Sparse Bayesian learning methods and statistical survival models

Ingvild Margrethe Helgøy

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2023

UNIVERSITY OF BERGEN

Sparse Bayesian learning methods and statistical survival models

Ingvild Margrethe Helgøy

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 02.06.2023

Year: 2023

Title: Sparse Bayesian learning methods and statistical survival models

Name: Ingvild Margrethe Helgøy

Print: Skipnes Kommunikasjon / University of Bergen

Preface

This thesis is submitted as a partial fulfillment of the requirements for the degree of Philosophiae Doctor (Ph.D.) at the University of Bergen. The advisory committee has consisted of Yushu Li (University of Bergen) and Hans J. Skaug (University of Bergen).

Acknowledgments

There are many who deserve to be thanked at the end of my work on this thesis. First and foremost, I thank my two supervisors, Yushu and Hans that have supported me through this endeavour. Yushu has been a thorough support and given me feedback on my work. I am very grateful for all the time you have spent discussing with me during my time as a PhD-candidate. The input and encouragement from Hans has been vital for the completion of this thesis.

I also give my thanks to my other colleagues and friends at the department of Mathematics for all the good memories, both during and outside of our studies. Further, I am very grateful to Volda University College for giving me time to finish my thesis after I was employed with them.

Finally, I would like to thank my friends, my parents and the rest of my family that have supported me during my time as a PhD-candidate. I especially want to thank Runar who has been there for me and given me help and support when I needed it the most.

Ingvild Margrethe Helgøy
January 2023

Samandrag

Med utviklinga av avanserte datainnsamlingsteknikkar og digitalisering, har omfanget av tilgjengelig data i ulike fagfelt auka enormt. Det blir difor stadig viktigare med effektive og nøyaktige algoritmar innan statistisk modellering, og denne avhandlinga inneheld bidrag til dette. Bidraga kan kategoriserast i to hovudtema: glisne (sparse) Bayesianske metodar og statistiske levetidsmodellar.

Glisne Bayesianske metodar har fått auka interesse dei siste åra, mellom anna på grunn av at dei produserer modellar som generaliserer godt og som er robuste mot data som inneheld mykje støy. Det finst mange ulike variantar av Bayesianske metodar, som til dømes metodar som brukar Markov Chain Monte Carlo. Denne avhandlinga fokuserer på den empirisk Bayesianske tilnærminga. Avhandlinga undersøker både nye glisne Bayesianske modellar og ein ny generell løysingsstrategi for desse modellane. I den nye løysingsstrategien brukar ein R-pakken, Template Model Builder (TMB), til å optimere modellparametrane ved å bruke automatisk derivasjon. Ein algoritme som forbetrar kjøretida for estimering av dei latente variablene i TMB, vert også presentert. Ved å bruke denne algoritmen oppnår ein tilnærma lik tidsbruk ved å bruke TMB som dei orginale algoritmane for desse Bayesianske modellane. Løysingsstrategien gjer det enkelt å justere modellane utan å måtte utleie nye komplekse algoritmar, og dette blir demonstrert ved å bruke den på nye glisne Bayesianske modellar. I tillegg til modellane som er utleia i det nye løysingsrammeverket, inneheld avhandlinga også ei analytisk løysing til ein ny modell som er relatert til Bayesiansk lasso. I motsetning til Bayesiansk lasso, gir den nye modellen glisne løysingar og kan også bli brukt til å løyse ikkje-lineære regresjonsproblem.

Det andre temaet for denne avhandlinga er statistiske levetidsmodellar. Her blir det presenterer ei ny multivariat fordeling for å modellere avhengige levetider, som blir kalla søskenfordelinga. Fordelinga blir definert ut frå levetida for søsken, der avhengighetsstrukturen blir indusert gjennom felles mor. Søskenfordelinga blir konstruert slik at komponentane som er knytt til mor er inkludert som latente variablar, og ein treng difor ingen informasjon om henne. Sjølv om fordelinga vert presentert som ei fordeling av leveår, kan den bli nytta meir generelt på avhengige komponentar. Vi beviser at den bivariate søskenfordelinga med konstante rater er Multivariate Totally Positive of order two (MTP2), som er ein sterk avhengig eigenskap og indikerer mellom anna ein positiv kovarians. Modellparametrane er fødsels- og dødsratene, i tillegg til dei individuelle

tidspunkta for død. Estimering av desse tidspunkta kan ikkje gjerast ved å bruke klassiske estimeringsprosedyrar, då rimelegheitsfunksjonen ikkje er deriverbar med hensyn til desse parametrane. For å løyse dette problemet blir ein iterativ estimeringsalgoritme utvikla, som gjev estimat på alle modellparametrar. Algoritmen blir testa både på simulerte og ekte data. Resultata viser at estimerte verdiar ligg tett opptil dei sanne verdiane ved testing på simulerte data.

