From the Unit of Biostatistics
Institute of Environmental Medicine
Karolinska Institutet, Stockholm, Sweden

TOPICS ON MATHEMATICAL STATISTICS FOR MEDICAL APPLICATIONS: SUMMARY MEASURES AND EXACT SIMULATION OF DIFFUSIONS

Celia García-Pareja

Karolinska Institutet

Stockholm 2019

All published papers reproduced with permission
Published by Karolinska Institutet
Printed by Universitetsservice US-AB 2019

Typeset by the author using LaTeX 2_ε, template courtesy of Dr. Andrea Discacciati

ISBN 978-91-7831-523-9

TOPICS ON MATHEMATICAL STATISTICS FOR MEDICAL APPLICATIONS: SUMMARY MEASURES AND EXACT SIMULATION OF DIFFUSIONS

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Celia García-Pareja

Principal supervisor:
Professor Matteo Bottai
Karolinska Institutet
Institute of Environmental Medicine
Unit of Biostatistics

Co-supervisor(s):
Professor Henrik Hult
KTH Royal Institute of Technology
Department of Mathematics
Division of Mathematical Statistics

Professor Anna Mia Ekström
Karolinska Institutet
Department of Public Health Sciences
Global and Sexual Health

Opponent:
Associate Professor Paul Jenkins
University of Warwick
Department of Statistics

Examination board:
Professor Ola Hössjer
Stockholm University
Department of Mathematics

Professor Paul Dickman
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Professor Niels Richard Hansen
University of Copenhagen
Department of Mathematical Sciences

A papá Jesús. Ésta sí, acabada.

Abstract

The first part of this thesis deals with exact simulation of multidimensional diffusion processes. The main contribution is the development of an exact rejection algorithm for sampling coupled Wright-Fisher diffusions. The algorithm's output provides a skeleton from the diffusion sampled at a random number of time points. To complete the simulation scheme, an exact simulation strategy for sampling from the corresponding multidimensional Wright-Fisher bridges is also presented. Besides the aforementioned results, which have interest on their own, sampling strategies for coupled Wright-Fisher diffusions are of importance to assess inferential methods that have applications to the estimation of evolutionary parameters such as selection or mutation of genetic traits over time. In particular, the coupled Wright-Fisher model tracks pairwise allele interactions across different loci over time. This model has applications in population genetics, for instance, to the analysis of interactions of networks of loci such as those encountered in the study of antibiotic resistance.

The second part of this thesis presents contributions in statistical methodology for summarizing probability distributions and dealing with commonly found problems in survival analysis settings. First, a novel summary measure for probability distributions is presented, along with a general estimation strategy based on quantile function estimators that allows for inclusion of covariates in a regression framework. Consistency and asymptotic normality results are also provided. This general framework allows for extension of the use of the measure in several scenarios such as life expectancy estimation, where observed variables are often censored. Results concerning the use of the measure in combination with the Cox proportional hazards and the accelerated failure time models are also provided.

Sammanfatning

Den första delen av avhandlingen behandlar exakt simulering av flerdimensionella diffusionsprocesser. Huvudresultatet består av en exakt simuleringsalgoritm för kopplade Wright-Fisher diffusioner. Algoritmen tillhandahåller utfall av diffusionsprocessen i slumpmässiga diskreta tidpunkter och kompletteras genom simulering av motsvarande bryggprocess för Wright-Fisher diffusioner. Förutom resultaten nämnda ovan, vilka är av oberoende intresse, så är simulering av kopplade Wright-Fisher diffusioner relevanta för estimering av evolutionära parametrar som till exempel selektions- eller mutationsintensitet. Den kopplade Wright-Fisher diffusionen är särskilt utvecklad för att beskriva parvis allelinteraktion vid olika loci över tid. Modellen har tillämpningar inom populationsgenetik för analys av interaktion av nätverk av loci som observeras inom studier av antibiotiker resistens.

Den andra delen av avhandlingen behandlar statistiska metoder för beskrivning av sannolikhetsfördelningar och vanliga problem inom överlevnadsanalys. Ett nytt mått för beskrivning av sannolikhetsfördelningar presenteras tillsammans med en generell estimeringsteknik baserad på kvantilfunktioner, vilken möjliggör inklusion av förklaringsvariabler i form av en regressionsmodell. Resultat för konsistens och asymptotisk normalitet bevisas. Det generella ramverk som presenteras möjliggör utvidgningar av måttet till estimering av förväntad livslängd där censurerade variabler är vanligt förekommande. Resultat för användning av måttet i kombination med Cox proportionella riskmodeller och accelererade feltidsmodeller är också inkluderade.

List of publications

I. Celia García-Pareja, Henrik Hult, and Timo Koski
Exact simulation of coupled Wright-Fisher diffusions
Submitted 2019

II. Celia García-Pareja, and Matteo Bottai
On mean decomposition for summarizing conditional distributions
Stat 2018; 7:e208

III. Celia García-Pareja, Michele Santacatterina, Anna Mia Ekström, and Matteo Bottai
Conditional life expectancy estimation by ordered fractions of population with censored data
Manuscript 2019

IV. Michele Santacatterina, Celia García-Pareja, Rino Bellocco, Anders Sönnerborg, Anna Mia Ekström, and Matteo Bottai
Optimal probability weights for estimating causal effects of time-varying treatments with marginal structural Cox models
Statistics in Medicine 2019; 38:1891–1902

The articles will be referred to in the text by their Roman numerals, and are reproduced in full at the end of the thesis.

Contents

Chapter 1

Introduction

> *"I like crossing the imaginary boundaries people set up between different fields—it's very refreshing. There are lots of tools, and you don't know which one would work. It's about being optimistic and trying to connect things."*
>
> —Maryam Mirzakhani

This introductory chapter is devoted to provide an overview of general concepts needed for a smoother understanding of the results presented in this thesis' constituent papers, and to put its contributions in a more general context.

1.1 Exact simulation of diffusion processes

Diffusion models appear in numerous applied fields, including engineering, biophysics, finance or biology. Such models, which describe the evolution of phenomena that change randomly over time, are governed by diffusion processes defined as the solution of a stochastic differential equation (SDE). In their vast majority, these processes have analytically intractable distributions and therefore can not be simulated naively. Development of simulation techniques for diffusion processes is therefore a widely studied field within applied probability.

Consider the scalar Itô diffusion $X = \{X_t, t \geq 0\}$, solution of the SDE

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t,\ X_0 = x_0,\ t \geq 0, \tag{1.1}$$

where $B = \{B_t, t \geq 0\}$ is a Brownian motion, and $b : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ are measurable functions assumed to satisfy

i) (*Locally Lipschitz*): For each $M > 0$ there exists a constant $K_M > 0$ such that

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq K_M|x - y|, \text{ for all } |x| \leq M,\ |y| \leq M,$$

ii) (*Bounded linear growth*): There exists a constant $K > 0$ such that

$$|b(x)|^2 + |\sigma(x)|^2 \leq K^2(1 + x^2), \text{ for all } x \in \mathbb{R}.$$

The regularity conditions i) and ii) guarantee that X is a weakly unique solution of (1.1), that is, all solutions of (1.1) have identical finite dimensional distributions, which is sufficient for simulation purposes.

In what follows, our interest lies on sampling X over a fixed time interval $[0, T]$ and we refer to the distribution law of X, say $\mathbb{Q}$, as the probability measure induced by X in the measurable space $(C, \mathscr{C})$, where C refers to continuous mappings from $[0, T]$ to $\mathbb{R}$ and $\mathscr{C}$ is the corresponding cylinder sigma algebra generated by the coordinate functions $h(t)$, with $t \in [0, T]$, where h is a typical element of C.

Realizations of X, i.e., its sample paths, are infinite dimensional and can only be represented for a finite subset of time points $\mathscr{T} = \{t_0, \ldots, t_k\} \subseteq [0, T],\ t_0 = 0,\ t_{j-1} < t_j, \forall j \in \{1, \ldots, k\}$, that is, a skeleton $\{X_{t_0}, \ldots, X_{t_k}\}$. In case the skeleton is simulated from the *exact* transition probability function $f(x, \cdot; s)$ that governs X, the sampling strategy is called exact. For example, in case X is a Brownian motion, i.e., satisfying

$$dX_t = dB_t,\ X_0 = x_0,\ t \geq 0,$$

its transition probability (density) function reads

$$f(x, \cdot; s) = \frac{1}{\sqrt{2\pi s}} \exp \frac{(y - x)^2}{2s}, \tag{1.2}$$

where we condition $x = B_{t_{j-1}}$, $y = B_{t_j}$ is the new skeleton point we aim to recover, and $s = t_j - t_{j-1}$ denotes the distance between two consecutive time points in $\mathscr{T}$. It is clear then, that a skeleton of a sample path of B can be recovered from (1.2) *exactly*.

