $\Delta$ and does not include any covariates. In particular, for the susceptible subjects, we assume $T_A \sim N(\mu, \sigma^2)$ and $\Delta \sim Exp(\lambda)$, with $\Delta$ independent of $T_A$, where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\lambda > 0$. Easily, $T_S = T_A + \Delta$ has density $f_{T_S}(t) = \lambda e^{\frac{\lambda^2\sigma^2}{2} + \lambda(\mu - t)} \Phi(t, \mu + \lambda\sigma^2, \sigma^2)$, where $\Phi(\cdot, \mu + \lambda\sigma^2, \sigma^2)$ is the cdf of a $N(\mu + \lambda\sigma^2, \sigma^2)$. Also, the conditional density of $T_S|T_A$ is $f_{T_S|T_A}(v|u) = \lambda e^{-\lambda(v-u)} I_{(u,\infty)}(v)$. Note that, marginally, $T_S$ follows an exponentially modified Gaussian (emg) distribution with parameters $(\mu, \sigma, \lambda)$. The contributions to the observed data likelihood are as follows (for the derivation please refer to Section 1.6). Using the notation introduced earlier, we have: (i) for a subject with an observed symptomatic detection

$$L_i = p \cdot f_{T_S}(t_S) \cdot \Big(1 - \frac{\Phi(l, \mu + \lambda\sigma^2, \sigma^2)}{\Phi(t_s, \mu + \lambda\sigma^2, \sigma^2)}\Big);$$

(ii) for a subject with an observed asymptomatic detection

$$L_i = p \cdot e^{\frac{\lambda^2\sigma^2}{2} + \lambda(\mu - d)} \cdot \Big(\Phi(d, \mu + \lambda\sigma^2, \sigma^2) - \Phi(l, \mu + \lambda\sigma^2, \sigma^2)\Big);$$

| $T_A$ | | $\Delta$ | Cure rate |
|---|---|---|---|
| $\mu$ | $\sigma$ | $\lambda$ | p |
| 64.9 (64.5,65.3) | 22.3 (17.6,27.0) | 1.62 (1.51,1.73) | 0.179 (0.172,0.186) |

Table 1.3: MLEs and 95% confidence intervals for the model parameters. Time is measured in years.

and (iii) for a subject without observed diagnosis

$$L_i = (1-p) + p\cdot\Big(\frac{f_{T_S}(c) - e^{\lambda(l-c)} f_{T_S}(l)}{\lambda} + 1 - \Phi(c,\mu,\sigma^2)\Big).$$

The probability of the conditioning event $\{T_S >$ Age at entry|Age at entry$\}$ is equal to

$$c_i = (1-p) + p\cdot\Big(1 - \Phi(f,\mu,\sigma^2) + \frac{f_{T_S}(f)}{\lambda}\Big).$$

Table 1.3 shows the estimates obtained from the maximization of the observed data likelihood with respect to the four model parameters $(\mu, \sigma, \lambda, p)$. The likelihood maximization is performed using the `R` function `maxLik`.[29] We reparameterized all models in such a way that the resulting parameter space becomes $\mathbb{R}^k$, where $k$ is the number of model parameters, i.e. with no constraints. In particular, we applied a logarithmic transformation to all parameters with a positivity constraint. The parameter $p$ is constrained to take values in the interval $[0, 1]$, and for that parameter we used a logistic reparametrization. Relying on invariance of maximum likelihood estimators one then obtains the estimates for the original parameters. Application of the Delta method (details not shown) allows one to compute their standard errors.

Note that, due to the left truncation and the right censoring in the observed data, the estimated latent proportion of women experiencing the disease in their lifetime is around 18% (recall that our model does not impose any constraint on the upper bound of the subjects'

lifespan). One may compare such rate to the estimated lifetime risk of breast cancer, that has been estimated as being one out of eight, or 12.5%.[66] Although the observed proportion of diagnoses in the sample was around 4%, the model reconstructs the frequency of many more lifetime diagnoses than those observed during our limited follow-up.

The start of the asymptomatic detectability is on average close to the age of 65 years, ranging between 20 and 110 with 95% estimated frequency. The numbers 20 and 110 are the values taken by the two consistent estimators for the percentiles 2.5% and 95% of the normal distribution of $T_A$. Consistency clearly follows from the continuous mapping theorem applied to the consistent MLEs. This is a wide interval; in Section 1.4.3 we will see that including some covariates in the model has the effect of reducing the marginal variability of $T_A$.

The model suggests that the sojourn time $\Delta$ is very short, lasting on average 7-8 months, with an exponential tail. This result is quite different from the estimates from previous studies, which suggest a mean sojourn time between 2 and 7 years [68]. The reason for such small estimate for the sojourn time is possibly the lack of detailed information on the examination results and on the kind of detection from our data (see our comment on this in Section 1.2). Indeed, we should also recall that $T_A$ and $T_S$ have been defined here starting from the dates of the observed diagnoses, while they capture the start of detectability. Thus a shorter sojourn time may be consistent with an over-estimation of $T_A$, and/or an under-estimation of $T_S$.

We now move to more flexible and informative models.

## 1.4 More flexible models

### 1.4.1 Approximate Bayesian Computation

As we have pointed out, the calculation of the observed data likelihood for latent processes with large amounts of missing data can be challenging even for relatively simple models. In

general, every small change to the model definition requires the observed likelihood function to be computed by hand and implemented. For example, the inclusion of a dependence structure between $T_A$ and $\Delta$ requires solving complicated integrals through numerical approximations that determine loss of accuracy, as well as a significant increase in the difficulty by the optimization algorithms in identifying the maximum likelihood estimates.

An estimation procedure that allowed one to quickly implement several different models would greatly increase the flexibility in modeling. This is possible by implementing a likelihood-free approach, where the observed likelihood function does not need to be calculated explicitly, nor maximized. A likelihood-free approach that seems particularly promising for disease history models is Approximate Bayesian Computation (ABC) [57].

The first step of ABC consists of setting prior distributions for the model parameters. One then samples a parameter vector from their prior distribution, and generates a dataset from the corresponding model. In the basic version of ABC, if the simulated data are "close enough" to the real data, that parameter combination is retained and included in the sample of parameter values that approximates the posterior distribution of the parameters given the data. Indeed, implementing this procedure a very large number of times (here $200,000$) and selecting only a very small proportion (called tolerance or retention rate) of samples, then allows one to approximate the parameters' posterior distribution.

It is also common to post-process the ABC output to improve the selected posterior sample by applying a so-called "regression adjustment." The idea is to regress each parameter (or to perform a multivariate regression with all the parameters as response vector) on the set of summary statistics and to apply a correction based on the difference between observed and simulated summaries [6, 37].

Measuring the distance between two datasets (observed vs. model generated in ABC) is not trivial: one should use informative summary statistics of the data, which reduce the dimensionality of the data but still retain the information needed to perform accurate

inference on the parameters. Indeed, only by using sufficient statistics and by conditioning on the event that they are identical (not close) in the observed data and in the model generated data, one would ensure that the sample of retained parameter values represents a sample from their exact posterior distribution [37].

There is a vast literature about the choice of summary statistics in ABC, and a variety of approaches have been proposed [14]. Most of the methods, however, do not propose any constructive procedure, but only suggest techniques to select a subset of summary statistics among a bigger set of proposals (subset selection methods) or to combine them to reduce the dimensionality (projection methods).

In our models we include the three binary covariates described in Section 1.2, which partition the subjects into eight groups. We consider a set of the same summary statistics computed on each of the eight groups.

In particular, we build "Metric 1" to measure the dissimilarity between the observed and a model-generated dataset, based on a total of 32 summary statistics (4 for each of the eight groups of women): proportion of observed diagnoses, proportion of observed symptomatic detections among the total number of observed detections, median age at asymptomatic detection, and median age at symptomatic detection. The distance between the two datasets is then defined as the $L^2$-distance between the standardized summary statistics of the two datasets. The standardization is performed by dividing each summary statistic by a robust estimate of its standard deviation (the median absolute deviation).

"Metric 2" refines "Metric 1" by also considering the entire distribution of the observed age at detection. This metric makes use of the classical statistical test for the comparison of two proportions and of the Kolmogorov-Smirnov test to assess if two observed samples can be considered to be generated by the same underlying distribution. We perform the first 16 tests to compare the proportions of observed detections and of symptomatic detections in each of the eight covariate groups. Then, we perform 16 additional tests to compare the

distributions of the age at asymptomatic and symptomatic detection, again for each group. We believe that the test statistics themselves or the corresponding p-values could provide a good measure of the distance between two objects (two proportions or two distributions, depending on the test). There are many ways to combine the test outputs (test statistics or p-values) into a distance function between the two datasets. In section 1.6 we briefly explore the relative performance of "Metric 1" and a version of "Metric 2" on two simulated datasets, and "Metric 1" seems to produce estimates of the parameter values which are closer to the true values.

In Section 1.4.3 we present the results obtained by using "Metric 1" while fitting different models to the motivating data. The retention (tolerance) rate is chosen through a leave-one-out cross-validation procedure, which is implemented and available in the `R` package `abc` [23]. We make use of local linear regression to correct the posterior samples by regression adjustment.

In the simulation process, we generate the screening examinations with the same schedule of the real screening program, and assuming a constant adherence rate of 0.6 to the prescribed examinations [58]. Hence, the screening parameters are fixed, and not object of inference. For the subjects belonging to the susceptible proportion, the disease history is then overlapped with the attended examinations to produce the observed age at detection (if it happens inside the interval of follow-up), the detection mode (symptomatic or asymptomatic), and the age at last negative examination. For the non-susceptible subjects, we identify the age at their last negative examination, if there is one, before the end of the follow-up. We thus obtain a dataset containing information that has similar structure to that of the observed data.

To make the simulated data as comparable to the observed ones as possible, we keep approximately the same distribution for the covariates. The approximation comes from the fact that one needs to generate a slightly larger sample of women because some of them will experience a symptomatic detection of the disease before the age at entry in the study, and therefore will be excluded from the effective sample. Through some simulations, we

estimated this proportion to be roughly 4% of women, so for each simulated sample we generated $78051/0.96 \approx 81305$ women. We assign the 78051 observed covariate vectors to the first 78051 women in the simulated sample, and take a random sample of the covariate vectors for the remaining $81305 - 78051 = 3254$ women.

Note that the ABC procedure described above, which is known as ABC-rejection algorithm, is very computationally demanding since only a small fraction of the generated samples are retained and contribute to the posterior distribution approximation. There exist many refinements of the ABC algorithm, aimed at reducing the inefficiency due to sampling from very uninformative prior distributions by exploiting the information of already accepted parameter values [40].

These refined algorithms, such as ABC-MCMC [41] and ABC-SMC [56], could bring a substantial computational gain but have the main drawback of not being easy parallelizable on multiple cores. Having the possibility to work on a server with many processors, we decided to implement the ABC-rejection procedure. For the implementation of all models we used the software `R` [51] on a server with 176 cores.

### 1.4.2 Models

Recall the three binary covariates described in Section 1.2, $X_1 =$"at least one birth", $X_2 =$"high level of education" and $X_3 =$"family history of cancer", all coded as 0 = no and 1 = yes. We posit models such that the "susceptible" proportion depends on the observed covariates $x = (x_1, x_2, x_3)$ through the logit link:

$$p(x) = \frac{e^{p_0+p_1x_1+p_2x_2+p_3x_3}}{1 + e^{p_0+p_1x_1+p_2x_2+p_3x_3}}.$$

For the susceptible (developing the disease) subjects, recall that the evolution of the disease is described by the time to its asymptomatic detectability $T_A$ and by its sojourn time

$\Delta$. We let the mean of $T_A$ depend on the covariates linearly, while the variance of $T_A$ is assumed constant across covariate groups.

The distribution of $\Delta$ is defined conditionally on the observed value of $T_A$, and it may reflect the effect of the covariates but only indirectly (see below). Note that any form of dependence between $T_A$ and $\Delta$ is easily manageable through ABC, since the simulated value of $T_A$ is already available when one generates the value of $\Delta$ from the distribution of $\Delta|T_A$.

We have explored several different models. We do not report all details, such as the prior distributions, for all of them here. Parameters associated with covariates had an uninformative prior distribution centered at zero. The prior distribution for the mean of $T_A$ in the baseline group, denoted with $\beta_0$, was chosen to be $\mathrm{N}(65, 10)$: indeed, from the literature and from the simple model in Section 1.3.1 (see MLEs in Table 1.3) , we expect a mean of 65 to be reasonable[13] but we still keep a variance large enough to let the data bring in relevant information on $\beta_0$. Similarly, $p_0$ represents the proportion of women who develop the disease in the baseline group and we assign to it a rather informative prior: $p_0 \sim \mathrm{logit}\big(\mathrm{Beta}(3,\ 21)\big)$ around the lifetime risk as described above. Indeed, the prior distribution corresponds to a woman in the baseline group has on average a probability of $3/(3+21) = 0.125$ of belonging to the diseased (susceptible) group.

Here below is the list of nine models. Note that the number of parameters to be estimated, indicated below between square brackets, is always equal to 4, for the non-susceptible proportion regression, plus 7 or 8 for the disease history.

1. **Normal + Exponential** [4+7=11 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \ \sim\ \mathrm{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,\ \sigma^2);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1 \ \sim\ \mathrm{Exp}(e^{\gamma_0 + \gamma_1 t_A}).$$

2. **Normal + Exponential (log-scale)** [4+7=11 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim logN\Big(\mu = m\big(\frac{s^2}{m^2}+1\big)^{-\frac{1}{2}},\ \sigma^2 = \log\big(\frac{s^2}{m^2}+1\big)\Big);$$

$$\tilde{\Delta} = log(T_S) - log(T_A) \mid \{T_A = t_A\}, \gamma_0, \gamma_1 \ \sim\ \text{Exp}(e^{\gamma_0+\gamma_1 t_A}),$$

where $m = E(T_A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $s^2 = Var(T_A)$. This parameterization is used to let the variance of $T_A$ (in the original scale) be independent of covariates, i.e. the same across groups.

3. **Bivariate Normal** [4+8=12 parameters]

$$(T_A, \Delta)|\ \beta_0, ..., \beta_3, \mu_\Delta, \sigma_1, \sigma_2, \rho \sim N_2(\mu, \Sigma),$$

where $\mu = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,\ \mu_\Delta)$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.

4. **Bivariate Normal (log-scale)** [4+8=12 parameters]

Let $\tilde{\Delta} = \log(T_S) - \log(T_A)$. We assume

$$(\log(T_A), \tilde{\Delta}) \mid \beta_0, ..., \beta_3, \mu_\Delta, \sigma_1, \sigma_2, \rho \sim N_2(\mu, \Sigma),$$

where $\mu = (m\big(\frac{s^2}{m^2}+1\big)^{-\frac{1}{2}},\ \mu_\Delta)$ and $\Sigma = \begin{bmatrix} \log\big(\frac{s^2}{m^2}+1\big) & \rho\log\big(\frac{s^2}{m^2}+1\big)^{\frac{1}{2}}\sigma_2 \\ \rho\log\big(\frac{s^2}{m^2}+1\big)^{\frac{1}{2}}\sigma_2 & \sigma_2^2 \end{bmatrix}$.

As in all the previous models, again here $m = E(T_A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ depends on the covariates, while $s^2 = Var(T_A)$ does not.

5. **Gamma + Weibull** [4+8=12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \ \sim \ \text{Gamma}\Big(\frac{(\mu(x))^2}{\sigma^2}, \ \frac{\mu(x)}{\sigma^2}\Big);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1, k \ \sim \ \text{Weibull}\big(\lambda(t_A), k\big),$$

where $E(T_A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $\lambda(t_A) = e^{\gamma_0 + \gamma_1 t_A}$ and $k$ has a prior distribution that includes one (corresponding to the exponential case).

6. **Gamma + piecewise Exponential** [4+8=12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \ \sim \ \text{Gamma}\Big(\frac{(\mu(x))^2}{\sigma^2}, \ \frac{\mu(x)}{\sigma^2}\Big);$$
$$\Delta \mid \{T_A = t_A\}, \lambda_1, \lambda_2, \lambda_3 \ \sim \ \text{Exp}(\lambda_1 \cdot 1(t_A \leq 55) + \lambda_2 \cdot 1(55 < t_A \leq 65) + \lambda_3 \cdot 1(t_A > 65),$$

where $E(T_A) = \mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

7. **Rescaled Beta + Exp** [4+7=11 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \ \sim \ 100 \cdot \text{Beta}(\alpha, \ \beta);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1 \ \sim \ \text{Exp}(e^{\gamma_0 + \gamma_1 t_A}),$$

where $E(T_A) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $\sigma^2 = 100^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

8. **Rescaled Beta + Weibull** [4+8=12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \ \sim \ 100 \cdot \text{Beta}(\alpha, \ \beta);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1, k \ \sim \ \text{Weibull}\big(\lambda(t_A), k\big),$$

where $E(T_A) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $\sigma^2 = 100^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, and $\lambda(t_A) = e^{\gamma_0 + \gamma_1 t_A}$. Here, too, $k$ has a prior distributions that includes one.

9. **Rescaled Beta + piecewise Exponential** [4+8=12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \ \sim \ 100 \cdot \text{Beta}(\alpha, \ \beta);$$
$$\Delta \mid \{T_A = t_A\}, \lambda_1, \lambda_2, \lambda_3 \ \sim \ \text{Exp}(\lambda_1 \cdot 1(t_A \leq 55) + \lambda_2 \cdot 1(55 < t_A \leq 65) + \lambda_3 \cdot 1(t_A > 65),$$

where $E(T_A) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $\sigma^2 = 100^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Note that both the normal and the gamma distributions have decreasing densities for older ages (with the gamma density decreasing more slowly, in addition to not imposing symmetry and not allowing for negative values). As the incidence is expected to always increase with age, these densities capture the phenomenon through their left tails. Note also that very limited data are available for older ages, due to right censoring which also includes death. The three models based on the rescaled beta density should provide a more realistic shape for the right tail of $T_A$.

In the next section we discuss the results of the ABC-based model selection procedure to choose among these models.

### 1.4.3 Model selection and results

To select the best model among the ones described above, we simulate 200,000 samples from each model [39, 57]. Using the metric described above, we compute the distance between each simulated sample and the observed one. Then, from the pooled set of samples produced by all the models, we select the samples that have the smallest distance from the observed data, keeping track of which model generated each sample. The resulting sample of parameter values and model index can be regarded as a sample from the approximate joint posterior distribution of the parameter and the model index. The number of retained samples generated by a specific model, divided by the total number of retained samples, thus represents an approximation of the posterior probability of that model. For a more detailed description of

this procedure, see Algorithm 1 in [39].

Since the initial number of samples (200,000) was the same for each model, we are assuming a uniform prior distribution over the nine models. Figure 1.3 shows the prior and approximate posterior probabilities of each model, and Table 1.4 contains the numerical values of the approximate posterior probabilities. Model 6 and Model 9 clearly show the highest (by far) posterior probabilities (0.295 and 0.267).

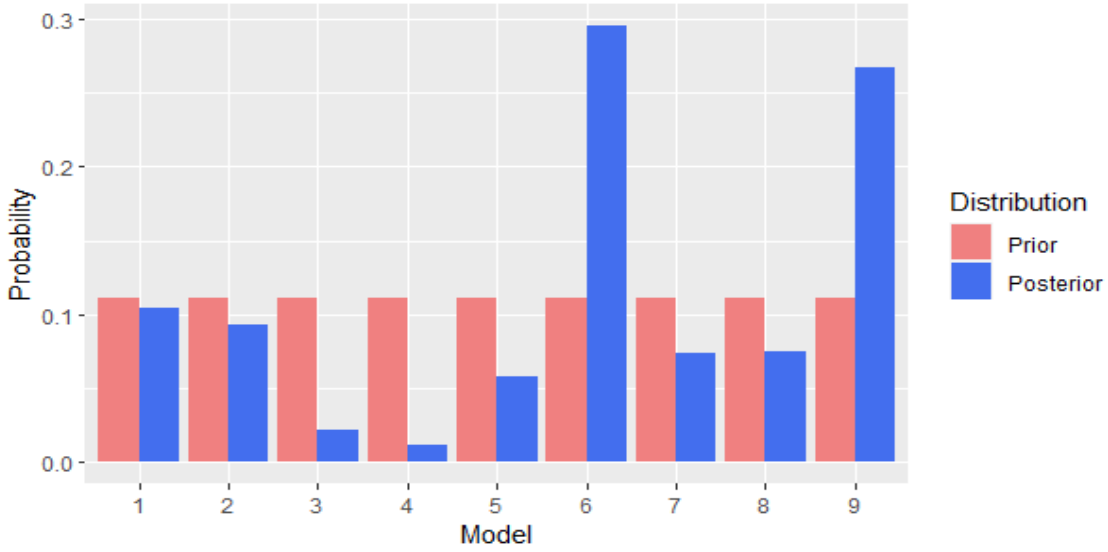



Figure 1.3: Prior and posterior probabilities of the nine models (global retention rate = 0.005).

| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Posterior probability | 0.104 | 0.094 | 0.022 | 0.011 | 0.058 | 0.295 | 0.074 | 0.075 | 0.267 |

Table 1.4: Posterior probabilities of the nine models (global retention rate = 0.005).

The ABC model choice procedure introduced above presents some quite severe potential pitfalls. Indeed, as it has been highlighted in [39], in many cases it may even fail to converge to a Dirac distribution on the true model as the size of the observed dataset grows to infinity. In other words, the so-called "curse of insufficiency" [39] is likely to occur, thus leading to

arbitrariness in the construction of the Bayes factor (and thus of the posterior probabilities of the models).

Given these concerns about the algorithm illustrated above, some alternative techniques to conduct model choice in the context of ABC have been proposed. We focus on an approach based on random forests that has been introduced in [50]. For an introduction to random forests, which are a machine learning tool consisting of the aggregation of simple classifiers (called trees) that can be used both for classification and regression purposes, we refer to Chapter 15 of [28]. In [50], model selection is reformulated as a classification problem, and is split into two steps.

The first step trains a random forest that predicts, for each possible value of the summary statistics, the model that best fits the data. In other words, the random forest is a classifier that associates to each vector of summary statistics a predicted model among the nine proposed. The training set is represented by the pooled set of simulations performed for the nine models. Once the classifier is trained, the predicted model for the set of observed summary statistics represents the selected model, i.e. the model that obtained the majority of votes among the classification trees of the random forest. For more details on this step we refer to Algorithm 2 in [50]. Table 1.5 shows that, given a trained random forest made of 1000 trees, Model 9 obtained the majority of votes (392) and it is, therefore, the model selected for having the best fit to the observed data.

| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Votes | 80 | 42 | 14 | 14 | 62 | 270 | 74 | 52 | 392 |

Table 1.5: Counts of votes for the nine models out of a total of 1000 trees composing the random forest.

In the second step, the posterior probability of the selected model is computed through a secondary random forest. The binary model prediction errors (Model 9 vs all the other models) are computed for each observation using the out-of-bag classifiers (see [28] for the

definition of out-of-bag classifier in a random forest). This secondary random forest, which is again trained on the pooled set of simulations performed for the nine models, performs a regression of the prediction error on the summary statistics. Lastly, the posterior probability of the selected model is computed as the random forest regression estimate associated to the vector of observed summary statistics. A detailed explanation of this second step can be found in Algorithm 3 of [50]. In our case, this procedure resulted in a posterior probability for Model 9 equal to 0.381.

The results from this alternative procedure for model selection disagree slightly with those from the simpler algorithm described at the beginning of this section. However, the two approaches agree on the best two models being Model 6 and Model 9. Given the motivation provided in the literature to consider the approach based on random forests more reliable (see [39]), we now describe the results of the ABC estimation procedure for Model 9, the "Rescaled Beta + piecewise Exponential" model.

The metric used to quantify the distance between two samples was based, for each covariate-defined stratum, on the proportion of diagnoses and of symptomatic detections, and on the median age at asymptomatic and symptomatic detection ("Metric 1", see also Section 1.4.1). We assumed the following independent prior distributions for the model parameters: $\beta_0 \sim \mathrm{N}(0.65, 0.05)$, $\beta_i \sim \mathrm{N}(0, 0.25)$, for $i = 1, 2, 3$, $\sigma \sim \mathrm{Unif}(0.02, 0.25)$, $\lambda_i \sim \mathrm{Unif}\big(0.1, 4\big)$, for $i = 1, 2, 3$, $p_0 \sim \mathrm{logit}\big(\mathrm{Beta}(3, 21)\big)$, $p_i \sim \mathrm{Unif}(-2, 2)$, for $i = 1, 2, 3$.

A retention (or tolerance) rate of 0.02 was chosen via a leave-one-out cross-validation procedure, by comparing the quality of several posterior estimates obtained using different tolerance rates. The posterior distributions shown in Figure 1.4 are thus based on a sample of $200,000 \times 0.02 = 4,000$ selected parameter values.

We note that the posterior distributions of the model parameters are much more concentrated than the prior distributions. The only exception is parameter $\lambda_1$, whose posterior distribution is still quite flat. This lack of posterior information is probably due to the small number of

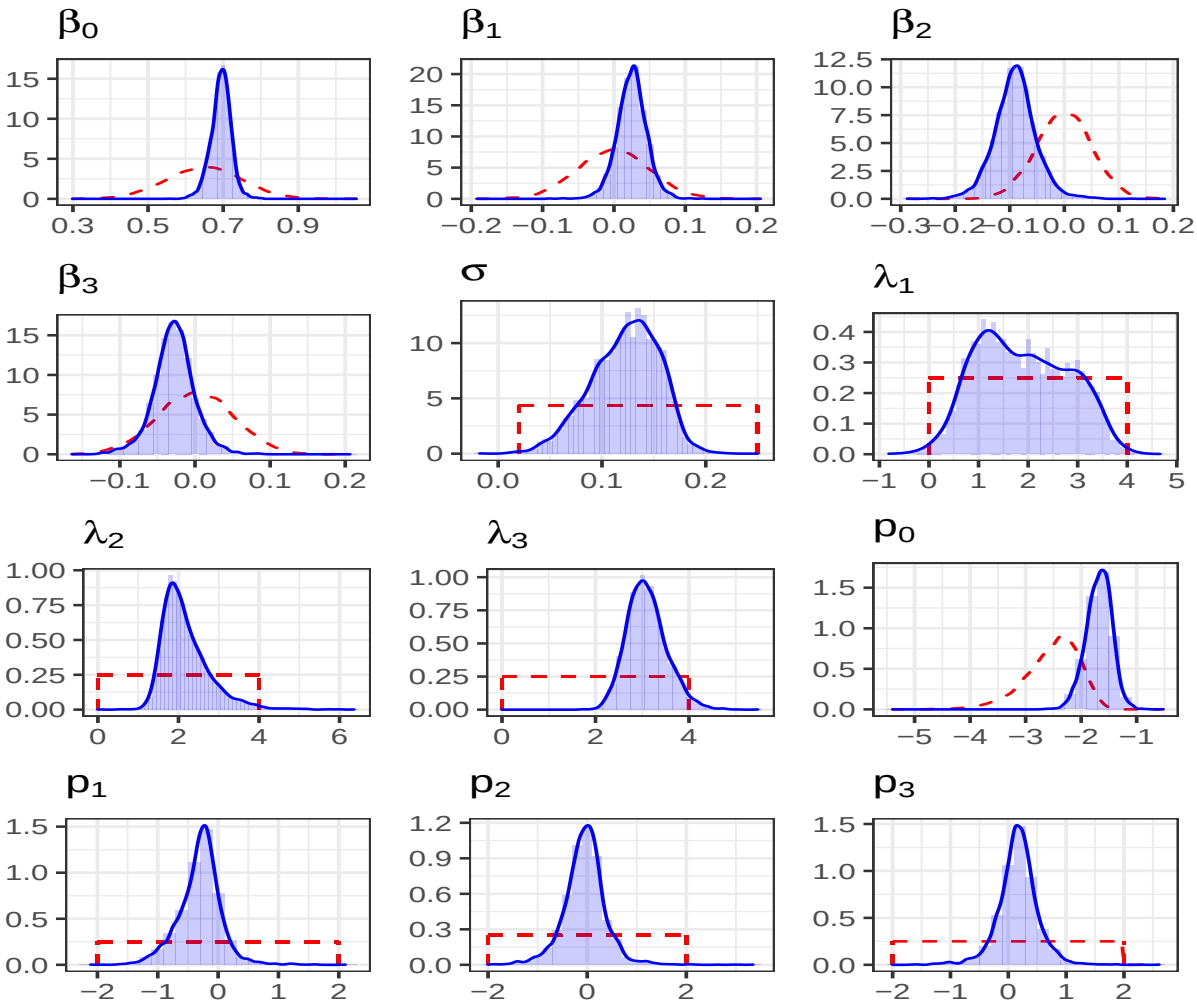



Figure 1.4: Prior (red dashed line) and local linear regression adjusted approximate posterior (blue histogram and solid line) densities for each parameter of the "Rescaled Beta + piecewise Exponential" model.

cases observed among women younger than 55 years old. Table 1.9 shows the posterior modes and the 95% intervals corresponding to the regions of the approximate posterior distributions that have the highest density (HPD intervals).

Some interesting observations on the effect of the covariates arise from the estimated posterior distributions: (i) women with at least one child tend to have a lower probability of ever experiencing breast cancer, and a later $T_A$ if they do (posterior distributions for $p_1$ and

| Parameter | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma$ | $\lambda_1$ |
|---|---|---|---|---|---|---|
| Mode | 0.698 | 0.028 | -0.087 | -0.027 | 0.135 | 1.203 |
| HPDI | (0.637, 0.744) | (-0.017, 0.065) | (-0.163, -0.009) | (-0.087, 0.026) | (0.058, 0.179) | (0.374, 3.582) |
| Parameter | $\lambda_2$ | $\lambda_3$ | $p_0$ | $p_1$ | $p_2$ | $p_3$ |
| Mode | 1.854 | 3.015 | -1.621 | -0.228 | 0.015 | 0.147 |
| HPDI | (1.218, 3.419) | (2.258 3.914) | (-2.171, -1.276) | (-1.118 0.318) | (-0.989, 0.726) | (-0.528, 0.872) |

Table 1.6: Posterior modes and the 95% highest posterior density (HPD) intervals.

$\beta_1$); (ii) having a family history of cancer has the opposite effects, according to the posterior distributions for $p_3$ and $\beta_3$; (iii) women with a high level of education experience breast cancer earlier than women with a lower education level, but this variable is probably not very relevant in modifying the susceptible proportion $p$ (posterior for $p_2$ almost symmetric around 0).

To gain a clearer idea on how covariates influence the mean of $T_A$, which is defined as $\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, we may combine the posterior distributions of $\beta_0, \beta_1, \beta_2$ and $\beta_3$ according to the covariate combination of each group. The resulting boxplots are shown in the left panel of Figure 1.5. We can see that covariates do indeed play an important role in determining $E(T_A)$, whose estimated posterior median ranges from a minimum of 58 to a maximum of 72 years old.

Similarly, combining the posterior distributions of $p_0, p_1, p_2$ and $p_3$, we can compute the posterior distribution of the susceptible proportion $\mathrm{p}(x)$ in the eight covariate groups. As we can see in right panel of Figure 1.5, the probability for a woman of developing breast cancer varies across groups. In particular, its median ranges from a minimum of about $11\% - 12\%$ for women in groups 5 and 7 (having at least one birth and with no family history of cancer) to a maximum of about $17\% - 18\%$ for women in groups 2 and 4 (without any birth and with family history of cancer).