Abstract

With the development of advanced data collection techniques and digitalization, the amount of available data in various fields has increased tremendously. The need for efficient and accurate algorithms for statistical modelling is therefore becoming more and more important, and this thesis contains contributions towards more efficient models. The contributions can be characterized in two main topics; sparse Bayesian learning methods and statistical survival models.

Sparse Bayesian learning methods have gained increased interest the recent years due to the favourable properties that they provide sparse models that generalize well and that they are robust to noisy datasets. While there are many approaches to Bayesian learning, such as Markov Chain Monte Carlo methods, this thesis focuses on the empirical Bayes approach. The thesis investigates both new sparse Bayesian models, and a new general solution strategy for these models. In the new solution strategy an `R` package, the Template Model Builder (TMB), is used to optimize the model parameters by applying automatic differentiation. An algorithm that speeds up the estimation procedure of the latent variables in TMB is also presented. Applying this algorithm obtains similar runtimes using TMB as compared to tailored algorithms of the sparse Bayesian models. The solution framework makes it easy to adjust the models without derivation of new complex algorithms, which is demonstrated by applying it to new sparse Bayesian models. In addition to the models derived in the new solution framework, the thesis also includes an analytical solution to a new model that is related to the Bayesian lasso. Opposed to the Bayesian lasso, the new model provides sparse solutions and can also be applied to solve nonlinear regression problems.

In the second topic of this thesis, a new multivariate distribution for modelling continuous lifetimes with positive dependence, named the sibling distribution, is presented. The distribution is defined in terms of the survival of siblings, where the dependency structure is induced through their shared mother. The sibling distribution is constructed such that the components related to the mother are included as latent variables, hence, no knowledge about the mother is required. Although it is presented as a distribution of lifetimes, it may be applied to any set of nonnegative components with positive dependence. We prove that the bivariate sibling distribution with constant rates is Multivariate Totally Positive of order two (MTP_2), which is a strong dependence property and implies among others things a positive covariance. The model parameters are the birth

and death rates, in addition to the individual death time points of the siblings. Estimates of the time points can however not be obtained by applying classical estimation procedures, as the likelihood is not differentiable with respect to these parameters. In order to solve this problem, an iterative estimation algorithm is derived, which provides estimates of all model parameters. The estimation algorithm is tested on both simulated and real data. The results show that the estimated values were close to the true values, when testing on simulated data.

List of papers

A **The sibling distribution for multivariate life time data**
Ingvild M. Helgøy, Hans J. Skaug
Accepted to Sankhya B
DOI: 10.1007/s13571-021-00259-w

B **A Bayesian Lasso based Sparse Learning Model**
Ingvild M. Helgøy, Yushu Li
Submitted

C **Sparse Bayesian Learning using TMB (Template Model Builder)**
Ingvild M. Helgøy, Hans J. Skaug, Yushu Li
Submitted

Contents

Part I

Background

Chapter 1

Introduction

One of the earliest applications of statistics was in the field of demography, which analyses the size, changes and structures of populations. A major contribution to modern demography was the discovery that the age distribution of populations with time-independent birth and death rates was stable, and from this result stable population models could be derived [1, 2, 3]. While Euler introduced the concept of a stable population model as early as 1760 [4], the practical use was not discovered at that point due to the limited data, which did not indicate stability. From the 1970s models with more general population dynamics were developed [5, 6, 7, 8], and Carr [9] extended the stable population model from a closed population model to a model that handles multi-regional populations and the emigration between regions. This was an important contribution that takes heterogeneity of populations into account. Today, the stable population model is not only relevant within demography [10, 11, 12, 13], but is also a central topic in population biology [14] and epidemiology [15]. This thesis builds on stable population theory and derives a new distribution that describes the life expectancy of siblings.