However, as mentioned above, the transition probability function of interest is often not known in closed form (as for instance in the problem presented in Paper I of this thesis), which poses difficulties for sampling exactly. Given the lack of more suitable techniques, the most common approach is to use time-discretized methods that sample from an *approximation* of the true unknown distribution of interest. For example, one can approximate (1.1) by an Euler-Maruyama discretization, yielding

$$dX_{t_j+s} = b(X_{t_{j-1}})s + \sigma(X_{t_{j-1}})\sqrt{s}Z,\ X_0 = x_0,\ j = 1, \ldots, k, \tag{1.3}$$

where Z follows a standard normal distribution. A skeleton of X is then sampled iteratively using (1.3). Approximation schemes, however, provide biased samples and are computationally expensive if one aims to minimize approximation errors (see, for example, Kloeden and Platen (1992) for a comprehensive overview).

In this context, recently proposed exact simulation methods constitute a desirable alternative and have become increasingly popular within the applied probability community. The family of exact rejection algorithms presented in Beskos and Roberts (2005), Beskos et al. (2006), Beskos et al. (2008), and publications therein, use Brownian candidates in a rejection sampling scheme to recover samples from the desired target distribution. However, in its most elementary form, the algorithm imposes bounding conditions on the drift of the target diffusion

and its derivative. Given that such assumption restricts the family of diffusions for which the sampling strategy is valid, the algorithms presented in Beskos et al. (2006) and Beskos et al. (2008) mean to provide sampling schemes under more relaxed conditions. Further extensions include also the algorithm provided in Casella and Roberts (2008) that permits simulation of killed diffusions and can be applied to double barrier problems. Moreover, Chen and Huang (2013) propose a localized exact algorithm that relaxes any boundedness assumptions by restricting the sampling to controlled smaller pieces of the target path that can be concatenated afterwards.

It is worth mentioning that other alternative techniques have also been proposed, such as the one in Blanchet and Zhang (2017) that presents an exact algorithm for simulation of multivariate diffusions based on tolerance enforced simulation and rough paths analysis. Their algorithm overcomes the more restrictive assumptions required in Beskos and Roberts (2005), Beskos et al. (2006) and Beskos et al. (2008), but has, however, infinite expected running time.

In the next section we overview the sampling strategy of the *basic* exact rejection algorithm, i.e., the one first presented in Beskos and Roberts (2005). For further details on its variants, we refer the reader to Beskos et al. (2006), Beskos et al. (2008) and subsequent publications.

1.1.1 Exact rejection algorithm

The main idea of the family of exact rejection algorithms relies on using a rejection scheme to recover samples from the distribution of the process solution of (1.1). As with all rejection sampling algorithms, one of the main challenges in the task lies on finding good candidate processes, which we comment in greater detail later on.

Before we proceed, we briefly recall the general rejection sampling scheme for sampling random variables. Let ν be some *target* probability distribution of a random variable Y from which we do not know how to sample, and that is absolutely continuous w.r.t. some other distribution μ called the *candidate* distribution of a random variable X, and from which we do know how to sample. The key of rejection sampling is to construct an event (the *decision event*) that occurs with probability ε proportional to the Radon-Nikodým derivative of ν w.r.t. μ (the *rejection probability*). Under these conditions, it can be proved that conditioned on a realization $X = x$ from μ, the occurrence of the decision event implies that $Y = x$ is as if sampled from ν. Thus, once the rejection probability is available, it only remains to sample from μ, and evaluate the proposed *decision event*. If conditioning on $X = x$ the event has occurred, $Y = x$ is accepted as a sample from ν.

Let now U be a uniform standard random variable, that is, $U \sim \mathcal{U}(0,1)$. Because U satisfies $\Pr(U \leq u) = u$ for any given threshold u, a simple decision event is that of a uniformly distributed variable U being smaller than the rejection probability ε evaluated at $X = x$. A general rejection sampling algorithm is shown in Algorithm 1.

Note that other than satisfying the absolute continuity assumption, a desirable property for μ is that it is *as similar as* possible to ν, so that ε is, on average, as large as possible, and less samples from the candidate μ are discarded.

Algorithm 1 General rejection sampling scheme

1 Simulate $X \sim \mu$.
2 Simulate $U \sim \mathcal{U}(0,1)$
3 **if** $U \leq \varepsilon(X) \propto \dfrac{d\nu}{d\mu}(X)$ **then**
4 **return** X (accepted as sample from ν)
5 **else**
6 Go back to Step 1.
7 **end if**

Going back to our original problem, we explain now the mechanics of the exact rejection algorithm proposed in Beskos and Roberts (2005). Let $Y = \{Y_t, t \geq 0\}$ be a uniquely weak solution of an SDE of the form of (1.1). Then, without loss of generality, one can apply the Lamperti transformation $Y_t \mapsto \eta(Y_t)$

$$X_t = \eta(Y_t) = \int_{\xi}^{Y_t} \frac{1}{\sigma(u)} du,$$

where ξ is an element of the state space of Y. Thus, for what remains of this section we consider the family of SDEs with unit diffusion coefficient

$$dX_t = \mu(X_t)dt + dB_t,\ X_0 = x_0,\ t \geq 0, \tag{1.4}$$

where

$$\mu(X_t) = \frac{b(\eta^{-1}(X_t))}{\sigma(\eta^{-1}(X_t))} - \frac{1}{2}\sigma'(\eta^{-1}(X_t)).$$

Let $\mathbb{Q}_{x_0}$ be the law of the target process X, and $\mathbb{P}_{x_0}$ the law of a Brownian motion B starting at $B_0 = x_0$, that will serve as our candidate process. Provided that Novikov's condition is satisfied, i.e.,

$$\mathrm{E}_{\mathbb{P}}\left[\exp\left\{\frac{1}{2}\int_0^T \mu^2(B_t)dt\right\}\right] < \infty,$$

the Radon-Nykodým derivative of $\mathbb{Q}_{x_0}$ w.r.t. $\mathbb{P}_{x_0}$ can be written by means of the Girsanov's transformation of measures, see Karatzas and Shreve (1998),

$$\frac{d\mathbb{Q}_{x_0}}{d\mathbb{P}_{x_0}}(B) = \exp\left\{\int_0^T \mu(B_t)dB_t - \frac{1}{2}\int_0^T \mu^2(B_t)dt\right\}. \tag{1.5}$$

If in addition, $\mu(\cdot)$ is assumed to be differentiable everywhere, we can rewrite (1.5) using Itô's lemma as

$$\frac{d\mathbb{Q}_{x_0}}{d\mathbb{P}_{x_0}}(B) = \exp\left\{A(B_T) - A(x_0) - \frac{1}{2}\int_0^T (\mu^2(B_t) + \mu'(B_t))dt\right\}, \tag{1.6}$$

where $A(x) := \int_0^x \mu(u)du$.

In order to use (1.6) as a rejection probability we need to ensure that the expression is a.s. bounded, which is likely to require some boundedness condition on $A(\cdot)$. In order to relax

this condition, we will slightly change the proposed candidate for a biased Brownian motion, $\overline{B} = \{B_t, t \in [0,T] : B_T \sim g\}$, where we set the density function

$$g(z) := \frac{1}{c} \exp\left\{ \frac{-(z-x_0)^2}{2T} + A(z) \right\}, \text{ with } c \text{ normalizing constant,}$$

and where we assume that $\exp\{-(z-x_0)^2/(2T) + A(z)\}$ is integrable for all $z \in \mathbb{R}$. Note that this condition on $A(\cdot)$ is milder than assuming it is bounded in its entire domain.

The process $\overline{B}$ is therefore a Brownian motion whose last point of a path in $[0,T]$, B_T, is distributed according to the density function $g(\cdot)$. The paths of $\overline{B}$ can be simulated by simply drawing first $B_T \sim g$ and then using the dynamics of a Brownian bridge from x_0 to B_T.