Once an approximation of the posterior distribution of the parameters is available, it is also possible to compute approximate predictive distributions for $T_A$ in each covariate group, as well as for $\Delta$ given the observed value of $T_A$. Given a specific covariate configuration, we

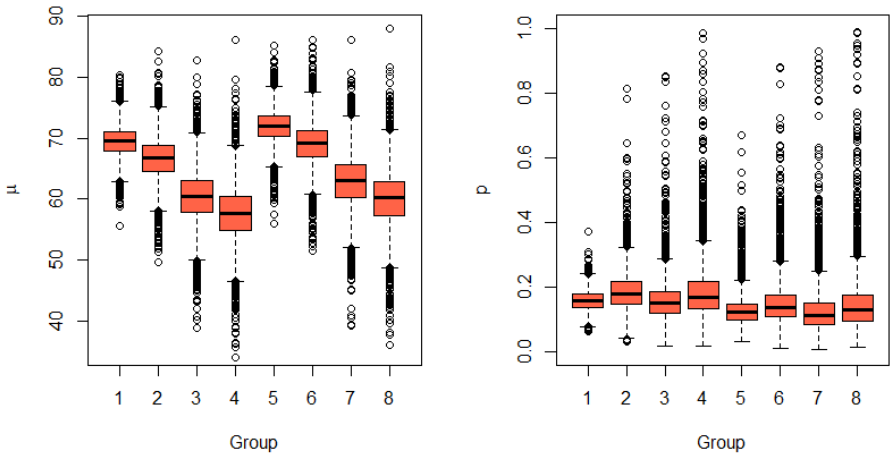



Figure 1.5: Approximate posterior distribution of the mean age at asymptomatic detectability $\mu(x)$ and of the susceptible proportion $\mathrm{p}(x)$ across covariate groups.

| Group | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 |
| 5 | 1 | 0 | 0 |
| 6 | 1 | 0 | 1 |
| 7 | 1 | 1 | 0 |
| 8 | 1 | 1 | 1 |

Table 1.7: $X_1$= At least one birth (0:No, 1:Yes); $X_2$= Education level (0:Low/Medium, 1:High); $X_3$= Family history of cancer (0:No, 1:Yes).

have a joint posterior sample for the mean and for the standard deviation of $T_A$, $\{(\mu_i, \sigma_i),\ i = 1, \ldots, 4000\}$. For each couple $(\mu_i, \sigma_i)$, we then draw a value of $t_A^i$ from the model, i.e. we generate

$$t_A^i \mid \mu_i, \sigma_i \overset{ind}{\sim} 100 \cdot \mathrm{Beta}\big(\mu_i,\ \sigma_i\big), \quad \text{for } i = 1, \ldots, 4000,$$

where $\mathrm{Beta}(\mu_i, \sigma_i)$ denotes a Beta random variable having mean $\mu_i$ and variance $\sigma_i^2$. The set of generated values $\{t_A^i,\ i = 1, \ldots, 4000\}$ then represents a sample from the ABC approximation of the predictive distribution of $T_A$ in that group [9].

We can repeat this procedure for each covariate group, obtaining the eight distributions shown by the boxplots in the left-hand side of Figure 1.6.

Similarly, the posterior sample of size 4,000 for $\lambda_1, \lambda_2$ and $\lambda_3$ can be used to generate a sample from the approximate predictive distribution of $\Delta$ given $T_A$ (see the right-hand side

of Figure 1.6), by using:

$$\delta_1^i \mid \{T_A \leq 55\}, \lambda_1^i \overset{ind}{\sim} \text{Exp}(\lambda_1^i), \quad \text{for } i = 1, \ldots, 4000;$$
$$\delta_2^i \mid \{55 < T_A \leq 65\}, \lambda_2^i \overset{ind}{\sim} \text{Exp}(\lambda_2^i), \quad \text{for } i = 1, \ldots, 4000;$$
$$\delta_3^i \mid \{T_A > 65\}, \lambda_3^i \overset{ind}{\sim} \text{Exp}(\lambda_3^i), \quad \text{for } i = 1, \ldots, 4000.$$

Note that these results suggest that $\Delta$ decreases when $T_A$ increases, which is in contrast with the medical literature [67]. However, such decrease is quite small in size.

Clearly, the predictive distributions of $T_A$ and $\Delta$ cannot be directly compared with the observed data. In Section 1.6.6 we provide an example where, under simplified assumptions, one can compute the distributions of the observed age at asymptomatic and symptomatic detection analytically. One way to explore the goodness of fit of these models would be to generate data from them and to compare such data to the observed data through some summaries. However, this is exactly how ABC has produced the estimated model parameters, so that the algorithm is indeed already based on a goodness-of-fit maximizing procedure (see also Section 1.4.4). Clearly, additional examination of the data generated from the estimated model could be entertained.

Relatedly, in the next section we analyze the effect of different screening policies (in terms of observed detections) given the estimated latent disease process.

### 1.4.4 Comparing alternative screening strategies

After estimating the parameters of the models, one can finally use this information to compare different screening strategies. In particular, knowing the distribution of $T_A$ and $\Delta$ can help identify an optimal screening strategy.

We now compare several screening strategies, which differ with respect to the gap between consecutive examinations, the proportion of attended examinations out of the total number

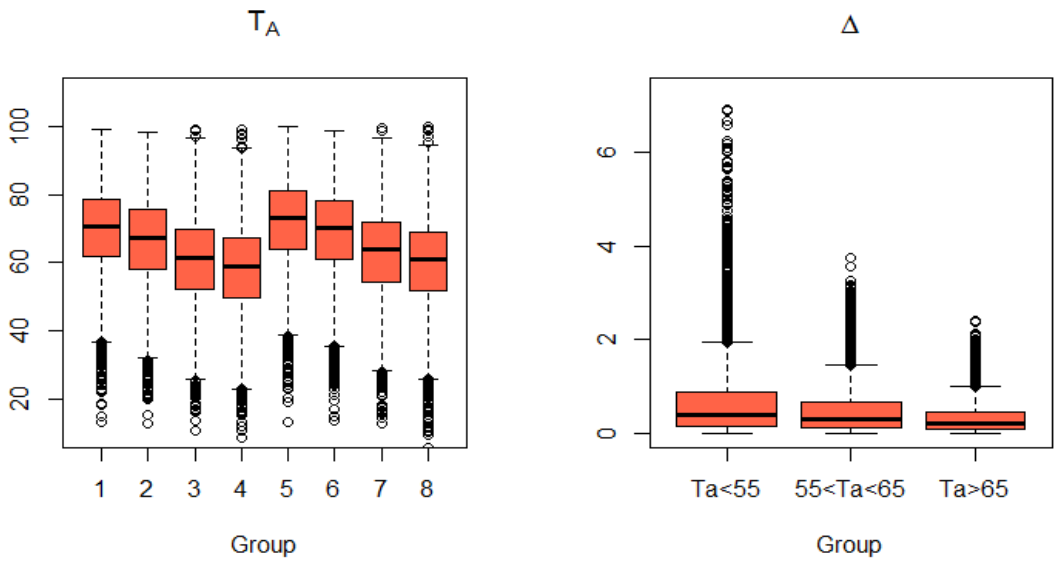



Figure 1.6: Predictive distributions for $T_A$ in each covariate group and for $\Delta$ given the observed value of $T_A$.

of invitations (adherence), and the screening age range. In particular, we start from the screening strategy offered in Lombardy (denoted by "Screening strategy 1") and we measure the effect of varying some of its features on the total number of observed detections during the screening age interval, the percentage of asymptomatic detections, and the median age at observed asymptomatic and symptomatic detection. The underlying assumption (as supported by many medical studies, see e.g. [46]), is that the moment in which a tumor is detected could make a difference on the outcome of the disease. Indeed, while here we have not discussed post-detection treatment and outcomes, detecting the disease earlier rather than when symptoms would have emerged, i.e. at a less advanced stage, should allow one to treat it with more success.

The set of six screening strategies that we have considered is shown in Table 1.8. All the screening strategies are applied to a sample of size 100,000 generated from the estimated predictive distributions for the "Rescaled Beta + piecewise Exponential" model. In the

simulated samples we assume an administrative follow-up interval that coincides with the screening interval (that is, 50-69 or 50-74 depending on the policy), except for a small proportion of about 5% of the subjects, for whom censoring for other causes occurs earlier.

As expected, reducing the gap between consecutive screening examinations from two years to one year results in an increase in the percentage of asymptomatic detections out of all detections by $72\% - 76\%$ (from 14.0% to 24.7% or from 18.0% to 30.9%), depending on adherence. Clearly, such an increase would come with a substantial increase in the cost of the program.

Another possibility to increase the percentage of tumors diagnosed before becoming symptomatic would be to increase the adherence to the screening program. From our results we estimate that increasing it from the current level of about 60% to an adherence of 80% would make the proportion of asymptomatic detections increase by $25\% - 29\%$. Thus, even without modifying the screening strategy, it seems crucial to find ways to raise the awareness of women on the importance of breast cancer screening. As adherence likely depends on subjects' covariates and is not constant over time, campaigns to encourage women to attend the screening examinations regularly should target categories of women who tend to adhere less [12].

Intensifying the screening examinations (either by reducing the gap or by increasing the adherence) does not seem to imply a relevant difference on the age at observed asymptomatic and symptomatic detections, but only on the total number of observed diagnoses.

Another observation concerns the effect of extending the end of the screening interval from the age of 69 to the age of 74 years old (this change has been recently implemented in the Lombardy screening program). The total number of tumors detected during the screening period (which is longer) increases by 30%. However, the proportion of asymptomatic detections slightly decreases by around $4\% - 6\%$. We can explain this result by recalling that tumors at older ages are (slightly) faster in becoming symptomatic according to our model, so screening

in the age range 69-74 is less "efficient" (produces slightly fewer asymptomatic detections) than screening at younger ages.

We should also point out that, despite the small values of the (latent) quantity $\Delta$ predicted by our model, the difference between the median age at observed asymptomatic and symptomatic detections is around 3 years, similar to the gap observed in the motivating data. Such observed difference seems to be due to the fact that women over 69 (or 74 with the new screening policy) are not screened, and therefore detections that occur after that age can only be symptomatic, making the median age at observed symptomatic detection increase.

This also shows, once again, that the data filtered by the partial observation mechanism do not give a clear picture of the underlying latent disease process in absence of a proper inferential model. Indeed, for more details on the results see Table 1.8.

| Screening Strategy | % Dx | % Asymp Dx | Median Age Asymp Dx | Median Age Symp Dx | Median Lead Time |
|---|---|---|---|---|---|
| (50-69, 2yrs, 60%) | 5.45% | 15.4% | 59.99 | 62.65 | 0.370 |
| (50-69, 2yrs, 80%) | 5.61% | 19.2% | 59.92 | 62.39 | 0.362 |
| (50-74, 2yrs, 60%) | 7.25% | 14.0% | 62.33 | 65.38 | 0.342 |
| (50-74, 2yrs, 80%) | 7.37% | 18.0% | 62.18 | 65.25 | 0.326 |
| (50-74, 1yrs, 60%) | 7.36% | 24.7% | 62.27 | 65.50 | 0.321 |
| (50-74, 1yrs, 80%) | 7.42% | 30.9% | 62.62 | 65.62 | 0.328 |

Table 1.8: Observed summary statistics on a sample of size 100,000 generated from the estimated "Rescaled Beta + piecewise Exponential" model under several different screening strategies. The screening strategies are defined by the screening age range, the gap between subsequent exams, and the overall adherence proportion.

## 1.5 Discussion

We have proposed several parametric models to describe the natural history of breast cancer, where the main events of interest are the start of asymptomatic detectability of the disease and the symptomatic detection ($T_A$ and $T_S$). The models differ in their parametric assumptions, but they all share a cure rate structure that takes into account that a fraction of women will never experience the disease in their lifetime. Estimating how long tumors stay in the latent phase between time $T_A$ and time $T_S$ (i.e. estimating the sojourn time $\Delta$) is of crucial importance for planning an efficient screening policy.

We have obtained the distribution of these random quantities by estimating the model parameters from data collected as part of a motivating study. While the results do seem to provide useful information, they should be handled with some care given the described lack of some information (and thus their reconstruction) in the available data.

Depending on the complexity of each model, we have employed a likelihood-based or likelihood-free estimation procedure. Given the complex missing data structure, it has shown to be very challenging and in most cases infeasible to obtain maximum likelihood estimates for the model parameters. The calculation and the maximization of the observed data likelihood rely on numerical algorithms, and even for relatively simple models they have been found to be computationally unstable. The numerical approximation of the Hessian matrix used to obtain standard errors for the parameters has also been found to be difficult.

Approximate Bayesian Computation (ABC) allowed us to perform both model selection and parameter estimation without having to maximize nor calculate explicitly the observed data likelihood function. However, we recall that inference based on ABC is subject to several levels of approximation: (i) the metric chosen to assess the dissimilarity between generated and observed data; (ii) the tolerance for acceptance of a generated parameter value; (iii) the use of Monte Carlo to estimate the posterior distributions; and (iv) the use of post processing adjustments [57].

We experimented with two different metrics to evaluate the distance between simulated and observed data and, based on some simulations, we chose one of them. One could try to refine the way of calculating the distance between the two datasets by using different statistical tests to measure the difference between the distributions of ages at observed diagnosis. Another possibility to quantify the distance between the datasets may be to consider the accuracy of a classification method implemented to distinguish between observed and simulated data [27].

The results from the model in Section 1.3 and the model selected in Section 1.4 are not directly comparable, since the MLEs obtained in Section 1.3 refer to a model that does not include covariates. However, Table 1.3 shows that the MLEs reflect an average across groups of the estimates found from the model with covariates, and a general agreement between the two models can be appreciated. In Section 1.6 we compare the results from the two models further.

Also, note that the time when asymptomatic detectability starts ($T_A$) depends on the accuracy of the technology used to perform the examination that, therefore, should be the same for all the visits included in the estimation procedure. An improvement in the examination technique could make $T_A$ move backwards and the length of the asymptomatic detectability interval increase.

The theoretical distribution of the observed age at asymptomatic and symptomatic detection can be computed analytically from a theoretical model, after superimposing the screening examinations. In Section 1.6.6, we obtain the form of the distributions of the observed age at detection (both symptomatic and asymptomatic) for such a model. The resulting expressions are rather complicated, and in most cases simulations are probably a more suitable tool to study the effect of the selection process on the observed detections under complex models and screening strategies.

Summing up, in this work we have highlighted that latent (realistic) models for disease

histories are challenging to develop and implement.

ABC is a very flexible and conceptually simple tool, that looks especially suitable in this setting where it is relatively easy to generate data even from models that have an intractable observed data likelihood.

We should point out that goodness-of-fit of the models here is evaluated conditionally on the choice of the prior distributions for the parameters of each model. Therefore, it is possible that a model is penalized by a poor choice of the prior or, on the contrary, that a model performs well thanks to a good prior choice. In particular, a change in the prior distributions may lead to a different result in ABC model choice. Note that model selection between two non-nested parametric models could also be performed by using Vuong's test [65]. However, Vuong's test is based on the ratio of the likelihood functions under the two models, and as a consequence also requires that one be able to compute them.

Our models have assumed perfect screening sensitivity and specificity. However, note that they can also be extended to estimate them from the data, and to take into account the dependence between the subject-specific adherence pattern and the latent disease process. These extensions could not be implemented on the motivating data, given that detailed information about screening invitations and examinations results was not available to us.

We have conducted a small experiment in this direction. We have extended the selected model by introducing an additional parameter for the sensitivity of the screeening examinations. The ABC estimation procedure based on the usual "Metric 1" did not work well. Despite using quite an informative prior for the new parameter (Beta with mean equal to 5/6), stability issues in the estimation of the susceptible proportions emerged. This could be due to the choice of metric distance (the summary statistics) or to the lack of information in the data. While in general sensitivity may be identifiable, these results suggest that the choice of the metric to be used in ABC may make the identifiability of some parameters more difficult. In addition, we have inserted in the data generation process a hard coded value of sensitivity

of 0.9, and that did not seem to change the posterior distributions of the other parameters of the model.

Moreover, access to data with a longer follow-up could allow to study the effect of screening and treatments on survival. In general, it will be of great interest to apply the models and methods that we have developed to other, similar datasets to confirm the information on the latent process that we have obtained here.

Changing the way in which event times are observed, for example by changing the screening schedule, does not impact the latent process. A possible way to check whether these models describe the latent process well could be to also use data collected under different screening policies or observed screening frequencies. For example, we know that the Covid-19 pandemic is causing a consistent drop in screening adherence. Therefore, it will be important to apply these models to data collected by screening programs during, and after this period.

In this work we did not mention overdiagnosis due to mammography screening, that is the detection of a breast cancer that would not be detected during the woman's lifetime in the absence of screening. In other words, an overdiagnosed cancer would have never become symptomatic, because of its very slow evolution, and would have never led to death. Many authors discussed this issue and proposed several methods to quantify the risk of overdiagnosis [7, 48, 4]. A possibility to extend our models to address such question might be to implement a cure rate structure on $\Delta$ for the in-screening detected cases or, equivalently, to assign a positive probability to the event $\{T_A < +\infty, T_S = +\infty\}$. Identifiability and estimability for such extended models, both in general and for even large sample sizes, are open questions that will need to be addressed.

# 1.6 Additional analyses and results

## 1.6.1 Observed data likelihood for the model in Section 1.3.1

In this Section, we present the derivation of the observed data likelihood function of the model studied in Section 1.3.1. As already mentioned, we have three kinds of data configurations:

(i) for a subject with an observed symptomatic detection at age $t_s$,

$$\begin{aligned}
L_i &= p \int_l^{t_s} f_{T_A,T_S}(x,t_s)dx = p\int_l^{t_s} f_{T_A}(x)f_{T_S|T_A}(t_S|x)dx \\
&= p \int_l^{t_s} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda(t_S-x)}dx \\
&= p\,\lambda\, e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-t_S)} \Big(\Phi(t_S,\mu+\lambda\sigma^2,\sigma^2) - \Phi(l,\mu+\lambda\sigma^2,\sigma^2)\Big) \\
&= p\, f_{T_S}(t_S)\Big(1 - \frac{\Phi(l,\mu+\lambda\sigma^2,\sigma^2)}{\Phi(t_S,\mu+\lambda\sigma^2,\sigma^2)}\Big) = p\,\Big(f_{T_S}(t_S) - e^{\lambda(l-t_S)} f_{T_S}(l)\Big),
\end{aligned}$$

where $l$ is the smallest possible value for the asymptomatic detectability, which can be the age at the last negative examination, if there is one, or the lower bound of the support of $T_A$;

(ii) for a subject with an observed asymptomatic detection at age $d$,

$$\begin{aligned}
L_i &= p \int_l^d \int_d^\infty f_{T_A,T_S}(x,y)dydx = p\int_l^d\int_d^\infty f_{T_A}(x)f_{T_S|T_A}(y|x)dydx \\
&= p \int_l^d \int_d^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda(y-x)}dydx = p \int_l^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{\lambda x}\int_d^\infty \lambda e^{-\lambda y}dydx \\
&= p \int_l^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{\lambda x} e^{-\lambda d}dx = p \int_l^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\lambda d + \frac{\lambda(\sigma^2\lambda+2\mu)}{2}} e^{-\frac{(x-(\mu+\sigma^2\lambda))^2}{2\sigma^2}}dx \\
&= p\, e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-d)}\Big(\Phi(d,\mu+\lambda\sigma^2,\sigma^2) - \Phi(l,\mu+\lambda\sigma^2,\sigma^2)\Big) = p\,\frac{f_{T_S}(d) - e^{\lambda(l-d)} f_{T_S}(l)}{\lambda},
\end{aligned}$$

where $l$ is the smallest possible value for the asymptomatic detectability, which can

be the age at the last negative examination, if there is one, or the lower bound of the support of $T_A$;

(iii) for a subject without an observed diagnosis at the censoring time $c$,

$$L_i = (1-p) + p\Big(\int_l^c \int_c^\infty f_{T_A,T_S}(x,y)dydx + \int_c^\infty \int_x^\infty f_{T_A,T_S}(x,y)dydx\Big),$$

where

$$\begin{aligned}\int_l^c \int_c^\infty f_{T_A,T_S}(x,y)dydx &= e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-c)}\Big(\Phi(c,\mu+\lambda\sigma^2,\sigma^2) - \Phi(l,\mu+\lambda\sigma^2,\sigma^2)\Big) \\ &= \frac{f_{T_S}(c) - e^{\lambda(l-c)}f_{T_S}(l)}{\lambda},\end{aligned}$$

and

$$\begin{aligned}&\int_c^\infty \int_x^\infty f_{T_A,T_S}(x,y)dydx = \int_c^\infty \int_x^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda(y-x)}dydx \\ &= \int_c^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \int_x^\infty \lambda e^{-\lambda(y-x)}dydx = \int_c^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 - \Phi(c,\mu,\sigma^2).\end{aligned}$$

Note that, this contribution is equal to $L_i = (1-p) + p \int_l^\infty \int_c^\infty f_{T_A,T_S}(x,y)dydx$. However, from the model assumptions, $T_A < T_S$ with probability one, and the integral $\int_c^\infty \int_c^x f_{T_A,T_S}(x,y)dydx$ takes the value zero.

Lastly, we compute the probability of the conditioning event (entry requirement for the motivating study):

$$\begin{aligned}\mathrm{c}_i &= P(T_S > \text{Age at entry} | \text{Age at entry}) \\ &= (1-p) \; + p \; P(T_S > \text{Age at entry} | \text{Diseased, Age at entry}).\end{aligned}$$

Denoting by $f$ the age at entry in the program, we have that

$$
\begin{aligned}
&P(T_S > \text{Age at entry} | \text{Diseased, Age at entry}) = 1 - \int_{-\infty}^{f}\int_{x}^{f} f_{T_A,T_S}(x,y)dydx \\
&= 1 - \int_{-\infty}^{f} f_{T_A}(x)\int_{x}^{f}\lambda e^{-\lambda(y-x)}dydx = 1 - \int_{-\infty}^{f}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\left(1-e^{\lambda(x-f)}\right) \\
&= 1 - \Phi(f,\mu,\sigma^2) + e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-f)}\Phi(f,\mu+\lambda\sigma^2,\sigma^2) = 1 - \Phi(f,\mu,\sigma^2) + \frac{f_{T_S}(f)}{\lambda}.
\end{aligned}
$$

The observed data likelihood is the product of the subjects' contributions, divided by the probability of the conditioning event: $L = \prod_{i=1}^{n}\frac{L_i}{c_i}$.

### 1.6.2 ABC vs MLE

We compare the ABC approximate posterior distributions to the MLEs obtained for the model described in Section 3.1 of the article, the "Normal + Exponential" model without covariates.

Figure 1.7 shows that the ABC approximate posterior distributions are concentrated on regions of the parameters' supports which are not far from the confidence intervals around the maximum likelihood estimates. Only the approximate posterior distribution for $\lambda$ seems to overestimate its magnitude, while those for $\mu$, $\sigma$ and $p$ show a substantial agreement with the MLEs.

Figure 1.8 shows that using the post-processing regression adjustment does not modify the ABC results significantly. Importantly, although confidence intervals and posterior density intervals are clearly different in their very meaning, the results from ABC suggest less precise inference compared to the MLEs. This seems to be a potential drawback in the use of the more flexible ABC approach, unless one is willing to impose more concentrated (and thus potentially misleading) prior distributions for the parameters.

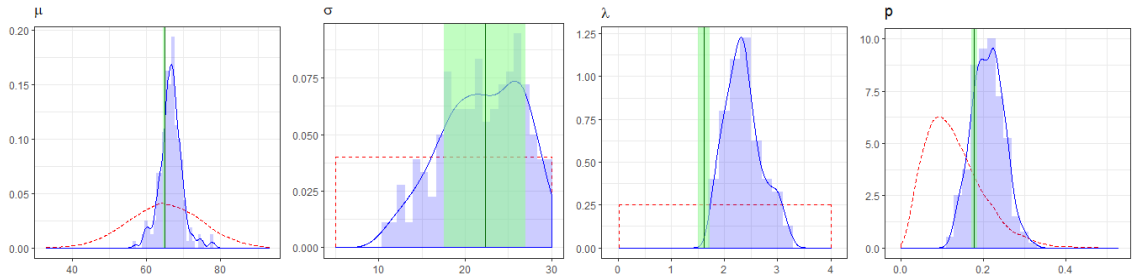



Figure 1.7: Prior (red dashed line) and unadjusted posterior (blue histogram and solid line) densities for each parameter of the "Normal + Exponential" model without covariates. The green vertical lines represent the MLEs, together with their 95% confidence intervals (green areas).

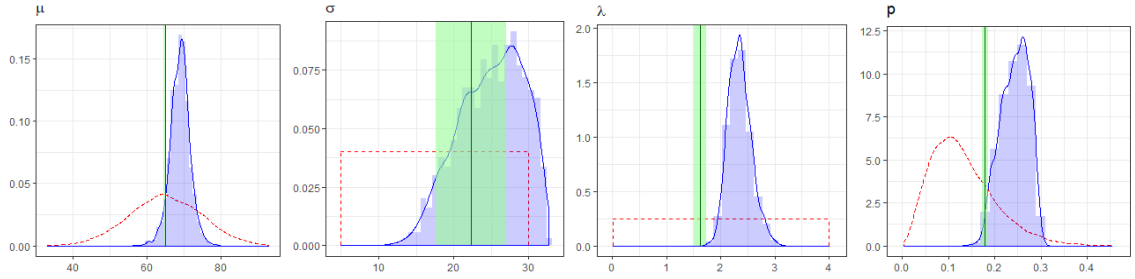



Figure 1.8: Prior (red dashed line) and local-linear-regression adjusted posterior (blue histogram and solid line) densities for each parameter of the "Normal + Exponential" model without covariates. The green vertical lines represent the MLEs, together with their 95% confidence intervals (green areas).

### 1.6.3 ABC estimation for the "Gamma + piecewise Exponential" model

We present the results of the ABC estimation procedure applied to Model 6, the "Gamma + piecewise Exponential" model, in order to compare them to the results of the "Rescaled Beta + piecewise Exponential" model. We present the local-linear-regression adjusted results obtained by using Metric 1. We assumed the following independent prior distributions for the model parameters: $\beta_0 \sim \mathrm{N}(65, 100)$, $\beta_i \sim \mathrm{N}(0, 25)$, for $i = 1, 2, 3$, $\sigma \sim \mathrm{Unif}(0, 25)$, $\lambda_i \sim \mathrm{Unif}\big(0.1, 4\big)$, for $i = 1, 2, 3$, $p_0 \sim \mathrm{logit}\big(\mathrm{Beta}(3, 21)\big)$, $p_i \sim \mathrm{Unif}(-2, 2)$, for $i = 1, 2, 3$. A retention (or tolerance) rate of 0.01 is chosen via a leave-one-out cross-validation procedure,

as performed by the `R` function `cv4abc`. This means that the posterior distributions are estimated from a sample of $200,000 \times 0.01 =, 2000$ retained parameter values.

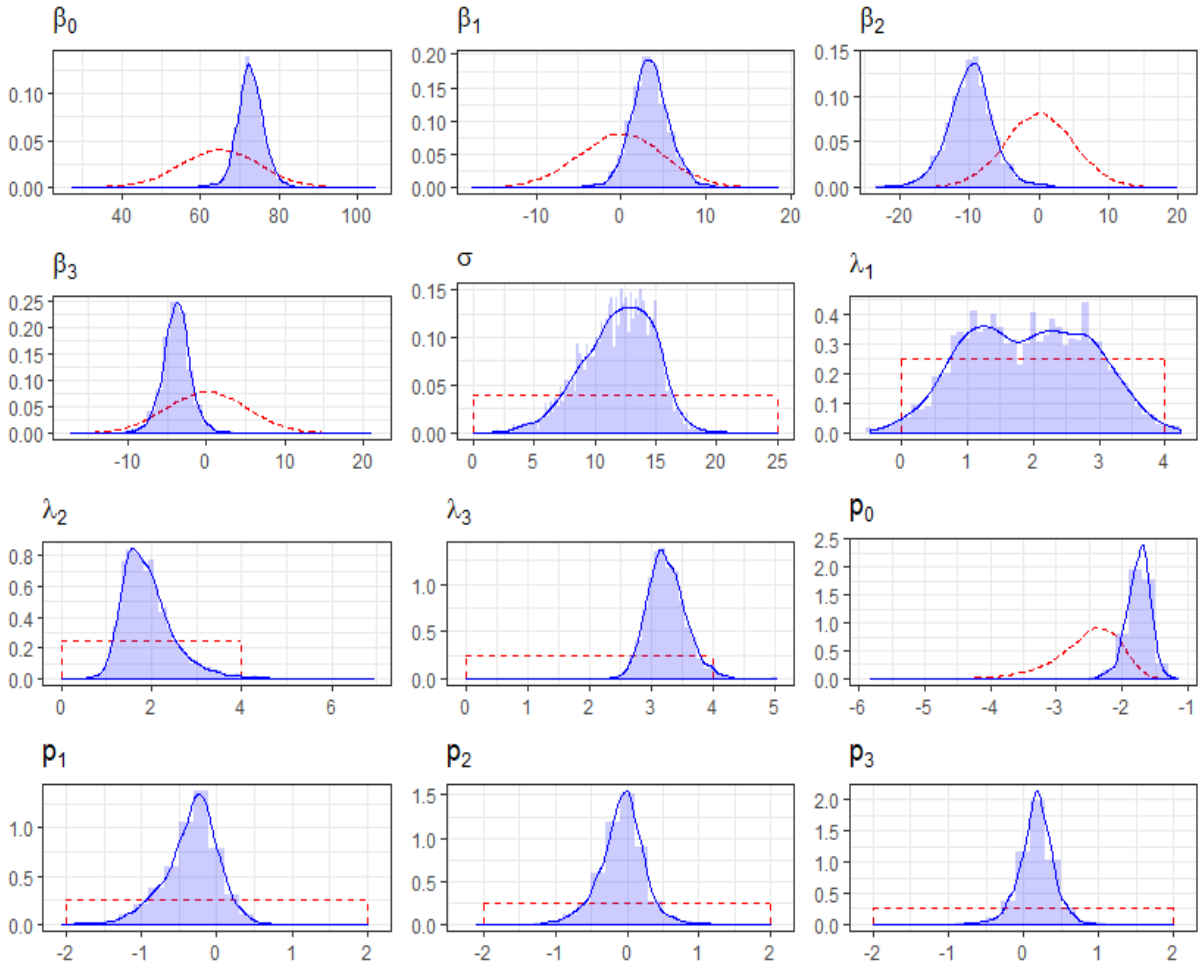



Figure 1.9: Prior (red dashed line) and local linear regression adjusted approximate posterior (blue histogram and solid line) densities for each parameter of the "Gamma + piecewise Exponential" model.

The posterior distributions of the model parameters are much more concentrated than the prior distributions (see Figure 1.9). The main observations about the effect of each covariate that arise from the estimated posteriors are very similar to those highlighted in Section 1.4.3 for the "Rescaled Beta + piecewise Exponential" model: (i) women with at least one child

tend to have a lower probability to experience breast cancer and a later $T_A$ when they do (posterior distributions of $p_1$ and $\beta_1$); (ii) having a family history of cancer has the opposite effect, according to the posterior distributions of $p_3$ and $\beta_3$; (iii) women with a high level of education experience breast cancer around 10 years earlier than women with a lower level, but they have a lower probability of getting diseased, according to the posterior distributions for $\beta_2$ and $p_2$.

