On complicated dependency structures

Nourhan Shafik
On complicated dependency structures

Nourhan Shafik

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of science, at a public examination held at Auditorium F239a, Otakaari 3 on 13 October 2023 at 12:00.

Aalto University
School of Science
Department of Mathematics and Systems Analysis
Supervising professor
Professor Paullina Ilmonen, Aalto University

Thesis advisors
Professor Lauri Viitasaari, Uppsala University
Docent Sirpa Heinävaara, Finnish Cancer Registry

Preliminary examiners
Professor Anna Kiriliouk, University of Namur
Docent Dario Gasbarra, University of Helsinki

Opponent
Professor Jukka Lempa, University of Turku
Abstract

Today, we collect enormous amount of data and the data comes in many forms. Modern models are able to capture complex dependencies and interactions. At the same time, parts of the data may be missing or contaminated. Analysts may be required to handle high dimensional or functional data, analyse complicated dependencies within the data, and possibly predict missing data at the same time.

In this dissertation, we consider dependency structures from different theoretical and applied perspectives. Firstly, we develop a new data driven method, based on Gaussian processes, to optimally predict missing data in the context of functional observations. Secondly, we analyse the rate of convergence of discretization of certain stochastic integrals involving Gaussian processes that possess non-trivial dependency structure. Finally, we analyse dependency structures in the context of applications by modelling cancer mortality and cost effectiveness of breast cancer screening under different screening policies.

Keywords Missing data, Dependency structure, Gaussian processes, Breast cancer screening
### Tekijä
Nourhan Shafik

### Väitöskirjan nimi
Monimutkaisista riippuvuusrakenteista

### Julkaisija
Perustieteiden korkeakoulu

### Yksikkö
Matematiikan ja systeemianalyysin laitos

### Sarja
Aalto University publication series DOCTORAL THESES 150/2023

### Tutkimusala
Matematiikka ja Tilastotiede

<table>
<thead>
<tr>
<th>Käsikirjoituksen pvm</th>
<th>Väitöspäivä</th>
<th>Kieli</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.08.2023</td>
<td>13.10.2023</td>
<td>Englanti</td>
</tr>
</tbody>
</table>

### Tiivistelmä
Tämä päivänä keräämme runsaasti monessa erilaisessa muodossa olevia aineistoja. Nykykaiset
menetelmät kykenevät mallintamaan monimutkaisia riippuvuusuhteita. Samaan aikaan osa
aineistosta saattaa puuttua tai olla viiottunutta. Tämän vuoksi analytyikko voi joutua
samanaikeisesti käsittelemään suuri- tai ääretönulotteista aineistoa, analysoimaan monimutkaisia
aineiston sisäisiä riippuvuuksia, ja mahdollisesti ennustamaan puuttuvia havaintoja.

Tässä työssä tarkastelemme riippuvuusrakenteita sekä teoreettisesti että sovellusten näkökulmasta.
Työssä kehitetään uusi Gausssiiin prosesseihin pohjaava menetelmä, jolla voidaan laskea
optimaalisia ennusteita funktionaalisten havaintojen puuttuville osille. Lisäksi työssä
analysoidaan tiettyjen stokastisten integraalien diskretointien suppenemisnopeutta.
Nämä integraalit sisältävät Gausssia prosesseja, joiden riippuvuusrakenne
poikkeaa tavanomaisesta. Työn soveltavassa osassa analysoidaan rintasyöpäkuolleisuutta ja
rintasyöpäseulontojen kustannustehokkuutta eri seulontastrategioilla.
Let's start with the best sentence that one should remember whenever he/she feels down in academia: if we don't feel stupid, it means we're not really trying. This sentence is stated by Martin A. Schwartz in his article about the importance of stupidity in scientific research which means that in the academic journey, you may feel stupid, but this is a good feeling and means that you are on the right track.

Now is the time to thank the people who always believe in me and support me. It is also the time when any words cannot express your feelings.

I deeply thank my supervisor Professor Pauliina Ilmonen for her constant presence and kind encouragement with her kind and bright words that always increase my confidence and strength. You are an amazing supervisor which gives an extraordinary push and motivation to others. You are also an excellent source of inspiration for me. I really hope you never lose your feelings and your sense of humor that make every meeting interesting and enjoyable.

Thank you, my advisor Professor Lauri Viitasaari, I am extremely astonished of your knowledge, I usually remember Pauliina saying Lauri is the best in calculating integrals. I am truly thankful for your guidance, support, patience and understanding. You are the best advisor ever. Thanks, Docent Sirpa Heinävaara for your support and guidance in cancer research. In addition, I wish to thank my great collaborators: Professor Germain Van Bever, Professor Tommi Sottinen, Professor Ehsan Azmoodeh and Docent Tytti Sarkeala.