Statistical analysis is often applied with the aim of describing or estimating a population from information of limited samples. The term population need not only refer to people, but has a much broader meaning; population in statistics is the entire group we want to study and can be any set of similar items or events such as objects, organizations, house prices, etc. The term population may therefore be used within many different subjects. However, no matter which subject, in order to obtain good statistical analysis, access to data is paramount.

Statistical data analysis has often been divided into two different ap-

proaches: the frequentist and the Bayesian. The main difference of the traditional frequentist and the Bayesian approach is how the parameters in the model are treated. In the frequentist approach, the parameters are non-random but unknown quantities. A common procedure to obtain estimates of the parameters in this framework is maximum likelihood estimation. The maximum likelihood estimation finds the values of the parameters that maximize a likelihood function. The method of maximum likelihood estimation is explained in more detail in Section 3.2. In the Bayesian approach the parameters are assumed to be random variables, and inference is mainly based on the posterior distribution of the parameters. The posterior distribution is the probability distribution of the parameters given the data. Bayesian ideas were presented by Thomas Bayes during the 18th century, however, the foundation for the modern Bayesian statistics was to a large extent developed a Century later by Pierre Simon Laplace [16, 17]. An introduction to the Bayesian framework is presented in Chapter 4.

Bayesian statistics has also seen an extensive development in the more recent years, and it has been used with great success to solve problems in various fields, such as, genetics where it has been used to discover the relationship between genetic variants and diseases [18, 19, 20], and machine learning with the development of Bayesian neural networks [21, 22, 23]. A challenge with the Bayesian framework is that the posterior distribution of the parameters may not be possible to calculate analytically as it often requires calculations of high-dimensional integrals. A class of popular simulation algorithms called Markov Chain Monte Carlo (MCMC) [24] has been developed to overcome this problem. These algorithms provide a sample from the posterior distribution, instead of computing integrals. MCMC methods have, however, some well known drawbacks, such as, often being time consuming and computationally costly.

The last decades, the amount of available data has increased exponentially in many fields, and in order to handle the large amount of data and extract the most important information, new sparse methods have been derived. An option which has gained increased attention is Sparse Bayesian Learning (SBL) [25]. Examples of these methods include the Relevance Vector Machine (RVM) [26, 27], the incremental relevance sample-feature machine [28] and the probabilistic feature selection and classification vector machine [29]. These methods can provide fast algorithms by approximating the posterior distribution by the empirical Bayes approach which is presented in Section 4.4. The SBL algorithms are in general faster than the MCMC methods and they produce sparse models where only a small

fraction of the model parameters are nonzero. Sparse models reduce the possibility of overfitting data and thereby achieve good generalization properties [30, 27]. The faster approximation and the sparsity of the models have resulted in successful applications of SBL in many different areas including image classification [31, 32, 33] and time series prediction [34, 35]. The SBL methods are particularly relevant within the field of compressive sensing [36, 37] where the signal has a low-rank representation, which favours sparse and noise robust models, and they have been used frequently in recent research in this field [38, 39, 40, 41].

1.1 Main contributions

The main contributions of this thesis are:

A new multivariate distribution for continuous lifetimes. The distribution, named the sibling distribution, is derived to model the life expectancy of siblings with a common mother. The distribution is presented in Paper A where the distribution is validated on both simulated and real data. When the birth and death rates are constants, we prove that the bivariate sibling distribution is MTP_2. This property implies that there is a strong dependency between the lifetimes of the siblings. We also show how the sibling distribution reduces to the Block-Basu class of distributions [42] under certain assumptions.

Novel applications of the open source R package, the Template Model Builder (TMB). In Paper C the focus is on applying the TMB [43] to sparse Bayesian models. While TMB has been very popular, it has to the best of our knowledge never been applied to solve these types of problems. A reason might be due to the large number of hyperparameters that makes the estimation procedure slow. One of the main contributions of Paper C is a tailored algorithm that speeds up the estimation procedures by orders of magnitude, making TMB a viable option to solve these type of problems. The TMB package is also applied in Paper A, where a challenge of the sibling distribution is to obtain estimates of the parameters, due to the unregularity of the distribution. By designing an iterative algorithm that utilize TMB we show how this issue can be overcome.

New sparse Bayesian models. The Bayesian lasso [44] is extended in Paper B to a new model that can be applied to more general nonlinear regression problems and, in addition, provides sparse solutions.