To gain a clear idea of how covariates influence the mean of $T_A$, which is given by $E(T_A) = \mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, we again combine the posterior distributions of $\beta_0, \beta_1, \beta_2$ and $\beta_3$ according to the covariate values combination of each group. The resulting boxplots are shown in the left panel of Figure 1.10. We can see how covariates play an important role in determining $E(T_A)$, whose median ranges from a minimum of 59 years old to a maximum of 76 years. The similarity with the left panel of Figure 1.5 is evident, showing an agreement of the two models on the estimate of $\mu(x)$. The right panel of Figure 1.10 is the analogous of the right panel of Figure 1.5 and shows how the probability for a woman of developing breast cancer varies across groups. Its posterior median ranges from a minimum of about 10% for women in groups 5 and 7 (having at least one birth and with no family history of cancer) to a maximum of about 20% for women in groups 2 and 4 (without any birth and with family history of cancer).

Similarly to what we have done in the previous section, we estimate the predictive distributions for $T_A$ in each covariate group, as well as for $\Delta$ given the observed value of $T_A$. Note how the distributions shown in Figure 1.11 are again consistent with the results obtained from the "Rescaled Beta + piecewise Exponential" model. Indeed, the set of boxplots on the left-hand side of the figure (predictive distributions for $T_A$) looks very similar to that shown in the left panel of Figure 1.6.

Similarly, the posterior sample of size 2,000 for $\lambda_1, \lambda_2$ and $\lambda_3$ can be used to generate a sample from the approximate predictive distribution of $\Delta$ given $T_A$ (see the right-hand side

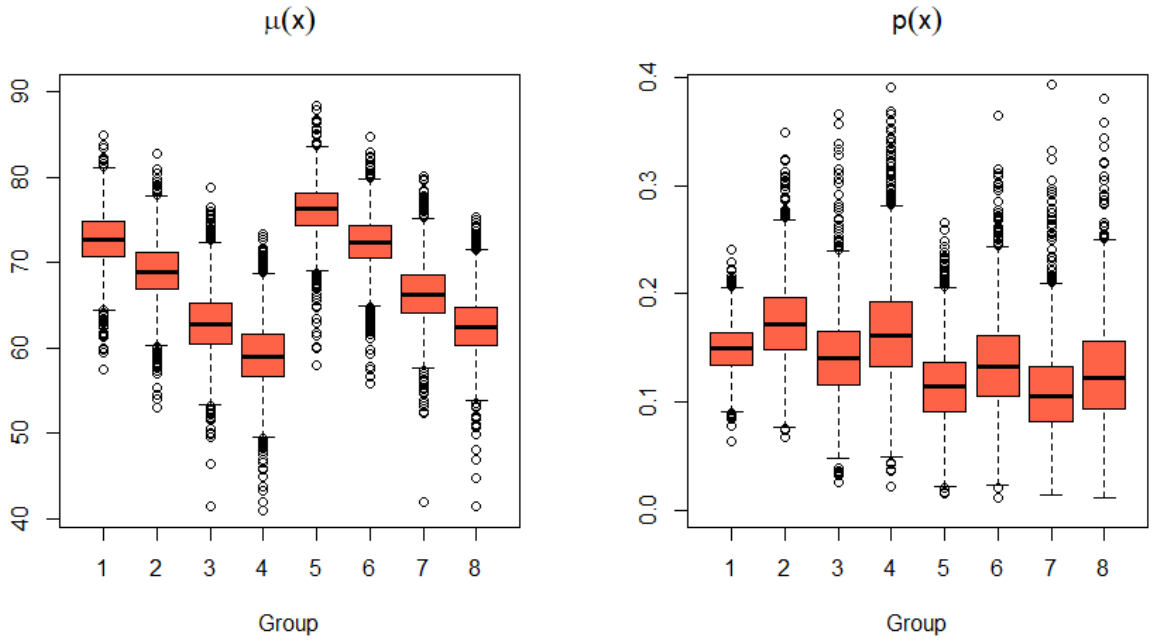



Figure 1.10: Approximate posterior distribution of the mean age at asymptomatic detectability $\mu(x)$ and of the susceptible proportion $p(x)$, across the eight covariate groups, for the "Gamma + piecewise Exponential" model.

of Figure 1.11). From these results the two models substantially agree in concluding that tumors with a later $T_A$ seem to evolve faster, and therefore to have a shorter $\Delta$, than tumors with earlier $T_A$.

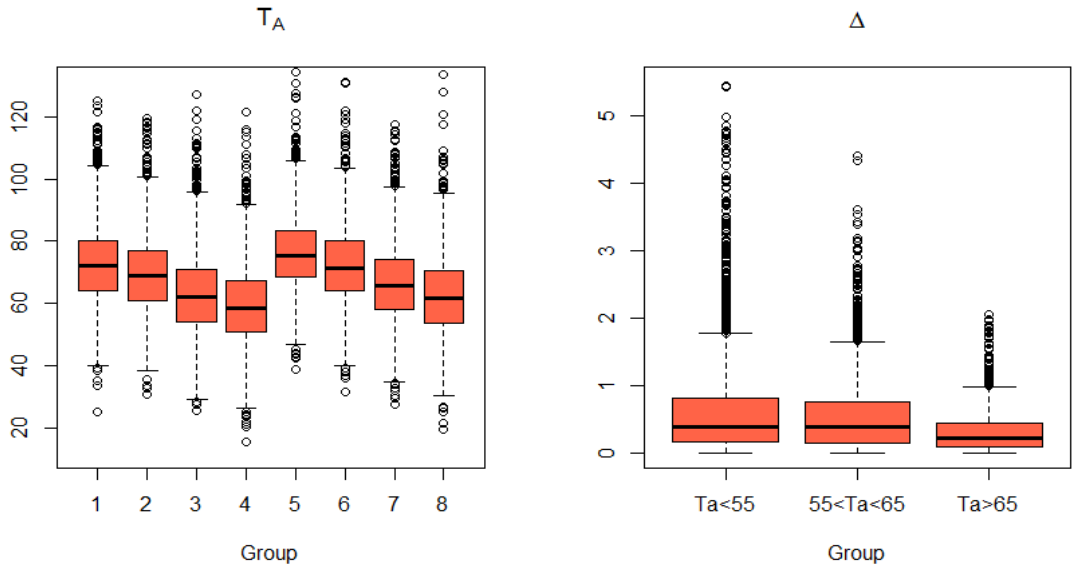



Figure 1.11: Predictive distributions for $T_A$ in each covariate group and for $\Delta$ given the observed value of $T_A$, for the "Gamma + piecewise Exponential" model.

### 1.6.4 Results under different data assumptions

In this Section we show the results obtained under an alternative set of assumptions on the data used to define symptomatic and asymptomatic detections.

Compared to the first set of assumptions, that we presented in the main body of the article, here we obtain a larger proportion of asymptomatic detections. Indeed, we checked if there was at least one screening examination within one year (previously we set this interval to be less than six months) prior to the diagnosis. If yes, then the last one before the diagnosis was assumed to have given a positive result and to have led to an asymptomatic detection. In this case the date of detection was defined as the date of that positive exam.

If, instead, there were no screening exams within one year of the date of diagnosis, we classified that detection as symptomatic, and we set the date of detection equal to the date of the most recent non-screening exam, if there were any within the six months prior to diagnosis. If no exams at all were recorded in the six months prior to diagnosis, then we set the date of the symptomatic detection back by a number of days equal to the average shift applied to the symptomatic detections which had that information (72 days).

Once the dates of detection were defined, we picked the last negative exam as the most recent exam performed at least one year before the detection. In other words, we imposed a larger minimum distance (one year instead of six months) between the last negative exam and the detection, as compared to the first set of assumptions.

By following these rules, we obtained 728 asymptomatic and 2306 symptomatic detections. The total number of diagnoses is of course unchanged (3034).

We performed ABC parameter estimation for the "Gamma + piecewise Exponential" model, using Metric 1 and applying the local-linear-regression adjustment to the posterior sample. Figure 1.12 shows the posterior distributions obtained retaining 0.3% (chosen via cross-validation) of the proposed parameter values. If one compares these distributions with those shown in Figure 1.4 of the article, he can notice that the differences are negligible for almost all parameters. The only small changes regard the posterior distributions of $\lambda_2$ and $\lambda_3$, which are now slightly shifted towards smaller values (and thus towards larger values of $\Delta$).

The resulting predictive distributions for $T_A$ are nearly identical to those presented in the article, while the predicted values for $\Delta$ are slightly larger than before (see Figure 1.13). The predicted mean sojourn times are now 12 months, 8 months and 5 months, for $T_A \leq 55$, $55 < T_A \leq 65$ and $T_A > 65$ respectively.

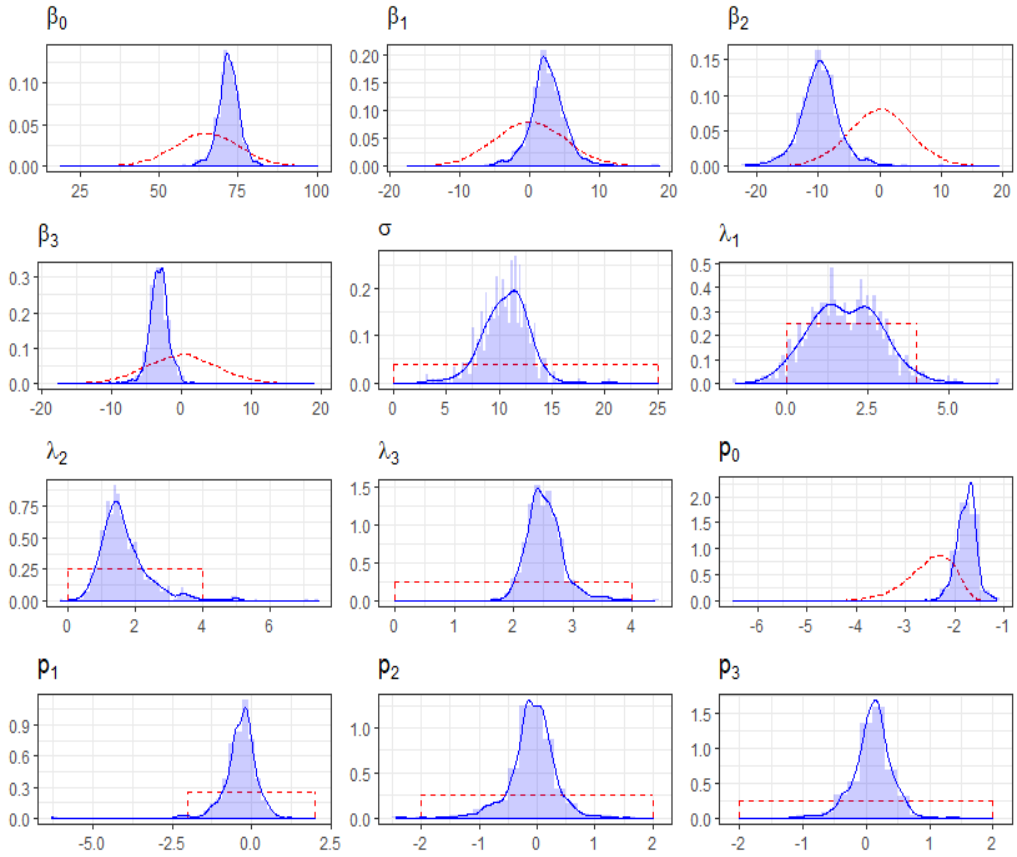



Figure 1.12: Prior (red dashed line) and local linear regression adjusted posterior (blue histogram and solid line) densities for each parameter of the "Gamma + piecewise Exponential" model.

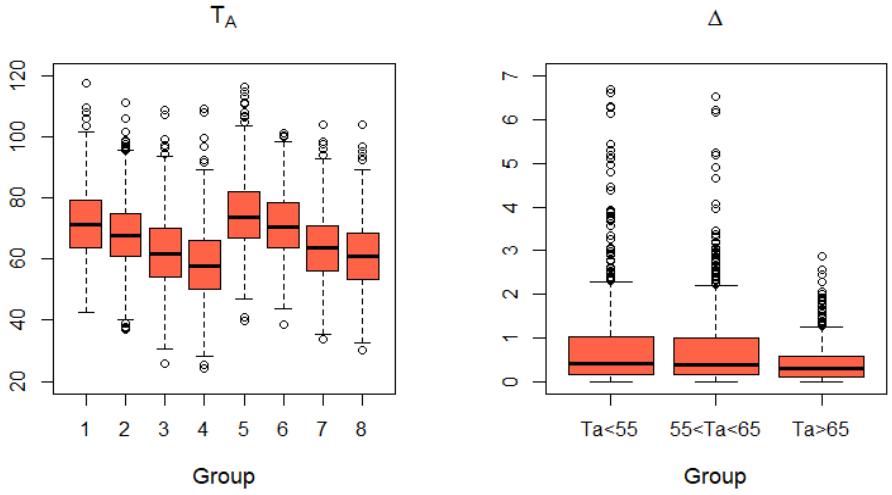



Figure 1.13: Predictive distributions for $T_A$ in each covariate group and for $\Delta$ given the observed value of $T_A$.

### 1.6.5 Sensitivity analysis for the ABC metric

We compare the results when using Metric 1 and Metric 2 on a "target" sample dataset simulated from the "Gamma + piecewise Exponential" model under some plausible parameter values. Metric 2 was built by considering the $L^2$-norm of the vector of the 32 test statistics, as described in the main body of the article, i.e. selecting those parameter combinations that lead to the minimum $L^2$-distance between the test statistics and the point $(0, \ldots, 0)$.

For each metric, we report the results obtained with and without the local-linear-regression adjustment. In each of the four procedures, the retention rate was chosen via a leave-one-out cross-validation procedure. Table 1.9 shows the posterior modes together with the 95% highest posterior density intervals (HPDI). The regression adjustment improves the performance of Metric 1 drastically, while it does not have a clear effect for Metric 2, which shows in both cases a poor fit to the (known) parameter values used to generate the data.

The regression-adjusted version of Metric 1 is strongly preferable over the other three options, showing a great ability to recover all the parameter values with precision. Figure 1.14 shows a graphical comparison between the resulting posterior distributions and the true parameter values. The plots show that the posterior distributions are located around the true value of each parameter, and that in most cases have a much smaller variance than the corresponding prior distribution.

This example confirmed our preference of using Metric 1 with regression adjustment in the analysis of the motivating data.

### 1.6.6 Distribution of the observed age at detection within a screening program: an analytical example

In this section, we show an analytical example of how the exact distributions of the observed age at asymptomatic and symptomatic detections can be computed.

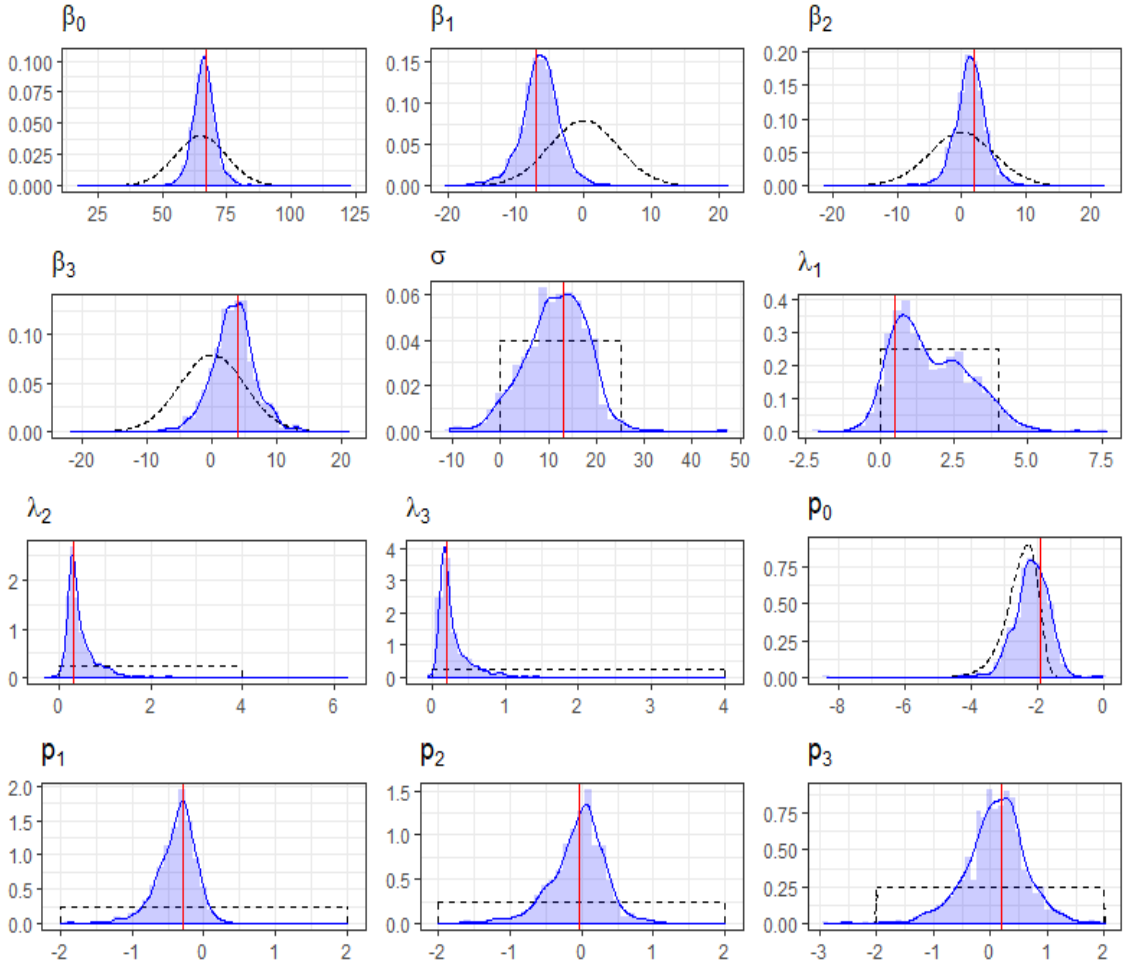



Figure 1.14: True values (red vertical lines), prior (black dashed lines), and local linear regression adjusted posterior (blue histograms and blue solid lines) densities obtained using Metric 1 and the "Gamma + piecewise Exponential" model.

| | True | Metric 1 | | Metric 2 | |
|---|---|---|---|---|---|
| | | Unadjusted | Adjusted | Unadjusted | Adjusted |
| $\beta_0$ | 67 | 65.25 (49.06,79.55) | 66.15 (57.53,74.75) | 64.72 (52.15,75.53) | 69.78 (67.89,71.58) |
| $\beta_1$ | -7 | -0.27 (-9.76,8.16) | -6.59 (-11.74,-0.21) | -1.67 (-9.30 ,7.57) | -13.73 (-16.08,-11.13) |
| $\beta_2$ | 2 | 0.28 (-8.83,8.34) | 1.24 (-3.22,5.92) | 0.61 (-9.32,8.48) | 5.87 (2.12,8.93) |
| $\beta_3$ | 4 | 0.46 (-8.94,8.89) | 4.26 (-2.86,10.22) | 0.98 (-8.39,8.65) | 5.49 (-0.48,10.75) |
| $\sigma$ | 13 | 19.65 (5.82,24.83) | 13.95 (-1.87,21.69) | 21.83 (10.46,24.98) | 16.35 (8.94,19.96) |
| $\lambda_1$ | 0.5 | 0.33 (0.10,3.75) | 0.81 (-0.20,4.19) | 0.66 (0.14,3.79) | 0.69 (0.00,2.87) |
| $\lambda_2$ | 0.3 | 0.16 (0.10,2.15) | 0.29 (-0.01,1.17) | 0.60 (0.11 ,2.25) | -0.34 (-0.37,-0.29) |
| $\lambda_3$ | 0.2 | 0.14 (0.10,1.60) | 0.17 (-0.01,0.79) | 0.47 ( 0.13 ,1.14) | 0.06 (0.04,0.10) |
| $p_0$ | -1.9 | -1.96 (-3.13,-1.04) | -2.23 (-3.20,-1.29) | -1.8 (-2.61,-0.98) | -1.87 (-2.23,-1.43) |
| $p_1$ | -0.3 | -0.29 (-1.96,1.16) | -0.29 (-0.99,0.10) | -0.31 (-1.75 ,0.97) | -0.25 (-0.84,0.68) |
| $p_2$ | -0.03 | -0.13 (-1.93,1.49) | 0.04 (-0.84,0.70) | 0.11 (-1.49 ,1.40) | 0.16 (-2.09,2.38) |
| $p_3$ | 0.2 | 0.12 (-1.67,1.77) | 0.27 (-0.97,1.24) | 0.21 (-1.19 ,1.71) | 1.19 (0.86,1.66) |

Table 1.9: Posterior modes and 95% highest posterior density intervals (HPDI).

Let $B$ indicate the calendar time of birth. We assume a homogeneous Poisson process for the births, so that conditionally on the number of events the times are distributed uniformly over a time interval, which we take as being $[\text{Bmin}, 0)$, i.e. $B \sim U(Bmin, 0)$.
Recall the usual definitions of the potential values $(T_A, T_S)$ associated with each individual in the population, where $T_A$ is the age at which asymptomatic detectability starts and $T_S$ is the age at which symptomatic detectability starts (and symptomatic detection occurs if disease is not detected asymptomatically prior to $T_S$). In addition, $T_D$ indicates the age at death in the absence of breast cancer. We assume stationarity with respect to birth time, and in particular:

$$\begin{aligned} T_A &= +\infty \text{ w.p. } (1-p_A) \text{ and } T_A \sim N(\mu_A, \sigma_A^2) \text{ w.p. } p_A \text{ (call the latter density } f_A^*); \\ T_S &= T_A + \Delta, \text{ such that } \Delta \sim Exp(\lambda) \text{ and } \Delta \perp\!\!\!\perp T_A; \\ T_D &\sim N(\mu_D, \sigma_D^2) \text{ and } T_D \perp\!\!\!\perp (T_A, T_S). \end{aligned}$$

We describe the effect of the selection process that leads to the observation of either $\widetilde{T}_A$ or of $T_S$, where $\widetilde{T}_A$ is the age at asymptomatic detection as determined from a screening

examination, as defined below.

Bias in the observed data arise from three sources: (i) Selection into the study (in particular, $B + T_S > 0$); (ii) Varying screening frequency and/or compliance as a function of age; and (iii) Right censoring due to ending the study. Here we focus on the first kind.

The sampling design is as follows: at calendar time zero subjects enter the study. The criteria for entry are being alive and not having shown symptoms of the disease yet. In other words, $\{\{B+T_D > 0\} \cap \{B+T_S > 0\}\}$. So subjects enter at the same time but with different ages, as described by the random variable $-B$.

Once in the study, subjects are monitored by screening examination every *gap* years, exactly. We let the screening examinations continue until death. Below we assume perfect compliance to the screening visits. Each screening examination clearly occurs only if the subject is alive at that time, i.e. if $B + T_D$ is greater than the calendar time of that screening examination. At each examination the subject can be found to be negative (indicating that $T_A$ has not occurred yet) or positive (indicating that $T_A$ is prior to the examinations schedule. In the latter case the new variable $\widetilde{T}_A$ is set equal to the time of the examination.

Notably, one reaches the examination time only if $T_S$ has not occurred yet. Indeed, if $T_S$ occurs prior to the examination time, then one would not observe $\widetilde{T}_A$ but rather $T_S$ itself.

For each observed $\widetilde{T}_A$ we expect the value to always be larger than the corresponding $T_A$ for that subject, due to the non-continuous monitoring performed by the screening. We also expect an indirect effect of the selection into the study, since conditioning on $T_S > 0$ carries with it a rather high (but not equal to one) probability that $T_A > 0$ as well. The screening process produces the observation of some $\widetilde{T}_A$ *instead of* the $T_S$ that would have been observed had screening not been performed. Overall, the overall effect of these phenomena is not clear-cut.

Our goal is to derive, under the assumptions above, the exact distribution of the two *observed* variables $\widetilde{T}_A$ and $T_S$.

**Distribution of the observed values of the variable $T_S$.**

We derive the distribution of the random variable $T_S|T_S$ is observed. Let $B = b$. For any $t > -b$, The event $\{T_S \in [t, t+dt), T_S \text{ obs}\}$ is equivalent to the event

$$A_t = \Bigg\{ T_S \in [t, t+dt), b + T_D > 0, b+T_S > 0, b + T_D > b + T_S, \\ b + T_A \in \left( \left\lfloor \frac{b+t}{gap} \right\rfloor gap, b + t \right) \Bigg\}, \tag{1.1}$$

which describes the fact that for $T_S$ to be observed and to be equal to $t$, the subject should have entered the study, be alive at calendar time $b + T_S$, and be such that the start of asymptomatic detectability of the disease should fall between the last screening examination before calendar time $b + t$ and time $b + t$. Indeed, screening examinations occur at calendar times $(j \cdot gap)$ for $j = 0, 1, 2, \ldots$, and $\left\lfloor \frac{u}{gap} \right\rfloor gap = j \cdot gap$ for $u \in [j \cdot gap, (j+1) \cdot gap)$. For ease of notation we call $l(u) = \left\lfloor \frac{u}{gap} \right\rfloor gap$.

The event $\{b + T_D > b + T_S\} \subseteq \{b + T_D > 0\}$, so that only the former event needs to be included. Note that the event $\{b + T_S > 0\}$ is such that the event in (1.1) is empty for $t < -b$, so that we only need consider its probability for $t > -b$.

Given $B = b$ (with $B$ independent of all other random variables) we have

$$\begin{aligned} P(A_t|B = b) &= P(T_D > t) P\left(T_S \in [t, t+dt), b + T_A \in (l(b+t), b+t)\right) \cdot \mathbb{1}\big(t > -b\big) \quad (1.2) \\ &= S_{T_D}(t) P\left(T_S \in [t, t+dt), T_A \in (l(b+t) - b, t)\right) \cdot \mathbb{1}\big(t > -b\big) \end{aligned}$$

The term $P(T_D > t) = S_{T_D}(t)$ corresponds to the first event in (1.1), and the fact that we consider all of these expressions as $dt$ tends to zero, so that $f_{T_S, T_S \text{obs}}(t)$ can be obtained from $(dt) f_{T_S, T_S \text{obs}}(t) \approx P(A_t)$.

Now, let us focus on the last term of (1.2). This can be written as a one-dimensional integral

$$\begin{aligned}
&P(T_S \in [t, t+dt), T_A \in (l(b+t)-b, t)) \approx (dt) \int_{l(b+t)-b}^{t} f_{T_A,T_S}(u,t)du \\
&= (dt) \int_{l(b+t)-b}^{t} f_{T_S|T_A}(t|u)\, p_A f^*_{T_A}(u)du = (dt) \int_{l(b+t)-b}^{t} f_{T_S|T_A}(t|u)\, p_A f^*_{T_A}(u)du \\
&= (dt) \int_{l(b+t)-b}^{t} f_{\Delta|T_A}(t-u|u)\, p_A f^*_{T_A}(u)du \overset{\perp\!\!\!\perp}{=} (dt) \int_{l(b+t)-b}^{t} f_{\Delta}(t-u)\, p_A f^*_{T_A}(u)du \\
&= (dt)\, p_A \int_{l(b+t)-b}^{t} \mathrm{e}^{-\lambda(t-u)} \frac{1}{\sqrt{2\pi}\sigma_A} \mathrm{e}^{-\frac{1}{2\sigma_A^2}(u-\mu_A)^2} du
\end{aligned} \tag{1.3}$$

Expanding the square in the exponent and collecting the terms allows one to recognize the kernel of the $N(\mu^*, \sigma_A^2)$ density, with $\mu^* = \mu_A + \lambda\sigma_A^2$. Call the associated cumulative distribution function $\Phi^*$. After some algebra we obtain

$$P(A_t|B=b) \approx (dt) S_{T_D}(t)\, p_A \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] [\Phi^*(t) - \Phi^*(l(b+t)-b)] \cdot \mathbb{1}\big(t > -b\big). \tag{1.4}$$

We now integrate with respect to the (uniform) birth time $B$:

$$\begin{aligned}
&P(T_S \in [t, t+dt), T_S\text{obs}) = \int_{Bmin}^{0} P(T_S \in [t, t+dt), T_S\text{obs}|B=b) f_B(b)db \\
&= \int_{Bmin}^{0} P(A_t|B=b) f_B(b)db \\
&= (dt) \int_{Bmin}^{0} S_{T_D}(t)\, p_A \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] [\Phi^*(t) - \Phi^*(l(b+t)-b)] \frac{1}{-Bmin} \cdot \mathbb{1}\big(b > -t\big) db \\
&= (dt) \frac{p_A\,[1 - \Phi_{T_D}(t)]}{-Bmin} \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] \int_{\max(Bmin, -t)}^{0} [\Phi^*(t) - \Phi^*(l(b+t)-b)]\, db,
\end{aligned}$$

so that taking limits as $dt \to 0$ yields the final form

$$f_{T_S}(t; T_S\text{obs}) = \frac{p_A\,[1 - \Phi_{T_D}(t)]}{-Bmin} \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] \int_{\max(Bmin, -t)}^{0} [\Phi^*(t) - \Phi^*(l(b+t)-b)]\, db \tag{1.5}$$

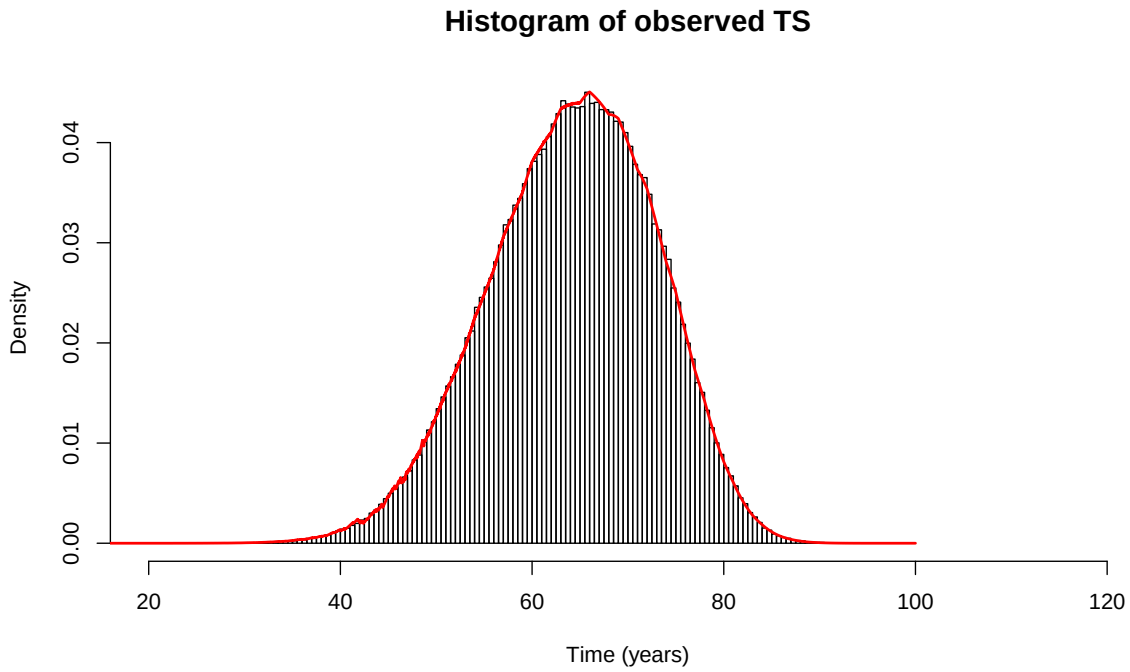



Figure 1.15: Sample of observed values $T_S$ in a simulated sample from 10 million initial subjects. Parameter values were $p_A = 0.15$, $\mu_A = 65$, $\sigma_A^2 = 100$, $\lambda = 1/3$, $\mu_D = 80$, $\sigma_D^2 = 25$, $gap = 3$, and $Bmin = -50$. The red curve represents the density function in (1.5)

which requires numerical integration of the normal cumulative distribution function $\Phi^*$. Finally, note from (1.1) that the distribution in (1.5) is actually the distribution of $T_S$ and $T_S$ observed. The desired distribution of $T_S$ conditionally on $T_S$ being observed is obtained by taking the ratio of (1.5) and the normalizing constant, which can also be obtained numerically.