I wish to thank the preliminary examiners Professor Anna Kiriliouk and Docent Dario Gasbarra for their careful examination of this thesis. Thank you, Professor Jukka Lempa, for agreeing to act as my opponent.

Thanks to my mother Mervat and my father Mohsen for their continuous support and encouragement. They always believe in me, and I have reached this point because of them. I believe that even when I thank them, it is less than what they did and always do for me. Dad, I really wished that you are here with me.

Thanks to my husband Ahmed, my partner in this journey. Thanks a lot
for being here, always by my side supporting, encouraging and lifting me up whenever I feel down. I really appreciate that.

Thanks to my little daughters: Farida and Leila, I hope I made you proud of me. Thank you my sister Salma for your sweet encouragement and believing in me and thanks to the best nephew ever Yehia. Thanks to my brother Mohamed for your help and support.

Thanks to my close friends here in Finland (my sisters whom my mother did not give birth to). You all are amazing friends. Thanks to my childhood friends for always being on my side. In addition, thank you my mother-in-law Nehad for your support.

I also wish to thank my grandmother Sadia and my uncle Magdy for their support in my journey.

Thanks to my team, they all have left a mark. They are always ready to help. Thanks to Sami, Marko, Paavo and Niko for your help and support in the beginning of my studies. Thanks to Jaakko, Aleksi and Natalia for having such good time, company, support and great atmosphere in the office.

I also want to thank all the staff at Aalto University School of Science, Department of Mathematics and systems.

In the end, yes, I have finished my thesis.

Helsinki, September 14, 2023,

Nourhan Shafik
Contents

Preface 1

Contents 3

List of Publications 5

Author's contributions 7

1. Introduction 9

2. Preliminaries 11
   2.1 Gaussian Processes 11
      2.1.1 Properties 12
      2.1.2 Fractional Brownian motion 13
      2.1.3 Prediction 13
   2.2 Stochastic Integration 14
      2.2.1 Riemann-Stieltjes integration 14
      2.2.2 Lebesgue-Stieltjes integration 15
   2.3 Breast cancer screening 17
      2.3.1 Modelling approaches 19

3. Summaries of the articles 21
   3.1 On optimal prediction of missing functional data 21
   3.2 On sharp rate of convergence for discretisation of integrals driven by fractional Brownian motions and related processes with discontinuous integrands 22
   3.3 On Cost-effectiveness of Breast cancer screening in Finland 22

4. Discussion and future prospects 25

References 27

Publications 31
This doctoral dissertation consists of an introduction, preliminaries and of the following publications.


Author’s contributions

Publication I: “On optimal prediction of missing functional data with memory”

The idea for the article came from L.V. and P.I. The preliminary simulations were conducted by N. S. First version of theoretical results was derived by N. S. under the supervision of L.V. The final simulations were conducted G.V.B. The theoretical results were polished by L.V. All the authors participated in writing and commenting the article.

Publication II: “On sharp rate of convergence for discretisation of integrals driven by fractional Brownian motions and related processes with discontinuous integrands”

The idea for the article came from E.A. and L.V. First version of theoretical results for fractional Brownian motion was derived by N. S. under the supervision of L.V. The theoretical results were extended and polished by L.V. All the authors participated in writing, commenting, and revising the article.

Publication III: “Flexible transition model for assessing cost effectiveness of breast cancer screening in Finland”

The idea for the article came from S.H. and T.S. Data analysis was conducted by N.S. under the supervision of S.H., P.I. and L.V. The model was built by S.H., P.I. and L.V. All the authors participated in writing, commenting, and revising the article.
1. Introduction

Nowadays, dealing with data has become essential in many fields. Due to massive storage capacity, lots of data can be stored and used. The types of data that we are dealing with are often complicated and high or even infinite dimensional, involving different types of dependencies. Dependency structures in data provide a rich source for data analysts in modelling modern day phenomena. Knowledge based decision making requires high quality data. However, analysts often have to face many different problems related to the quality of data. One of the problems that analysts have to deal with is missing data. Knowledge of underlying dependency structures may help in dealing with missing data as using just the average could result in losing information.

Data analysts today face the problem of dealing with various types of complicated data and dependency structures. This thesis tackles this problem. The goal is to provide new methods, tools, for analyzing modern data. In particular, in the theoretical part of the thesis, the aim is to derive tools for predicting missing parts of observations and to consider convergence rates of certain stochastic integrals. The applied part of the thesis is devoted to analyzing cost effectiveness of cancer screening. The main goal in this thesis is to not only apply existing tools, but in particular, to provide new tools for researchers to utilize in their own analyses. Theoretical part of the thesis is built on nowadays rather well-developed Gaussian analysis that is applied in an innovative way in the thesis. The applied part involves on building a new modelling approach utilizing context knowledge on cancer.