The new model in Paper B, called BLS (Bayesian Lasso Sparse), is compared to other well known sparse Bayesian models like the Relevance Vector Machine (RVM) by Tipping [26] and the Fast Laplace (FLAP) by Babacan et al. [45]. The methods studied in Paper B can, however, only achieve sparsity with respect to either the samples or the dimension, but not both simultaneously. On the other hand, sparse Bayesian methods that can do both sample and dimension reduction are studied in Paper C. This paper shows how we can modify sparse Bayesian models, in a simple manner, without having to write an extensive new algorithm by taking advantage of the simple structure of the TMB package. This is exemplified by presenting a novel extension of the original RVM method that can also perform dimension reduction.

1.2 Outline

This thesis consists of two parts. The first part provides an introduction to the statistical framework, methods and models that are used in Part II. Part II presents the main scientific contributions which consists of three papers. The reminder of Part I has the following structure:

Chapter 2 gives a brief introduction to mathematical demography and the stable population model.

Chapter 3 presents the concept of supervised learning and the extended linear regression framework.

Chapter 4 presents the Bayesian framework, and gives an introduction to sparse Bayesian learning models.

Chapter 5 introduces new theory for extending the BLS method to a classification setting.

Chapter 6 gives a summary of Papers A, B, and C.

Chapter 2

Mathematical demography

This chapter presents theory and relevant background for the sibling distribution which is developed in Paper A. The sibling distribution is a multivariate distribution that models the lifetimes of siblings. The distribution is derived by assuming a stable population model, which is characterised by an age distribution that is stable and independent of the time. The derivation of the sibling distribution requires a framework that defines vital rates and specification of the underlying population model. The following sections present an excerpt from the stable population theory and classical results from mathematical demography, which lay the mathematical foundation for the sibling distribution developed in Paper A.

Mathematical demography is a wide research area that originates from many different fields such as biology, mathematics, statistics and actuarial science [4, 46, 2]. A central topic in mathematical demography is how to analyse the dynamics of a population, such as population growth and decline.

2.1 Malthusian population theory

There exist a variety of different population models. One of the earliest and most well known models was presented in 1789 in *An Essay on the Principle of Population* where Malthus claimed that populations grow geometrically [47]. The model by Malthus is a simple, yet powerful population model, which is useful for introducing the basic concepts of demographic processes. The changes in a population is described by the time-dependent growth rate, which we denote by r_t. The changes in the population at time

t is then given by

$$\frac{\mathrm{d}N_t}{\mathrm{d}t} = r_t N_t, \tag{2.1}$$

where N_t is the population size at time t. It is assumed that N_t is a continuous function of the time. Continuous populations models are realistic assumptions for populations that are sufficiently large and who reproduce continuously in time [46]. Assuming a constant growth rate r, the solution of Equation (2.1) is an exponential function:

$$N_t = N_0 e^{rt},$$

where N_0 denotes the population size at time $t = 0$. This Malthusian population model assumes that there is no lack of resources, as the model allows an unlimited growth of the population when the growth rate is positive. While this may seem like an unrealistic model, the model can often capture the initial growth of a population well.

2.2 Birth and death rates

The random variables in the sibling distribution, presented in Paper A, are the life times of the siblings. In order to model the life span of the siblings, we therefore need theory from survival analysis, which we explore in this section. In the sibling distribution, we also account for the fertility and mortality of the mother. The individual birth and death rates are consequently the main components of the sibling distribution.

Two central concepts in survival theory are the survival function and the hazard rate. In general, the survival function gives the probability that the event of interest has not yet happened by time t. Thus, the survival function is not exclusively applied to situations related to survival or death, but can also be applied to a much larger set of situations. The survival function is given by

$$l(t) = P(T > t), \tag{2.2}$$

where the random variable T represents the survival time. The survival function is closely connected to the hazard rate, denoted by $h(t)$. Given that T is a continuous random variable, one assumes the probability that the event of interest occurs in a small time interval, $[t, t + dt)$, to be $h(t)dt$ [48]. The hazard rate can be defined as a limit by using conditional

probabilities:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{2.3}$$

and can in general be any nonnegative function opposed to the survival function which starts at 1 and usually declines toward zero [4].

For the sibling distribution, the survival function in Equation (2.2) will be interpreted as a function of age and not time. Assume that A is a random variable that denotes the age of death. The survival time for a given individual in the population is then given by $l(a) = P(A > a)$.