Let $\mathbb{Z}$ be the law of $\overline{B}$, then

$$\frac{d\mathbb{Z}_{x_0}}{d\mathbb{P}_{x_0}}(B) = \frac{c^{-1}\exp\{-(B_T-x_0)^2/(2T) + A(B_T)\}}{(\sqrt{2\pi T})^{-1}\exp\{-(B_T-x_0)^2/(2T)\}} \propto \exp\{A(B_T) - A(x_0)\}. \tag{1.7}$$

Combining (1.7) with the expression in (1.6) yields

$$\frac{d\mathbb{Q}_{x_0}}{d\mathbb{Z}_{x_0}}(\overline{B}) = \frac{d\mathbb{Q}_{x_0}}{d\mathbb{P}_{x_0}}(B)\frac{d\mathbb{P}_{x_0}}{d\mathbb{Z}_{x_0}}(\overline{B}) \propto \exp\left\{ -\int_0^T \frac{1}{2}(\mu^2(B_t) + \mu'(B_t))dt \right\}.$$

Assume now that the function $\phi(x) := \frac{1}{2}[\mu^2(x) + \mu'(x)]$ is bounded by constants K^- and K^+, that is, $K^- \leq \phi(x) \leq K^+$, then

$$\varepsilon(\overline{B}) = \exp\left\{ -\int_0^T (\phi(B_t) - K^-)dt \right\} \leq 1, \tag{1.8}$$

is our rejection probability.

Recalling the general rejection sampling scheme, now we ought to construct an event that occurs with probability $\varepsilon(\overline{B})$, and then, given a sample from the candidate $\overline{B}$, evaluate whether it has occurred or not. Nonetheless, it is clear by the integral in (1.8) that for evaluating it exactly (that is, without involving any numerical approximation), we would require to store infinitely many points of a sampled path of $\overline{B}$ in $[0,T]$. The next subsection is devoted to explain in detail how this can be done by only sampling the candidate at finitely many points.

Rejection probability: exact evaluation

Let N be the number of points of an homogeneous spatial Poisson process Φ with unit intensity that lie in $G \subset \mathbb{R}^2$, a bounded region on the real plane. Then the probability that $N = n$ is

$$\Pr(N = n) = \frac{(|G|)^n}{n!}\exp\{-|G|\}, \text{ where } |G| \text{ denotes area of } G.$$

Consider now $G = \{(x,y) \in H : y \leq (\phi(x) - K^-),\ H = [0,T] \times [0, K^+ - K^-]\}$. Then, given a realization of the candidate $\overline{B}$,

$$\Pr(N = 0 | \overline{B}) = \exp\left\{ - \int_0^T (\phi(B_t) - K^-) dt \right\},$$

is the probability that no points of Φ lie on G, i.e., below the graph of $t \mapsto \phi(B_t) - K^-$. We refer to this event as ω_Φ.

Note that ω_Φ happens exactly with probability $\varepsilon(\overline{B})$ and that it can be evaluated by storing the candidate path only at finitely many points. The procedure is the following. First, we draw a sample from the Poisson process Φ, that provides a collection of points $\{(t_j, \psi_j) : j = 1, \ldots, J\}$ distributed in the area H according to the law of Φ. To evaluate whether ω_Φ has occurred, we only need to sample $\overline{B}$ at the time points $(t_j)_{j=1}^J$, and check whether the coordinates $(\psi_j)_{j=1}^J$ lie above or below the graph of $\phi(B_t) - K^-$. Figure 1.1 shows two examples in which we would reject (left panel) or accept (right panel) the drawn skeleton from $\overline{B}$ depending on whether ω_Φ had occurred or not.

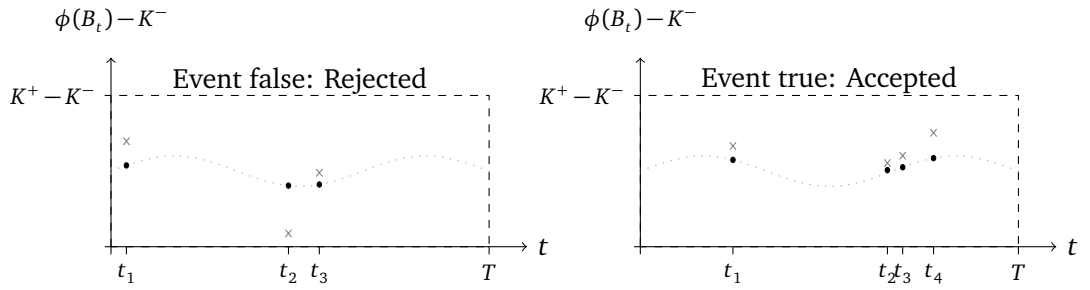

Figure 1.1: Evaluation of the decision event after drawing a candidate sample $\overline{B}$. On the left, the sample is rejected because some of the points drawn from Φ lie below the graph of $\phi(B_t) - K^-$. On the right, the sample is accepted because all points lie above.

In order to use the acceptance-rejection mechanism that we just described to obtain samples from $\mathbb{Q}_{x_0}$, the assumptions imposed on the drift $\mu(\cdot)$ of (1.3) are rather strong. Namely, it requires $K^- \leq \phi(x) \leq K^+$. This condition is necessary to ensure the points sampled from the Poisson process Φ will lie on a bounded region H such that contains the graph of $\phi(x)$.

Further extensions of the exact rejection algorithm with biased Brownian candidates are mainly devoted to relax this boundedness condition. Roughly speaking, they consider subintervals within $[0,T]$ where the minimum and maximum of $\phi(x)$ are well defined, and provide the boundaries of the region in $\mathbb{R}^2$ where to sample Φ.

The algorithm

Algorithm 2 presents the rejection sampling scheme that we just described. As a first step, the last point on the candidate path of the biased Brownian motion is sampled, followed by the sampling from the Poisson process Φ in the bounded region $H = [0,T] \times [0, K^+ - K^-]$. Then, the candidate path is simulated at the random time points obtained when sampling from Φ and

finally the decision event is evaluated. In case all sampled points from Φ lie above the graph of $\phi(\cdot)$ the candidate skeleton is accepted. Otherwise, the routine starts again until acceptance. It is important to mention that while in some cases the acceptance step might require several iterations of the algorithm, that is, in cases where the Brownian motion might not be a *good* candidate process and the average rejection probability is high, Algorithm 2 is ensured to end in finite time.

Algorithm 2 Exact algorithm for simulating skeletons of paths $(X_t)_{t\in[0,T]}$ of a diffusion process with law $\mathbb{Q}_{x_0}$

1 Simulate $B_T \sim g$
2 Simulate Φ, the Poisson process on $[0,T]\times[0,K^+-K^-]$
3 Given $\Phi=\{(t_j,\psi_j): j=1,\dots,J\}$, simulate $\overline{B}\sim\mathbb{Z}_{x_0}$ at times $\{t_1,\dots,t_J\}$.
4 **if** $\phi(B_{t_j})-K^-\leq\psi_j,\ \forall j$ **then**
5 **return** $\{(t_j,B_{t_j}),\ \forall j\}\cup\{(T,B_T)\}$
6 **else**
7 Go back to Step 1.
8 **end if**

Note that Algorithm 2 provides a skeleton of X but drawn at a random collection of time instances $(t_j)_{j=1}^J$. However, once the skeleton is accepted, other points of the path can be recovered using the corresponding Brownian bridges, and with no further reference to the target law $\mathbb{Q}_{x_0}$ needed.

Disadvantages of the presented approach include situations in which the Brownian motion (or its biased counterpart) are not appropriate candidate processes, which calls for other exact simulation strategies that provide new candidates and take advantage of the exact rejection algorithm approach. Moreover, because the Lamperti transformation can not be generalized for multidimensional SDEs, extending Algorithm 2 to the multidimensional case is only possible for diffusions with identity diffusion matrix. However, as shown in Paper I, if other candidate processes are available, there are cases in which such extension is indeed possible.