The main part of the thesis is three scientific articles. In the first article, we utilize the rich theory of Gaussian processes in reconstructing the missing parts of functional observations. In the second article, we study sharp rate of convergence for discretisation of integrals with respect to fractional Brownian motions and related processes. In the third article, we consider dependency structures behind breast cancer survival and the effect of screening.

Gaussian processes play an important role in many fields such as ma-
Introduction

...chine learning and finance, and they play a crucial role in this dissertation as well. Gaussian processes are useful for modelling since they can capture general dependency structures but at the same time, provide sufficient structure for analysis. This is especially visible in the first article, where we reconstruct missing observations for functional data by using the rich theory of Gaussian modelling at our disposal.

Understanding stochastic integration is of paramount important both from mathematical point of view as well as from the applications point of view. Moreover, from numerical perspective one needs to understand how stochastic integrals can be discretised. While the problem is rather well understood in the case of the Brownian motion, the most simple Gaussian process, it is much more subtle in the case of processes that have more complicated dependency structures. In particular, this is a topic of active research in the case of fractional Brownian motions and other related Gaussian processes that have non-trivial dependency structures. We complement the known results on the problem in the second article by providing a sharp rate of convergence for certain discretisations of integrals by exploiting Gaussianity and known facts about the underlying dependency.

Finally, understanding dependency structures is of paramount importance in decision making. In the third article, we analyze dependencies in real data related to breast cancer screening. Indeed, it is well known that organized screening reduces breast cancer mortality on certain age groups and that brings up the question whether it would be beneficial to extend screening to new age groups as well. The third article in this thesis tackles this question.

The theoretical results solve important open problems in the field. The applied results provide crucial information for policy makers.

To summarize the key findings of this thesis:

1. We provide a flexible data driven optimal solution for the missing functional observations with and without assuming Gaussianity.

2. We derive a new sharp result on rate of convergence for the discretisation of stochastic integrals involving general Gaussian processes.

3. We develop a new model for assessing the cost-effectiveness of breast cancer screening under different screening strategies.

The rest of the thesis is organized as follows. Section 2 provides a short introduction to Gaussian processes giving preliminaries for understanding the first and the second articles, and information about breast cancer and screening giving preliminary for being able to read the third article. In section 5, we present summaries of the three articles.
2. Preliminaries

In the following chapters, we go through the preliminaries related to Gaussian processes and to breast cancer screening.

2.1 Gaussian Processes

Gaussian processes form an important and widely applied class of stochastic processes. For details on Gaussian processes, the reader is referred to [1], [18], [15], [26] and [27]. Throughout this thesis, stochastic processes $X_t$ are indexed over an interval $T = [0, 1]$.

A stochastic process $\{X_t\}$ is called Gaussian if for any time points $(t_1, ..., t_n)$ the vector $X = (X_{t_1}, X_{t_2}, ..., X_{t_n})$ is an $n$-dimensional Gaussian random vector.

One of the most important Gaussian processes is Brownian Motion. A real-valued stochastic process $B = (B_t)_{t \geq 0}$ is a standard Brownian motion if it satisfies the following conditions [20]:

(i) **Starts at zero**: $B_0 = 0 \in \mathbb{R}$.

(ii) **Independent increments**: For $0 \leq t_1 < t_2 < ... < t_n$, increments $B_{t_{j+1}} - B_{t_j}$ are independent and $B_{t_{j+1}} - B_{t_j} \sim \mathcal{N}(0, t_{j+1} - t_j)$.

(iii) **Almost sure continuity**: $\mathbb{P}$-almost every sample path $t \to B_t(\omega)$ is continuous: $\mathbb{P}[\omega \in \Omega | t \to B_t(\omega) \text{ is continuous}] = 1$.

For a Gaussian process, a mean function is defined as $E\{X_t\}$, and covariance function of $X$ is defined as $C(s, t) = \text{Cov}(X_s, X_t)$ for all $s, t \in T$.

A Gaussian process is uniquely determined by the mean and covariance function. That is, a Gaussian process gives rise to the mean and covariance function and conversely, for each mean function $\mu$ and each positive semi-definite function $C$ there exists a Gaussian process with mean $\mu$ and covariance $C$. In particular, Brownian motion could be defined as a
Gaussian process with zero mean and covariance $C(s, t) = \min(s, t)$.

### 2.1.1 Properties

We start this section by recalling two important properties of Gaussian processes that are essential for the results of this thesis:

- A Gaussian process is Hölder continuous (of any order $< \frac{3}{2}$) if and only if
  
  $$E(|Y_t - Y_s|^2) \leq C|t - s|^\eta$$

  for some constant C, see [3]. This property can be used to define stochastic integrals with respect to a Gaussian process in Article 2.