The age-specific death rate is defined as the instantaneous death rate at age a and will be denoted by $\phi(a)$. The death rate can be defined from the hazard rate in Equation (2.3) where the event of interest is the occurrence of death [4]. The death rate and its relation to the survival function $l(a)$ can be expressed by

$$\begin{aligned}
\phi(a) &= \lim_{\Delta a \to 0} \frac{P(a \leq A < a + \Delta a | A \geq a)}{\Delta a} \\
&= \lim_{\Delta a \to 0} \frac{l(a) - l(a + \Delta a)}{l(a) \Delta a} \\
&= -\frac{l'(a)}{l(a)}.
\end{aligned}$$

Thus, using the initial condition that $l(0) = 1$ one obtain

$$-\log l(a) = \int_0^a \phi(u) \, \mathrm{d}u,$$

and it follows that an expression for the survival function can be given as

$$l(a) = e^{-\int_0^a \phi(u) \, \mathrm{d}u}. \tag{2.4}$$

Notice that when the death rate is age-independent, i.e. $\phi(a) = \phi$, the conditional probability of living at least x additional years is found from Equation (2.4) as $l(a + x)/l(a) = \exp(-\phi x)$.

The age specific birth rate is defined such that an average individual produces $\beta(a) \, \mathrm{d}a$ children during the age interval $[a, a + \mathrm{d}a)$ per unit of time. Since the event of giving birth may occur more than once, a common approach is to express the birth process by a Poisson process [4]. Let M be the random variable that represents the total number of offspring, and $M(a)$ the number of offspring obtained at age a. The probability of a birth in the time interval $[a, a + h)$ is given by $\beta(a)h + o(h)$ and the

probability that more than one birth occur is given by $o(h)$. For simplicity, let $P_m(a) = P(M(a) = m)$, that is the probability of having m offspring at age a. The birth process can then be expressed by the inhomogeneous Poisson process:

$$\frac{\mathrm{d}P_0(a)}{\mathrm{d}a} = -\beta(a)P_0(a), \tag{2.5}$$

$$\frac{\mathrm{d}P_m(a)}{\mathrm{d}a} = -\beta(a)P_m(a) + \beta(a)P_{m-1}(a), \qquad m \geq 1. \tag{2.6}$$

The probability of having no offspring at age zero is $P_0(0) = 1$ and the probability of having m offspring at age zero is $P_m(0) = 0$. By using these initial conditions, the solution of the system of differential Equations (2.5)-(2.6) is

$$P_m(a) = \frac{e^{-\int_0^a \beta(u)\,\mathrm{d}u}}{m!}\left[\int_0^a \beta(u)\,\mathrm{d}u\right]^m, \tag{2.7}$$

which gives the probability of having m children at age a [4]. Equation (2.7) is used in the sibling distribution to include the probability that the mother had m offspring during her lifetime.

2.3 The stable age distribution

This section presents the stable age distribution for continuous population models. The stable theory has also been extended to discrete population models by Caswell [49]. This is not presented here as the sibling distribution assumes a continuous population model. The stable population theory provides a convenient mathematical framework for studying the effect of fixed rates on population dynamics, estimating rates or compose new models like we have done for the sibling distribution.

A stable age distribution exists when the age-specific birth and death rates do not depend on the time or have been constant over a considerable time period. A population that obtains constant birth and death rates will converge to a stable age distribution over time, although it is not necessary stable at the current time point [46]. One of the first theories on stable population dynamics was derived by Lotka [1, 2] and Lotka and Sharpe [3] at the start of the 20th century. The theory assumed a closed population defined as a population with no immigration nor emigration. The growth rate is then determined solely by the birth and death processes.

Lotka and Sharpe [3] derived the theory for the stable population model by first considering a model for the total number of births at time t, denoted by $B(t)$, known as the "renewal equation"

$$B(t) = G(t) + \int_0^t B(t-a)l(a)\beta(a)\,\mathrm{d}a. \tag{2.8}$$

The renewal equation consists of two components. The first component, $G(t)$, represents the births of females who were alive at time $t = 0$. The second component is the number of births of females who were born after time $t = 0$. The number of females of age a that were born after time $t = 0$ can be found from the number of all newborns at time $t-a$, given by $B(t-a)$. Among the total number of newborns, the number that survives to time t and reaches the age a is given by $B(t-a)l(a)$. Taking the birth rate $\beta(a)$ into account, results in the number of births of females of age a, that were born after time $t = 0$. The number of births of females that were born after time $t = 0$ is therefore found from the integral of $B(t-a)l(a)$ between $a = 0$ and $a = t$, which is the second component of Equation (2.8).