Figure 1.15 shows a sample output of the simulations perfomed in R. In particular, the density function in (1.5) is superimposed on the histogram of the observed values of $T_S$ from a simulate sample from the model.

**Distribution of the observed variable $\widetilde{T}_A$.**

We define the new random variable $\widetilde{T}_A$ which indicates the time from birth until an observed asymptomatic diagnosis, i.e. a detection that has occurred at one of the planned screening examinations. Receiving such a diagnosis occurs when $T_A$ is prior to the screening examination, while $T_S$ is after it. In addition, the subject should be in the study, and not have died prior to the screening examination. If these events do not happen, then $\widetilde{T}_A$ is not observed (N/A). The distribution of the observed $\widetilde{T}_A$ can be obtained in closed form, and computed without numerical integration except for the (readily available) cumulative distribution function of a normal random variable, and for the normalizing constant.
Note that $\widetilde{T}_A$ can only be such that $B + \widetilde{T}_A$ is equal to one of the calendar times at which screening examinations are offered, i.e. calendar times $j \cdot gap$ for $j = 0, 1, 2, \ldots$. Hence

$$P(\widetilde{T}_A \in [t, t+dt), \widetilde{T}_A \text{obs}) = \sum_{j=0}^{+\infty} P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = j \cdot gap, \widetilde{T}_A \text{obs}). \tag{1.6}$$

Let us first focus on $j \geq 1$. For a fixed $j$ the following events are identical:

$$\{B + \widetilde{T}_A = j \cdot gap\} = \left\{ \left\lfloor \frac{B + T_A}{gap} \right\rfloor = j-1 \right\} \cap \{B + T_S > j \cdot gap\}.$$

Since $B + t = j \cdot gap$, $j < \lfloor \frac{t}{gap} \rfloor$ must hold, so that the sum in (1.6) only needs to run until $\lfloor \frac{t}{gap} \rfloor$.

$$\begin{aligned} &\left\{ B + \widetilde{T}_A \in [t, t+dt), \widetilde{T}_A \text{obs} \right\} = \\ &\left\{ B + T_S > 0, B + T_D > 0, \left\lfloor \frac{B + T_A}{gap} \right\rfloor = j-1, B + T_S > j \cdot gap, B + T_D > j \cdot gap, B + t = j \cdot gap \right\} \end{aligned} \tag{1.7}$$

where the first two events on the right hand side can be dropped as they are captured in their intersection with later events. Now, let $B = b$. Recall that $B$ is independent of all other variables. Then, we can compute the conditional probability

$$\begin{aligned}
&P\,(\frac{B+T_S}{gap} > j, \frac{B+T_D}{gap} > j, \left\lfloor \frac{B+T_A}{gap} \right\rfloor = j-1 \,\middle|\, B = b\Big) \\
&= P\left(\frac{b+T_A+\Delta}{gap} > j, \left\lfloor \frac{b+T_A}{gap} \right\rfloor = j-1, b+T_D > j\cdot gap, b = j\cdot gap - t\right) \\
&= P\left(\frac{b+T_A}{gap} > j - \frac{\Delta}{gap}, j-1 \leq \frac{b+T_A}{gap} < j, b+T_D > j\cdot gap, b = j\cdot gap - t\right) \\
&= P\left(\max\left(j-1, j-\frac{\Delta}{gap}\right) \leq \frac{b+T_A}{gap} < j, b+T_D > j\cdot gap, b = j\cdot gap - t\right) \\
&= P\left(\max\left(j-1, j-\frac{\Delta}{gap}\right) \leq \frac{b+T_A}{gap} < j\right) P(T_D > -b + j\cdot gap)\cdot \mathbb{1}\big(b = j\cdot gap - t\big).
\end{aligned} \tag{1.8}$$

Let $Y = (b+T_A)/gap$ and $T = \Delta/gap$. From our assumptions it follows that $Y \perp\!\!\!\perp T$, and that

$$Y \sim N\left(\frac{\mu_A + b}{gap}, \frac{\sigma_A^2}{gap^2}\right) w.p.p_A, and + \infty w.p.(1-p_A); T \sim Exp(\lambda\cdot gap).$$

The expression in (1.8) is therefore equal to

$$\begin{aligned}
&S_{T_D}(-b + j\cdot gap)\, P(\max(j-1, j-T) \leq Y < j)\cdot \mathbb{1}\big(b = j\cdot gap - t\big) \\
&= S_{T_D}(-b + j\cdot gap)\cdot \mathbb{1}\big(b = j\cdot gap - t\big) p_A \int_{j-1}^{j} \left[\int_{j-y}^{+\infty} f_{T|Y}(t|y)dt\right] f_Y(y)dy \\
&= S_{T_D}(t)\cdot \mathbb{1}\big(b = j\cdot gap - t\big) p_A \int_{j-1}^{j} S_T(j-y) f_Y(y)dy \\
&= S_{T_D}(t)\cdot \mathbb{1}\big(b = j\cdot gap - t\big) p_A \int_{j-1}^{j} \mathrm{e}^{-\lambda gap(j-y)} \frac{gap}{\sqrt{2\pi}\sigma_A} e^{-\frac{gap^2}{2\sigma_A^2}\left(y - \frac{\mu_A+b}{gap}\right)^2} dy \\
&= S_{T_D}(t)\cdot \mathbb{1}\big(b = j\cdot gap - t\big) p_A\, \mathrm{e}^{-\lambda\cdot j\cdot gap} \exp\left[-\frac{gap^2}{2\sigma_A^2}\left(\left(\frac{\mu_A+b}{gap}\right)^2 - \widetilde{\mu}_b^2\right)\right]\left[\widetilde{\Phi}_b(j) - \widetilde{\Phi}_b(j-1)\right],
\end{aligned} \tag{1.9}$$

where $\widetilde{\Phi}_b$ is the cdf of the $N\left(\widetilde{\mu}_b, \frac{\sigma_A^2}{gap^2}\right)$ distribution, with $\widetilde{\mu}_b = \frac{\sigma_A^2}{gap}\lambda + \frac{\mu_A+b}{gap}$.

Hence we finally obtain

$$\begin{aligned}
&P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = j \cdot gap, \widetilde{T}_A \text{obs}) \\
&\quad = P\left(\frac{B+T_S}{gap} > j, \frac{B+T_D}{gap} > j, \left\lfloor \frac{B+T_A}{gap} \right\rfloor = j-1 \,\middle|\, B = b\right) \cdot f_B(b) \\
&\quad = \frac{S_{T_D}(t)}{(-\text{Bmin})} p_A \, \mathrm{e}^{-\lambda \cdot j \cdot gap} \, \exp\left[-\frac{gap^2}{2\sigma_A^2}\left(\left(\frac{\mu_A + j \cdot gap - t}{gap}\right)^2 - \widetilde{\mu}^2_{(j \cdot gap - t)}\right)\right] \\
&\qquad \cdot \left(\widetilde{\Phi}_{(j \cdot gap - t)}(j) - \widetilde{\Phi}_{(j \cdot gap - t)}(j-1)\right) \cdot \mathbb{1}\big(j \cdot gap < t < j \cdot gap - \text{Bmin}\big).
\end{aligned} \tag{1.10}$$

Let us now turn to the case in which the disease is detected during the first screening examination after entry into the study, i.e. the case of $B + \widetilde{T}_A = 0$. This corresponds to the first element ($j = 0$) of the series in (1.6). Recall that for a subject to find herself in this situation she must have $B + T_A < 0$ and $B + T_S > 0$ (and be alive at calendar time zero). The probability of such event is therefore equal to

$$\begin{aligned}
&P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = 0, \widetilde{T}_A \text{obs}) \\
&\quad = P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = 0, \widetilde{T}_A \text{obs} | B = b) \cdot f_B(b) \\
&\quad \approx P(b + \widetilde{T}_A = 0, b + T_A \leq 0, b + T_S > 0, b + T_D > 0) \cdot \mathbb{1}\big(b = -t\big) \frac{\mathbb{1}\big(b \in (\text{Bmin}, 0)\big)}{|\text{Bmin}|} \\
&\quad = \frac{S_{T_D}(t)}{|\text{Bmin}|} P(T_A \in (-\infty, -b], \Delta > -b - T_A) \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \\
&\quad \overset{\Delta \perp\!\!\!\perp T_A}{=} p_A \, \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \int_{-\infty}^{t} \left[\int_{t-u}^{+\infty} f_\Delta(\delta) d\delta\right] f_{T_A}(u) du \\
&\quad = p_A \, \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \int_{-\infty}^{t} S_\Delta(t-u) f_{T_A}(u) du,
\end{aligned} \tag{1.11}$$

Therefore, with a bit of rearranging,

$$\begin{aligned}
&P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = 0, \widetilde{T}_A\text{obs}) \\
&\quad = p_A \, \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-u)} \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left[-\frac{1}{2\sigma_A^2}\,(u-\mu_A)^2\right] du \\
&\quad = p_A \, \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \mathrm{e}^{-\lambda t}\, \Phi^*(t) \exp\left[-\frac{1}{2\sigma_A^2}\left(\mu_A^2 - (\mu^*)^2\right)\right] \\
&\quad = p_A \, \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \mathrm{e}^{-\lambda t}\, \Phi^*(t) \exp\left(\frac{\lambda^2\sigma_A^2}{2} + \lambda\mu_A\right),
\end{aligned} \tag{1.12}$$

where $\Phi^*$ is the cdf of the $N(\mu^*, \sigma_A^2)$ random variable, with $\mu^* = \mu_A + \lambda\sigma_A^2$. We now put all the terms together to obtain

$$\begin{aligned}
P(\widetilde{T}_A \in [t, t+dt), \widetilde{T}_A\text{obs}) \approx (dt)\, p_A \, \frac{S_{T_D}(t)}{|\text{Bmin}|} \Bigg\{ & \mathrm{e}^{-\lambda t}\, \Phi^*(t) \exp\left(\frac{\lambda^2\sigma_A^2}{2} + \lambda\mu_A\right) \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \\
& + \sum_{j=1}^{\lfloor \frac{t}{gap} \rfloor} \left[ \mathrm{e}^{-\lambda \cdot j \cdot gap} \exp\left[ -\frac{gap^2}{2\sigma_A^2} \left( \left( \frac{\mu_A + j\cdot gap - t}{gap} \right)^2 - \widetilde{\mu}^2_{(j\cdot gap - t)} \right) \right] \cdot \right. \\
& \left. \cdot \left[ \widetilde{\Phi}_{(j\cdot gap - t)}(j) - \widetilde{\Phi}_{(j\cdot gap - t)}(j-1) \right] \cdot \mathbb{1}\big(t \in (j\cdot gap, j\cdot gap - \text{Bmin})\big) \right] \Bigg\}.
\end{aligned} \tag{1.13}$$

Figure 1.16 shows the sample output of simulations performed in R. In particular, the density function in (1.13) is superimposed on the histogram of the observed values of $\widetilde{T}_A$ from one simulated sample from the model.

Note that, conditionally on $B = b$, the distribution of $\widetilde{T}_A$ is discrete. In particular, if $\widetilde{T}_A$ is observed, then it takes the countable number of values $-b + j\cdot gap$ for $j = 0, 1, 2, \ldots$, plus the additional value $N/A$ (which we may also set as $+\infty$. Conditionally on $B = b$, these

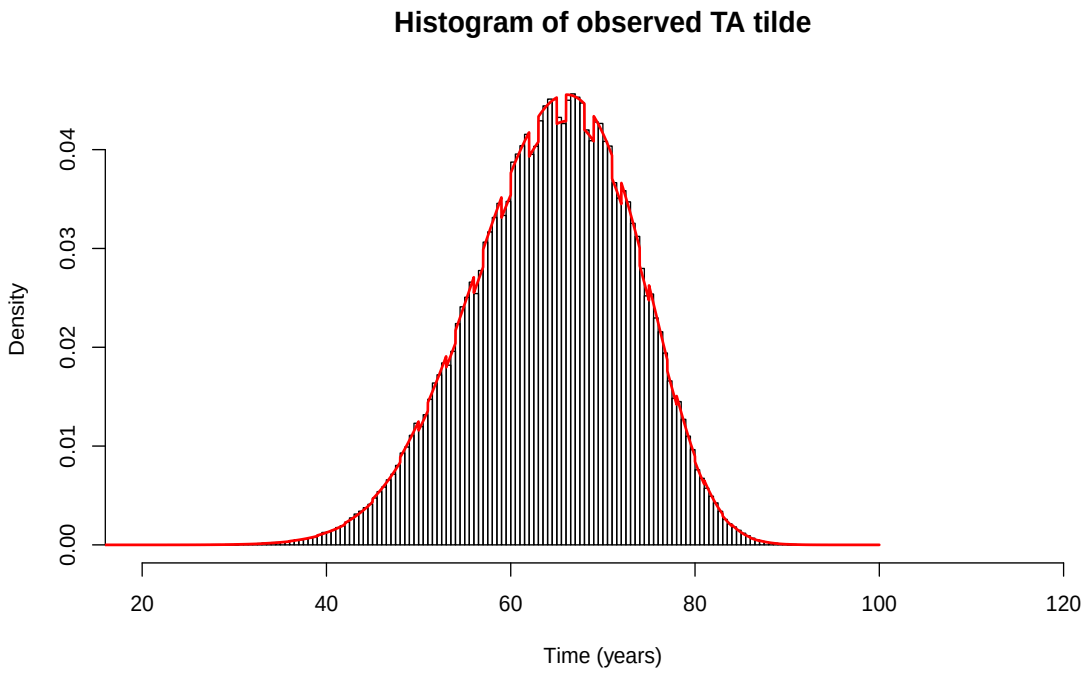



Figure 1.16: Sample of observed values $\widetilde{T}_A$ in a simulated sample from 10 million initial subjects. Parameter values were $p_A = 0.15$, $\mu_A = 65$, $\sigma_A^2 = 100$, $\lambda = 1/3$, $\mu_D = 80$, $\sigma_D^2 = 25$, $gap = 3$ and $Bmin = -50$. The red curve represents the density function in (1.13).

values are taken with the following probabilities:

$$\widetilde{T}_A = \begin{cases} -b & P(b+T_A<0, b+T_S>0, b+T_D>0) \\ -b+j\cdot gap & P(\lfloor \frac{b+T_A}{gap} \rfloor = j-1, \frac{b+T_S}{gap}>j, \frac{b+T_S}{gap}>j) \text{forj} = 1,2,\ldots \\ N/A & k. \end{cases}$$

Hence the probabilities of the non-$N/A$ values are obtained as described above from the distributions of $(T_a, \delta)$ and $T_D$.

The conditional probability $k = P(\widetilde{T}_A = N/A|B=b)$ can be obtained as one minus the series of the probabilities of the other values taken by the random variable. The marginal probability $P(\widetilde{T}_A = N/A)$ is then

$$P(\widetilde{T}_A = N/A) \;=\; \int_{Bmin}^{0} P(\widetilde{T}_A = N/A) f_B(b)db = \int_{Bmin}^{0} P(\widetilde{T}_A = N/A|B=b) f_B(b)db,$$

that, as we have mentioned, can be obtained numerically.

### Marginal probability of being in study

Lastly, we obtain the marginal probability that a randomly selected member of the population is included into the sample. Indeed,

$$\begin{aligned} P(B+T_A+\Delta>0, B+T_D>0) &= \int_{Bmin}^{0} P(B+T_A+\Delta>0, B+T_D>0|B=b) f_B(b)db \\ &= \int_{Bmin}^{0} P(b+T_A+\Delta>0, b+T_D>0) f_B(b)db \\ &= \int_{Bmin}^{0} P(b+T_A+\Delta>0) P(b+T_D>0) f_B(b)db, \end{aligned}$$

where

$$
\begin{aligned}
P(b+T_A+\Delta>0) &= 1-P(\Delta<-b-T_A) = 1-p_A\int_{-\infty}^{-b}\int_0^{-b-t} f_\Delta(\delta)d\delta f_{T_A}(t)dt_A \\
&= 1-p_A\int_{-\infty}^{-b}\left[1-e^{-\lambda(-b-t)}\right] f_{T_A}(t)dt \\
&= 1-p_A\,\Phi_{T_A}(-b)+p_A\,e^{\lambda b}\int_{-\infty}^{-b} e^{\lambda t} f_{T_A}(t)dt \\
&= 1-p_A\,\Phi_{T_A}(-b)+p_A\,e^{\lambda b}\int_{-\infty}^{-b} e^{\lambda t}\frac{1}{\sqrt{2\pi}\sigma_A}e^{-\frac{1}{2\sigma_A^2}(t-\mu_A)^2}dt \\
&= 1-p_A\,\Phi_{T_A}(-b)+p_A\,e^{\lambda b}\exp\left(\frac{\sigma_A^2\lambda^2}{2}+\mu_A\lambda\right)\int_{-\infty}^{-b}\frac{1}{\sqrt{2\pi}\sigma_A}e^{-\frac{1}{2\sigma_A^2}(t-(\mu_A+\sigma_A^2\lambda))^2}dt \\
&= 1-p_A\,\Phi_{T_A}(-b)+p_A\,e^{\lambda b}\exp\left(\frac{\sigma_A^2\lambda^2}{2}+\mu_A\lambda\right)\Phi_Z\left(\frac{-b-(\mu_A+\sigma_A^2\lambda)}{\sigma_A}\right) \\
&= 1-p_A\,\Phi_Z\left(\frac{-b-\mu_A}{\sigma_A}\right)+p_A\,\exp\left(\frac{\lambda b+\sigma_A^2\lambda^2}{2}+\mu_A\lambda\right)\Phi_Z\left(\frac{-b-(\mu_A+\sigma_A^2\lambda)}{\sigma_A}\right),
\end{aligned}
$$

and therefore

$$
\begin{aligned}
P(B+T_A+\Delta>0, B+T_D>0) &= \int_{Bmin}^{0} P(b+T_A+\Delta>0)P(b+T_D>0)\,f_B(b)\,db \\
&= \int_{Bmin}^{0} P(b+T_A+\Delta>0)P(T_D>-b)\,f_B(b)\,db \\
&= \int_{Bmin}^{0} P(b+T_A+\Delta>0)\left(1-\Phi_{T_D}(-b)\right)f_B(b)db \\
&= \frac{-1}{Bmin}\int_{Bmin}^{0} P(b+T_A+\Delta>0)\left[1-\Phi_Z\left(\frac{-b-\mu_D}{\sigma_D}\right)\right]db,
\end{aligned}
$$

which can be easily calculated numerically in R.

# Chapter 2

# An exploration of ABC and dissimilarities[2]

## 2.1 Introduction

As we have already pointed out in Chapter 1, the choice of the summary statistics and of the distance function are fundamental issues when performing inference through Approximate Bayesian Computation (ABC). Ideally, the summary statistics should be sufficient statistics for the considered model. However, in practice it is often the case that sufficient statistics are unavailable or unknown, and as a consequence other (hopefully) *informative* data summaries should be used.

Recall that, given a parameter $\boldsymbol{\theta} \in \Theta$ and a parametric model $f(\boldsymbol{y}|\boldsymbol{\theta})$, a statistic $\boldsymbol{S}(\boldsymbol{y})$ is sufficient for $\boldsymbol{\theta}$ if and only if the conditional distribution of the data given the statistic, $f(\boldsymbol{y}|\boldsymbol{S}(\boldsymbol{y}))$, does not depend on $\boldsymbol{\theta}$. Equivalently, sufficiency can also be defined in a Bayesian sense, by requiring that the posterior distribution of the parameter obtained by conditioning on the statistic is the same as the one obtained by conditioning on the full data: $\pi(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y}) =$

[2]Joint work with Marco Bonetti and Raffaella Piccarreta

$\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Following this definition, an empirical way to check if a statistic is "close" to sufficiency could be the following. Consider a (small) set of values of $\boldsymbol{\theta}$, e.g. a sample from its prior distribution, $\{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K}\}$. For each $k = 1, \ldots, K$, generate *many* samples from the model $f(\boldsymbol{y}|\boldsymbol{\theta_k})$. Note that there is no need to know the explicit form of the likelihood function of the model, but it is enough to be able to generate data from it (as it happens when applying ABC). From all samples we want to see whether the conditional distribution of the sample given $\boldsymbol{S}(\boldsymbol{y}) = \boldsymbol{s}$ (for different values $\boldsymbol{s}$) does not depend on $\boldsymbol{\theta}$. In the univariate case (i.e. when $s \in \mathbb{R}$), for values $s$ equal to, say, the deciles of the marginal distribution of $S$, one can extract the samples with similar values $s$, and estimate a few moments of the observed data for different values of $\boldsymbol{\theta}$. In this way, we could in principle measure how much, for each $s$, the estimated moments vary across the values of $\boldsymbol{\theta}$. Limited variation across a range of values of $\boldsymbol{\theta}$ would suggest that $\boldsymbol{S}$ is "close" to sufficiency, and that it may be used reliably in ABC. However, such an algorithm would clearly suffer from the curse of dimensionality, and it would not be feasible in practice in most cases.

To appreciate the effect of using a "not very sufficient" statistic in ABC, consider the following simple example. The example will illustrate how the goodness of the approximation of ABC varies with respect to summary statistics with increasing information content.

Consider the iid sample $Y_1, \ldots, Y_n \sim f_Y(y; \theta)$, and let $S = S(\boldsymbol{Y})$ be a possibly multidimensional statistic, function of that sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. For simplicity consider $\theta \in \mathbb{R}$. Let $I_Z(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \log\left(g_Z(Z;\theta)\right)\right)$ be Fisher's information of a statistic $Z \sim g_Z(z;\theta)$. It is well known that

$$I_S(\theta) \leq I_{\boldsymbol{Y}}(\theta) = n\, I_{Y_1}(\theta)$$

and that $S$ is sufficient for $\theta$ iff $I_S(\theta) = I_{\boldsymbol{Y}}(\theta) = n\, I_{Y_1}(\theta)$ (see, e.g. [45]). We wish to relate the goodness of the approximation of the posterior distribution obtained with ABC when using a statistic $T$ to the fraction of the total information contained in the statistic. Consider

the sequence of summary statistics defined as the partial samples $\boldsymbol{S}_j = (Y_1, \ldots, Y_j)^T$ for $j = 1, \ldots, n$, so that

$$I_{Y_1}(\theta) = I_{\boldsymbol{S}_1}(\theta) < I_{\boldsymbol{S}_2}(\theta) < \ldots < I_{\boldsymbol{S}_n}(\theta) = n\, I_Y(\theta)$$

i.e. such that they contain an increasing fraction of the total amount of information.

If the sample mean is sufficient for $\theta$, one may replace the statistics $\boldsymbol{S}_j$ by the statistics $T_j = \frac{1}{j}\sum_{h=1}^{j} Y_h$, for $j = 1, \ldots, n$ since clearly $I_{T_j}(\theta) = I_{\boldsymbol{S}_j}(\theta)$. The statistics $T_1, T_2, \ldots, T_n$ therefore contain fractions of the total information $I_{\boldsymbol{Y}}(\theta)$ equal to $(\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} = 1)$. We use these statistics $T_j$ in the retention criterion for ABC.

Specifically, suppose that one wants to infer the mean $\mu$ of a normal random variable with known variance equal to 1, where the observed data $y_1, \ldots, y_n$, with $n = 100$, are independently generated from a $N(2, 1)$. We clearly do not need to resort to ABC for such a problem, but it will be a nice setup to experiment with. We simulate $K = 10000$ samples to perform an ABC estimation for $\mu$, i.e. for $k \in \{1, \ldots, K\}$, we generate $X_{k1}, \ldots, X_{kn} \overset{iid}{\sim} N(\mu, 1)$, conditionally on the value $\mu_k$ sampled from the prior distribution $\mu \sim N(0, 4)$. Let $T_j$, for $j = 1, \ldots, n$, be the set of summary statistics as defined above (the partial means), to be computed on each of the $K$ generated samples. Based on each of the $n$ summary statistics $T_j$ above and using the $L^2$-distance, we obtain $n$ different ABC approximate posterior distributions for $\mu$: $\widehat{\pi}_{ABC}(\mu|T_j)$, for $j = 1, \ldots, n$.

We then quantify and plot (see Figure 2.1) the distance between $\widehat{\pi}_{ABC}(\mu|T_n)$ and $\widehat{\pi}_{ABC}(\mu|T_j)$ using three different metrics:

(i) $\left|\frac{\widehat{E}(\mu|T_n) - \widehat{E}(\mu|T_j)}{\widehat{E}(\mu|T_n)}\right|$;

(ii) $\left|\frac{\widehat{sd}(\mu|T_n) - \widehat{sd}(\mu|T_j)}{\widehat{sd}(\mu|T_n)}\right|$;

(iii) $d_{K-S}\big(\widehat{\pi}_{ABC}(\mu|T_n), \widehat{\pi}_{ABC}(\mu|T_j)\big) = \left\|\widehat{F}_n(\mu|T_n) - \widehat{F}_j(\mu|T_j)\right\|_\infty$.

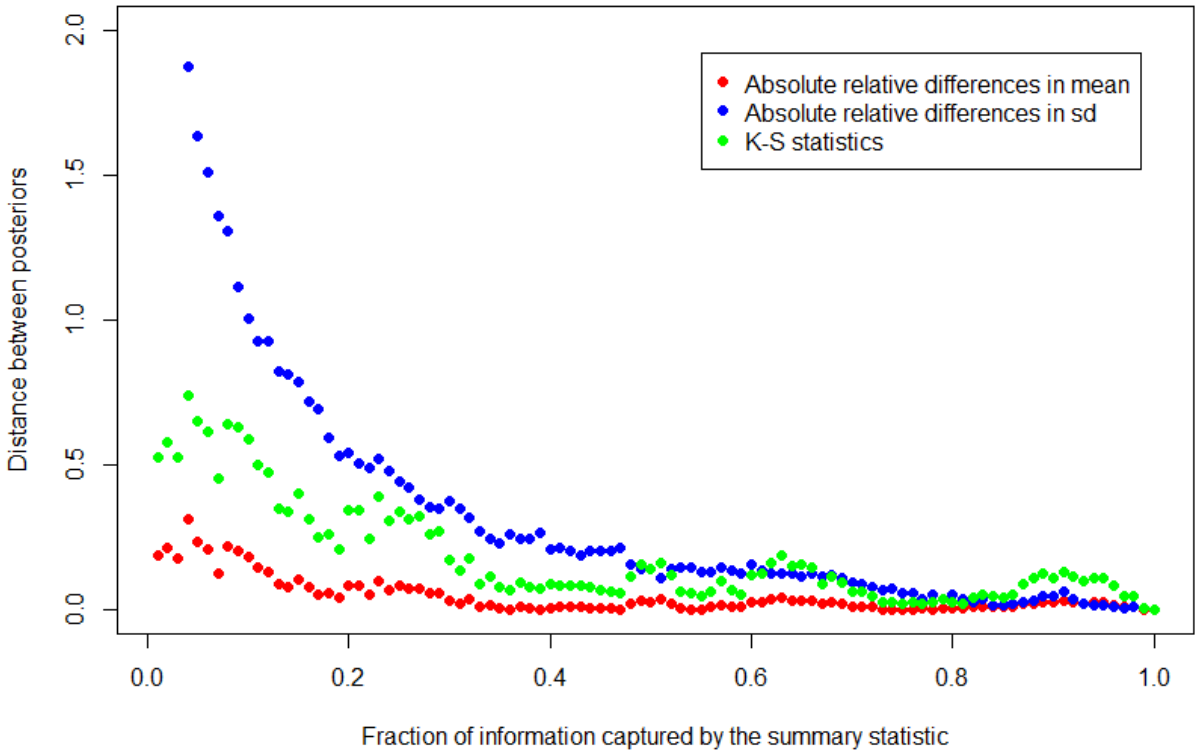



Figure 2.1: Distances between approximated posterior distributions obtained from ABC with a sequence of summary statistics having increasing fraction of information.

As expected, Figure 2.1 shows a clearly decreasing trend in the distances between the posterior distributions as the amount of information captured by the summary statistics increases. This experiment shows that even in this simple setting, using an inappropriate statistic in ABC may produce samples quite far from the true posterior distribution of the parameter.

For completeness, one should not forget that the *non-sufficiency* of the summary statistics is just one of the several levels of approximation occurring when performing ABC [57]. Indeed, unless the summary statistics are discrete (enough), in practice ABC retains parameter

values that have produced a non-perfect but "close" match between the summaries of the observed and generated data (and, notably, most of the generated datasets are discarded in the process). Also, a Monte Carlo error arises from using a sample from $\pi(\boldsymbol{\theta}|\boldsymbol{s})$ to estimate it. Lastly, as for any statistical analysis, every model is always an approximation to the real data-generating process.

However, the choice of the summary statistic on which to base retention of the generated parameter values probably remains the most delicate aspect of ABC.

In the literature on ABC, several techniques to select summary statistics or to reduce the dimensionality of the available summaries have been proposed ([14], [57] and [27]). For example, the authors of [27] introduced the idea of employing classification accuracy to measure the distance between two datasets. Indeed, the more two datasets are "similar" to each other, and the harder it will be for one to classify their observations as belonging to one or the other. Therefore, a distance function can be defined to be inversely proportional to the misclassification error in a test set of a classifier trained to distinguish between observed and generated data. As a selection criterion in ABC, the retained parameter values in that case would be those that yield almost chance-level discriminability (misclassification error equal to 0.5).

We now suggest a novel direction to define a dissimilarity measure between two datasets, based in turn on the collection of the pairwise dissimilarities between observations in the two datasets.

In Section 2.2 we present some available techniques that can be used to such goal. The statistics described in that section do not refer to a specific kind of data, and they can be applied to a variety of settings. Indeed, the choice of the pairwise dissimilarity function is left unspecified.

Section 2.3 is devoted to the description of a new estimation technique, alternative to ABC, but still likelihood-free and built from the same dissimilarity-based metric. The

proposed estimator is based on calibration ideas, and it is defined as the solution of a minimization problem.

In Section 2.4 we illustrate these ideas through the application of the new dissimilarity-based estimation procedure to a simple model: the bivariate normal case. We close with some discussion in Section 2.6.

## 2.2 Dissimilarity-based criteria

Consider an iid sample $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})^T$, where each $\boldsymbol{y_i}$ is possibly multivariate and contains information relative to the $i$-th individual.