- Any Gaussian process with integrable variance function admits a Fredholm representation
  
  $$X_t = \int_0^T K(t, s) dW_s$$

  where $K$ is a deterministic kernel, $W$ is a Wiener process, and stochastic integral is a so-called Wiener integral (for details, see e.g. [40, 41]). This property allows to predict missing data in Article 1.

Below, we present some other properties of stochastic processes.

**Definition 2.1.1.** [Stationarity] A process $X = (X_t)_{t \geq 0}$ is called stationary if we have for every $h, t \geq 0$

$$X_{t+h} \overset{\text{law}}{=} X_t.$$

**Definition 2.1.2.** [Stationary increments] A process $X = (X_t)_{t \geq 0}$ has stationary increments if we have for every $h, t \geq 0$

$$X_{t+h} - X_t \overset{\text{law}}{=} X_h - X_0.$$

**Definition 2.1.3.** [Self-similarity] A real valued stochastic process is H-self-similar process if for every $a > 0$ we have

$$(X_{at})_{t \geq 0} \overset{\text{law}}{=} (a^H X_t)_{t \geq 0}.$$

A Brownian motion is Hölder continuous of any order $< \frac{1}{2}$, has a deterministic kernel $K(t, s) = I_{[0,t]}(s)$, has stationary increments, and is $\frac{1}{2}$-self similar.
2.1.2 Fractional Brownian motion

The fractional Brownian motion, a generalisation of the Brownian motion, appeared already in the works of Kolmogorov [21, 22] though the term fractional Brownian motion was coined later in [30], and it is possibly the simplest Gaussian process that can be used to model dependencies, with applications ranging from finance to meteorology.

Fractional Brownian motion $B^H$, where $H \in (0, 1)$, is a real-valued Gaussian process with a zero mean and a covariance

$$C(s, t) = \frac{t^{2H} + s^{2H} - |t - s|^{2H}}{2}.$$ 

Note that for $H = \frac{1}{2}$, one recovers the standard Brownian motion.

The parameter $H$ is called the Hurst index, and it describes the path regularity and self-similarity property of the process. Indeed, a fractional Brownian motion is Hölder continuous of any order $< H$, and is $H$-self similar that is another wanted feature for many applications, for details see [7]. Like Brownian motion, fractional Brownian motion has also stationary increments. On the contrary, the increments are not independent that allows to model dependencies. This property can be considered as the main step in Article 1 for its use for prediction of missing parts of functional observations. Similarly, Hölder continuity can be used to define stochastic integrals with respect to fractional Brownian motion. Such stochastic integrals with respect to fractional type processes are studied in Article 2. Here, fractional type processes refer to Gaussian processes having similar properties than the fractional Brownian motion.

2.1.3 Prediction

Predicting the unknown is one of the most important tasks in statistics, as it provides support e.g. for decision makers. Formally, the task is to give educated guesses based on the available information. In Article 1, we studied this problem in the context of functional data. That is, the data is assumed to be realizations of stochastic processes.

The available information is mathematically encoded into sigma-algebras and filtration. If we have a sequence of random variables $X_1, X_2, \ldots$, then we can define the natural filtration as the collection $F_n$, where $F_n$ is the information stored in $X_1, \ldots, X_n$. Mathematically, this means that random variables $X_1, \ldots, X_n$ are measurable with respect to sigma-algebra $F_n$. In practice, this simply means that we have observed the values of random variables $X_1, \ldots, X_n$.

In the prediction, the task is to give the best estimate for the unknown random variable $Z$ based on the given information, encoded as a sigma-algebra $F_n$. Now if $Z$ would be measurable with respect to $F_n$, then we know the value of $Z$ already so there is nothing to predict. More generally, if $F_n$
is the natural information produced by random variables $X_1, \ldots, X_n$, then
the task is to predict unknown $Z$ as a function $h(X_1, \ldots, X_n)$ of the known
variables. Optimal prediction in the $L^2$-sense is given by the conditional
expectation.

**Definition 2.1.4.** [Conditional expectation] Let $Y$ be an integrable random
variable. Then the conditional expectation $E[Y|F_n]$ is defined as the unique
$F_n$-measurable random variable satisfying

$$E[E[Y|F_n]1_A] = E[Y1_A]$$

for all events $A \in F_n$.

Conditional expectations are natural to use in prediction as they mini-
mimise the variance of the error. For details, see [20]. In the case of Gaussian
random variables, conditional expectations remain Gaussian.