The solution of Equation (2.8), when omitting the first term $G(t)$, is known as the characteristic equation or the *Euler-Lotka* equation [46]. As $t \to \infty$ the term $G(t) \to 0$ since the females that contribute to $G(t)$ will no longer be alive. Lotka and Sharpe [3] studied the solution of this simpler problem by assuming constant exponential growth for the births: $B(t) = e^{rt}$, from which it follows that $B(t-a) = e^{r(t-a)}$. Equation (2.8) is then equal to

$$1 = \int_0^\infty e^{-ra}l(a)\beta(a)\,\mathrm{d}a,$$

which is called the characteristic equation for the growth rate r. The characteristic equation has exactly one real solution for r, which is determined by the survival function and the age-specific birth rate [see 46].

By using the assumption that births grow exponentially at a constant rate, it follows that the number of females of age a at time t is given by $B(t-a)l(a) = e^{r(t-a)}l(a)$. The integral of this component is the total population at time t, and dividing by this component gives the proportion of the population of age $a + \mathrm{d}a$. This results in the stable age distribution:

$$f(a) = \frac{e^{-ra}l(a)}{\int_0^\infty e^{-ra}l(a)\,\mathrm{d}a}. \tag{2.9}$$

Equation (2.9) is independent of the time t and the proportion of the individuals of age $a + \mathrm{d}a$ will therefore remain constant as long as the

growth rate, r, does not change. The sibling distribution, presented in Paper A, assumes that the age of the mother is given by the stable age distribution in Equation (2.9).

From the stable population theory it also follows that the average birth rate, β, is given by

$$\beta = \left[\int_0^\infty e^{-ra} l(a) \, \mathrm{d}a \right]^{-1},$$

thus the stable age distribution in Equation (2.9) can also be expressed as $f(a) = \exp(-ra)\beta l(a)$. It can be shown that the growth rate can be calculated from $r = \beta - \phi$, where ϕ is the average death rate [4].

2.4 The sibling distribution

It is the framework of the stable population model and survival theory that makes it possible to specify the sibling distribution. The survival of the siblings can be estimated from the survival function in Equation (2.4) given the death rate $\phi(a)$. The individual survival times alone are however not sufficient to describe the sibling distribution, as the dependency of the lifetimes is related to their common mother. The sibling distribution is constructed such that no information is required about the mother. The only thing we know is that she must have been alive at a certain time point and that she had m offspring during her lifespan. The unknown time point when the mother must have been alive is taken to be the reference time point of the distribution where we set the time $t = 0$ (see Figure 2.1). The mother's age at $t = 0$ is denoted by the random variable A_0 and the assumption is that she is randomly selected among all females alive at $t = 0$, such that the density of A_0 is given by the stable age distribution in Equation (2.9). The birth and death time points of the mother are latent variables in this distribution. The birth time of the mother can be expressed by $Y_0 = -A_0$ and the time of death, T_0, can be modelled from the survival function in Equation (2.4). Conditioning on A_0 and T_0, the length of the mother's lifespan is $x_0 = a_0 + t_0$. The probability of having M offspring during the life span is found from Equation (2.7) by setting $a = x_0$. The death time points of the siblings $t_{1:m} \in R^m$ are the parameters of this distribution, in addition to the birth and death rates.

The main components of the sibling distribution can therefore be derived from Equations (2.4), (2.7) and (2.9). The general expression for the multivariate sibling distribution is formulated in terms of the age-specific

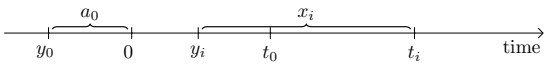

Figure 2.1: Birth and death times of the mother (y_0, t_0), and corresponding times for the i'th offspring (y_i and t_i). The age of the mother at the reference point $t = 0$ is denoted by a_0.

birth and death rates. The explicit expression for the bivariate sibling density in Paper A is, however, calculated by assuming constant birth and death rates, i.e. $\beta(a) = \beta$ and $\phi(a) = \phi$. Figure 2.2 shows different scenarios of the bivariate sibling density where $\phi = 1$ and where β varies.