Other candidate processes: unsuitability of the Brownian motion

It has been shown that situations in which the target diffusion process has a finite entrance boundary, the Brownian motion results in a poor candidate process, see Jenkins (2013). In this context, other diffusions that mimic the behavior of the target around such boundary can be proposed instead, so that rejection rates are tolerable. One such candidate is a Bessel process because of its entrance boundary at 0 (at least for a certain characterization of its drift), and the possibility to sample it exactly. Another desirable characteristic of the Bessel process is that samples from the corresponding Bessel bridge can also be recovered, which makes it an ideal candidate option for the exact rejection scheme.

Generalization of the exact rejection algorithm, however, is hindered by the general lack of other sampling strategies that allow to recover candidate skeletons without any approximation error. In a recent publication, Jenkins and Spanò (2017) propose a modification of the alter-

nating series method to sample from neutral (one-dimensional) Wright-Fisher diffusions (and their bridges), which in turn, can be used as candidates in a rejection scheme to sample from a wider family of Wright-Fisher diffusions. In brief, the Wright-Fisher diffusion has boundaries at 0 and 1, which depending on the drift can be either entrance, exit or regular, making Brownian motion or the Bessel process unsuitable candidates. Paper I of this thesis presents the modified alternating series sampling strategy in detail and provides an extension of the exact rejection algorithm for sampling from a family of multidimensional Wright-Fisher diffusions.

1.1.2 Simulation of diffusion processes in population genetics

The field of population genetics has gained importance over the last decades due to its role in the study and interpretation of genetic data (Charlesworth and Charlesworth (2017)). For example, as in the well-known genome-wide association studies that focus on the study of variations found in the human genome which are in turn associated with risk of disease, see, for instance, Ehret et al. (2011) or Maurano et al. (2012). Increased availability of this kind of data has called for the development of complex statistical models, in which diffusion processes play a central part.

Methods that estimate parameters from these models are extensively present in the mathematical literature, see Bollback et al. (2008), Malaspinas et al. (2012) or Schraiber et al. (2016), among others. Inferences on such models, however, are often hindered by the lack of reliable and efficient simulation techniques, needed for sampling from known models to be used as ground truth data.

In particular, simulation methods for Wright-Fisher diffusions, which describe the evolution of allele types frequencies over time and are of great interest in population genetics, have been extensively explored. Examples include time discretization methods that are based on the standard Euler-Maruyama approximation scheme but ensure that the simulated paths do not leave the state space $[0, 1]$, see Dangerfield et al. (2012), Schraiber et al. (2013), Neuenkirch and Szpruch (2014)). For some specific models, spectral expansion representations for the transition functions have been derived, and these can be approximated, for instance, by truncating the series expansion, see Song and Steinrücken (2011), Steinrücken et al. (2013), or inserting asymptotic distributional approximations derived from coalescent theory, see Griffiths (1984), Jewett and Rosenberg (2014). Other numerical approximation methods include Williamson et al. (2005) or Schraiber et al. (2013).

Exact simulation techniques, however, are more desirable alternatives, because approximation errors are often difficult to quantify and have to be assessed experimentally. A multidimensional version of the above mentioned Wright-Fisher diffusions are those that take into account possible interactions between different populations, as encountered for example in studies of interacting genes' networks, see Skwark et al. (2017). These diffusions have an additive term on the drift, the coupling term, that accounts for such interactions. Suitable simulation methods for these nested or *coupled Wright-Fisher* models (Aurell et al. (2019)) are of great importance for validating new estimation techniques.

1.2 Summary measures for probability distributions

The conditional mean is the default central tendency measure chosen by many researchers. The most popular regression technique to model conditional means is undoubtedly Ordinary Least Squares (OLS). OLS estimation is computationally simple and provides an optimal linear predictor of the outcome. Efficiency of the OLS estimator, however, is often compromised in observational experiments, when the model errors might depend on the predictors, see, for instance, Seber and Lee (2012). Furthermore, inference on central tendency measures might not carry enough information about the underlying conditional distribution of the data.

Quantile regression (Koenker and Bassett Jr (1978)) presents an advantageous alternative and has been used in numerous studies, see for example, Burgette et al. (2011), Fenske et al. (2011). Quantiles render a picture of the entire underlying distribution rather than summarizing it in a single number. Furthermore, estimation of regression quantiles makes no distributional assumptions about the error term, is more robust to model misspecifications and less sensitive to outliers (Koenker (2005)). Nonetheless, a set of regression quantiles might still lack relevant information for many applications, where sought results rely on average behavior.

Several other measures have also been explored. In Newey and Powell (1987), they propose the asymmetric least squares estimators, or so-called expectiles. Similarly to quantiles, expectiles provide information about the whole distribution of the outcome but they are computationally simpler to obtain and easier to make inference on. Based on the M-estimators presented in Huber et al. (1964), Breckling and Chambers (1988) suggested the M-quantiles. M-quantiles aim to estimate location measures from the distribution, while offering the robustness inherited from the M-estimator. A major disadvantage of these methods is the lack of interpretability of the respective measures, which many authors have intended to relate with quantiles, see Jones (1994), Abdous and Remillard (1995). As a consequence, expectiles and M-quantiles have become less attractive for the practical use.

Regression-based estimation methods that investigate the underlying distribution of the data are still to be further explored. When estimating regression quantiles, for instance, those located in low density regions of the conditional distribution are harder to estimate and tend to yield poorer inferences, while inference on high-density quantiles tends to be significantly more precise. While the OLS estimator represents a suitable summary measure in many scenarios, its estimation relies on global properties of the distribution and thus, it is extremely sensitive to the presence of outliers, see Cook (1977). In order to solve this existing gap, Paper II of this thesis proposes a summary measure in the spirit of combining the best of both worlds.

1.2.1 Quantile regression

Results presented in Paper II are shown for a general estimation framework. However, all practical examples and discussions are based on quantile regression estimates. This subsection provides a brief overview on this particular model. For a comprehensive view, see Koenker and Bassett Jr (1978), Koenker and Machado (1999), Koenker and Hallock (2001), Koenker and

Xiao (2002), and Koenker (2005).

Let Y be a random variable in $\mathbb{R}$ with cumulative distribution function $F(\cdot)$ such that $F(y) = \Pr(Y \leq y)$. Then, its quantile function $Q(\cdot)$ is a left-continuous function defined as

$$Q(p) := F^{-1}(p) = \inf\{y \in \mathbb{R} : F(y) \geq p\},$$

for $p \in (0,1)$. Note that both F and Q completely characterize the distribution of Y. If we consider now a collection of fixed covariates or explanatory variables of interest $x_1, \ldots, x_m$, the quantile regression model assumes that there exists a $m+1$ vector of regression coefficients $\beta^T = (\beta_0, \ldots, \beta_m)$, where T refers to a transposed column vector, such that

$$Q(p|x) = x^T \beta, \tag{1.9}$$

with $x^T = (1, x_1, \ldots, x_m)$. The conditional quantile function of Y given x is then

$$Q(p|x) := F^{-1}(p|x) = \inf\{y \in \mathbb{R} : F(y - x^T\beta) \geq p\}.$$

One of the main advantages of quantile regression is that the error function is distribution-free. This makes the model robust to distributional assumptions, unlike for example, the case of OLS.

The estimation strategy presented in Koenker and Bassett Jr (1978) proposes to recover regression quantiles as the solution of a certain optimization problem derived from a simple decision theory problem. We first present how does this work in the univariate case, where we want to recover sample quantiles.

Consider a loss defined as the following piece-wise linear function

$$\rho_p(y) = \begin{cases} y(p-1) \text{ for } y < 0, \\ yp \text{ otherwise }, \end{cases}$$

or more compactly

$$\rho_p(y) = y(p - I(y < 0)),$$

where I denotes the indicator function. The aim is then to find $\hat{y}$ such that minimizes the expected loss, that is,

$$\mathrm{E}[\rho_p(Y - \hat{y})] = (p-1)\int_{-\infty}^{\hat{y}} (y - \hat{y}) dF(y) + p\int_{\hat{y}}^{\infty} (y - \hat{y}) dF(y).$$

Differentiating w.r.t. $\hat{y}$ one obtains

$$0 = (1-p)\int_{-\infty}^{\hat{y}} dF(y) - p\int_{\hat{y}}^{\infty} dF(y) = F(\hat{y}) - p, \tag{1.10}$$

which shows that $\rho_p(\cdot)$ is an unbiased estimating equation for the p-th quantile of Y. Note

that in case Y is continuous the solution $\hat{y}$ is unique, whereas in case Y is discrete there are a range of values solution of (1.10) from which we choose the smallest in order to satisfy the convention of Q being left-continuous.