### 2.2 Stochastic Integration

Stochastic integration has many applications in various areas of science.
For example, partial differential equations and differential systems de-
scribe many aspects of our daily life, and their stochastic counterparts
incorporate random outside forces or measurement errors into the system.
However, in typical situations the randomness arises e.g. from the Brow-
nian motion that is non-differentiable, and consequently the equation is
understood as an integral equation instead. Moreover, even the concept of
integral is subtle, and there are various ways to define stochastic integrals
with respect to given processes with different properties. Here we present
only few concepts of stochastic integration that are used in this thesis. The
following definitions belong to the class of so-called pathwise integrals,
meaning that concepts are *a priori* analytical and in the context of pro-
cesses, they are considered to hold for almost all paths of the associated
processes.

#### 2.2.1 Riemann-Stieltjes integration

One natural approach to define the integral $\int_0^T X_s dY_s$ is through the limit
of Riemann-Stieltjes sums. That is, let $(\pi_n)_{n=1}^\infty = \{0 = t_0^n < \cdots < t_{k(n)}^n = T\}$
be a partition with $|\pi_n| = \max_{j=1,\ldots,k(n)} |t_j^n - t_{j-1}^n| \to 0$ as $n \to \infty$. Then the
Riemann-Stieltjes integral is defined as

$$\int_0^T X_s dY_s = \lim_{n \to \infty} \sum_{t_j^n \in \pi_n \cap [0,T]} X_{\xi_j^n}(Y_{t_j^n} - Y_{t_{j-1}^n}), \quad \xi_j^n \in [t_{j-1}^n, t_j^n]$$

provided that the limit exist and is independent of the choice of the middle
point $\xi_j^n$. In the context of stochastic processes, the Riemann-Stieltjes
integral is referred as pathwise integral if the limit exists almost surely. A well-known sufficient condition for the existence of the Riemann-Stieltjes integral is that $X$ is continuous and $Y$ has bounded variation (or vice versa, $X$ has bounded variation and $Y$ is continuous). That is, if $Y$ satisfies

$$Var_p(Y) = \sup_{\pi_n} \sum_{t^n_j \in \pi_n \cap [0,T]} |Y^n_{t^n_j} - Y^n_{t^n_{j-1}}|^p < \infty$$

for $p = 1$, where supremum is taken over all possible partitions $\pi_n$. We denote by $W_p([0,T])$ the class of functions $Y$ such that $Var_p(Y) < \infty$. The following extension is due to [44] and can be used for stochastic integration if $X$ and $Y$ are suitably Hölder regular.

**Theorem 2.2.1.** Suppose $X \in W_p([0,T])$ and $Y \in W_q([0,T])$ for $1 \leq p, q < \infty$ such that $\frac{1}{p} + \frac{1}{q} > 1$. Assume further that $X$ and $Y$ do not share any common points of discontinuities. Then the Riemann-Stieltjes integral $\int_s^t X_u dY_u$ exists for any subinterval $[s,t] \subset [0,T]$.

Note that if $X$ and $Y$ are Hölder continuous of sufficient order, then they automatically have finite $p$-variations and consequently, Young integration can be applied. In some cases the limit of the Riemann-Stieltjes sums depend on the chosen point $\xi^n_{t_j}$. For example, this is the case for the standard Itô integral where one considers only forward type sums (note however that the limit in Itô integration is in probability, and as such Itô integral is not a pathwise integral). In the forward integral, the integral is defined as the limit

$$\lim_{n \to \infty} \sum_{t^n_j \in \pi_n \cap [0,T]} X^n_{t^n_j} (Y^n_{t^n_j} - Y^n_{t^n_{j-1}}), \quad \xi^n_{t_j} \in [t^n_{j-1}, t^n_j]$$

provided it exists.

### 2.2.2 Lebesgue-Stieltjes integration

The integral $\int_0^T X_s \, dY_s$ can be defined in the Lebesgue-Stieltjes sense, if the "differential" $dY_s$ can be understood as a measure. In particular, if $Y$ is a non-decreasing function, then the derivative $dY_s$ exists as a measure. More generally, if $Y \in W_1([0,T])$, i.e. has bounded variation, then $Y$ can be written as the difference of two non-decreasing functions and consequently, derivative of $Y$ can be regarded as a signed measure (that is, a difference of two positive measures). Note that if $X$ is continuous and $Y$ is of bounded variation, then both Riemann-Stieltjes integral and Lebesgue-Stieltjes integrals exists and their value coincides. Similarly as Young integration extends Riemann-Stieltjes integrals into functions of bounded $p$-variations, Lebesgue-Stieltjes integrals can be generalised by means of fractional integrals and derivatives.
Definition 2.2.1. [Riemann-Liouville fractional integrals] For a function $f \in L^1([a, b])$ and $t \in (a, b)$, the (one-sided) Riemann-Liouville fractional integrals of order $\beta \in (0, 1)$ are defined as
\[
(I^\beta_{a+}f)(t) = \frac{1}{\Gamma(\beta)} \int_a^t f(u)(t-u)^{\beta-1}du
\]
and
\[
(I^\beta_{b-}f)(t) = \frac{(-1)^{-\beta}}{\Gamma(\beta)} \int_t^b f(u)(u-t)^{\beta-1}du,
\]
where $\Gamma$ denotes the Gamma function.