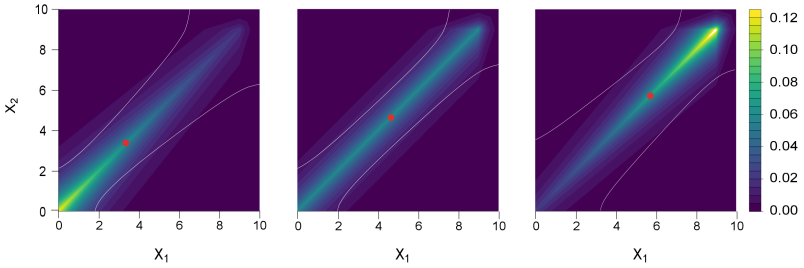

Figure 2.2: The bivariate sibling density $f(x_1, x_2|t_1, t_2)$ with parameters $\phi = 1$, $\beta = 0.8, 1.0, 1.2$ (left to right) and $(t_1, t_2) = (4, 4)$. The red dots show the expected value. The white curve shows the contour $c(x_1, x_2) = 1$ defined in Equation (3.15) in Paper A, which is a local dependency measure between X_1 and X_2.

Chapter 3

Supervised learning and regression

In this chapter we start by briefly introducing supervised learning and the extended linear regression framework, which is a natural starting point for the methods that are developed and studied in Papers B and C. The chapter also includes a short introduction to maximum likelihood estimation and regularization in the frequentist framework.

3.1 Supervised learning

Statistical learning problems are often divided into two main categories; supervised and unsupervised learning problems. Supervised learning requires a labelled training dataset that comprises observations of the input variables along with their corresponding output variables. The training set is used to learn a function that maps an input variable to an output variable. The goal is to learn a function that also generalizes well, that is, a function that maps new unseen input variables to reasonable outputs. The two main categories of supervised learning are regression [50] and classification [51]. Unsupervised learning, on the other hand, is used on unlabelled data and attempts learn the underlying structure based solely on the data, as there are no corresponding output observations. Examples of popular unsupervised learning methods are Principal Components Analysis (PCA) [52] and clustering [53]. In this thesis, only supervised learning problems are treated.

Assume that we have a set of training data that consists of N observations, $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where the input variable consists of D predictors (or features), $\boldsymbol{x}_i \in \mathbb{R}^D$, and the output variable is one-dimensional, $y_i \in \mathbb{R}$. In general, one may have more than one output variable, but in this thesis we assume that the y_i's are one-dimensional. The general assumption in supervised learning is that there is a relationship between the output variable and the input variable, which can be modelled by

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \tag{3.1}$$

where ϵ_i is a random error term. The function f is a fixed, but in general unknown function of $\boldsymbol{x}_i$. The learning task is to find an estimate $\hat{f}$ by using the training data. When searching for a function $\hat{f}$, we are mainly interested in how well the estimated function predicts future observations that are not used to train the model. Thus, given a new test observation $(y^*, \boldsymbol{x}^*)$, we commonly seek to find a function that minimizes

$$E(y^* - \hat{f}(\boldsymbol{x}^*))^2, \tag{3.2}$$

where E denotes the expectation. Equation (3.2) is the expected test Mean Squared Error (MSE), where a squared loss function is a measure of the distance between the test observation y^* and the estimated prediction $\hat{f}(\boldsymbol{x}^*)$. It can be shown that Equation (3.2) can be decomposed as

$$E(y^* - \hat{f}(\boldsymbol{x}^*))^2 = \mathrm{Var}(\hat{f}(\boldsymbol{x}^*)) + \Big[\mathrm{Bias}(\hat{f}(\boldsymbol{x}^*))\Big]^2 + \sigma^2,$$

where $\sigma^2 = \mathrm{Var}(\epsilon^*)$ is known as the irreducible error [54]. The first term describes how sensitive the function is to changes in the training data. A function with high variance will change significantly if it is estimated from a different training set, i.e., the function is overfitting the training data. On the contrary, if the function is not flexible enough to capture the general trend from the training data it will obtain a high bias. When choosing a predictive function, $\hat{f}$, it is therefore a trade-off between minimizing the variance and minimizing the bias term. A classical example of the bias-variance trade-off is to let $\hat{f}$ be a polynomial function of order q. Increasing the order will in general reduce the bias, however it will increase the variance term, $\mathrm{Var}(\hat{f}(\boldsymbol{x}^*))$, and potentially overfit the data. The goal is therefore to search for a function that obtains both low variance and bias.