Let now $Y_1, \ldots, Y_n$ be i.i.d. samples from Y. Then, we can replace F by the empirical distribution function $F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$ yielding

$$\min_{\hat{y}\in\mathbb{R}} \int_{\mathbb{R}} \rho_p(y-\hat{y}) dF_n(y) = \min_{\hat{y}\in\mathbb{R}} \sum_{i=0}^{n} \rho_p(y_i-\hat{y}) = \min_{\hat{y}\in\mathbb{R}} \sum_{i=0}^{n} (y_i-\hat{y})(p - I(y_i < \hat{y})), \quad (1.11)$$

where $\hat{y}$ is the p-th sample quantile, $\hat{y} = \hat{Q}(p)$.

Similarly to the extension from the sample mean to least squares optimization in the case of the mean and OLS, one can extend (1.11) to an estimating equation for regression quantiles, i.e.,

$$\min_{\beta\in\mathbb{R}^{m+1}} \sum_{i=0}^{n} \rho_p(y_i - x^T\beta). \quad (1.12)$$

Note that for $p = 0.5$ (the median), (1.12) coincides with the absolute least squares estimator optimization problem.

Finally, we state asymptotic normality results for sample quantiles and its natural extension to regression quantiles. The first result is due to Mosteller (1946), while the latter is presented in Koenker and Bassett Jr (1978).

Consider the case where Y is continuous. Then, F is continuous with continuous density function f. Let $\{\hat{Q}_n(p_1), \ldots, \hat{Q}_n(p_m)\}$, with $0 < p_1 < \ldots < p_m < 1$ be a sequence of unique sample quantiles estimated from samples of size n each. If f is positive at $Q(p_i)$ for $i = 1, \ldots, m$, then as $n \to \infty$

$$\sqrt{n}(\hat{Q}_n(p_1) - Q(p_1), \ldots, \hat{Q}_n(p_m) - Q(p_m)) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \Sigma),$$

where $\xrightarrow{\mathscr{D}}$ denotes convergence in distribution and Σ has typical element

$$\sigma_{ij} = \frac{p_i(1-p_j)}{f(Q(p_i))f(Q(p_j))}, \text{ for } i \leq j. \quad (1.13)$$

Let now $\{\hat{\beta}_n(p_1), \ldots, \hat{\beta}_n(p_m)\}$, with $0 < p_1 < \ldots < p_m < 1$ be a sequence of regression quantiles estimated from model (1.9) and i.i.d. samples of size n each. If F is continuous with continuous and positive f at $Q(p_i)$ for $i = 1, \ldots, m$, and as $n \to \infty$, $n^{-1}X^TX$ is a positive definite matrix of covariate variables, then

$$\sqrt{n}(\tilde{\beta}_n(p_1) - \tilde{Q}(p_1), \ldots, \tilde{\beta}_n(p_m) - \tilde{Q}(p_m)) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \Sigma \otimes (X^TX)^{-1}), \quad (1.14)$$

where $\tilde{\beta}_n(p_i) = \hat{\beta}_n(p_i) - \beta(p_i)$, $\tilde{Q}(p_i)^T = (Q(p_i), 0, \ldots, 0)$ for $i = 1, \ldots, m$, Σ has typical element as in (1.13) and $\otimes$ denotes the Kronecker product.

It is worth mentioning the inherent difficulty of estimating the asymptotic variance in practice, due to the need of evaluating the unknown probability density function. Usual estimating

strategies are either bootstrap or sandwich estimators.

1.2.2 Time-to-event variables and survival models

Time-to-event or survival analysis deals with real-valued non-negative random variables that represent time until a certain event occurs. Models for survival are widely used in biostatistics, where interest lies in time until the occurrence of a medical event, for example, time to recovery after an intervention, time to relapse after surgery or, in many instances, time to death.

Time-to-event outcomes deserve a separate statistical treatment because observed data for estimating such outcomes often suffers from a particular type of missingness, called censoring. Roughly speaking, censoring prevents the observation of the time at which the event of interest occurs but still preserves the partial information of that time being larger than the time at which the censoring event is observed. This is typically the case in clinical trials or observational studies when patients are followed-up with the goal of observing an event of interest but some other event occurs before that observation is possible. Simple cases being a patient dropping out or the finalization of the study occurring before the event is observed.

Paper III and Paper IV of this thesis deal with problems commmonly found in survival analysis. This section provides a brief overview of survival models used in the above mentioned publications and their properties. For a comprehensive treatment on the subject, see, for example, Kalbfleisch and Prentice (2011). Further interesting reading are the works of Odd Aalen and his take on survival from a counting process point of view, see for example Aalen (1978) and Aalen and Johansen (1978), and for a detailed description see Andersen et al. (2012).

Let T be a non-negative continuous real-valued random variable that denotes time, and has cumulative distribution function $F_T(\cdot)$ and survival function $S_T(\cdot) = 1 - F_T(\cdot)$. Let now C be a censoring variable such that instead of T, we observe $Y = \min(T, C)$, and consider $x^T = (x_1, \ldots, x_m)$ a vector of covariates of interest.

Given a set of observed i.i.d. variables $Y_1, \ldots, Y_n$, and their corresponding fixed set of covariates, the interest lies on estimating the conditional survival function of T given x, that is, $\hat{S}_T(t|x)$. In what follows it is assumed that T and C are independent given x, that is, we consider only scenarios of no competing risks.

Accelerated failure time model

The accelerated failure time model assumes that the effect of the set of covariates x is multiplicative on T or, equivalently, additive on $Z = \log T$, that is, it assumes

$$\log T = \mu + x^T \beta + \sigma W, \tag{1.15}$$

where $\beta^T = (\beta_1, \ldots, \beta_m)$ is a vector of regression coefficients, W follows an unspecified error probability distribution, and μ and σ are location and scale parameters, respectively.

The name accelerated time model refers to the fact that under (1.15) the conditional survival

function of T is actually

$$S(t|x) = S_0(t\exp(-x^T\beta)),$$

where $\exp(-x^T\beta))$ is called the deceleration factor. Therefore, under the accelerated time failure model the conditional survival function accelerates (or decelerates) survival time w.r.t. the baseline survival $S_0(t)$ (or reference group), by a factor that depends on the covariates.

There are several approaches to the estimation of the parameters appearing in (1.15), all of them providing large sample properties of the proposed estimators, see, for instance, Robins and Tsiatis (1992), Jin et al. (2003), Zeng and Lin (2007). For an approach based on linear rank statistics and their censored counterparts, see Kalbfleisch and Prentice (2011).

Cox proportional hazards model

The Cox proportional hazards model assumes a multiplicative effect of the vector of covariates x on a baseline hazard $\lambda_0(t)$, which is assumed common for the entire population. Recall that the hazard function of T, $\lambda(t)$, is related with the survival function through

$$S(t) = \exp\left\{-\int_0^t \lambda(s|x)ds\right\} \text{ with } \int_0^t \lambda(s|x)ds = \Lambda(t),$$

and where $\Lambda(t)$ is the cumulative hazard function of T. Thus, the Cox model reads

$$\lambda(t|x) = \lambda_0(t)\exp\left(x^T\beta\right),$$

or, in its cumulative form

$$\Lambda(t|x) = \Lambda_0(t)\exp\left(x^T\beta\right),$$

where $x^T = (x_1,\ldots,x_m)$ is a vector of fixed covariates, $\beta^T = (\beta_1,\ldots,\beta_m)$ is a vector of regression coefficients and $\exp(\beta_i)$ is the hazard ratio between subgroups in the population that differ (by one unit) in covariate x_i.