As recalled above, given the (sufficient) statistic $S = \boldsymbol{S}(\boldsymbol{y})$, in ABC one retains the values $\theta$ that generate a sample with value of $S$ *close* to $\boldsymbol{S}(\boldsymbol{y}_{obs})$, with $\boldsymbol{y}_{obs}$ being the observed data. Below we focus, for a fixed value of $\boldsymbol{\theta}$, on the *whole* distribution $f_S(\boldsymbol{s}; \boldsymbol{\theta})$ of $S(\boldsymbol{Y})$. In the Bayesian setup, the distribution would be $f_{S|\theta}(\boldsymbol{s}|\theta)$, but that is not relevant here.
More generally, we define, for the observed $\boldsymbol{y}_{obs}$ and for a value $\boldsymbol{y}$ of the random vector $\boldsymbol{Y} \sim f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$, a quantity

$$T(S(\boldsymbol{y}); S(\boldsymbol{y}_{obs})),$$

that can be used to measure how far the two samples $\boldsymbol{y}$ and $\boldsymbol{y}_{obs}$ are. Note that this includes the case of $S$ being the identity transformation, i.e. the case $T(\boldsymbol{y}; \boldsymbol{y}_{obs})$. Our "Metric 2", which was introduced in Chapter 1, is an example of such a case (for other examples see e.g. [10] and [47]). Typically, in ABC one uses as a choice for $T$ the form

$$T(S(\boldsymbol{y}); S(\boldsymbol{y}_{obs})) = d\left(S(\boldsymbol{y}), S(\boldsymbol{y}_{obs})\right)$$

where $d(\boldsymbol{s}_1, \boldsymbol{s}_2)$ is a dissimilarity measure between its two arguments. For example, one may

set $d$ to be the $L^2$-norm in Euclidean space

$$d(\boldsymbol{s}_1, \boldsymbol{s}_2) = ||\boldsymbol{s}_1 - \boldsymbol{s}_2||.$$

"Metric 1", used for the analysis presented in Chapter 1, belongs to this family of discrepancy measures. Note also that the metrics based on summary statistics represent a subset in the larger set of all the possible criteria to measure discrepancy between datasets.

ABC then retains the values of $\boldsymbol{\theta}$ that have generated a sample $\boldsymbol{y}$ such that $T(\boldsymbol{y}, \boldsymbol{y}_{obs}) \approx 0$, i.e. such that $S(\boldsymbol{y}) \approx S(\boldsymbol{y}_{obs})$. The selected parameter values, that represent a sample from the approximated posterior distribution, can then be *adjusted* to account for the discrepancy between simulated and observed statistics.

Indeed, it is common to post-process the ABC output to improve the selected posterior sample by applying a so-called "regression adjustment." The idea is to regress each parameter (or to perform a multivariate regression with all the parameters as response vector) on the set of summary statistics and to apply a correction based on the difference between observed and simulated summaries.[6, 37].

Here, our aim is to introduce a new wide set of possibilities for the definition of $T(y; y_{\text{obs}})$. Depending on the kind of data, one can define a one-dimensional dissimilarity measure between two observations $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$, $d_{ij} = D(\boldsymbol{y_i}, \boldsymbol{y_j})$. For example, the dissimilarity might be computed as the $L^2$-distance, or by other distances defined on vector spaces (with the triangle inequality not being strictly necessary). The distribution of the dissimilarity between two randomly selected observations in a population is called the *interpoint distance distribution (IDD)* (see, e.g., [16] and [17]).

We propose to use the set of all pairwise dissimilarities $\{d_{ij}\}_{i,j=1,\dots,n}$ to define a function T that can be used to measure the distance between two datasets, for it to be used in ABC. One can distinguish among three kinds of dissimilarities:

i) within-group dissimilarities within the observed data, i.e. all the pairwise dissimilarities between observed subjects' data;

ii) within-group dissimilarities within the generated data i.e. all the pairwise dissimilarities between generated subjects' data;

iii) between-group dissimilarities, i.e. all the pairwise dissimilarities between an observed and a generated subject's data.

The first approach that we explore is to define $T(y; y_{\text{obs}})$ from the Wilcoxon-Mosler statistic (WM), see [44]. The idea is to contrast the within-group dissimilarities to the between-group dissimilarities through a rank-based test statistic. Specifically, the set including all the dissimilarities (both the within- and the between-group) is sorted in ascending order and ranked, and the test statistic WM is defined as

$$\text{WM} = \sum_{i \in \{obs\}} \sum_{j \in \{gen\}} \text{rank}\left(d_{ij}\right), \tag{2.1}$$

i. e. it is the sum of the ranks of the between-group dissimilarities.

Similarly to Wilcoxon's test, when observed and model-generated data are similar, the ranks of the between-group dissimilarities should be placed at random, compatibly with the dependence structure imposed by the use of the dissimilarities. In particular, under such null scenario, which corresponds to the use of the true parameter (and model) to generate the data, the pairwise dissimilarities are identically distributed - although not independent. One can show that under permutation of the $m + n$ group labels the statistic WM has expected value

$$E(\text{WM}) = \frac{mn}{2}\Big[\binom{m+n}{2} + 1\Big]. \tag{2.2}$$

The second moment is also known and an approach to testing with WM has been developed using permutation distribution inference, by repeatedly calculating WM on samples constructed

by permuting the group labels ([44]).

For our purposes, we do not need to perform testing of hypotheses, but only to quantify the distance between two datasets. In particular, we define the distance as the squared difference between the observed value of WM and its expected value under the null hypothesis that the observed and simulated data were generated by the same parameter value (Eq. (2.2)), i.e.

$$T(y; y_{\text{obs}}) = \Big[\text{WM} - E(\text{WM})\Big]^2.$$

Another approach to quantify the similarity between the observed and simulated data is based on the comparison of the estimated cumulative distribution function of the within-group dissimilarities with the cumulative distribution function of the between-group dissimilarities. The cumulative distribution function $F_D(\cdot)$ of the dissimilarity $D$ can be estimated from an iid sample of observations by (see [15])

$$\widehat{F_D}(d) = \frac{2}{n(n-1)} \sum_{i \neq l} 1(d_{il} \leq d), \tag{2.3}$$

for any possible value $d$.

From (2.3), we can compute the estimated cumulative distribution function for the within-group dissimilarities of the observed data, which we denote by $\widehat{F}_w(\cdot)$.

Similarly, one may estimate the cumulative distribution function of the dissimilarity between a subject belonging to the observed data and another subject belonging to the model-generated data. We denote it by $\widehat{F}_b(\cdot)$. The distance between these two estimated distribution functions can be calculated using one of many available distances, such as the Kolmogorv-Smirnov (KS) one. Thus we may use

$$T(y; y_{\text{obs}}) = KS\big(\widehat{F}_w, \widehat{F}_b\big) = \left\| \widehat{F}_w - \widehat{F}_b \right\|_\infty. \tag{2.4}$$

If the model-generated data are *similar* to the observed data, the distribution of the between-group dissimilarities will be similar to the distribution of the within-group dissimilarities in the observed data, and therefore the KS distance will be small.

In the next section we elaborate further on the use of these metrics in ABC. In turns out that this yields a natural alternative estimator for $\boldsymbol{\theta}$.

## 2.3 A new ABC-inspired estimator

Recall the definitions from the previous section. For any given $\boldsymbol{\theta}$ and $\boldsymbol{y}_{obs}$, there exists a whole distribution of the random variable $T(\boldsymbol{Y}, \boldsymbol{y}_{obs})$, and we now seek to exploit it. The idea is that if $T$ is a *good* measure of how far the random vector $\boldsymbol{Y}$ is from the fixed $\boldsymbol{y}_{obs}$ sample, then its distribution over a range of values of $\boldsymbol{\theta}$ can be used to estimate the true value of $\boldsymbol{\theta}$.

Specifically, we may estimate the quantile of order $\tau$, with $\tau \in (0, 1)$, of the conditional distribution of T given the parameter value $\boldsymbol{\theta}$, that is characterized as

$$q_\tau(T \mid \boldsymbol{\theta}) = \text{argmin}_a E[\rho_\tau(T - a) \mid \boldsymbol{\theta}],$$

where $\rho_\tau(\cdot)$ is the so-called *check function* and it is given by

$$\rho_\tau(z) = \begin{cases} \tau z, & \text{if } z > 0 \ , \\ -(1-\tau)z, & \text{otherwise} \, . \end{cases} \tag{2.5}$$

This is known as quantile regression (for an introduction, see [32]). In particular, we perform quantile regression marginally for each parameter component $\theta$ (note that one may also apply the multivariate version of quantile regression). As proposed in [69], we fit a local

linear quantile regression curve, where the estimated quantile and its first derivative are obtained through the following minimization problem:

$$(\widehat{q}_\tau(\theta), \widehat{q'}_\tau(\theta)) = \text{argmin}_{a,b} \sum_{i=1}^{n} \rho_\tau\Big(T_i - a - b(\theta_i - \theta)\Big) K\Big(\frac{\theta - \theta_i}{h}\Big), \tag{2.6}$$

where $K(\cdot)$ denotes a symmetric probability density function (the *kernel*) and $h$ is the bandwidth, the parameter that determines the amount of smoothing applied to the curve. Locally, estimating $a$ is equivalent to estimating $q_\tau(\theta)$ and estimating $b$ is equivalent to estimating $q'_\tau(\theta)$. Note that this estimator is just one of the many possibilities for nonparametric quantile regression. Methods for the optimal choices of the function $K(\cdot)$ and of the bandwidth $h$ have been proposed (see [69]), but they are not the focus of this work.

Given the estimated conditional quantile of order $\tau$ of the statistic $T$ given $\theta$, we define a new estimator $\widehat{\theta}$ of $\theta$ as the solution of the following minimization problem:

$$\widehat{\theta}_\tau(\boldsymbol{y}_{obs}) \quad = \quad \underset{\theta \in \Theta}{\text{argmin}}\, \widehat{q}_\tau\,(\theta; \boldsymbol{y}_{obs}). \tag{2.7}$$

This construction is motivated by the ABC method that retains values of $\theta$ that have produced samples close to $\boldsymbol{y}_{obs}$. Indeed, ideally one would only retain parameter values that produce $\boldsymbol{y}$'s which have the *smallest* distance from $\boldsymbol{y}_{obs}$. The calibration approach that we are suggesting exploits this idea by looking for the parameter value that produces samples closest to $\boldsymbol{y}_{obs}$. Such estimate is obtained from examination of the distribution of $T(\boldsymbol{y}; \boldsymbol{y}_{obs})$ when $\boldsymbol{y}$ is generated from a range of values of $\theta$. In particular, we reuse all samples generated by ABC, but the prior distribution of $\theta$ is only used to produce values of $\theta$, without any Bayesian interpretation.

One may choose $\tau$ to be relatively close to zero (say, $\tau = 0.1$ or $0.05$) or perhaps $\tau = 0.5$ (i.e. the conditional median), and estimate $\widehat{q}_\tau\,(\theta; \boldsymbol{y}_{obs})$ through quantile regression from the ABC-produced samples (thus taking advantage of the computational effort already spent to

perform ABC), or from samples generated over a grid of values of $\theta$.
Note that as an estimator, $\widehat{\theta}_\tau = \widehat{\theta}_\tau(\boldsymbol{Y}_{obs})$ with $\boldsymbol{Y}_{obs} \sim f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta^*)$, with $\theta^*$ the true value of $\theta$. Also, $\widehat{\theta}_\tau$ clearly depends on the choice of $\tau$.
As a second possibility, one may estimate $\theta$ by the following minimization problem:

$$\widehat{\theta}_E(\boldsymbol{y}_{obs}) \quad = \quad \underset{\theta \in \Theta}{\operatorname{argmin}} \, \widehat{E}\left(T\left(\boldsymbol{Y}, \boldsymbol{y}_{obs}\right)\right), \tag{2.8}$$

where the conditional expectation $\widehat{E}\left(T\left(\boldsymbol{Y}, \boldsymbol{y}_{obs}\right)\right)$ is estimated by a regression model. Note that here, too, as an estimator this is $\widehat{\theta}_E = \widehat{\theta}_E(\boldsymbol{Y}_{obs})$, with $\boldsymbol{Y}_{obs} \sim f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta^*)$, with $\theta^*$ the true value of $\theta$. This second definition does not require that one specify a value $\tau$.

Note that, if the objective function to be minimized in (2.7) or (2.8) is such that it is regular enough, i.e. such that it admits one minimum, it can possibly be obtained by setting the (partial) derivatives of the objective function with respect to $\boldsymbol{\theta}$ equal to zero. If, for simplicity, we focus on $\theta \in \mathbb{R}^1$, then the estimator $\widehat{\theta}_\tau$ can be obtained by solving the estimating equation

$$\frac{\partial}{\partial \theta} \widehat{q}_\tau(\theta; \boldsymbol{y}_{obs}) = 0. \tag{2.9}$$

Asymptotic properties, such as consistency and asymptotic normality, of the proposed estimators clearly need to be investigated. We do not further explore this direction, but we believe that useful theoretical results can be found in the framework of the theory of estimating equations (see [43]).

## 2.4 The bivariate normal model

We now present an example of application of a dissimilarity-based metric in the framework of ABC, and we illustrate how the new estimator introduced in Section 2.3, can be computed.

Let us consider the problem of estimating the mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ of a bivariate normal distribution with known covariance matrix $\Sigma$. Let $\boldsymbol{y}_{\text{obs}} = (y_{1i}, y_{2i})_{i=1,\dots,n}$ be a sample of $n$ observations from the distribution, which we denote by $N_2(\boldsymbol{\mu}, \Sigma)$.

We wish to explore whether an estimation procedure may be effective at recovering the true parameter values $\boldsymbol{\mu}$.

We focus on the Wilcoxon-Mosler criterion introduced above, i.e. we consider $T$ to be equal to the squared difference between the Wilcoxon-Mosler statistic WM between $\boldsymbol{y}$ and $\boldsymbol{y}_{obs}$ (defined by equation (2.1)) and its expected value when the two samples (generated and observed) are generated by the same distribution:

$$T = \left[ WM - \frac{n^2}{2} \left( \binom{2n}{2} - 1 \right) \right]^2 . \tag{2.10}$$

First, we set $\boldsymbol{\mu} = (1, 2)^T$ and $\Sigma = \text{diag}(2, 3)$ and generate a sample of size $n = 100$. We set prior distributions for the unknown parameter $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and generate $K = 10,000$ datasets from the model, each one corresponding to a value of $\boldsymbol{\mu}$ sampled from the prior distribution (and keeping $\Sigma = \text{diag}(2, 3)$ fixed). We assume that, a priori, $\mu_i \sim N(0, 4)$, $i = 1, 2$, and we produce $K$ simulated datasets of size $n = 100$. We define the pairwise distance between observations as the $L^2$-distance. For each $k = 1, \dots, K$, the Wilcoxon-Mosler statistic is then computed.

Working in the framework of the traditional ABC, one should retain only those parameter values that generated the samples with the smallest distance from the observed sample. The choice of the fraction of values to keep, that represents a sample from an approximated version of the posterior distributions of the parameters, is typically perfomed through a data-driven procedure, such as leave-one-out cross-validation [28].

Briefly, cross-validation works by randomly dividing the data into $k$ folds. Each time all data except one fold are used to estimate the model and then a prediction error is computed on the remaining portion of the data (one fold). In leave-one-out cross-validation, each fold

is made of only one observation. In the setting of ABC, observations are represented by generated samples. Each learning set is created by taking all the $K$ samples except one, the test set being the sample left out. The posterior of the model parameters is obtained via the ABC procedure using each time $K-1$ generated samples and a prediction error is computed on the sample left out. Taking an average of the prediction errors across samples, and repeating the same procedure for several values of the retention rate, one can select the optimal fraction of values to keep to approximate the posterior distributions of the model parameters.

Recall that, our interest is in exploring how well the Wilcoxon-Mosler distance performs in discriminating among the parameter values proposed by the prior distribution. Figure 2.2 shows the relationship between the values of the parameters and the distance from the observed data, marginally for the two components of $\boldsymbol{\mu}$. We first use a sample of size $n = 100$.

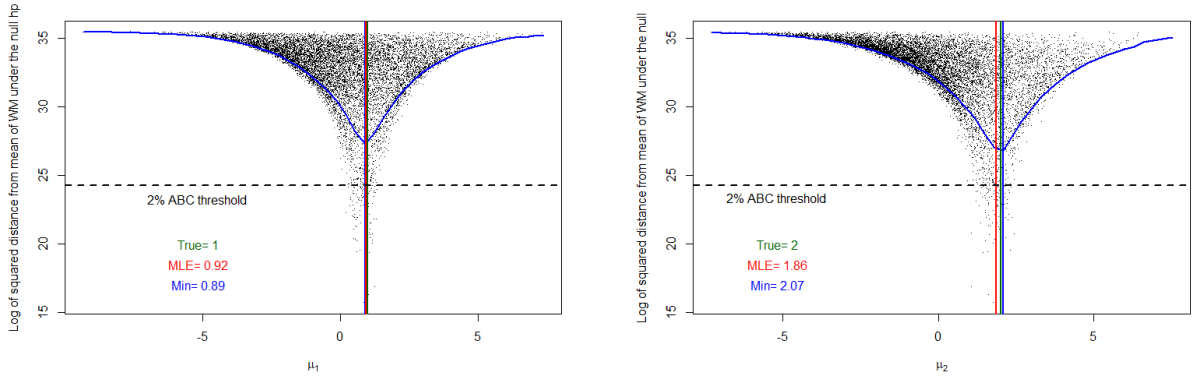



Figure 2.2: Samples from the prior distribution of the two mean components, and Wilcoxon-Mosler distance (in log-scale) of each generated dataset from the observed data. The blue curve is the estimated conditional quantile of order $\tau = 0.1$.

The blue curves in the plots of Figure 2.2 represent the estimates of the conditional quantiles of order $\tau = 0.1$. The two plots of Figure 2.2 suggest that the metric based on the Wilcoxon-Mosler statistic is quite informative about the unknown parameter $\boldsymbol{\mu}$. The

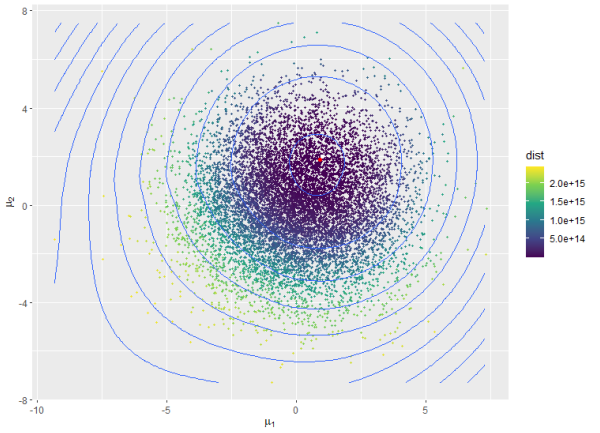
minimum distances are indeed reached for parameter values that are very close to the true ones ($\mu_1 = 1, \mu_2 = 2$). Note that one may be tempted to use such plots to assess the value of the statistic used in ABC to retain parameter values. However, one should be careful in such interpretation, since ABC only focuses on the smaller (in theory zero) values of the object function.

As shown by the blue vertical line, the suggested estimator for each parameter component is the value for which the minimum of the quantile regression curves is reached, as described in Section 2.3.

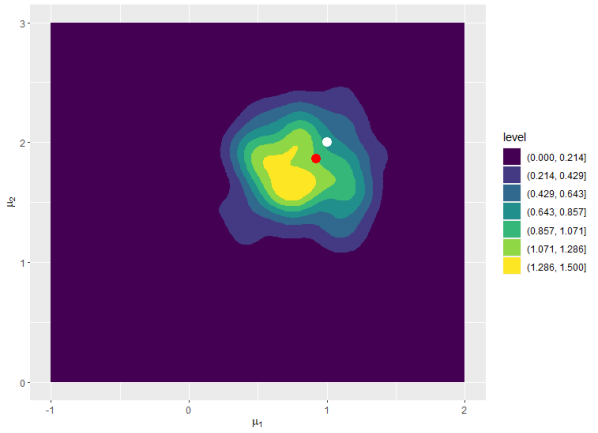
The plots in Figure 2.2 also show the values of the true parameter values (green vertical lines and text) and of the maximum likelihood estimates (red vertical lines and text). Note that some of the lines are not visible too clearly because of the overlapping with the others. Figure 2.3a shows the measure T, together with its level curves, as a function of $\mu_1$ and $\mu_2$ simultaneously, and Figure 2.3b the level curves of the joint posterior distribution for the paratemers obtained through ABC when retaining the best 2% of the values. Lastly, Figure 2.3c shows the prior (red curves) and ABC-posterior distributions (blue histograms and curves), marginally for the two parameter components. Moreover, given that for this model we can easily compute conjugate posterior distributions, they are also plotted (green curves). Note that, as expected, the exact posterior distributions have a smaller spread than the ABC ones. However, even the approximate posterior distributions are quite effective at recovering the true parameter values.

We now repeat the experiment, but increasing the sample size from $n = 100$ to $n = 1000$. As shown in Figure 2.4, the point estimates obtained from minimizing the quantile regression curves (that also in this case is the first decile) are again quite close to the true parameter values and to the MLEs.

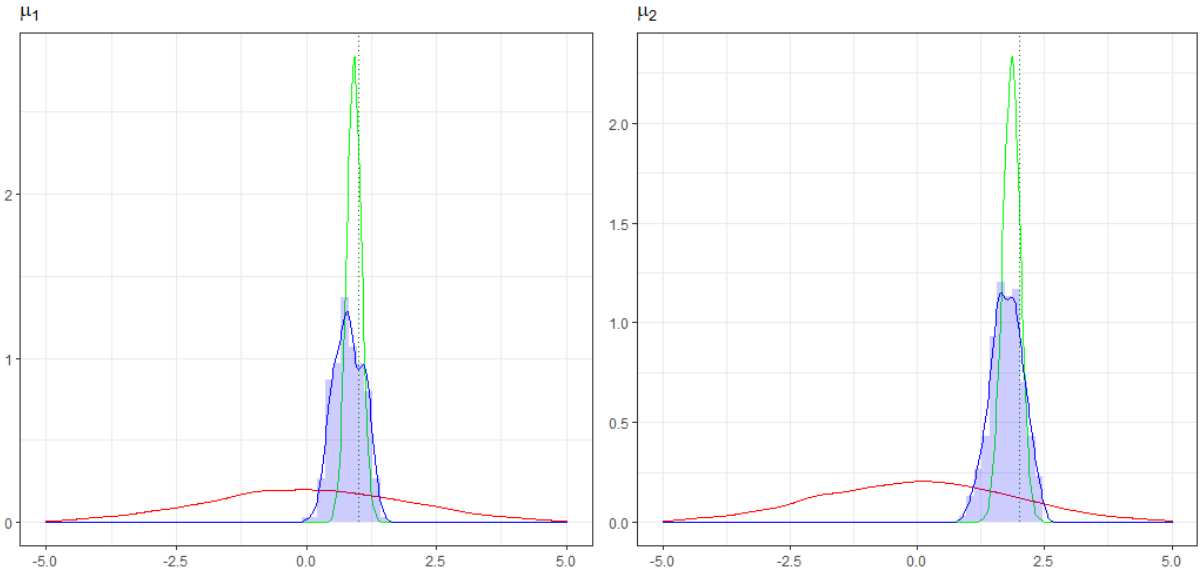
A comparison of Figures 2.2 and 2.4 suggest that - not surprisingly - T seems to contain more information on $\theta$ when $n$ is larger.




(a) Bivariate sample from the prior distribution of $\mu$ colored according to the Wilcoxon-Mosler distance, whose level curves are also plotted.




(b) Joint posterior distribution obtained from ABC with retention rate of 0.02. The white dot shows the true data generating parameter and the red one the MLE.




(c) Marginal prior (red curves), conjugate posterior distributions (green cruves) and posterior (blue histograms and curves) distributions obtained from ABC with retention rate of 0.02. The green vertical dotted lines show the true data generating parameter, $\mu_1 = 1$ and $\mu_2 = 2$.

Figure 2.3: Posterior distributions for the bivariate normal model with known covariance matrix.

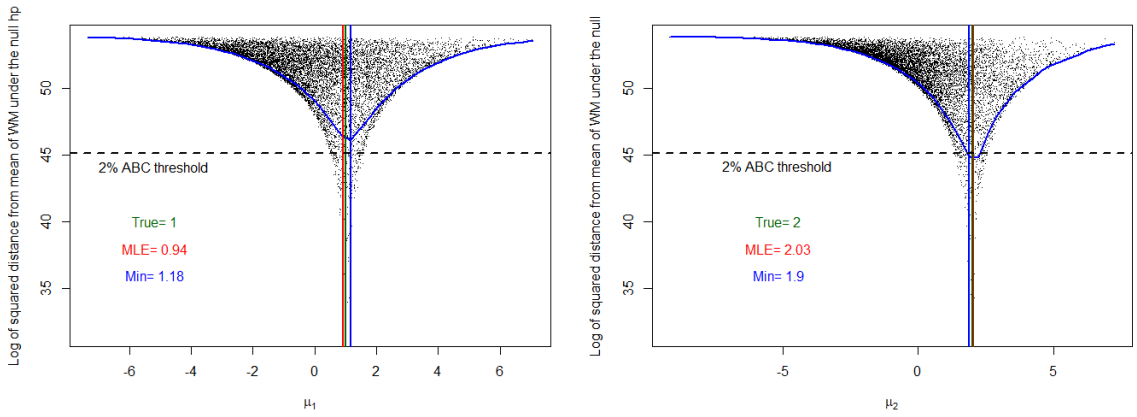
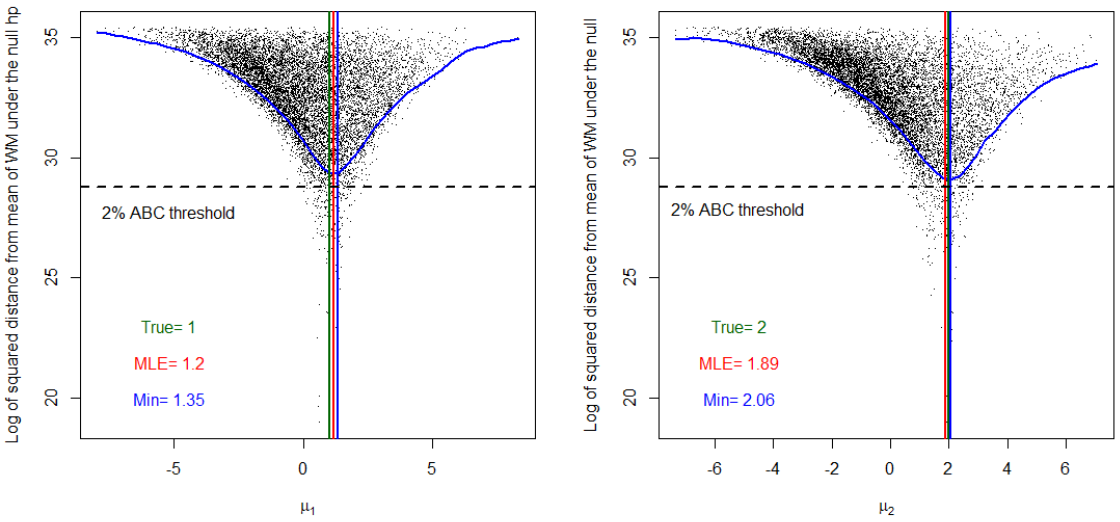



Figure 2.4: Case $n = 1000$. Prior samples for the two mean components and Wilcoxon-Mosler distance (in log-scale) of each generated dataset from the observed data. The blue curve is the estimated conditional quantile of order $\tau = 0.1$.

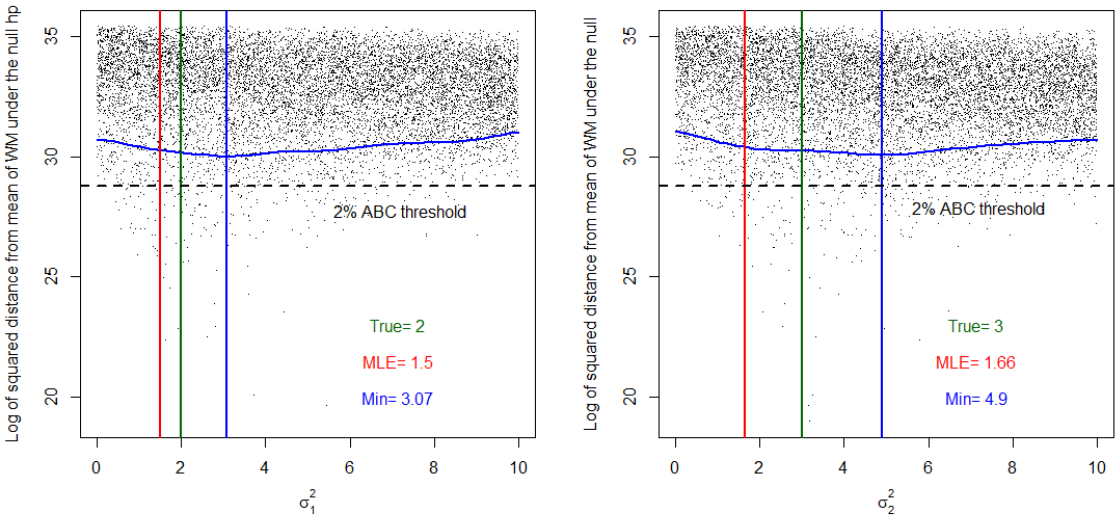
We may estimate the second derivative of the quantile regression curve at its minimum. To obtain such value, we can fit a quadratic curve to approximate the quantile curve, so that the coefficient of the quadratic term may be interpret as the local curvature of the function. This coefficient slightly varies from $\mu_1$ to $\mu_2$, and it also depends on the smoothing parameter $h$ used in the local linear quantile regression, but its order of magnitude seem to depend mostly on $n$. In particular, for $n = 100$ the order of magnitude of the second derivative is $10^{13}$, while for $n = 1000$ it increases to $10^{21}$. Recall that the plots are in log-scale, but the curves are estimated (and the derivatives are calculated) on the original scale. This suggests that, as expected, the amount of *information* provided by the observed data increases with the sample size, and that the T-based criterion may capture that fact.

We now turn to the case in which variances are not assumed to be known (the covariance is set to zero). We keep the same prior distributions for the mean parameters $\mu_1$ and $\mu_2$, and set $\sigma_1^2 \sim \text{Unif}(0, 10)$ and $\sigma_2^2 \sim \text{Unif}(0, 10)$ a priori, independently of each other.

The same procedure described above, based on pairwise dissimilarities and on the Wilcoxon-Mosler distance, performs still promisingly to estimate the mean $\boldsymbol{\mu}$ but, as expected, it fails




(a) Prior samples for the two mean components and Wilcoxon-Mosler distance of each generated dataset from the observed data




(b) Prior samples for the two variance components and Wilcoxon-Mosler distance of each generated dataset from the observed data.

Figure 2.5: Results for the bivariate normal model with unknown variances.

completely in recovering the variances $\sigma_1^2$ and $\sigma_2^2$. The point estimates produced by the minimization of the quantile regression curves are, indeed, quite far both from the true values $\sigma_1^2 = 2$ and $\sigma_2^2 = 3$, and from the maximum likelihood estimates $\widehat{\sigma}_1^2 = 1.50$ and $\widehat{\sigma}_2^2 = 1.66$. See Figure 2.5 for a graphical representation, analogous to Figure 2.2 of the previous example. From the bottom panels of the figure, it is evident that the estimated quantile curve is almost flat, suggesting that the Wilcoxon-Mosler statistic does not provide relevant information on $\sigma_1^2$ and $\sigma_2^2$. Figure 2.6 is the analogous of Figure 2.3a.

This should not come as a surprise, since the Wilcoxon-Mosler distance is sensitive to location changes but not to scale changes, since the ranks of the $L^2$-distance are scale invariant. More generally, as pointed out is [44] (p. 149), the use of Euclidean distances is such that the Wilcoxon test statistic is not invariant against affine linear transformations, but it is invariant against affine orthogonal transformations.

This shows that the choice of the dissimilarity is indeed very important to capture information on the parameter. In the next section we explore this point further, with an example implemented on discretized data and with different dissimilarities.