The Riemann-Liouville fractional derivatives of order $\beta \in (0, 1)$, denoted by $(D^\beta_{a+})$ and $(D^\beta_{b-})$, are defined as the left-inverses of the corresponding integrals $(I^\beta_{a+})$ and $(I^\beta_{b-})$. For more details, see [35] and references therein. The generalised Lebesgue-Stieltjes integral is given in terms of fractional derivatives, defined as follows.

Definition 2.2.2. Set $f_{a+}(x) = f(x) - f(a+)$ and $g_{b-}(x) = g(x) - g(b-)$. If $f_{a+} \in L^p([a, b])$ and $g_{b-} \in L^q([a, b])$ for some $\beta \in (0, 1)$ and $p, q \geq 1$ satisfy $\frac{1}{\beta} + \frac{1}{q} \leq 1$, then we can define the generalised Lebesgue-Stieltjes integral over an interval $[a, b]$ as
\[
\int_a^b fdg = \int_a^b (D^\beta_{a+}f_{a+})(s)(D^{1-\beta}_{b-})g(s)ds - f(a+)(g(b-) - g(a+)).
\]

It can be shown that the value of the integral does not depend on the choice of $\beta$ in the sense that for every value $\beta$ for which the integral is well-defined, it has the same value. Generalised Lebesgue-Stieltjes integration works admirably well with the fractional Besov/Sobolev spaces. We define the (weighted) fractional Besov space $W^{\beta, \infty}_T = W^{\beta, \infty}_{T}([0, T])$ as the space of all real-valued measurable function $f : [0, T] \rightarrow \mathbb{R}$ for which
\[
\|f\|_{\beta, \infty} = \sup_{0 < t < T} \left( \frac{|f(T) - f(t)|}{(T-t)^{\beta}} + \int_t^T \frac{|f(t) - f(s)|}{|t-s|^{1+\beta}} ds \right) < \infty.
\]

Similarly, we define the (weighted) fractional Besov space $W^{\beta, 1}_0 = W^{\beta, 1}_{0}([0, T])$ as the space of all real-valued measurable function $f : [0, T] \rightarrow \mathbb{R}$ for which
\[
\|f\|_{\beta, 1} = \int_0^T \frac{|f(t)|}{t^{\beta}} dt + \int_0^T \int_0^T \frac{|f(t) - f(s)|}{|t-s|^{1+\beta}} dsdt < \infty.
\]

Then it is known that if $f \in W^{\beta, 1}_0$ and $g \in W^{1-\beta, \infty}_T$, the generalised Lebesgue-Stieltjes integral exists. For details, see [45] and for some recent developments, see e.g. [16, 17]. In particular, such an approach can be used to define pathwise stochastic integral $\int_0^T f(B^H_s)dB^H_s$, where $B^H$ is a fractional Brownian motion with $H > \frac{1}{2}$ or similar fractional type process, and $f$ is a one-sided derivative of a convex function in which case $f$ is a non-decreasing and hence the derivative $f'$ exists as a measure.
2.3 Breast cancer screening

Breast cancer is the most common cancer type among women in Western countries [19]. For example, for women in Finland, the estimated lifetime probability of breast cancer is 13% [11]. Several known risk factors affect the probability of breast cancer. It is well-known that breast cancer risk depends on age, the use of post-menopausal hormones, tobacco, and alcohol [43]. Also, the risk may depend on genetic factors [31].

Treatment and prognosis of breast cancer depend heavily on the cancer stage. Early detection is therefore crucial. Fortunately, breast cancer mortality can be reduced by screening. Several studies on the effectiveness of breast cancer screening, as well as on stage-specific breast cancer costs have been published (see for example, [5, 13, 14, 24, 32, 33, 37, 39]). Before year 2022 the European Union recommended to offer mammographic screening to all women aged 50-69 years and most European countries follow this recommendation (see for example: [4, 8, 28]). New EU recommendations suggest extending breast cancer screening to younger and older age groups [9, 10].