One of the most praised properties of the Cox model lies on the fact that in order to estimate the β regression coefficients, the shape of the baseline hazard $\lambda_0(\cdot)$ can be left unspecified. Indeed, given a set of time censored observations $Y_1,\ldots,Y_n$, estimation of the regression coefficients β can be done maximizing the following partial likelihood

$$L(\beta) = \prod_{\substack{i=1\\ \{i:C_i=1\}}}^{n} L_i(\beta) = \prod_{\substack{i=1\\ \{i:C_i=1\}}}^{n} \frac{\exp\left(x_{(i)}^T\beta\right)}{\sum\limits_{\substack{j=1\\ \{j:Y_j\geq Y_i\}}}^{n} \exp\left(x_{(j)}^T\beta\right)}, \tag{1.16}$$

where $C_i = 1$ refers to those individuals that have not been censored, the index j refers to those individuals still at risk after observing Y_i, and the common baseline hazard term $\lambda_0(t)$ has been simplified. In Cox (1975) it is shown that large sample properties (e.g., asymptotic normality) of maximum likelihood estimators also apply when the partial likelihood is used

instead, as, for example, in (1.16).

The large sample properties of existing estimators for both the accelerated time and the Cox proportional hazards models provide an appropriate framework in which to apply the summary measure presented in Paper II to survival analysis problems in combination with these models.

Chapter 2

Aims of the thesis

Appropriate analytical tools are at the center of advances in biological and medical research. Availability of large and complex datasets along with intricate research questions, require more sophisticated and tailored analysis techniques, which call for the development of new mathematical tools. This doctoral thesis aims to provide some of these tools. In particular, the specific goals are

- Develop an exact rejection algorithm suitable for sampling from a family of multidimensional diffusions.
- Provide a tool for sampling from coupled Wright-Fisher diffusions, which have relevant applications in population genetics.
- Develop a summary measure in a regression framework, the conditional compound expectation, study its properties in a general estimation setting and advantages in terms of interpretation.
- Contribute to the advance of survival analysis methods by 1) presenting the use and advantages of the conditional compound expectation in combination with widely used survival models and, 2) proposing an alternative set of inverse probability weights in the estimation of causal effects of a time-varying treatment to improve over estimates obtained with currently available techniques.

Chapter 3

Contributions

The present chapter provides an overview of the contributions included in the constituent papers of this thesis.

3.1 Paper I

Paper I presents the first exact rejection algorithm for a family of multidimensional diffusions with non-unit diffusion coefficient. In particular, this paper is devoted to the development of an exact rejection algorithm for coupled Wright-Fisher diffusions.

The paper starts providing background on the mechanics of exact rejection algorithms for one-dimensional diffusions, and an overview on coupled Wright-Fisher diffusions. A detailed characterization of the coupling term is provided in Proposition 2.1.

Then, it follows a detailed exposition on the sampling strategy for the candidate processes, that is, for neutral multidimensional Wright-Fisher diffusions. The strategy is based on a modification of the alternating series method for sampling from discrete distributions. Although this sampling scheme has been shown before, see Jenkins and Spanò (2017), we provide an improved algorithm (Algorithm 3) that includes suggested changes that improve practical performance.

The main result of the paper is presented in Theorem 4.1, where the suitability of the multidimensional Wright-Fisher diffusions as candidate processes is proven by means of a Girsanov transformation of measures. Afterwards, the proposed exact rejection algorithm is presented (Algorithm 4) and results on the complexity of the proposed algorithm are also provided (Proposition 4.1).

Analogously to the algorithms presented in Chapter 1 of this thesis, our exact rejection algorithm's output is a skeleton of the target path, provided only at a random number of time points. To complete the simulation of paths of coupled Wright-Fisher diffusions at any desired time points, a sampling scheme for multidimensional neutral Wright-Fisher bridges is also presented (Lemma 6.1, Proposition 6.2, and Algorithm 5).

Finally, simulation results for two illustrative examples, namely, with two and four loci and two allele types each, are also shown, and are consistent with the complexity results provided in

Proposition 4.1. A qualitative comparison of the histogram generated for the two dimensional case in comparison with the diffusion's stationary density is also presented.

3.2 Paper II

This paper presents a novel summary measure in a regression framework, the conditional compound expectation, that offers a compromise between ordinary least squares and quantile regression.

After giving an overview of existing summary measures, the concept of compound expectation of a random variable Y is presented. In brief, the compound expectation is the average of Y over different subsets of its domain. These subsets are defined by means of a grid of proportions so that they can be directly identified with specific quantiles. This formulation allows to provide average values over different groups that are determined by their order within the distribution of Y. For example, if Y measures the score of an intelligence quotient test, and we define a group delimited by the proportions $\{0.8, 1\}$, the compound expectation of Y over that interval provides the average score for the 20% smartest individuals. This formulation proves of great use in several applications, e.g., resource-allocation and intervention-evaluation problems, that are illustrated by means of two real-data examples.

The compound expectation is then extended to a regression framework, in which one can assess the effect of a set of covariates over the averages mentioned above. Exploiting the relation between the expectation and the quantile function of Y, the paper proceeds on proposing a general estimation strategy based on an unspecified estimator for the underlying conditional quantile function. Results on unbiasedness (Proposition 1), consistency (Proposition 2) and asymptotic normality (Proposition 3) of the conditional compound expectation (CCE) estimator are provided w.r.t. the same properties of the conditional quantile function estimator. Finally, a bound on the variance of the CCE estimator in terms of the variance of the conditional quantile function estimator is also provided (Proposition 4).

A discussion on the grid of proportions is also included. An interesting observation derived from Proposition 4 is that it provides a certain control over the variance of the CCE estimator, which is exemplified by means of a simulation study. The conclusion is that regions of the domain where the variance of the conditional quantile function estimator is lower might allow a finer grid, while still maintaining good variances on the CCE estimator. On the contrary, regions of the domain where the variance of the conditional quantile function estimator is higher might suggest to set a coarser grid in order to maintain the same level of variance on the derived CCE estimator.

3.3 Paper III

This paper provides new estimators for life expectancy based on the conditional compound expectation (CCE) presented in Paper II.

The problem of estimating mean survival time (or life expectancy when the survival outcome of interest is death) in the presence of random censoring is presented, and the CCE is suggested as an advantageous alternative summary measure in these scenarios. Then, the CCE is compared to the restricted mean, first, in terms of interpretation of the provided estimates and then, to highlight its advantages in terms of groups' comparison.

Finite sample properties of CCE estimators are shown in a simulation study, where the underlying model for the data is assumed to be first an accelerated failure time model, and then a Cox proportional hazards model. Simulation results show that CCE estimators for subsets of the data where less censoring events have been observed are more precise than those with higher presence of censoring. Finally, an illustrative example with real data is also provided.

3.4 Paper IV

In this paper we present the use of optimal probability weights, see Santacatterina and Bottai (2018), in the context of the estimation of the causal effect of a time-varying treatment on a survival outcome.

The paper first gives an overview on drawbacks inherent in current state-of-the-art inverse probability weights' estimators, more specifically, in situations where the positivity assumption is violated, and truncation techniques are often used. This is followed by a brief account on marginal structural Cox models, which are later used in the empirical examples where we estimate the causal effect of a time-varying treatment on death. The proposed set of optimal probability weights are obtained as the solution of a quadratic optimization problem, stated in Section 3.

By means of a comprehensive simulation study, the paper shows that the set of optimal probability weights outperforms those obtained by the usual truncation technique, in the sense that they provide less biased and more precise estimates for the causal treatment effect of interest. Scenarios in which the positivity assumption is strongly and weakly violated are also considered, yielding similar results. Finally, an illustrative example with real data is also provided.

Chapter 4

Future research

Based on the contributions presented in this thesis, future research includes:

- *Exact rejection algorithms in combination with rare event simulation*. Possible problems of interest in rare event simulation are those defined as the underlying diffusion process hitting or remaining contained in a specified bounded set. Examples of these can be found in many applications, e.g., population genetics, see Iorio and Griffiths (2004a), Iorio and Griffiths (2004b), finance, see Casella and Roberts (2008), physics, see Del Moral and Garnier (2005), or engineering, see Blom et al. (2007). In this context, time-discretized approximations induce further errors because samples of the path in between considered times may lie outside the desired boundaries. Such errors are difficult to quantify and are often deemed negligible. A possible alternative is to propose exact rejection algorithms that sample from the true conditional distribution and that can be later embedded in suitable rare event simulation algorithms.

- *Exact simulation of multidimensional Wright-Fisher bridges when the mutation parameters are* 0. The simulation scheme for multidimensional neutral Wright-Fisher bridges provided in Paper I assumes positive mutation parameters. An interesting extension would be to explore sampling schemes in cases where some of these parameters were actually 0.