## 2.5 An example with discrete data and different dissimilarities

Still within the context of a bivariate normal model, we now discretize the observed and model generated data, and experiment with a few different dissimilarity measures. We discretize the data increasingly, to assess whether the WM-based criterion is likely to lose its premise with very poor data and/or dissimilarities.

We keep the same model definition, true parameter values and prior distributions as defined at the beginning of Section 2.4. We discretize the data according to a grid of values from $-20$ to 20 and using three different spacing steps, 0.5, 2 and 5, to gradually increase

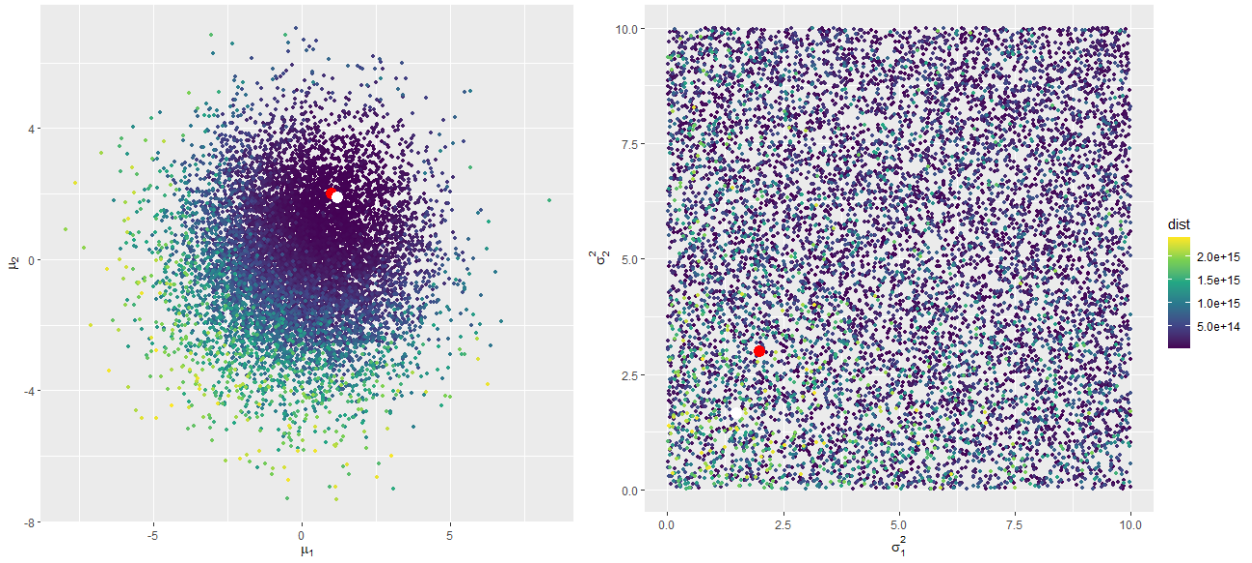



Figure 2.6: Bivariate sample from the prior distributions of $(\mu_1, \mu_2)$ and $(\sigma_1, \sigma_2)$ colored according to the Wilcoxon-Mosler distance of each corresponding generated dataset from the observed data. The red dots indicate the true parameter value.

the loss of information from the original data.

For each set of discretized data, we explore three dissimilarity measures. Given a couple of observations $\boldsymbol{y_j} = (y_{1j}, y_{2j})$, and $\boldsymbol{y_k} = (y_{1k}, y_{2k})$, discretized according to one of the grids, they are defined as follows:

1. Squared Euclidean distance: $d_E(\boldsymbol{y_j}, \boldsymbol{y_k}) = \sum_{h=1,2} (y_{hj} - y_{hk})^2$;
2. Manhattan distance: $d_M(\boldsymbol{y_j}, \boldsymbol{y_k}) = \sum_{h=1,2} |y_{hj} - y_{hk}|$;
3. "Rough" 0/1/2 distance: $d_R(\boldsymbol{y_j}, \boldsymbol{y_k}) = 2 - \sum_{h=1,2} \mathbb{1}(y_{hj} = y_{hk})$.

These distances are clearly ordered from more- to less- informative.

The distance between observed and model generated datasets is computed, as in the previous section, through the Wilcoxon-Mosler statistic.

Figures 2.7, 2.8 and 2.9 show the results for the three sets of discretized data, for the Euclidean, Manhattan and 0/1/2 dissimilarity-based metric, respectively. In each of the Figures, plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5). We can clearly see that the plots deteriorate, for any of the three considered dissimilary-based metrics, when the data are discretized with the roughest grid. For those cases the estimation procedure would clearly be hopeless. However, one would be able to spot that fact from those very graphical displays. On the other hand, when the space between the points of the grid is set to 0.5, and even 2, the plots seem to still be informative.

From this simple experiment, it seems that the degree of discreteness of the data may play a more relevant role than the choice of the dissimilarity measure in determining the likely performance of the estimation procedure.

Note that in such cases, the use of ABC with the same metric would be unlikely to produce any meaningful result.

## 2.6 Discussion

While this work is only a proof-of-concept, it does seem to suggest that the new estimation procedure that we have proposed might hold promise to work well, and its properties and generalization to more complex models deserve to be investigated.

Summarizing, to implement this approach in realistic models one needs: (i) the ability to generate data from the model; (ii) a metric to measure the dissimilarity between observed and generated data; (iii) an estimate of the conditional quantile of the distribution of the dissimilarity given the possibly high-dimensional $\theta$; and (iv) the ability to minimize that function.

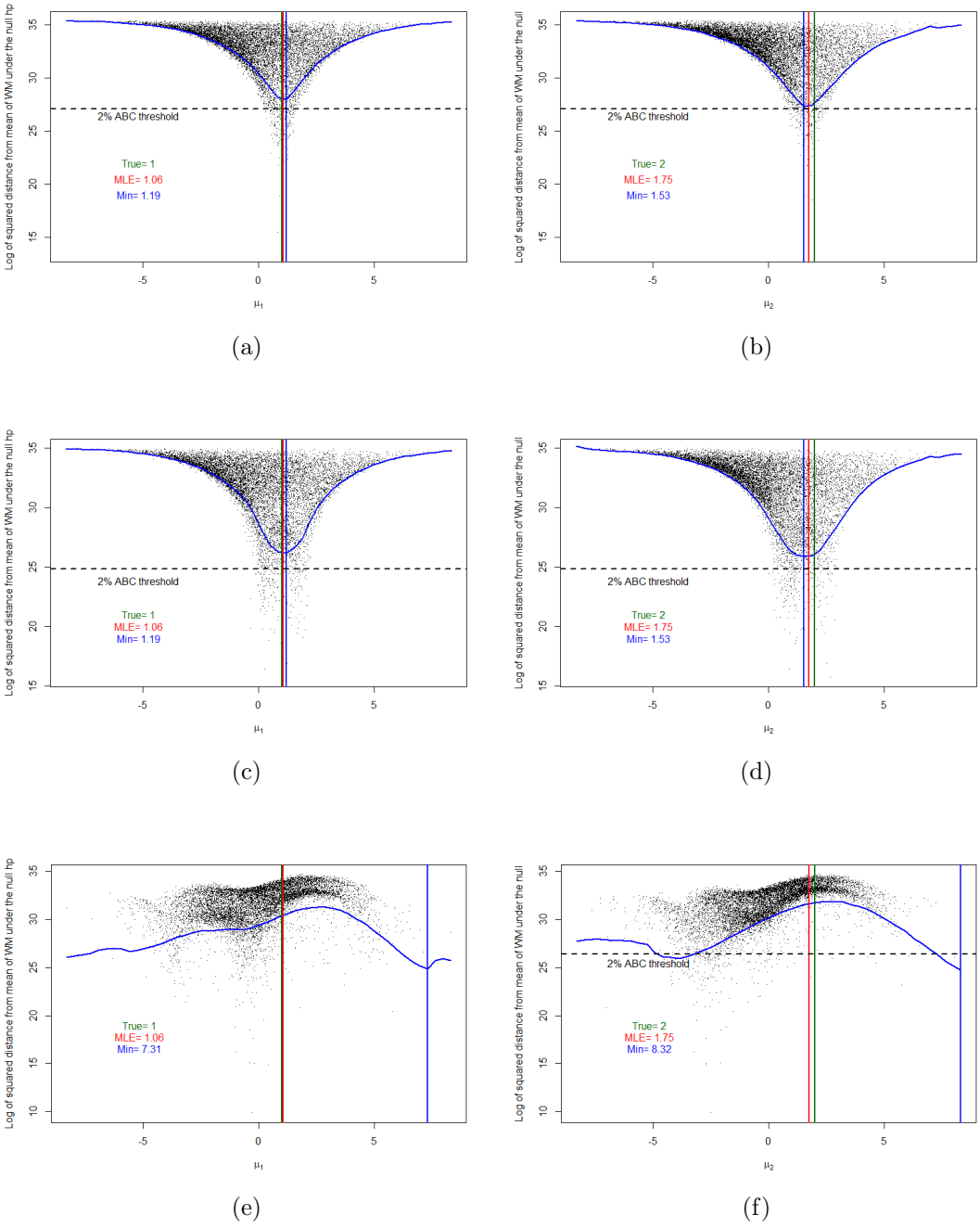



Figure 2.7: Results from the Euclidean distance. Plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5).

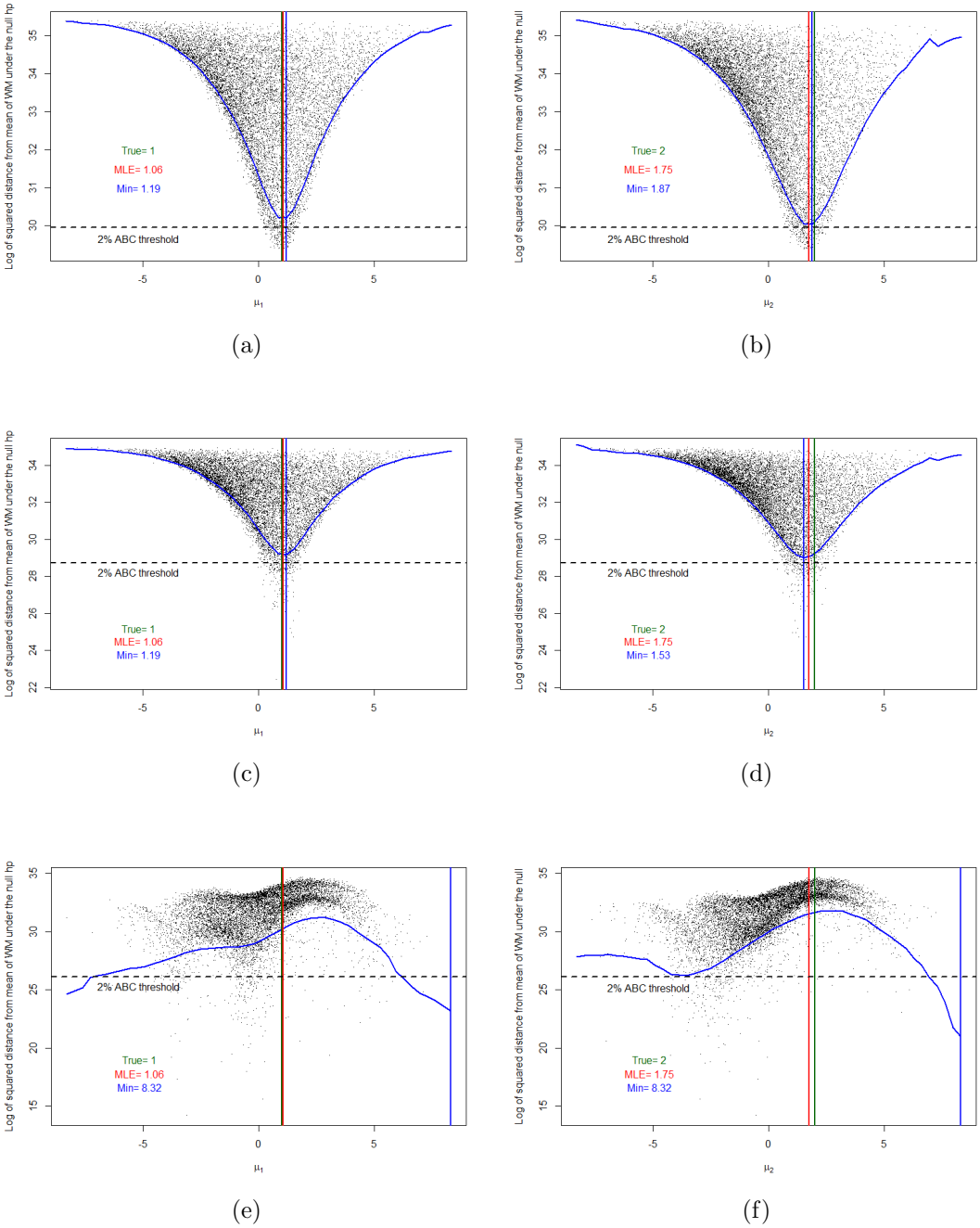



Figure 2.8: Results from the Manhattan distance. Plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5).

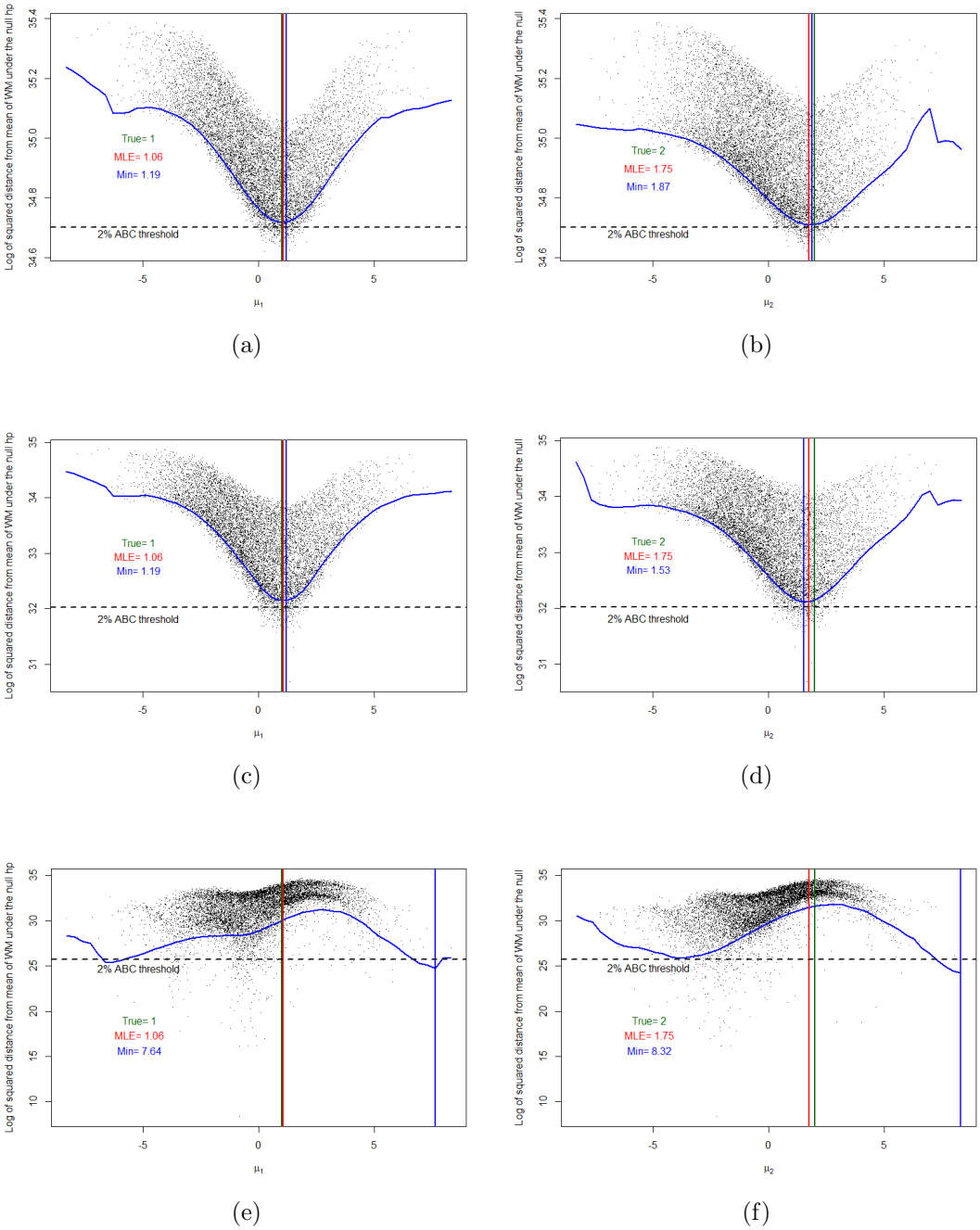



Figure 2.9: Results from the rough 0/1/2 distance. Plots (a) and (b) corresponds to the finest grid of discretization (with spacing equal to 0.5), plots (c) and (d) to the intermediate one (with spacing equal to 2), and plots (e) and (f) to the roughest grid (with spacing equal to 5).

An interesting setting for such methods may be models for longitudinal data with intractable likelihood function. For example, [60] considers discretely observed continuous-time multi-state models, such that for each subject a sequence of states is observed at discrete points and there is not additional information about the transition times. The study focuses on Markov and semi-Markov models, where the likelihood function is not available in closed form, and proposes summary statistics for conducting ABC-based inference. Similarly to what we have done Figure 2.3c, a dissimilarity-based metric may be an alternative approach when implementing ABC for this kind of models.

Indeed, very importantly, pairwise dissimilarities among observations, and a dissimilarity-based measure of the distance between two datasets can be defined for any kind of data, depending on their nature. One interesting direction that will be explored in the future is the extension of the methods proposed in this chapter to the setting of longitudinal data, i.e. data such that for each subject a sequence of states is observed over discrete time. Several dissimilarity measures among sequences can be defined, for both continuous and discrete state spaces. Among them, one can test the performance of measures that exploit the continuous nature of the data, such as correlation, but also measures that work with values discretized in bins. This could be a first attempt to move towards categorical longitudinal data and test the ability of these methods to recover the true parameter values in that setting. To quantify the distance between sequences of discrete states, a very common choice is Optimal Matching (OM). It is based on the effort needed to transform a sequence into another. The OM alignment technique was first introduced in molecular biology to study proteins and DNA sequences [53], and was then extended to sociology [2]. Therefore, one may try to construct a metric for ABC for sequence data based on the OM.

.

# Chapter 3

# Optimal estimation of the sparsity index in Poisson size-biased sampling[3]

## 3.1 Introduction

Size bias arises naturally from many common sampling designs. For example, when the sampling unit is the individual and the population consists of clusters of individuals, larger groups have a higher probability of being sampled than smaller groups. Therefore, if one is interested in studying parameters related to the group size, such as the average group size, or the density of individuals per unit area, it is essential to adjust for such bias.

Two examples that motivated this work arise from very different applied problems. The first one is related to the study of the family history of cancer: larger households have higher probability to manifest at least one case of cancer and to be, therefore, included into the Cancer Registry. The second motivating example is a study [5] (in progress) on the plague of 1630 in northern Italy, where the authors modeled the process of isolating infected subjects into the plague ward (*lazzaretto*). The internment was decided on a household basis.

[3]Joint work with Marco Bonetti and Marcello Pagano (Harvard School of Public Health)

Whenever a member of a household resulted infected, all the members of the same household had to be admitted to the plague ward. In that context, therefore, the household size plays a central role in shaping the individual risk of getting infected and being confined in the ward, and the fraction of the total population that is confined in the ward depends on the distribution of the households' sizes. As it will be clearer from the next paragraphs, the model studied in this chapter is closely related to the problem of the admission to the plague ward.

Another common example of size bias is the so-called "visibility bias", that occurs in aerial censuses of wildlife populations (see [22], [49] and [52]). In [22], the authors propose a model for estimating the mean group size and the population density of animals in a given study area, partitioned in quadrats. They assume that each animal has the same probability of being observed and that, whenever an individual is observed, its entire group is observed. The implication of these assumptions is that, by ignoring the biased-sampling design, one would overestimate the average group size. Indeed, this sampling design produces a sample not from the original distribution of group size, but rather from a weighted version of that distribution, where the weights grow with the group size. Denoting by $\beta$ the probability of sighting an animal, the sampling weight assigned to a group of size $x$ would be $w(x) = 1 - (1 - \beta)^x$ (see [49]). When the probability $\beta$ is small, the weights can be approximated by $w(x) \approx \beta x$, that corresponds to the common length-biased sampling setting, up to a normalizing constant.

Many other examples of size bias can be found in other applications, such as galaxy surveys [55], line transect sampling [24], forest inventory and estimation of tree basal area [26], histological analysis of cell size/type distributions [36], and more. Some of these problems involve discrete distributions, while others involve continuous measurements, but they all have in common the use of weighted distributions to correct for the sampling bias.

Most of the available literature on size bias and weighted distributions focuses on maximum likelihood estimation for the model parameters. An exception is [54], where the authors

consider three families of continuous distributions (i.e. gamma, Weibull and log-normal), and derive the uniformly minimum variance unbiased estimators (UMVUE) for the mean group size.

In this article we focus on a discrete problem. We assume that groups have size $Y = 1+H$, where $H \sim Poisson(\lambda)$, so that, for $\lambda > 0$,

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{(y-1)}}{(y-1)!}, \text{for} y = 1, 2, \ldots$$

In particular, $E(Y) = 1 + \lambda$.

We sample $X_1, X_2, \ldots, X_n$ *i.i.d.* group sizes ($n \geq 1$) from the population through random extraction of *individuals* from the population.

Hence, the probability that a group of size $x$ is chosen is proportional to $x$ (length-biased sampling):

$$P(X = x) \propto x\, P(Y = x)\,,$$

so that

$$P(X = x) \;=\; \frac{1}{k}\, x\, P(Y = x) = \frac{1}{\lambda + 1}\, \frac{x\, e^{-\lambda}\lambda^{x-1}}{(x-1)!}, \quad x = 1, 2, \ldots, \tag{3.1}$$

since

$$\begin{aligned} k^{-1} \;&=\; \sum_{x=1}^{\infty} [x\, P(Y = x)] = \sum_{x=1}^{\infty} \frac{x\, e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \sum_{y=0}^{\infty} \frac{(y+1)\, e^{-\lambda}\lambda^{y}}{y!} \\ &=\; \sum_{y=0}^{\infty} \frac{y\, e^{-\lambda}\lambda^{y}}{y!} + \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!} = \lambda + 1. \end{aligned}$$

This shows that the (size-biased) distribution of $X$ in (3.1) belongs to the one-dimensional exponential family, with minimal sufficient and complete statistic $S = \sum_{i=1}^{n} X_i$ (see, e.g.,

[45]). Note that

$$P(X=x) \;=\; \frac{x^2}{\lambda(\lambda+1)}\,\frac{e^{-\lambda}\lambda^x}{x!} \;=\; \frac{x^2}{\lambda(\lambda+1)}\,P(H=x), \quad x=0,1,2,\ldots$$

so that $X$ follows a weighted Poisson distribution with weights $w(x,\lambda) = x^2\,[\lambda(\lambda+1)]^{-1}$.

In other words, the length-biased version of a 1-translated Poisson is a weighted Poisson distribution with quadratic weights. Properties of weighted Poisson distributions, and their connection to over dispersion and under dispersion have been extensively studied (see e.g. [20], [19] and [33]). In those articles, statistical inference is performed through maximization of the likelihood function.

Here, we develop optimal exact inference based on the uniformly minimum variance unbiased estimator (UMVUE), and compare it to inference based on likelihood maximization, with a focus on small sample sizes.

In Section 3.2 we discuss the maximum likelihood estimator (MLE) for the parameter $\lambda$. In Section 3.3 we turn to inference on the "sparsity" parameter $\mu$, defined as $\mu = \frac{1}{1+\lambda}$, both through the MLE and the UMVUE, for which we develop an exact calculation algorithm, as well as an approximate algorithm based on the characteristic function. In Section 4 we describe the results of a simulation exercise designed to compare the two estimators.

## 3.2 Estimating $\lambda$ by maximum likelihood

We show that the sample mean $T_1 = \frac{1}{n}\sum_{i=1}^n X_i$ is biased for $E(Y) = 1+\lambda$. Of course, this is to be expected since larger groups are more likely to be sampled. Indeed,

$$\begin{aligned} E(T_1) &= E(X) = \frac{1}{\lambda+1}\sum_{x=1}^{\infty}\left[\frac{x^2\, e^{-\lambda}\lambda^{x-1}}{(x-1)!}\right] = \frac{1}{\lambda+1}\sum_{y=0}^{\infty}\left[\frac{(y+1)^2\, e^{-\lambda}\lambda^{y}}{y!}\right] \\ &= \frac{1}{\lambda+1}\sum_{y=0}^{\infty}\left[\frac{(y^2+1+2y)\, e^{-\lambda}\lambda^{y}}{y!}\right] = \frac{1}{\lambda+1}\left[E(Y^2)+1+2E(Y)\right] \\ &= 1+\lambda\left(\frac{\lambda+2}{\lambda+1}\right) = 1+\lambda+\frac{\lambda}{\lambda+1}. \end{aligned}$$

The estimator $T_1$ is also the UMVUE for $1+\lambda+\frac{\lambda}{\lambda+1}$, since it is unbiased for it and it is function of the minimal sufficient and complete statistic $S$. Moreover, by well-known results on the exponential family, $1+\lambda+\frac{\lambda}{\lambda+1}$ is the only function of $\lambda$ (up to linear transformations) for which the UMVUE achieves the Cramer-Rao lower bound (CRLB), see e.g. [45].

From (3.2), $E(T_1)$ is strictly greater than $(1+\lambda)$, so that $T_1$ is a biased estimator for $E(Y)$. Also, the absolute bias increases with $\lambda$, while the relative bias decreases with $\lambda$ (see Table 3.1). Note that in particular for small values of $\lambda$ the relative bias is clearly non-negligible.

| $\lambda$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $1+\lambda$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $E(X)$ | 2.50 | 3.67 | 4.75 | 5.80 | 6.83 | 7.86 | 8.88 |
| Absolute Bias | 0.50 | 0.67 | 0.75 | 0.80 | 0.83 | 0.86 | 0.88 |
| Relative Bias | 25% | 22% | 19% | 16% | 14% | 12% | 11% |

Table 3.1: Absolute and relative bias of $T_1$ in estimating $E(Y) = 1+\lambda$.

Given an iid sample $\underline{x} = (x_1, \ldots, x_n)^T$, the likelihood function for $\lambda$ takes the following form

$$\mathcal{L}(\lambda) = \frac{e^{-\lambda n}}{(\lambda+1)^n}\lambda^{(\sum_{i=1}^n x_i - n)}\prod_{i=1}^{n}\frac{x_i}{(x_i-1)!},$$

and the corresponding log-likelihood is

$$\ell(\lambda) = -\lambda n - n\log(\lambda+1) + \log(\lambda)\left(\sum_{i=1}^{n} x_i - n\right) + \sum_{i=1}^{n}\log\left(\frac{x_i}{(x_i-1)!}\right).$$

The maximum likelihood estimator for $\lambda$ can be obtained as the only positive solution to the normal equation $\frac{d\ell(\lambda)}{d\lambda} = 0$, that is

$$\widehat{\lambda} = \frac{\overline{X} - 3 + \sqrt{(\overline{X} - 3)^2 + 4(\overline{X} - 1)}}{2}, \tag{3.2}$$

with $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Differentiating the log-likelihood function twice shows that $\widehat{\lambda}$ is the only maximum on $(0, +\infty)$, except when $\overline{X} = 1$ (i.e. $X_i = 1$, for every $i = 1, \ldots, n$). In this case the MLE is not defined, since the likelihood function is monotonically decreasing on $(0, +\infty)$.

## 3.3 Estimation of the sparsity parameter $\mu$

We now turn to the estimation of the transformed parameter $\mu = (\lambda + 1)^{-1}$, which we call the group size distribution's *sparsity* parameter. Indeed, unbiased estimator for $\mu$ can be obtained, and used to construct an optimal unbiased estimator.

### 3.3.1 Maximum likelihood estimation of $\mu$

By invariance, the MLE for $\mu$ is $\widehat{\mu} = \frac{1}{1+\widehat{\lambda}}$, where $\widehat{\lambda}$ is the maximum likelihood estimator obtained in Section 3.2. The analytical calculation of $E(\widehat{\mu})$ is not trivial. However, as shown numerically in Section 3.5, $\widehat{\mu}$ is a biased estimator of $\mu$. Recall that, asymptotically, $\mathrm{Var}(\widehat{\lambda}) \approx \frac{1}{n\mathcal{I}(\lambda)}$, where

$$\mathcal{I}(\lambda) = -E\Big(\frac{\partial^2}{\partial\lambda^2}\ell(\lambda)\Big) = \frac{\lambda^2 + 2\lambda + 2}{\lambda(\lambda+1)^2} = \frac{1}{\lambda} + \frac{1}{\lambda(\lambda+1)^2}.$$

$[n\mathcal{I}(\lambda)]^{-1}$ is the Cramer-Rao lower bound (CRLB($\lambda$)) for the variance of any unbiased estimator of $\lambda$ (see, e.g. [35]). The asymptotic variance of the estimator $\widehat{\mu}$ can be approximated

through the delta method as

$$\widehat{\mathrm{Var}}(\widehat{\mu}) = \widehat{\mathrm{Var}}(\widehat{\lambda}) \cdot \left[\frac{d}{dt}\Big(\frac{1}{1+t}\Big)\right]^2_{|_{t=\widehat{\lambda}}} = \frac{1}{(1+\widehat{\lambda})^4} \cdot \frac{1}{n\mathcal{I}(\widehat{\lambda})} = \frac{\widehat{\lambda}}{n\,(1+\widehat{\lambda})^2\,(\widehat{\lambda}^2+2\widehat{\lambda}+2)},$$

and, given the asymptotic normality of maximum likelihood estimators [35],

$$\frac{\widehat{\mu}-\mu}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\mu})}} \xrightarrow{\mathrm{d}} Z \sim N(0,1), \text{ as } n \to \infty. \tag{3.3}$$

Testing procedures and confidence intervals can then be constructed easily. In particular, we consider two possible ways to build confidence intervals for the MLE. The first one is based on the asymptotic normal distribution of $\widehat{\mu}$ (see equation (3.3)) and it is therefore symmetric around $\widehat{\mu}$. The second possibility consists in constructing a confidence interval, denoted by $(L, U)$, for the MLE for $\lambda$, based on its asymptotic distribution and then applying the transformation $\mu = \frac{1}{1+\lambda}$ to obtain the interval for $\mu$, i.e. $(\frac{1}{1+U}, \frac{1}{1+L})$. Simulations show that the first kind of interval has a smaller average width compared to the second one, for any given coverage level, and thus in the following we shall present the results obtained by the first construction.