In Finland, nationwide biennial mammographic breast cancer screening of 50-59-year-old women started in 1992. From Figure 2.1, one can see the effect of screening on mortality. Even though the incidence rate is increasing, mortality is decreasing. All municipalities have been responsible to offer breast cancer screening and some of them have offered mammo-

![Figure 2.1. Age-standardised incidence and mortality per 100 000 women in Finland in different years. Data is obtained from NORDCAN [2].](image-url)
graphic screening also to wider target ages [36]. The nationwide target
group in Finland is 50-69-years-old women. The age group was introduced
gradually and adopted nationwide in 2017. Participation to screening is
currently about 82% and varies very little within the target age [11].

The goal of breast cancer screening is to prevent death from the dis-
ease by means of early detection. Benefits should be larger than harms.
Reducing mortality should not be the only target. Also, costs, overdiag-
noses, psychological effects of screening and treatment related morbidities
should be considered. New screening programs should be evaluated using
cost-effectiveness modeling [28].

Finnish Cancer Registry collects and stores nationwide individual data
on all breast cancer cases in Finland. In addition, their Mass Screening
Registry database contains individual level data on breast cancer screen-
ing from invitations, participations, test results, referral examinations,
and findings for close to 1,4 million women from year 1992 and onwards.
That provides useful resources for research and the databases are utilized
in Article 3, where health-economic dimensions related to the current
breast cancer screening strategies in Finland and alternative target ages
of screening are evaluated.

The study presented in Article 3 initiated from discussions related to
new EU recommendations that suggest extending breast cancer screening
target age groups [9, 10]. As it can be seen from Figure 2.2, the breast
cancer incidence rates in Finland are high for over 70-year-old women
and low for under 50-year-old women. However, one should note that it is
important that cancer is detected as early as possible to save more lives
and to reduce mortality rates.
Preliminaries

Figure 2.2. Breast cancer incidence per 100 000 women in Finland in years 2016–2020. Data is obtained from NORDCAN [2].

2.3.1 Modelling approaches

Decision-analytic modelling is important for healthcare economic evaluations such as measuring the cost-effectiveness of cancer screening. Randomized clinical trials are often not possible for analysing the effect of breast cancer screening as they are costly, time consuming, and may involve ethical issues. Thus one usually relies on deterministic or simulation based models. [38]

Cost effectiveness of breast cancer screening in Finland was studied in 1999 [25]. The effectiveness of screening was simply measured by life-years saved due to screening. The study was based on modelling the effect of screening on the cumulative relative mortality. However, this study is already outdated as cancer incidence and treatment are rapidly changing.

There are several techniques for modelling the cost-effectiveness of breast cancer screening. The applied models include models based on decision trees, probabilistic approaches that might involve modelling the natural progression of cancer, Markovian type models, or discrete event simulation models, see, e.g. [6, 12, 23, 29, 34, 38, 42].
3. Summaries of the articles

The main contributions of this thesis are the following three scientific articles that are summarised below.

3.1 On optimal prediction of missing functional data

While in the machine learning era we are familiar with big data, sometimes parts of the data are missing, and one has to handle with missing data. Interpreting the missing data is not an easy task, as the missing random data typically depends on the observed data in a very complicated way, due to complex dependency structures. On the contrary, in the presence of independence one cannot gain anything from the observed data. That is, one gains accuracy of predictions under strong dependency, but at the same time this makes analysis much more difficult.

The article studies the problem of predicting missing data from the observed in the functional data context. More precisely, we assume that the observations are functions (of time $t$), but some intervals are missing and needs to be predicted by using the observed parts. The observed functions are considered to be of the form $Y = f(X)$, where $f$ is a bijective (possibly unknown) function and $X$ is a Gaussian process with (possibly unknown) covariance structure $C$. This provides a relatively rich class of functions by allowing rather arbitrary transformation $f$, but at the same time we have the rich theory of Gaussian processes at our disposal.

First main results of the article consider the case where observations arise from Gaussian processes with known covariance function. We prove that in this case, by using the Fredholm representation of the Gaussian process as an integral with respect to a standard Brownian motion, one can obtain optimal prediction (that is, the conditional expectation given the information what is observed) by solving certain integral equations. We also provide methodology how such equations can be solved in practice by
Summaries of the articles

considering the case where the Gaussian process is not observed at certain intervals, but merely on certain time points. By exploiting the results on the Gaussian case, we then derive methodology to obtain optimal prediction in the general case \( Y = f(X) \), with unknown function \( f \) and unknown covariance of \( C \). For this purpose we assume that we have observed \( Y \) many times independently, allowing to estimate covariance and using empirical measures to estimate \( f \). We prove consistency for the estimator of the predictor and analyse the corresponding rate of convergence.