- *Automatized search of the grid of proportions*. Paper II suggests estimation of the conditional compound expectation when the grid of proportions is given. One possible extension includes automatized searches that satisfy certain criteria. For example, exploiting the result provided in Proposition 4 and given an estimator for the conditional quantile function, one can define the grid such that the variance of the compound expectation estimator remains constant across components, or such that the variance does not exceed a certain threshold.

- *Conditional compound expectation with other survival models*. An interesting feature of the conditional compound expectation is that, similary to quantile regression, it allows for estimation of different covariate effects along the conditional distribution of the

data. Estimation of the conditional compound expectation in survival scenarios under the accelerated failure time and Cox proportional hazards models in Paper III does not exploit this feature. Cosidering models such as the Cox proportional hazards with time-dependent covariates or flexible parametric approaches would take advantage of this property and provide more informative estimates.

- *Robustness of the Cox proportional hazards model to stochastic perturbations*. Although the Cox proportional hazards model is widely used for the analysis of real data, ensuring its robustness to practical violations of the proportional hazards' assumption remains still a challenge. This question can be addressed by simulating data from a perturbed Cox model, where this perturbation is a multiplicative effect on the baseline hazard, sampled from a suitable bounded stochastic noise. Once the simulated data is available, one can estimate the model's coefficients assuming an unperturbed Cox model and compare the results. Simulation of these bounded stochastic noises poses an interesting problem in itself, because it can lead to new exact rejection algorithms for bounded stochastic processes.

References

Aalen, O. 1978. Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6(4): 701–726.

Aalen, O., and S. Johansen. 1978. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 5(3): 141–150.

Abdous, B., and B. Remillard. 1995. Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics* 47(2): 371–384.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding. 2012. *Statistical models based on counting processes*. Springer Science & Business Media.

Aurell, E., M. Ekeberg, and T. Koski. 2019. Networks of loci by a multilocus Wright-Fisher model with selection and mutation, and a conjecture by Motoo Kimura. *arXiv preprint arXiv:1906.00716* .

Beskos, A., O. Papaspiliopoulos, and G. O. Roberts. 2006. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli* 12(6): 1077–1098.

———. 2008. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability* 10(1): 85–104.

Beskos, A., and G. O. Roberts. 2005. Exact simulation of diffusions. *Ann. Appl. Probab.* 15(4): 2422–2444.

Blanchet, J., and F. Zhang. 2017. Exact simulation for multivariate Itô diffusions. *arXiv preprint arXiv:1706.05124* .

Blom, H. A. P., G. J. Bakker, and J. Krystul. 2007. Probabilistic reachability analysis for large scale stochastic hybrid systems. In *2007 46th IEEE Conference on Decision and Control*, 3182–3189.

Bollback, J. P., T. L. York, and R. Nielsen. 2008. Estimation of 2Nes from temporal allele frequency data. *Genetics* 179(1): 497–502.

Breckling, J., and R. Chambers. 1988. M-quantiles. *Biometrika* 75(4): 761–771.

Burgette, L. F., J. P. Reiter, and M. L. Miranda. 2011. Exploratory quantile regression with many covariates: an application to adverse birth outcomes. *Epidemiology* 22(6): 859–866.

Casella, B., and G. O. Roberts. 2008. Exact Monte Carlo simulation of killed diffusions. *Advances in Applied Probability* 40(1): 273–291.

Charlesworth, B., and D. Charlesworth. 2017. Population genetics from 1966 to 2016. *Heredity* 118(1): 2–9.

Chen, N., and Z. Huang. 2013. Localization and exact simulation of Brownian motion-driven stochastic differential equations. *Mathematics of Operations Research* 38(3): 591–616.

Cook, R. D. 1977. Detection of influential observation in linear regression. *Technometrics* 19(1): 15–18.

Cox, D. R. 1975. Partial likelihood. *Biometrika* 62(2): 269–276.

Dangerfield, C. E., D. Kay, S. MacNamara, and K. Burrage. 2012. A boundary preserving numerical algorithm for the Wright-Fisher model with mutation. *BIT Numerical Mathematics* 52(2): 283–304.

Del Moral, P., and J. Garnier. 2005. Genealogical particle analysis of rare events. *Ann. Appl. Probab.* 15(4): 2496–2534.

Ehret, G. B., et al. 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478: 103–109.

Fenske, N., T. Kneib, and T. Hothorn. 2011. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association* 106(494): 494–510.

García-Pareja, C., and M. Bottai. 2018. On mean decomposition for summarizing conditional distributions. *Stat* 7: e208.

García-Pareja, C., H. Hult, and T. Koski. 2019a. Exact simulation of coupled Wright-Fisher diffusions.

García-Pareja, C., M. Santacatterina, A. M. Ekström, and M. Bottai. 2019b. Conditional life expectancy estimation by ordered fractions of population with censored data.

Griffiths, R. C. 1984. Asymptotic line-of-descent distributions. *Journal of Mathematical Biology* 21(1): 67–75.

Huber, P. J., et al. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35(1): 73–101.

Iorio, M. D., and R. C. Griffiths. 2004a. Importance sampling on coalescent histories. I. *Advances in Applied Probability* 36(2): 417–433.

———. 2004b. Importance sampling on coalescent histories. II: Subdivided population models. *Advances in Applied Probability* 36(2): 434–454.

Jenkins, P. A. 2013. Exact simulation of the sample paths of a diffusion with a finite entrance boundary. *arXiv preprint arXiv:1311.5777* .

Jenkins, P. A., and D. Spanò. 2017. Exact simulation of the Wright–Fisher diffusion. *Ann. Appl. Probab.* 27(3): 1478–1509.

Jewett, E. M., and N. A. Rosenberg. 2014. Theory and applications of a deterministic approximation to the coalescent model. *Theoretical Population Biology* 93: 14–29.

Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying. 2003. Rank-based inference for the accelerated failure time model. *Biometrika* 90(2): 341–353.

Jones, M. C. 1994. Expectiles and M-quantiles are quantiles. *Statistics & Probability Letters* 20(2): 149–153.

Kalbfleisch, J. D., and R. L. Prentice. 2011. *The statistical analysis of failure time data*. John Wiley & Sons.

Karatzas, I., and S. E. Shreve. 1998. *Brownian motion and stochastic calculus*. 2nd ed. Springer New York.

Kloeden, P. E., and E. Platen. 1992. *Numerical solution of stochastic differential equations*. Springer.

Koenker, R. 2005. *Quantile regression*. Cambridge University Press.

Koenker, R., and G. Bassett Jr. 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society* 46(1): 33–50.

Koenker, R., and K. F. Hallock. 2001. Quantile regression. *Journal of Economic Perspectives* 15(4): 143–156.

Koenker, R., and J. A. F. Machado. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* 94(448): 1296–1310.

Koenker, R., and Z. Xiao. 2002. Inference on the quantile regression process. *Econometrica* 70(4): 1583–1612.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin. 2012. Estimating allele age and selection coefficient from time-serial data. *Genetics* 192(2): 599–607.

Maurano, M. T., et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099): 1190–1195.

Mosteller, F. 1946. On some useful "inefficient" statistics. *The Annals of Mathematical Statistics* 17(4): 377–408.

Neuenkirch, A., and L. Szpruch. 2014. First order strong approximations of scalar SDEs defined in a domain. *Numerische Mathematik* 128(1): 103–136.

Newey, W. K., and J. L. Powell. 1987. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society* 55(4): 819–847.

Robins, J., and A. A. Tsiatis. 1992. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* 79(2): 311–319.

Santacatterina, M., and M. Bottai. 2018. Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association* 113(523): 983–991.

Santacatterina, M., C. García-Pareja, R. Bellocco, A. Sonnerbörg, A. M. Ekström, and M. Bottai. 2019. Optimal probability weights for estimating causal effects of time-varying treatments with marginal structural cox models. *Statistics in Medicine* 38(10): 1891–1902.

Schraiber, J. G., S. N. Evans, and M. Slatkin. 2016. Bayesian inference of natural selection from allele frequency time series. *Genetics* 203(1): 493–511.

Schraiber, J. G., R. C. Griffiths, and S. N. Evans. 2013. Analysis and rejection sampling of Wright-Fisher diffusion bridges. *Theoretical Population Biology* 89: 64–74.