We now turn to the unbiased estimation of the sparsity parameter $\mu$. We first show that the estimator $T_2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{X_i}$ is unbiased for $\mu$, by showing that $E(X^{-1}) = \mu$. Indeed,

$$E\left(\frac{1}{X}\right) = \frac{1}{\lambda+1}\sum_{x=1}^{\infty}\left[\frac{1}{x}\cdot\frac{x\,e^{-\lambda}\lambda^{x-1}}{(x-1)!}\right] = \frac{1}{\lambda+1}\sum_{y=0}^{\infty}\left(\frac{e^{-\lambda}\lambda^{y}}{y!}\right) = \frac{1}{\lambda+1} = \mu.$$

We note that for $n \geq 2$, $\mathrm{Var}(T_2)$ is strictly greater than the Cramer-Rao lower bound (CRLB). Indeed, since $T_2$ is not a function of the minimal sufficient and complete statistic $S = \sum_{i=1}^n X_i$, it cannot be the UMVUE for $\mu$, and it therefore has variance strictly greater than the variance of the UMVUE (see, e.g. [45]). As a consequence, $\mathrm{var}(T_2) > \mathrm{var}(\text{UMVUE}) \geq \mathrm{CRLB}(\mu)$, where

the last inequality holds by definition of UMVUE. The CRLB($\mu$) for $\mu$ coincides with the asymptotic variance of the MLE $\widehat{\mu}$, and it is equal to

$$\mathrm{CRLB}(\mu) = \frac{\lambda}{n} \cdot \frac{(\lambda+1)^2}{(\lambda^2+2\lambda+2)} \cdot \frac{1}{(\lambda+1)^4} = \frac{\lambda}{n} \cdot \frac{1}{(\lambda^2+2\lambda+2)(\lambda+1)^2}.$$

Simple calculations yield

$$\mathrm{Var}(T_2) = \frac{1}{n}\Big[E\Big(\frac{1}{X_1^2}\Big) - E\Big(\frac{1}{X_1}\Big)^2\Big] = \frac{1-(1+\lambda)\,e^{-\lambda}}{n\lambda(\lambda+1)^2},$$

and one can verify numerically that

$$\frac{\mathrm{Var}(T_2)}{\mathrm{CRLB}(\mu)} = \frac{\Big(1-(\lambda+1)e^{-\lambda}\Big)(\lambda^2+2\lambda+2)}{\lambda^2} > 1.$$

The Rao-Blackwell and the Lehmann-Scheffé theorems can be used to construct the UMVUE for $\mu$. The optimal estimator is the conditional expected value of *any* unbiased estimator of $\mu$, when the conditioning is with respect to $S$ (see, e.g., [45]).
We pick as the initial unbiased estimator the estimator $X_1^{-1}$, and construct the UMVUE $T_3 = E(X_1^{-1}|S=s)$. We note that even the variance of the UMVUE $T_3$ is *not* equal to the CRLB. Indeed, using the Lehmann-Scheffé theorem, manipulation of the log-likelihood function shows that only linear transformations of $\sum_{i=1}^n X_i$ estimate their expected value efficiently (see, e.g., [45]).

In the following Section we focus on the calculation of $T_3$, and on constructing confidence intervals for $\mu$ based on it.

## 3.4 Computation of the umvue of $\mu$ ($T_3$)

### 3.4.1 A first exact algorithm to compute $T_3$

To compute $T_3$ we need the conditional probability distribution of $X_1$ given $S$. For the generic value $x \in \{1, 2, \ldots\}$ and $s \in \{1, 2, \ldots\}$,

$$\begin{aligned} P\left(X_1 = x | S = s\right) &= \frac{P\left(X_1 = x\right) P\left(S = s | X_1 = x\right)}{P(S = s)} \\ &= \frac{P\left(X_1 = x\right) P\left(\sum_{i=2}^{n} X_i = s - x | X_1 = x\right)}{P(S = s)} \\ &= \frac{P\left(X_1 = x\right) P\left(\sum_{i=2}^{n} X_i = s - x\right)}{P(S = s)}, \end{aligned} \tag{3.4}$$

where the last equality follows from the fact that $X_1, X_2, \ldots, X_n$ are independent, which implies in particular that $X_1$ and $(X_2, \ldots, X_n)$ are independent. Note that the conditional probability above is non-zero only as long as $x$ and $s$ are such that $x \geq 1$, $s \geq n$, and $s - x \geq n - 1$.

One can then proceed by computing, for the observed value of $S = s$, the numerator of (3.4) for all values $x \in \{1, 2, \ldots, s - n + 1\}$.

Note that $P(X_1 = x)$ is known, so that the only term that needs evaluation is $P(\sum_{i=2}^{n} X_i = s - x)$. For any given $s$, the number of possible values of $x$ is finite. Indeed,

$$P\left(\sum_{i=2}^{n} X_i = s - x\right) = \sum_{\{(x_2, \ldots x_n) \,:\, x_2 + x_3 + \ldots + x_n = s - x\}} \left[P(X = x_2) P(X = x_3) \ldots P(X = x_n)\right],$$

and this can be calculated by nested loops:

$x_2 \in \{1, \ldots, (s - x) - (n - 2)\}$

$x_3 \in \{1, \ldots, (s - x - x_2) - (n - 3)\}$

$\vdots$

$$x_{n-1} \in \{1, \ldots, (s - x - x_2 - \ldots - x_{n-2}) - 1\}$$

and with $x_n = s - (x + x_2 + \cdots + x_{n-1})$. The normalizing constant at the denominator of (3.4) can be obtained easily and exactly (up to numerical precision). Note that this algorithm works for $n \geq 3$; for the case $n = 2$ direct enumeration ca be used to produce all conditional probabilities.

The conditional expected value can then be computed from the conditional distribution, thus yielding the estimate of the UMVUE.

### 3.4.2 An improved exact algorithm to compute $T_3$

The computational cost of this algorithm is too heavy to recommend for practical use, especially for use in simulations to study the performance of the UMVUE.

However, one can take advantage of the *i.i.d.* nature of the $X_i$'s (in particular, of their exchangeability) and construct a much faster algorithm. Specifically, let $X_1 = x$ and $S = s$. Since all other $X_i$ variables are at least equal to one, and since $x + x_2 + \ldots + x_n = s$ holds (with $s \geq n$), then it must be $x \in \{1, 2, \ldots, s - (n - 1)\}$.

For each value $X_1 = x$ one can then generate all possible ordered configurations of values of $X_2, \ldots, X_n$ (including all configurations that contain tied values) such that $x + x_2 + \ldots + x_n = s$, but with $x_2 \leq x_3 \leq \ldots \leq x_n$. Noting that the new constraint $x_2 + \ldots + x_n = s - x$ must be satisfied, this can be achieved with the following alternative nested loops construction:

$$x_2 \in \left\{1, \ldots, \left\lfloor \tfrac{s-x}{n-1} \right\rfloor\right\}$$

$$x_3 \in \left\{x_2, \ldots, \left\lfloor \tfrac{s-x-x_2}{n-2} \right\rfloor\right\}$$

$\vdots$

$$x_{n-1} \in \left\{ x_{n-2}, \ldots, \left\lfloor \tfrac{s-x-x_2-\ldots-x_{n-2}}{n-(n-2)} \right\rfloor \right\}$$

and with $x_n = s - (x + x_2 + \cdots + x_{n-1})$. It can be checked that such $x_n$ will also satisfy $x_n \geq x_{n-1}$. Indeed, since $x_n + x_{n-1} = s - x - \sum_{i=2}^{n-2} x_i$ and, by construction, $x_{n-1} \leq \left\lfloor \frac{s-x-\sum_{i=2}^{n-2} x_i}{2} \right\rfloor$, it must hold that $x_n \geq \left\lfloor \frac{s-x-\sum_{i=2}^{n-2} x_i}{2} \right\rfloor$. The probability of each combination of ordered values should be multiplied by a factor to reflect the fact that the sequence was ordered. Indeed, letting $\mathbf{x}_{[1]} = (x_2, \ldots, x_n)^T$ and $A_{x-s} = \{\mathbf{x}_{[1]} : \sum_{i=2}^{n} x_i = s - x\}$,

$$\begin{aligned}
P\left(\sum_{i=2}^{n} X_i = s - x\right) &= \sum_{\{\mathbf{x}_{[1]} \in A_{s-x}\}} P\left(X_2 = x_2, \ldots, X_n = x_n\right) \\
&= \sum_{\{\mathbf{x}_{[1]} \in A_{s-x} : x_2 \leq x_3 \leq \ldots \leq x_n\}} \; \sum_{\{\mathbf{y}_{[1]} : \sigma(\mathbf{y}_{[1]}) = \mathbf{x}_{[1]}\}} P\left(X_2 = y_2, \ldots, X_n = y_n\right) \\
&= \sum_{\{\mathbf{x}_{[1]} \in A_{s-x} : x_2 \leq x_3 \leq \ldots \leq x_n\}} P\left(X_2 = y_2, \ldots, X_n = y_n\right) \sum_{\{\mathbf{y}_{[1]} : \sigma(\mathbf{y}_{[1]}) = \mathbf{x}_{[1]}\}} 1 \\
&= \sum_{\{\mathbf{x}_{[1]} \in A_{s-x} : x_2 \leq x_3 \leq \ldots \leq x_n\}} \left[k(x_2, \ldots, x_n) P\left(X_2 = y_2, \ldots, X_n = y_n\right)\right]
\end{aligned}$$

where $k(x_2, \ldots, x_n) = \#\{\mathbf{y}_{[1]} : \sigma(\mathbf{y}_{[1]}) = \mathbf{x}_{[1]}\}$, with $\sigma(\mathbf{y}_{[1]})$ the ordered sequence obtained from the elements of the vector $\mathbf{y}_{[1]} = (y_2, \ldots, y_n)^T$.

Given the discrete nature of $X$, the constant $k(x_2, \ldots, x_n)$ depends on the number of repeated values in each ordered sequence $(x_2, \ldots, x_n)^T$, i.e. on its absolute frequency distribution. If the sequence contains $k$ different values, repeated $v_1, \ldots, v_k$ times respectively (with $\sum_{i=1}^{k} v_i = n - 1$), then the number of configurations $\mathbf{y}_{[1]}$ such that $\sigma(\mathbf{y}_{[1]}) = \mathbf{x}_{[1]}$ is

$$k(x_2, \ldots, x_n) = \frac{(n-1)!}{v_1! v_2! \ldots v_k!}.$$

As earlier, multiplication by $P(X_1 = x_1)$ and normalization then yields Eq. (3.4), from which the Rao-Blackwellized estimator can be computed. We provide the `R` [51] code written to

compute the UMVUE by this latter algorithm using version 4.0.1 of the software.

### 3.4.3 An efficient approximate algorithm to compute $T_3$ using the characteristic function

The algorithm proposed in Section 3.4.2 provides the exact calculation of the UMVUE, but its computational cost makes it still not practical to use for large values of $n$ and $s = \sum_{i=1}^{n} x_i$. A way to reduce the computational cost of obtaining the UMVUE for $\mu$ can be by computing the distribution of $S = \sum_{i=1}^{n} X_i$, which appears in equation (3.4), in closed form. Since we are dealing with a sum of independent random variables, working with characteristic functions seems the most convenient way to proceed. After some calculations, we have that the characteristic function of $S$ is given by

$$\begin{aligned}
\phi_S(t) &= \Big(\sum_{x=1}^{+\infty} e^{itx} \frac{1}{1+\lambda} \frac{x\lambda^{x-1}e^{-\lambda}}{(x-1)!}\Big)^n = \Big((-i)\frac{\partial}{\partial t}\frac{e^{it}}{1+\lambda}\sum_{x=0}^{+\infty} e^{itx}\frac{\lambda^x e^{-\lambda}}{x!}\Big)^n \\
&= \Big((-i)\frac{\partial}{\partial t}\frac{e^{it}}{1+\lambda}\cdot e^{\lambda(e^{it}-1)}\Big)^n = \Big(\frac{e^{it}}{1+\lambda}\cdot e^{\lambda(e^{it}-1)} + \frac{e^{it}}{1+\lambda}\cdot e^{\lambda(e^{it}-1)}(\lambda e^{it})\Big)^n \\
&= \Big(\frac{e^{it}}{1+\lambda}\cdot e^{\lambda(e^{it}-1)}\cdot(1+\lambda e^{it})\Big)^n. \qquad (3.5)
\end{aligned}$$

Unfortunately, by the inversion theorem, we obtain only an integral form of the probability mass function of $S$, that is not solvable in closed form.

However, an approximation of the distribution of $S$ can be computed as the inverse discrete Fourier transform of the characteristic function in equation (3.5) thanks to the inverse fast Fourier transform (ifft) algorithm [11]. We rely on the implementation of this algorithm in the `R` package `pracma` [18]. As shown in Figure 3.1, the algorithm leads to approximations of the distribution of $S$ that are very similar to the empirical distributions obtained from a sample of size $30,000$. However, the computational gain in perfoming ifft instead of generating such large samples is remarkable.

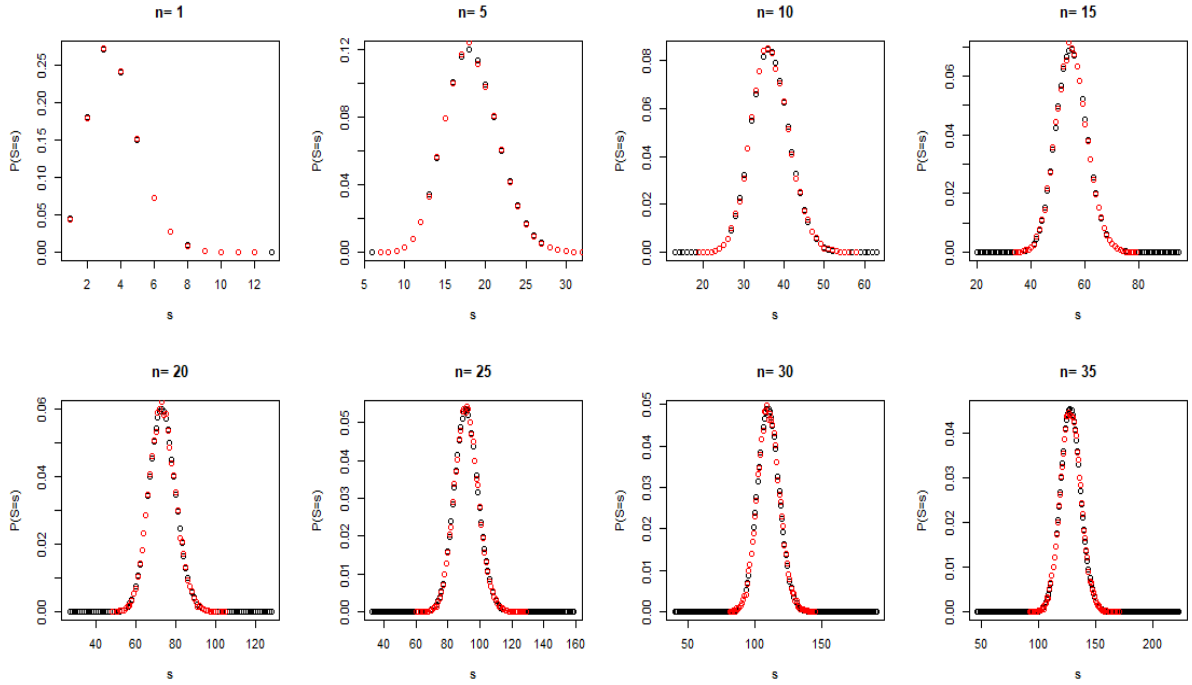



Figure 3.1: Probability mass function of S for $\lambda = 2$ and several values of $n$. The black points represent the distribution approximated by ifft and the red ones the empirical distribution computed from a sample of size $k = 30,000$.

We need to compute the conditional distribution of $X_1|S$, that is, by definition of sufficient statistic, independent of $\lambda$. This means that any value of $\lambda$ could be used, in principle, to approximate the numerator and the denominator by the ifft algorithm, as long as the same value is used in both approximations. Nevertheless, for the sake of numerical stability, it is convenient to perform the calculations using a *plausible* value of $\lambda$ for each combination of $n$ and $s$. Some empirical observations suggested that a good adaptive choice for $\lambda$ is the MLE, whenever it belongs to the interval $[0.5, 20]$, or the values 3 or 20 whenever the MLE exceeds the extremes of the interval. With this choice of $\lambda$ we were able to compute the UMVUE for any relevant combination of $s$ and $n$. Section 3.7 contains two tables showing the estimates corresponding to several different combinations of $n$ and $s = \sum_{i=1}^{n} x_i$. Table 3.4 contains

the estimates produced by the exact algorithm, used for small values of $n$, while Table 3.5 contains the estimates produced by the approximated ifft algorithm proposed in this section, for some larger values of $n$.

### 3.4.4 Constructing confidence intervals for $\mu$ from $T_3$

We now propose a way to construct an exact confidence interval for the UMVUE, by inverting the level-$\alpha$ bilateral test for $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. Numerically, if a confidence interval is defined with the limits quoted to two digits of accuracy, this is achieved by performing the test for a grid of values in $(0, 1)$, i.e. $\mu_0 \in \{0.10, 0.11, \ldots, 0.95\}$, and retaining only those values for which the test did not reject the null hypothesis. In other words, the lower bound of the interval is the smallest value $\mu_0$ for which the test did not reject $H_0$ and the upper bound is the largest value of $\mu_0$ for which the test did not reject $H_0$. Of course, if done sequentially from left or right, there is no need to evaluate this at all points, only up to the limits of the confidence interval. Note that, due to the randomization on the tails and to the variability in the simulation process, it can happen that, for some values of $\mu_0$ belonging to the confidence interval, the test did reject $H_0$. However, this happens very rarely, so that the definition of the interval given above seems still consistent to us. The test is randomized on both tails, so that the probability of the type I error is precisely equal to $\alpha$. The distribution of the test statistic, that is the UMVUE $E\big(\frac{1}{X_1}|S\big)$, under $H_0$ can approximated by a sample of size $10,000$ from its exact distribution or, to speed up the computation time, by the approximated distribution obtained by the ifft algorithm.

The procedure described above guarantees exact confidence intervals having coverage $1 - \alpha$, up to the approximation error due to the discretization of the parameter space into a grid of values. As we will see in the simulation study presented in Section 3.5, the proposed grid ($\{0.10, 0.11, \ldots, 0.95\}$) is not fine enough to reach the required coverage precisely, and this approximation error becomes more relevant for intervals with a smaller width. For this

reason, we ran additional simulations using a finer grid having binwidth equal to 0.002, that produces intervals whose actual coverage is very close to the nominal confidence levels, for most values of $n$ and $\mu$. We present the results of these simulations in the next section.

## 3.5 Simulation study: MLE vs UMVUE for $\mu$

Table 3.2 compares the estimated mean and standard deviation for the unbiased estimator $T_2$, the UMVUE and MLE for $\mu$, computed from 1,000 simulations, with $n$ varying from $n = 3$ to $n = 35$. The values of the UMVUE are obtained by the exact algorithm for $n \leq 10$ and by the ifft algorithm for $n > 10$. As expected, both $T_2$ and the UMVUE are on average very close to the true value $\mu = 0.25$, no matter how small the value of $n$, while the MLE has a bias that decreases when $n$ increases. The distribution of the UMVUE has the smallest standard deviation.

For a graphical comparison of the UMVUE and the MLE see Figure 3.2. The upper plot of Figure 3.2 shows the biases of the two estimators as a function of the sample size $n$. As we know from theory, as $n$ grows the bias of the MLE tends to zero

The lower plot of Figure 3.2 shows the overall effect of variance and bias for the two estimators, by comparing their mean squared errors (MSE). The y-axis on the right side of the two plots measures the relative reduction in variance and MSE of the UMVUE compared to the MLE (green diamond-shaped points). It is clear that for small $n$ the UMVUE should be preferred to the MLE.

We now compare the coverage and the average length of the exact confidence interval for the UMVUE with those of the asymptotic confidence interval based on the MLE. Table 3.3 shows the results obtained for several different values of $\mu$ and $n$, when using the finest of the two grids described in Section 3.4.4 (grid of equispaced points between 0 and 1 having binwidth equal to 0.002).

| n | Mean UMVUE | sd UMVUE | Mean $T_2$ | sd $T_2$ | Mean MLE | sd MLE |
|---|---|---|---|---|---|---|
| 3 | 0.2497 | 0.0641 | 0.2497 | 0.0744 | 0.2644 | 0.0691 |
| 4 | 0.2499 | 0.0546 | 0.2499 | 0.0647 | 0.2608 | 0.0578 |
| 5 | 0.2503 | 0.0484 | 0.2504 | 0.0578 | 0.2589 | 0.0507 |
| 6 | 0.2500 | 0.0441 | 0.2502 | 0.0530 | 0.2572 | 0.0459 |
| 7 | 0.2501 | 0.0406 | 0.2501 | 0.0488 | 0.2562 | 0.0420 |
| 8 | 0.2497 | 0.0377 | 0.2496 | 0.0455 | 0.2550 | 0.0389 |
| 9 | 0.2502 | 0.0358 | 0.2502 | 0.0432 | 0.2549 | 0.0368 |
| 10 | 0.2499 | 0.0337 | 0.2499 | 0.0408 | 0.2542 | 0.0345 |
| 11 | 0.2500 | 0.0322 | 0.2499 | 0.0390 | 0.2538 | 0.0329 |
| 12 | 0.2500 | 0.0306 | 0.2499 | 0.0371 | 0.2535 | 0.0312 |
| 13 | 0.2500 | 0.0295 | 0.2501 | 0.0360 | 0.2532 | 0.0300 |
| 14 | 0.2500 | 0.0283 | 0.2499 | 0.0345 | 0.2530 | 0.0288 |
| 15 | 0.2499 | 0.0274 | 0.2499 | 0.0335 | 0.2527 | 0.0278 |
| 16 | 0.2500 | 0.0265 | 0.2499 | 0.0323 | 0.2526 | 0.0269 |
| 17 | 0.2500 | 0.0257 | 0.2499 | 0.0312 | 0.2525 | 0.0260 |
| 18 | 0.2499 | 0.0249 | 0.2499 | 0.0304 | 0.2523 | 0.0253 |
| 19 | 0.2500 | 0.0243 | 0.2500 | 0.0296 | 0.2522 | 0.0246 |
| 20 | 0.2501 | 0.0237 | 0.2502 | 0.0290 | 0.2522 | 0.0240 |
| 21 | 0.2500 | 0.0231 | 0.2501 | 0.0283 | 0.2520 | 0.0234 |
| 22 | 0.2501 | 0.0226 | 0.2500 | 0.0276 | 0.2520 | 0.0228 |
| 23 | 0.2501 | 0.0221 | 0.2501 | 0.0270 | 0.2519 | 0.0224 |
| 24 | 0.2498 | 0.0215 | 0.2497 | 0.0263 | 0.2515 | 0.0217 |
| 25 | 0.2500 | 0.0212 | 0.2501 | 0.0260 | 0.2517 | 0.0214 |
| 26 | 0.2500 | 0.0206 | 0.2500 | 0.0253 | 0.2516 | 0.0208 |
| 27 | 0.2500 | 0.0203 | 0.2500 | 0.0249 | 0.2515 | 0.0205 |
| 28 | 0.2500 | 0.0199 | 0.2500 | 0.0244 | 0.2515 | 0.0201 |
| 29 | 0.2500 | 0.0196 | 0.2500 | 0.0241 | 0.2515 | 0.0198 |
| 30 | 0.2499 | 0.0193 | 0.2499 | 0.0235 | 0.2513 | 0.0194 |

Table 3.2: Mean and standard deviation for the three estimators for $\mu$ computed from 1,000 simulations performed using $\mu = 0.25$ ($\lambda = 3$) and several different values of $n$.

The empirical coverage of the UMVUE-based intervals is in most cases very close to the nominal confidence level of 0.95. Since we performed 1000 simulations, we expect the empirical coverage to fall in the interval $(0.936, 0.964)$ in the 95% of the cases.

However, the effect of the discretization of the parameter space into the grid is still visible for large values of $n$ and small values of $\mu$, for which the empirical confidence level falls below

| | $\mu$ ($\lambda$) | 0.3 (2.33) | 0.4 (1.67) | 0.5 (1) | 0.6 (0.67) | 0.7 (0.43) |
|---|---|---|---|---|---|---|
| n=4 | UMVUE | 0.940 (0.26) | 0.954 (0.36) | 0.968 (0.42) | 0.966 (0.45) | 0.987 (0.45) |
| | MLE | 0.946 (0.27) | 0.945 (0.37) | 0.939 (0.43) | 0.901 (0.47) | 0.927 (0.47) |
| n=6 | UMVUE | 0.956 (0.22) | 0.956 (0.29) | 0.967 (0.35) | 0.967 (0.38) | 0.964 (0.39) |
| | MLE | 0.949 (0.22) | 0.947 (0.30) | 0.943 (0.36) | 0.913 (0.39) | 0.925 (0.40) |
| n=8 | UMVUE | 0.951 (0.18) | 0.967 (0.25) | 0.959 (0.30) | 0.963 (0.34) | 0.947 (0.34) |
| | MLE | 0.960 (0.19) | 0.937 (0.25) | 0.920 (0.31) | 0.909 (0.34) | 0.908 (0.35) |
| n=10 | UMVUE | 0.937 (0.16) | 0.944 (0.23) | 0.953 (0.27) | 0.956 (0.30) | 0.959 (0.31) |
| | MLE | 0.951 (0.17) | 0.955 (0.23) | 0.935 (0.28) | 0.931 (0.31) | 0.924 (0.32) |
| n=15 | UMVUE | 0.952 (0.13) | 0.946 (0.18) | 0.953 (0.22) | 0.950 (0.25) | 0.949 (0.26) |
| | MLE | 0.956 (0.14) | 0.960 (0.18) | 0.935 (0.23) | 0.957 (0.25) | 0.945 (0.26) |
| n=20 | UMVUE | 0.940 (0.11) | 0.946 (0.15) | 0.962 (0.19) | 0.946 (0.22) | 0.963 (0.22) |
| | MLE | 0.945 (0.12) | 0.943 (0.16) | 0.956 (0.20) | 0.943 (0.22) | 0.945 (0.23) |
| n=25 | UMVUE | 0.924 (0.10) | 0.945 (0.14) | 0.964 (0.17) | 0.950 (0.19) | 0.947 (0.20) |
| | MLE | 0.948 (0.11) | 0.944 (0.14) | 0.955 (0.18) | 0.959 (0.20) | 0.946 (0.20) |
| n=30 | UMVUE | 0.935 (0.09) | 0.961 (0.13) | 0.941 (0.16) | 0.946 (0.17) | 0.946 (0.18) |
| | MLE | 0.937 (0.10) | 0.949 (0.13) | 0.938 (0.16) | 0.946 (0.18) | 0.939 (0.19) |
| n=35 | UMVUE | 0.925 (0.08) | 0.946 (0.12) | 0.941 (0.14) | 0.955 (0.16) | 0.941 (0.17) |
| | MLE | 0.957 (0.09) | 0.955 (0.12) | 0.955 (0.15) | 0.942 (0.17) | 0.937 (0.18) |

Table 3.3: Coverage and average width (within brackets) of confidence intervals computed from $1,000$ simulations where the theoretical coverage $1-\alpha$ was set to 0.95.

the nominal level. Indeed, the approximation error due to the choice of the grid has a bigger impact when the average width of the intervals is small.

For both UMVUE- and MLE-based intervals, the average width decreases with $n$ and increases with $\mu$, as larger values of $\mu$ are associated with larger variability. In general, MLE-based intervals are slightly wider then UMVUE-based intervals. The empirical coverage of the MLE-based intervals is in most cases close to the nominal confidence level, except for some combinations of $\mu$ and small values of $n$ (e.g. $\mu = 0.6, 0.7$, $n = 4, 6, 8, 10$). Therefore, this motivates the use of UMVUE-based inference, in particular when the sample size $n$ is small and the average group size is believed to be small ($\mu$ large).

## 3.6 Discussion

We have introduced algorithms to compute the UMVUE of the sparsity index, and a procedure to build confidence intervals for $\mu$ based on that estimator. Through simulations, we showed that the inference based on the proposed estimator is more precise than the classical likelihood-based inference when the sample size $n$ is small.

If, in addition to constructing the confidence interval, one is also interested in estimating the variance $\mathrm{Var}(T_3) = \mathrm{Var}\big(E(X_1^{-1}|S)\big)$ of the UMVUE, then the following algorithm may be used. By construction of $T_3$, there exists a function $h(\cdot)$ such that $T_3 = h(S)$. We wish to compute

$$\mathrm{Var}\,(T_3) = E[h(S)^2] - E[h(S)]^2 = \sum_{s=n}^{+\infty} \big(h(s)^2 P(S=s)\big) - \left(\sum_{s=n}^{+\infty} h(s) P(S=s)\right)^2. \qquad (3.6)$$

The procedure described in the previous Section allows us to compute $h(s) = E(X_1^{-1}|S=s)$ for $s \geq n$. Thus, we only need to derive the distribution of $S$, which however depends on the unknown parameter $\lambda$. As an approximation, we may replace $\lambda$ by a consistent estimator $\widetilde{\lambda}$ such as the MLE computed on the available data, and produce a sample of size $M$ from the distribution of $S$. In particular, for $i = 1, \ldots, M$ we generate $(X_{i1}, \ldots X_{in})$ and compute $S_i = \sum_{i=1}^{n} X_{ij}$. Since $S$ is discrete, we can then estimate $P(S=s)$ with the sample proportion $\frac{\#\{i:\, S_i = s\}}{M}$. Since $M$ is finite, the sums in equation (3.6) will be truncated at $s = \max\{S_i :\, i = 1, \ldots, M\}$. The numerical calculation of $T_3$ as described in Section 3.4, together with (3.6), then yields an approximation for $\mathrm{Var}\,(T_3)$.

The estimation procedure proposed in this article can be easily extended to the case $H \sim \mathrm{NB}(r, p)$, i.e. when $H$ follows a negative binomial distribution and has the following probability mass function:

$$P(H=h) = \binom{h+r-1}{r-1} \cdot (1-p)^h\, p^r, \quad h = 0, 1, 2, \ldots,$$

where the parameter $r$ is fixed and $p$ is the object of inference. Note that the case $r = 1$ corresponds to the geometric distribution. As before, we consider $Y = 1 + H$, and then we define $X$ as the size-biased version of $Y$, i.e.

$$P(X = x) = \frac{x \cdot \binom{x+r-2}{r-1} \cdot (1-p)^{x-1}\, p^{r+1}}{p + r \cdot (1-p)}, \quad x = 1, 2, 3 \ldots$$

Also in this case $X$ belongs to the one-dimensional exponential family, and it is not hard to show that $S = \sum_{i=1}^{n} X_i$ is the complete minimal sufficient statistic. Therefore, whenever we have an unbiased estimator for some function $g(\cdot)$ of the unknown parameter $p$, we can Rao-Blackwellize it by taking its expected value conditioning on S to obtain the UMVUE for $g(p)$. For example, when $r = 1$ (geometric case), we have that

$$E\Big(\frac{1}{X}\Big) = \sum_{i=1}^{n} \frac{1}{x} \cdot \frac{x \cdot \binom{x+r-2}{r-1} \cdot (1-p)^{x-1}\, p^{r+1}}{p + r \cdot (1-p)} = \sum_{i=1}^{n} (1-p)^{x-1}\, p^2 = p,$$

and $E\big(\frac{1}{X}|S\big)$ is then the (unique) UMVUE for $p$. Note that the computational algorithm is essentially unchanged, except for the substitution of the generating distribution of $X$.