3.2 On sharp rate of convergence for discretisation of integrals driven by fractional Brownian motions and related processes with discontinuous integrands

The article studies the rate of convergence for discretisation of stochastic integrals with equidistant time intervals. More precisely, we consider stochastic integral of type \( \int_{0}^{1} f(X_s) dX_s \) where \( f \) is a function of bounded variation and \( X_s \) is the fractional Brownian motion with Hurst index \( H > \frac{1}{2} \) or a closely related Gaussian process that is Hölder continuous of order above \( \frac{1}{2} \). The innovative aspect of the article compared to the existing literature lies in the assumptions on the coefficient \( f \) that is allowed to possess discontinuities, making stochastic Riemann-type integration a cumbersome problem. However, by using a change of variable the integral can be written as \( F(X_1) - F(X_0) \) where \( F \) is a primitive of \( f \). Together with some convex analysis and explicit Gaussian analysis, this allows to study sharp estimates on the rate of convergence for the discretisation of the integral.

The main result of the article states that this rate is proportional to \( n^{1-2H} \), corresponding to the known rate in the case of smooth coefficient function \( f \). The result is rather surprising due to the fact that, in the case of the standard Brownian motion, introducing jumps to the coefficient function \( f \) drops the rate to be proportional to \( n^{-\frac{3}{2}} \), compared to the rate \( n^{-\frac{1}{2}} \) for the smooth case.

3.3 On Cost-effectiveness of Breast cancer screening in Finland

Breast cancer has become widespread cancer in Western countries. It is the most common cancer among Western women. One of the most successful methods in reducing breast cancer mortality is organized screening with mammography, as early detection has a significant effect on prognosis. European Union has given recommendations to screen women aged 50-69 years and this recommendation is followed well in most European countries. However, significant amount of breast cancer cases occurs in
women who are younger than 50 or older than 69. New studies and recommendations support extending screening to new age groups. Before extending screening, it is important to consider the benefits and harms of screening as the costs and benefits are not the same in different countries and regions. Thus, cost-effectiveness modeling should be conducted in order to support local policy making.

In cost-effectiveness modeling, it is customary to model the outcome and costs in a cohort under different screening strategies and no-screening strategy is used as a reference. Reliable modeling relies on reliable data, and it would be ideal to use data where individuals were randomly assigned to different screening strategy groups. However, this type of data is rarely available, and modeling has to be adjusted according to availability of data. In this article, we propose a new approach to cost-effectiveness modeling of breast cancer screening. The model is suitable for situations where there is already a long ongoing screening program and when historical data is incomplete. It is based on estimating the breast cancer stage distributions under different screening strategies and instead of no screening strategy, the ongoing screening strategy is used as the baseline. The model is flexible as it enables to apply different estimating procedures for estimating the treatment costs, screening costs, stage distributions, and outcomes. In particular, if randomized data is available, estimation can be based on that.

The new model is applied for assessing the costs-effectiveness of extending the current biennial breast cancer screening program to younger and older age groups in Finland. Our modeling suggests that if screening would be extended to younger age groups, the incremental costs per life years gained would be low, and if screening would be extended to older age groups, lives would be saved, but incremental costs per life-years gained would be high.
In the first article, we predict missing functional observations optimally in the mean square sense. Exploiting Gaussian analysis, we prove that this can be done by solving certain integral equations. In a more general setting, the method relies on analysing empirical distributions and covariance structures of the underlying processes. The results are significant as they solve an important problem in modelling functional observations. The strength of the results lie on flexibility and data-drivenness of the provided approach. On the other hand, as a fully data-driven method, the approach requires significant amount of computing power.

In the second article, we derive sharp rate of converge for discretisation of certain complicated stochastic integrals involving Gaussian processes. The obtained rate provides significant improvement to the rates presented in the literature previously. The results are significant. The studied stochastic integrals are known to be mathematically challenging objects to analyse. A convergence rate for discretisations enable error analysis, e.g. in simulations and real data applications. The strength of the results lie on the fact that the obtained rate is sharp, allowing exact analysis. The method however is limited to stochastic integrals driven by Gaussian processes.

In the third article, we build a cost-effectiveness model to assess the effect of different breast cancer screening strategies. These results are significant as knowledge of cost efficiency is crucial for decision making. The lack of randomized data limited the analysis. However, the strength of the study lies on the fact that while guiding decision making related to breast cancer screening in Finland, the same modelling approach could be applied for different populations and under different screening scenarios and estimation procedures.

Future prospects involve studying the effect of discretization grid and other parameter choices in more detail in the approach provided in the first article, extending convergence rate results of the second article beyond Gaussian processes with Hölder restrictions such as processes that could be applied in finance, and modifying the analysis presented in the third
Discussion and future prospects

article to involve separate screening protocols for known risk groups.
References