Seber, G. A., and A. J. Lee. 2012. *Linear regression analysis*. John Wiley & Sons.

Skwark, M. J., et al. 2017. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS genetics* 13(2): e1006508.

Song, Y. S., and M. Steinrücken. 2011. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* 190(3): 1117–1129.

Steinrücken, M., Y. R. Wang, and Y. S. Song. 2013. An explicit transition density expansion for a multi-allelic wright–fisher diffusion with general diploid selection. *Theoretical Population Biology* 83: 1–14.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences* 102(22): 7882–7887.

Zeng, D., and D. Y. Lin. 2007. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association* 102(480): 1387–1396.

Acknowledgements

"It is invaluable to have a friend who shares your interests and helps you stay motivated."

—Maryam Mirzakhani

This incredible journey has brought me the opportunity to cross many people's paths. This is my attempt to thank them for what we have shared and for their contribution to the completion of this thesis.

Thanks to my supervisor **Matteo Bottai** for welcoming me into the Unit of Biostatistics in the first place and then offering the possibility of starting a PhD. Moving to Stockholm certainly changed my life, both professionally and personally. Thank you for introducing me to the world of quantile regression and research in general.

I wish to thank my co-supervisor **Henrik Hult** for the opportunity to work at KTH and expand my horizons. Thank you for the stimulating supervision meetings, and for all that I have learned while working with you. Thanks for sharing your mathematical expertise, for the right pointers, and for bringing intuitive views to our problems without overlooking the details. Your honesty and clarity have always been refreshing and most welcome. Thank you for believing in me, for always being supportive and encouraging in my initiatives, and for striving to find the time (even when there really wasn't) to work together.

Thanks to my co-supervisor **Anna-Mia Ekström** for always welcoming me to the group meetings at the Public Health Department, and for always offering your help, in research and otherwise. Thanks also to **Anders Sönnerborg**, my co-supervisor for the first half of the PhD, for all your support and suggestions.

My deep gratitude to **Anna Puig Puig**, my PhD mentor, for believing in me in the early days and supervising my first ever research project. I have not only had the pleasure of working with you, but I also feel honoured to call you my friend. I wish to thank also **Josep Ginebra**, for all your recommendation letters and for your invariant support even after all these years.

I wish to thank **Lena Palmberg** for your invaluable help throughout my time as a PhD student. Thank you for always taking the time for meeting with me, for your quick answers to my many e-mails and for your support and understanding.

A deep thanks also to **Erin Gabriel**, for your contagious passion for research, your many useful comments and your limitless will to help.

Thank you to **Alberto D'Onofrio** and all the people at the International Prevention Research Institute in Lyon, for your warmth during my stay. I'm looking forward to continuing our collaboration. Grazie anche a **Antonella**, my Sicilian colleague, for the nice chats and morning and afternoon bus rides while in Lyon.

A special mention goes to **Joakim Jaldén**, whose commitment, integrity and advice have always been a reference for me.

My deep appreciation to **Pol del Aguila Pla**, colleague and TikZ master, for being always up for the challenge and for your contagious resilience and enthusiasm. Thank you for the countless hours of problem solving, and for sharing many sleepless working nights. The outcome of this thesis would have surely not been the same without our fruitful discussions and your constant support. Thank you also for your feedback and comments on paper drafts and chapters of this thesis.

Thank you my dear **Michele** for sharing what, in other times, we thought an unachievable dream and for the chance of outgrowing it together. Thank you for the almost uncountable shared lunch times, for always being supportive, in the good and the bad times, for our endless conversations about research and life, for the laughs and the tears. This journey would have certainly not been the same without knowing you'd be at the office every morning when I arrived. Of course, not forgetting to mention **Alice** and **PK**. Thanks for our awesome Thanksgiving dinners and pizza marathons altogether.

I wish to thank also my colleagues at the Unit of Biostatistics and IMM, **Michele**, **Paolo**, **Ulf**, **Jonas**, **Daniel**, **Mike**, **Erin**, **Xin**, **Qing**, **Andrea**, **Yang** and **Cecilia** for all the good times. Thanks to **Silvia Columbu** for all your help during my first months in Stockholm. Of course, thanks to **Paolo Frumento**, a.k.a. "the Post-doc", for the joyful challenge that has been sharing all these years with you.

Thanks also to all my colleagues at the Department of Mathematics at KTH, **Emil**, **Giampaolo**, **Martina**, **Lena**, **Michele**, **Federico**, **Gerard**, **Alexander**, **Daniel**, and all the others, for your warmth and for always making me feel like one more of the group. Thanks to **Martina** and **Adam** (a.k.a. "the locals"), **Calle**, **Marcus**, and **Hannah** for our discussions on the blackboard, and also to **Adam**, **Peter**, **Hans** and **Aleksa**, for the good times on the 4th floor. A special thanks to my KTH office mate **Boris**, for the nice chats, shared laughs and pictures, and for teaching me that Christmas doesn't have to be necessarily on Dec 25th.

I wish to thank also the great **Sandra Schöning** at Danscompagnyet and our incredible team of tap dance partners: **Ingrid**, **Sofia**, **Eva**, **Lotta**, **Katarina**, **Andreas**, **Tanya**, and the rest of our group, for our weekly unmissable lessons that have represented a most welcome oasis

in the midst of the (not always so calm) PhD life. I will certainly miss you and hope to be back soon!

My deep gratitude an appreciation to my Swedish family **Adeline**, **Paola**, **Nasren**, **Kris**, **Paolo** and **Ana Luisa** (or as someone once said "the Portuguese version of me"), for all the birthdays, National dagar, Midsommar and cosy and memorable celebrations and dinners altogether. These shared years by your side have meant the world to me, and I will always keep them as one of my dearest memories. Thank you **Adeline** for always being there for us, and **Paola** for your immense sensitivity that always manages to move me to tears. Thank you **Paolo** for the time spent in preparing fun statistical puzzles. Gràcies també al **Jordi** i la **Maria** per ser un trosset de casa enmig d'Escandinàvia.

Thank you **Elba**, **Laia** and **Marta** because everything started in that basement bar, where we would spend hours eating spaguetti and talking about our future PhDs. Certainly, not everything turned out as we envisioned it back then (I'd even dare to say fortunately), but yet, here we are! I love you dearly and I'm really proud of you all. Thank you for all the priceless moments, and for being my friends.

Moltes gràcies a la **Georgina**, l'**Agustí**, l'**Aina** i la **Nala** per fer-me sentir des del principi una més de la família. Gràcies **Georgina** per venir sempre amb la maleta plena i comprendre les nostres abscències, i a tu **Agustí**, pels teus ànims i confiança.

A **Mamá y Papá**, aquí la tenéis! Mi tesis doctoral. Sería injusto no decir que es en parte vuestra. Gracias por enseñarme que lo importante es encontrar algo que te apasione, y por siempre haberme hecho sentir que era capaz de hacer cuanto me propusiera. El camino ha sido tortuoso, sí, pero como dijo Loli: "La base es buena". Gracias, de corazón, por todo lo que habéis hecho y hacéis cada día por mí. A **Iris**, gracias por venir a Estocolmo aquella Semana Santa donde nos encontramos la una a la otra, y por dejarme ser tu hermana mayor. Gracias por todo lo que compartimos, que es mucho, y por la certeza de saber que siempre estás ahí. Gràcies **Albert** per donar sempre un toc de color que, certament, de vegades ajuda. Ja saps que el Mälaren t'està esperant. Gracias también a **tía Rosa**, por estar siempre pendiente de nosotros, por tus mails entrañables que invariablemente me arrancan una lágrima (o dos), y también, por tu generosa contribución en mi primer tiempo en Estocolmo.

And last, but certainly, not least, a tu **Pol**, amor de la meva vida, marit, company, i sens dubte, el millor amic que mai he tingut, per ajudar-me a recordar qui sóc i del que sóc capaç, per malacostumar-me a viure en un món, el nostre, on no existeix el malentès, on qualsevol cosa que ens proposem sembla (i fins ara, ho ha estat) possible. Gràcies per la teva paciència infinita i per la teva ajuda incansable, i sobretot, per donar-me els anys més feliços de la meva vida i per voler compartir amb mi (i els nostres) tots els que vindran.

This work was partially supported by **Karolinska Institutet**'s funding for doctoral students (KID-funding).