A natural extension of this work might be to a more general class of sampling weights. Let $\alpha > 0$ and, as before, consider $Y = 1 + H$, with $H \sim Poisson(\lambda)$. We can define the distribution of $X$ as

$$P(X = x) \;\; \propto \;\; x^{\alpha} \cdot P(Y = x) = \frac{x^{\alpha}\, e^{-\lambda}\, \lambda^{x-1}}{\mu_{\alpha}\, (x-1)!}, \quad x = 1, 2, 3, \ldots,$$

where $\mu_{\alpha}$ is the normalizing constant. The minimal sufficient statistic is, again, $S = \sum_{i=1}^{n} X_i$. Note also that, for $\alpha$ integer, $\mu_{\alpha}$ can be computed as a function of the moments of $H \sim$

$Poisson(\lambda)$:

$$\mu_\alpha \;=\; \sum_{x=1}^{+\infty} \frac{x^\alpha\, e^{-\lambda}\, \lambda^{x-1}}{(x-1)!} = \sum_{y=0}^{+\infty} \sum_{k=0}^{\alpha} \binom{\alpha}{k} y^k \frac{e^{-\lambda}\, \lambda^y}{y!} = \sum_{k=0}^{\alpha} \binom{\alpha}{k} E(H^k).$$

In the case of $\alpha = 1$, that corresponds to the sampling scheme considered in this article, we find, as expected, $\mu_1 = E(H^0) + E(H^1) = 1 + \lambda$. Also, from $E\big(\frac{1}{X^\alpha}\big) = \frac{1}{\mu_\alpha}$, it follows that $E\big(\frac{1}{X^\alpha}|S\big)$ is the UMVUE for $\frac{1}{\mu_\alpha}$, for any $\alpha > 0$. The algorithms that we have introduced in Section 3.4 can be therefore also applied to this more general case.

As a last comment, now let $H \sim Poisson(\lambda)$ and $Y = k + H$, for any $k \in \mathbb{N}$. This corresponds to a minimum number of $k$ members of the groups. Then,

$$P(Y = y) = \frac{y!}{(y-k)!\, \lambda^k} \cdot P(H = y), \quad \text{for } y \geq k. \tag{3.7}$$

Note that this provides an immediate calculation of the factorial moment of order $k$ of the Poisson random variable $H$ as

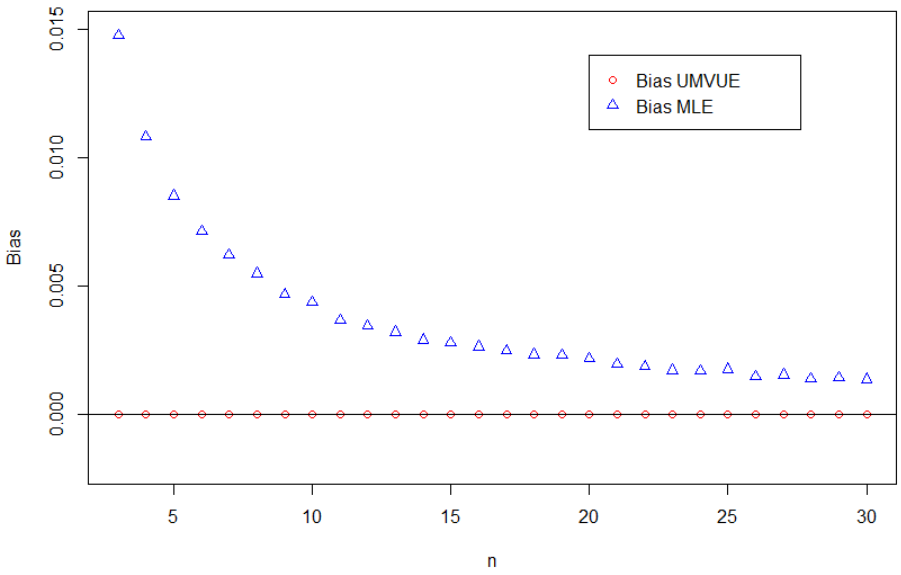
$$E\Big(\frac{H!}{(H-k)!}\Big) = \sum_{y=0}^{+\infty} \frac{y!}{(y-k)!}\, P(H = y) = \lambda^k \sum_{y=0}^{+\infty} P(Y = y) = \lambda^k.$$

The random variable $Y$ (see equation 3.7), and any weighted version of it (i.e. any $X$ such that $P(X = x) \propto x^\alpha P(Y = x)$), are again in the class of weighted Poisson distributions, and therefore with complete minimal sufficient statistic given by $S = \sum_{i=1}^n X_i$. Whenever an unbiased estimator for a parameter of interest is available, an algorithm similar to the ones that we have introduced in Section 3.4 can be therefore used to find the UMVUE.

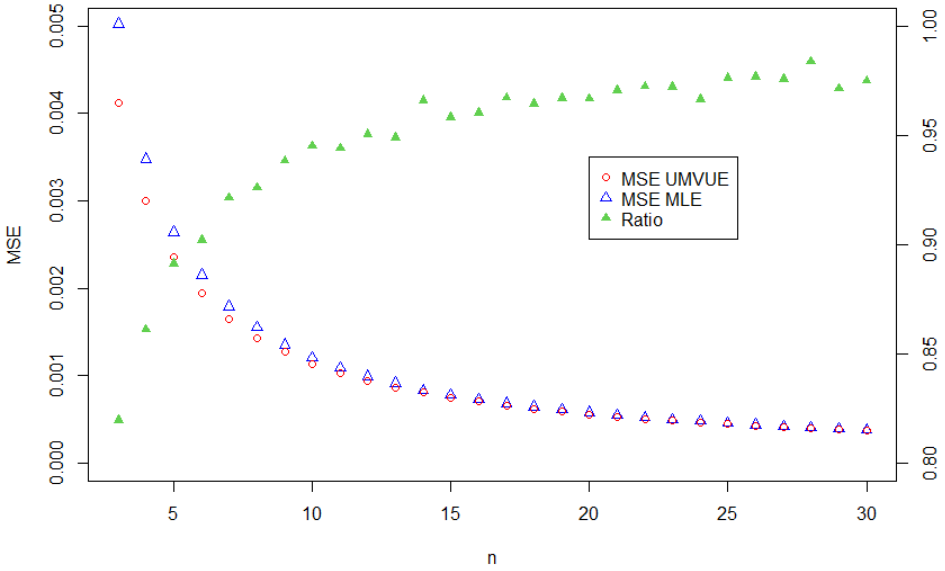
## 3.7 Tables of Rao-Blackwellized estimates for the sparsity index

| | n | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| s | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 1.000 | | | | | | | | |
| 3 | 0.750 | 1.000 | | | | | | | |
| 4 | 0.571 | 0.833 | 1.000 | | | | | | |
| 5 | 0.455 | 0.697 | 0.875 | 1.000 | | | | | |
| 6 | 0.375 | 0.589 | 0.767 | 0.900 | 1.000 | | | | |
| 7 | 0.318 | 0.506 | 0.675 | 0.811 | 0.917 | 1.000 | | | |
| 8 | 0.276 | 0.440 | 0.597 | 0.731 | 0.841 | 0.929 | 1.000 | | |
| 9 | 0.243 | 0.389 | 0.533 | 0.662 | 0.772 | 0.862 | 0.938 | 1.000 | |
| 10 | 0.217 | 0.347 | 0.479 | 0.602 | 0.710 | 0.802 | 0.879 | 0.944 | 1.000 |
| 11 | 0.196 | 0.313 | 0.433 | 0.549 | 0.654 | 0.746 | 0.825 | 0.892 | 0.950 |
| 12 | 0.179 | 0.285 | 0.395 | 0.504 | 0.605 | 0.695 | 0.774 | 0.843 | 0.903 |
| 13 | 0.165 | 0.261 | 0.363 | 0.464 | 0.560 | 0.649 | 0.727 | 0.797 | 0.858 |
| 14 | 0.152 | 0.241 | 0.335 | 0.429 | 0.521 | 0.607 | 0.684 | 0.754 | 0.815 |
| 15 | 0.142 | 0.224 | 0.310 | 0.399 | 0.486 | 0.569 | 0.645 | 0.713 | 0.775 |
| 16 | 0.132 | 0.208 | 0.289 | 0.372 | 0.455 | 0.534 | 0.608 | 0.676 | 0.738 |
| 17 | 0.124 | 0.195 | 0.271 | 0.349 | 0.427 | 0.503 | 0.575 | 0.650 | 0.703 |
| 18 | 0.117 | 0.183 | 0.254 | 0.328 | 0.402 | 0.474 | 0.544 | 0.610 | 0.670 |
| 19 | 0.110 | 0.173 | 0.240 | 0.309 | 0.379 | 0.448 | 0.516 | 0.582 | 0.640 |
| 20 | 0.105 | 0.164 | 0.227 | 0.292 | 0.359 | 0.425 | 0.490 | 0.556 | 0.611 |
| 21 | 0.100 | 0.155 | 0.215 | 0.277 | 0.340 | 0.404 | 0.466 | 0.529 | 0.585 |
| 22 | 0.095 | 0.148 | 0.204 | 0.263 | 0.323 | 0.384 | 0.444 | 0.503 | 0.559 |
| 23 | 0.091 | 0.141 | 0.195 | 0.250 | 0.308 | 0.366 | 0.424 | 0.480 | 0.538 |
| 24 | 0.087 | 0.135 | 0.186 | 0.239 | 0.294 | 0.349 | 0.405 | 0.461 | 0.515 |
| 25 | 0.083 | 0.129 | 0.178 | 0.228 | 0.281 | 0.334 | 0.388 | 0.441 | 0.492 |
| 26 | 0.080 | 0.124 | 0.170 | 0.219 | 0.269 | 0.320 | 0.372 | 0.425 | 0.475 |
| 27 | 0.077 | 0.119 | 0.163 | 0.210 | 0.258 | 0.307 | 0.357 | 0.408 | 0.456 |
| 28 | 0.074 | 0.114 | 0.157 | 0.202 | 0.248 | 0.295 | 0.343 | 0.392 | 0.438 |
| 29 | 0.071 | 0.110 | 0.151 | 0.194 | 0.238 | 0.284 | 0.330 | 0.377 | 0.423 |
| 30 | 0.069 | 0.106 | 0.146 | 0.187 | 0.230 | 0.273 | 0.318 | 0.365 | 0.407 |
| 31 | 0.067 | 0.103 | 0.141 | 0.180 | 0.222 | 0.264 | 0.307 | 0.350 | 0.393 |
| 32 | 0.064 | 0.099 | 0.136 | 0.174 | 0.214 | 0.255 | 0.296 | 0.338 | 0.381 |
| 33 | 0.062 | 0.096 | 0.132 | 0.169 | 0.207 | 0.246 | 0.286 | 0.327 | 0.370 |
| 34 | 0.060 | 0.093 | 0.128 | 0.163 | 0.200 | 0.238 | 0.277 | 0.317 | 0.356 |
| 35 | 0.059 | 0.090 | 0.124 | 0.158 | 0.194 | 0.231 | 0.268 | 0.306 | 0.346 |
| 36 | 0.057 | 0.088 | 0.120 | 0.153 | 0.188 | 0.224 | 0.260 | 0.298 | 0.334 |
| 37 | 0.056 | 0.085 | 0.116 | 0.149 | 0.182 | 0.217 | 0.252 | 0.288 | 0.324 |
| 38 | 0.054 | 0.083 | 0.113 | 0.145 | 0.177 | 0.211 | 0.245 | 0.280 | 0.316 |
| 39 | 0.053 | 0.081 | 0.110 | 0.141 | 0.172 | 0.205 | 0.238 | 0.272 | 0.306 |
| 40 | 0.051 | 0.079 | 0.107 | 0.137 | 0.168 | 0.199 | 0.231 | 0.265 | 0.299 |
| 41 | 0.050 | 0.077 | 0.104 | 0.133 | 0.163 | 0.194 | 0.225 | 0.258 | 0.290 |
| 42 | 0.049 | 0.075 | 0.102 | 0.130 | 0.159 | 0.189 | 0.219 | 0.251 | 0.283 |

Table 3.4: Rao-Blackwellized estimates of $\mu$ given the sample size $n$ and $s = \sum_{i=1}^{n} x_i$ computed by the exact algorithm.




(a)




(b)

Figure 3.2: Biases (a) and MSEs (b) of the two estimators for $\mu$ as functions of the sample size $n$, when $\lambda = 3$ ($\mu = 0.25$).

| | n | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| s | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 11 | 1.000 | | | | | | | | | |
| 12 | 0.955 | 1.000 | | | | | | | | |
| 13 | 0.911 | 0.958 | 1.000 | | | | | | | |
| 14 | 0.870 | 0.918 | 0.962 | 1.000 | | | | | | |
| 15 | 0.831 | 0.880 | 0.925 | 0.964 | 1.000 | | | | | |
| 16 | 0.794 | 0.844 | 0.889 | 0.930 | 0.967 | 1.000 | | | | |
| 17 | 0.759 | 0.809 | 0.855 | 0.897 | 0.934 | 0.969 | 1.000 | | | |
| 18 | 0.725 | 0.776 | 0.823 | 0.865 | 0.903 | 0.938 | 0.971 | 1.000 | | |
| 19 | 0.694 | 0.745 | 0.792 | 0.834 | 0.873 | 0.909 | 0.942 | 0.972 | 1.000 | |
| 20 | 0.665 | 0.715 | 0.762 | 0.805 | 0.845 | 0.881 | 0.914 | 0.945 | 0.974 | 1.000 |
| 21 | 0.638 | 0.687 | 0.734 | 0.777 | 0.817 | 0.854 | 0.888 | 0.919 | 0.948 | 0.975 |
| 22 | 0.617 | 0.661 | 0.707 | 0.750 | 0.790 | 0.827 | 0.862 | 0.894 | 0.923 | 0.951 |
| 23 | 0.594 | 0.636 | 0.681 | 0.724 | 0.765 | 0.802 | 0.837 | 0.869 | 0.899 | 0.927 |
| 24 | 0.567 | 0.614 | 0.660 | 0.700 | 0.740 | 0.777 | 0.812 | 0.845 | 0.875 | 0.904 |
| 25 | 0.546 | 0.592 | 0.640 | 0.676 | 0.716 | 0.754 | 0.789 | 0.822 | 0.853 | 0.881 |
| 26 | 0.522 | 0.569 | 0.613 | 0.654 | 0.694 | 0.731 | 0.767 | 0.800 | 0.830 | 0.859 |
| 27 | 0.502 | 0.547 | 0.594 | 0.633 | 0.672 | 0.710 | 0.745 | 0.778 | 0.809 | 0.838 |
| 28 | 0.484 | 0.529 | 0.576 | 0.613 | 0.658 | 0.689 | 0.724 | 0.757 | 0.788 | 0.818 |
| 29 | 0.467 | 0.512 | 0.554 | 0.594 | 0.633 | 0.669 | 0.704 | 0.737 | 0.768 | 0.798 |
| 30 | 0.451 | 0.493 | 0.535 | 0.577 | 0.613 | 0.651 | 0.684 | 0.717 | 0.749 | 0.778 |
| 31 | 0.437 | 0.478 | 0.518 | 0.562 | 0.595 | 0.629 | 0.665 | 0.698 | 0.730 | 0.759 |
| 32 | 0.422 | 0.462 | 0.502 | 0.541 | 0.578 | 0.619 | 0.646 | 0.680 | 0.711 | 0.741 |
| 33 | 0.409 | 0.449 | 0.486 | 0.529 | 0.561 | 0.601 | 0.637 | 0.663 | 0.694 | 0.723 |
| 34 | 0.396 | 0.436 | 0.471 | 0.510 | 0.546 | 0.583 | 0.615 | 0.648 | 0.677 | 0.706 |
| 35 | 0.385 | 0.421 | 0.461 | 0.494 | 0.532 | 0.568 | 0.600 | 0.631 | 0.659 | 0.690 |
| 36 | 0.371 | 0.408 | 0.446 | 0.481 | 0.519 | 0.553 | 0.581 | 0.615 | 0.647 | 0.673 |
| 37 | 0.362 | 0.398 | 0.432 | 0.470 | 0.501 | 0.538 | 0.568 | 0.600 | 0.631 | 0.653 |
| 38 | 0.351 | 0.386 | 0.421 | 0.458 | 0.488 | 0.523 | 0.552 | 0.584 | 0.615 | 0.643 |
| 39 | 0.341 | 0.375 | 0.410 | 0.445 | 0.478 | 0.509 | 0.543 | 0.574 | 0.600 | 0.625 |
| 40 | 0.331 | 0.367 | 0.398 | 0.432 | 0.465 | 0.495 | 0.528 | 0.558 | 0.591 | 0.614 |
| 41 | 0.322 | 0.355 | 0.388 | 0.422 | 0.454 | 0.484 | 0.516 | 0.543 | 0.574 | 0.606 |
| 42 | 0.315 | 0.347 | 0.378 | 0.412 | 0.440 | 0.473 | 0.501 | 0.532 | 0.564 | 0.590 |
| 43 | 0.307 | 0.337 | 0.370 | 0.400 | 0.432 | 0.461 | 0.489 | 0.520 | 0.546 | 0.575 |
| 44 | 0.298 | 0.328 | 0.361 | 0.391 | 0.420 | 0.449 | 0.481 | 0.507 | 0.535 | 0.562 |
| 45 | 0.292 | 0.321 | 0.351 | 0.381 | 0.410 | 0.439 | 0.468 | 0.495 | 0.523 | 0.552 |
| 46 | 0.284 | 0.314 | 0.343 | 0.371 | 0.402 | 0.429 | 0.458 | 0.485 | 0.515 | 0.540 |
| 47 | 0.278 | 0.307 | 0.334 | 0.364 | 0.391 | 0.421 | 0.447 | 0.475 | 0.501 | 0.531 |
| 48 | 0.271 | 0.299 | 0.326 | 0.355 | 0.384 | 0.412 | 0.438 | 0.464 | 0.489 | 0.518 |
| 49 | 0.265 | 0.292 | 0.319 | 0.346 | 0.374 | 0.401 | 0.429 | 0.455 | 0.483 | 0.507 |
| 50 | 0.260 | 0.285 | 0.312 | 0.339 | 0.367 | 0.393 | 0.419 | 0.445 | 0.471 | 0.496 |
| 51 | 0.254 | 0.280 | 0.305 | 0.332 | 0.358 | 0.384 | 0.410 | 0.438 | 0.464 | 0.487 |
| 52 | 0.248 | 0.273 | 0.300 | 0.324 | 0.352 | 0.377 | 0.403 | 0.429 | 0.453 | 0.477 |
| 53 | 0.243 | 0.268 | 0.293 | 0.318 | 0.344 | 0.370 | 0.396 | 0.419 | 0.444 | 0.470 |
| 54 | 0.237 | 0.263 | 0.287 | 0.312 | 0.336 | 0.363 | 0.386 | 0.411 | 0.435 | 0.461 |
| 55 | 0.233 | 0.257 | 0.282 | 0.306 | 0.329 | 0.354 | 0.379 | 0.402 | 0.426 | 0.451 |

Table 3.5: Rao-Blackwellized estimates of $\mu$ given the sample size n and $s = \sum_{i=1}^{n} x_i$ computed by the approximated algorithm based on the inverse fast Fourier transform.

# Conclusions

To avoid repeating ourselves, we refer to each chapter for the conclusions, and add only a couple of additional relevant points here.

In Chapter 1 we have developed and implemented several parametric models to describe the latent process of the insurgence and evolution of breast cancer, from asymptomatic to symptomatic. The estimation of this kind of models presents many challenges, mainly because the quantities of interest are almost never observable precisely, yet the real scientific question clearly refers to such latent process. Our focus was on breast cancer, however these models, that belong to the class of multi-state models, can be useful tools to study the latent evolution of many other diseases. While likelihood-based inference is not feasible for complex models, generating data from them typically remains manageable, if one is able to filter them by the partial observation mechanism accurately to obtain data reasonably similar to the observed data. For this reason, ABC looks as a very attractive tool in this setting. Yet, how to perform reliable likelihood-free inference on such models is an open and very interesting question, and we have addressed it in Chapter 1. In that work we have proposed a possible way to summarize this kind of data and to measure the distance between datasets so that ABC could be implemented. However, there is still work to be done to explore alternative metrics for ABC in this setting and to evaluate their performance.

Chapter 2 is meant as a first step in this direction. We introduced the idea of working with pairwise dissimilarity measures to construct metric functions for ABC. Based on the

same idea, we introduced a new estimator, that is computed without using the likelihood of the model. Some examples, all related to the estimation of the mean in a bivariate normal model, were considered to test the perfomance of the proposed estimation procedures. The results appear promising and suggest to further explore the use of dissimilarity-based metrics to *efficiently* summarize data and perform likelihood-free inference. In particular, future research work will involve the study of different metrics in the setting of longitudinal data, where for each subject one observes a sequence of states over discrete time.

An interesting opportunity to validate our results on the breast cancer natural history will be to estimate the models on new data collected by the Norwegian Breast Cancer Screening Program [61]. Those data are more complete in terms of information about examinations and diagnoses, and they should allow for more detailed analyses of the latent disease process. We successfully completed the formal application procedure to receive these data as part of a joint project with the Norwegian colleagues.

# Bibliography


[1] Aalen, OO. "Understanding disease processes". In: *Statistics in Medicine* (2010), pp. 1159–1160.

[2] Abbott, A. "Sequence Analysis: New Methods for Old Ideas". In: *Annual Review of Sociology* 21 (1995), pp. 93–113.

[3] Abrahamsson, L and Humphreys, K. "A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density". In: *Statistical Methods in Medical Research* 25.4 (2013), pp. 1620–1637.

[4] Ahn, HS, Kim, HJ, and Welch, HG. "Korea's Thyroid-Cancer "Epidemic"–Screening and Overdiagnosis". In: *New England Journal of Medicine* 371.19 (2014), pp. 1765–1767.

[5] Alfani, G, Bonetti, M, and Fochesato, M. "Pandemics and socio-economic status. Evidence from the plague of 1630 in northern Italy". In: *Working paper* (2021).

[6] Beaumont, MA, Zhang, W, and Balding, DJ. "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4 (2002), pp. 2025–2035.

[7] Beckmann, K et al. "Estimates of over-diagnosis of breast cancer due to population-based mammography screening in South Australia after adjustment for lead time effects". In: *Journal of Medical Screening* 22.3 (2015), pp. 127–135.

[8] Bergqvist, O. *Calibration of Breast Cancer Natural History Models Using Approximate Bayesian Computation*. 2020. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-273605.

[9] Bernardo, JM and Smith, AFM. "Bayesian Theory". In: *Bayesian Theory* (2008), pp. 1–595.

[10] Bernton, E 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.

[11] Beyene, J. "Uses of the fast Fourier transform (FFT) in exact statistical inference". PhD thesis. 2001.

[12] Bhargava, S et al. "Lower attendance rates in immigrant versus non-immigrant women in the norwegian breast cancer screening programme". In: *Journal of Medical Screening* 25.3 (2018), pp. 155–161.

[13] Bidoli, E et al. "Worldwide Age at Onset of Female Breast Cancer: A 25-Year Population-Based Cancer Registry Study". In: *Scientific Reports* 9.1 (2019), pp. 1–8.

[14] Blum, MGB et al. "A comparative review of dimension reduction methods in approximate bayesian computation". In: *Statistical Science* 28.2 (2013), pp. 189–208.

[15] Bonetti, M and Pagano, M. "The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection." In: *Statistics in Medicine* 24.5 (2005), pp. 753–73.

[16] Bonetti, M et al. "Parametric models for interpoint distances and their use in biosurveillance". In: *Proc Amer Stat Assoc, Biometrics* (2003).

[17] Bonetti, M et al. "The Distribution of Interpoint Distances". In: *Bioterrorism: Mathematical Modeling Applications in Homeland Security*. 2003. Chap. 4, pp. 87–106.

[18] Borchers, HW. *pracma: Practical Numerical Math Functions*. R package version 2.3.3. 2021. URL: `https://CRAN.R-project.org/package=pracma`.

[19] Castillo, J del and Pérez-Casany, M. "Overdispersed and underdispersed Poisson generalizations". In: *Journal of Statistical Planning and Inference* 134.2 (2005), pp. 486–500.

[20] Castillo, J del and Perez-Casany, M. "Weighted Poisson distributions for overdispersion and underdispersion situations". In: *Annals Institute of Statistical Mathematics* 50.3 (1998), pp. 567–585.

[21] Chen, B, Yi, GY, and Cook, RJ. "Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process". In: *Statistics in Medicine* 29.11 (2010), pp. 1175–1189.

[22] Cook, RD and Martin, FB. "A model for quadrat sampling with "visibility bias"". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 345–349.

[23] Csilléry, K, François, O, and Blum, MGB. "Abc: An R package for approximate Bayesian computation (ABC)". In: *Methods in Ecology and Evolution* 3.3 (2012), pp. 475–479.

[24] Drummer, TD and Mcdonald, LL. "Size Bias in Line Transect Sampling". In: *Biometrics* 43.1 (1987), pp. 13–21.

[25] Gøtzsche, PC and Jørgensen, K. "Screening for breast cancer with mammography". In: *Cochrane Database of Systematic Reviews* (6 2013).

[26] Gove, JH. "Moment and maximum likelihood estimators for Weibull distributions under length- and area-biased sampling". In: *Environmental and Ecological Statistics* 10.4 (2003), pp. 455–467.

[27] Gutmann, MU et al. "Likelihood-free inference via classification". In: *Statistics and Computing* 28.2 (2018), pp. 411–425.

[28] Hastie, T, Tibshirani, R, and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[29] Henningsen, A and Toomet, O. "maxLik: A package for maximum likelihood estimation in R". In: *Computational Statistics* 26.3 (2011), pp. 443–458.

[30] Hu, P and Zelen, M. "Planning clinical trials to evaluate early detection programmes". In: *Biometrika* 84.4 (1997), pp. 817–829.

[31] Isheden, G and Humphreys, K. "Modelling breast cancer tumour growth for a stable disease population". In: *Statistical Methods in Medical Research* 28.3 (2019), pp. 681–702.

[32] Koenker, R. *Quantile Regression*. Cambridge University Press, 2005.

[33] Kokonendji, CC, Mizère, D, and Balakrishnan, N. "Connections of the Poisson weight function to overdispersion and underdispersion". In: *Journal of Statistical Planning and Inference* 138.5 (2008), pp. 1287–1296.

[34] Kopans, D et al. "A simple model of breast carcinoma growth may provide explanations for observations of apparently complex phenomena". In: *Cancer* 97 (June 2003), pp. 2951–9.

[35] Lehmann, EL and Casella, G. *Theory of Point Estimation*. Second. New York, NY, USA: Springer-Verlag, 1998.

[36] Lenz, M et al. "Estimating real cell size distribution from cross-section microscopy imaging". In: *Bioinformatics* 32.17 (2016), pp. i396–i404.

[37] Lintusaari, J et al. "Fundamentals and recent developments in approximate Bayesian computation". In: *Systematic Biology* 66.1 (2017), e66–e82.

[38] Little, RJA and Rubin, DB. *Statistical analysis with missing data*. Second. New York: Wiley, 2002.

[39] Marin, JM et al. "Likelihood-free Model Choice". In: *ArXiv* (2015), pp. 1–21. arXiv: 1503.07689.

[40] Marjoram, P. "Approximation Bayesian computation". In: *OA Genetics* 1.1 (2013), pp. 1–5.

[41] Marjoram, P et al. "Markov chain Monte Carlo without likelihoods". In: *Proc Natl Acad Sci U S A* 100.26 (2005), pp. 15324–8.

[42] Marmot, MG et al. "The benefits and harms of breast cancer screening: an independent review". In: *The Lancet* 380.9855 (2012), pp. 1778–1786.

[43] McLeish, DL and Small, CG. *The theory and applications of statistical inference functions. Lecture notes in statistics V.44*. New York, NY: Springer, 1988.

[44] Mosler, K. *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach.* New York: Springer-Verlag, 2002.

[45] Mukhopadhyay, N. *Probability and statistical inference*. Dekker, 2000.

[46] Nelson, HD et al. "Effectiveness of breast cancer screening: Systematic review and meta-analysis to update the 2009 u.s. preventive services task force recommendation". In: *Annals of Internal Medicine* 164.4 (2016), pp. 244–255.

[47] Pacchiardi, L et al. "Distance-learning For Approximate Bayesian Computation To Model a Volcanic Eruption". In: *Sankhya B* 83 (2021), pp. 288–317.

[48] Pathirana, T et al. "Lifetime risk of prostate cancer overdiagnosis in Australia: Quantifying the risk of overdiagnosis associated with prostate cancer screening in Australia using a novel lifetime risk approach". In: *BMJ Open* 9.3 (2019), pp. 1–7.

[49] Patil, GP and Rao, CR. "Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families". In: *Biometrics* 34.2 (1978), pp. 179–189.

[50] Pudlo, P et al. "Reliable ABC model choice via random forests". In: *Bioinformatics* 32.6 (Nov. 2015), pp. 859–866.

[51] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[52] Samuel, MD and Pollock, KH. "Correction of Visibility Bias in Aerial Surveys Where Animals Occur in Groups". In: *The journal of wildlife management* 45.4 (2012), pp. 993–997.

[53] D Sankoff and JB Kruskal, eds. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, 1983, pp. 1–44.

[54] Scheaffer, RL. "Size-Biased Sampling". In: *Technometrics* 14.3 (1972), pp. 635–644.

[55] Schmidt, F et al. "Size bias in galaxy surveys". In: *Physical Review Letters* 103.5 (2009), pp. 1–4.

[56] Sisson, S. A., Fan, Y., and Tanaka, Mark M. "Sequential Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1760–1765.

[57] Sisson, SA, Fan, Y, and Beaumont, M. *Handbook of Approximate Bayesian Computation*. New York: Chapman and Hall/CRC, 2018.

[58] Struttura Promozione della Salute e Screening. *Gli screening oncologici in Lombardia. Report dati 2015 (survey 2016) e dati 2016 (survey 2017 prima parte)*. Tech. rep. Regione Lombardia, 2017. URL: https://www.regione.lombardia.it/wps/wcm/connect/5cfe19ff-e3c7-4b1e-a336-f1e1b9670c42/Report_screening_2016_Luglio2017.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-5cfe19ff-e3c7-4b1e-a336-f1e1b9670c42-n0c4p8H.

[59] Sweeting, MJ, Farewell, VT, and De Angelis, D. "Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C". In: *Statistics in Medicine* 29.11 (2010), pp. 1161–1174.

[60] Tancredi, A. "Approximate Bayesian inference for discretely observed continuous-time multi-state models". In: *Biometrics* 75 (2019), pp. 966–977.

[61] The Research Council of Norway. *Research-based evaluation of the Norwegian Breast Cancer Screening Program*. Tech. rep. 2015.

[62] U.S. Preventive Services Task Force. *Modeling Report: Collaborative Modeling of U.S. Breast Cancer Screening Strategies.* Tech. rep. 2016. URL: https://www.uspreventiveservicesorg/Page/Document/modeling-report-collaborative-modeling-of-us-breast-cancer-1/breast-cancer-screening1.

[63] Van Oortmarssen, GJ, Boer, R, and Habbema, JD. "Modelling issues in cancer screening". In: *Statistical Methods in Medical Research* 4.1 (1995), pp. 33–54.

[64] Ventura, L et al. "Mammographic breast cancer screening in Italy: 2011-2012 survey". In: *Epidemiologia e prevenzione* 39.3 Suppl 1 (2015), pp. 21–29.

[65] Vuong, QH. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". In: *Econometrica;Journal of the Econometric Society* 57.2 (1989), pp. 307–333.

[66] Waks, AG and Winer, EP. "Breast Cancer Treatment: A Review". In: *JAMA* 321.3 (Jan. 2019), pp. 288–300.

[67] Weedon-Fekjær, H et al. "Breast cancer tumor growth estimated through mammography screening data". In: *Breast Cancer Research* 10.3 (2008), pp. 1–13.

[68] Weedon-Fekjær, H et al. "Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results". In: *Journal of Medical Screening* 12.4 (2005), pp. 172–178.

[69] Yu, K and Jones, MC. "Local Linear Quantile Regression". In: *Journal of the American Statistical Association* 93.441 (1998), pp. 228–237.