3.2 Maximum likelihood estimation

Maximum Likelihood Estimation (MLE) is a common approach for deriving estimators, and it can be applied when searching for an estimate of f in Equation (3.1). Assume that we have a set of training data where the output vector $\boldsymbol{y} = y_1, \ldots, y_N$ contains independent samples that come from a parametric distribution with density $p(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the vector containing the total set of parameters. The frequentist approach assumes that the parameters in $\boldsymbol{\theta}$ are fixed, but unknown quantities that can be estimated from the data. MLE estimates the parameters by maximizing the likelihood function that is defined as

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\theta}), \tag{3.3}$$

where the last equality follows because of the independence of the data. A maximum likelihood estimate of the parameters in $\boldsymbol{\theta}$ are the values where $L(\boldsymbol{\theta}|\boldsymbol{y})$ reaches its maximum as a function of $\boldsymbol{\theta}$. Thus, MLE is a sensible estimator because it finds the estimate for which the observed sample is most likely.

In most cases, it is easier to find the maximum of the natural logarithm of $L(\boldsymbol{\theta}|\boldsymbol{y})$, denoted by $l(\boldsymbol{\theta}|\boldsymbol{y})$. The logarithm is a strictly increasing function, and therefore has the same solution to the maximization problem. The maximum of $l(\boldsymbol{\theta}|\boldsymbol{y})$ can be found by first calculate its derivative with respect to each parameter and solve $\frac{\partial}{\partial_{\theta_i}} l(\boldsymbol{\theta}|\boldsymbol{y}) = 0$.

3.3 Extended linear regression

Maybe the most well known supervised learning method is obtained by assuming that the function $f(\boldsymbol{x})$ in Equation (3.1) is linear:

$$f(\boldsymbol{x}) = \beta_0 + \sum_{p=1}^{P} \beta_p x_p,$$

where β_0 is a bias term and $\beta_1, \ldots \beta_P$ are the coefficients. The problem of estimating f reduces to finding estimates for the coefficients. This is known as linear regression. The simple structure of the linear regression is useful when modelling data where the response can be predicted by a linear function, however, when the relationship between the input variables and the output variable is nonlinear, more complex learning models may

be needed. However, many of the more advanced learning approaches can be interpreted and deduced from a generalization of the linear regression framework. This will be the case for the statistical learning models that are presented in Papers B and C, in this thesis.

We now consider more general functions, and let f be a nonlinear function that can be represented by a linear combination of a set of basis functions $\phi_m(\boldsymbol{x})$:

$$f(\boldsymbol{x}) = w_0 + \sum_{m=1}^{M} w_m \phi_m(\boldsymbol{x}), \tag{3.4}$$

where w_0 is a bias term and $w_1, \ldots, w_M$ are the weights. Thus, the function f is still linear with respect to the weight parameters, while the basis functions $\phi_m(\boldsymbol{x})$ are in general nonlinear. The basis functions are usually determined beforehand based on characteristics of the dataset, hence, the regression problem reduces to find the optimal values of the weights in order to estimate $f(\boldsymbol{x})$.

We will make the common assumption that the error term in Equation (3.1) is normally distributed with variance σ^2 and mean zero. Using the decomposition of $f(\boldsymbol{x})$ into basis functions from Equation (3.4) we can then rewrite Equation (3.1) as:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} = \mathcal{N}(0, \sigma^2 \boldsymbol{I}), \tag{3.5}$$

where $\boldsymbol{w} = (w_0, \ldots, w_M)^\top$ is the vector containing all weights, $\boldsymbol{I}$ is the $N \times N$ identity matrix and $\boldsymbol{\Phi}$ is the $N \times (M+1)$ design matrix with elements $\boldsymbol{\Phi} = [\boldsymbol{1}, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_M]$, where $\boldsymbol{\phi}_m = (\phi_m(\boldsymbol{x}_1), \ldots, \phi_m(\boldsymbol{x}_N))^\top$.

From the model specification given in Equation (3.5), we have that the density $p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)$ is Gaussian with mean $\boldsymbol{\Phi}\boldsymbol{w}$ and variance σ^2. When applying the MLE procedure from Section 3.2, we obtain the following estimate for $\boldsymbol{w}$:

$$\hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}. \tag{3.6}$$

Note that if one choose linear basis functions, the expression in Equation (3.6) reduces to the well known normal equations obtained in linear regression. From Equation (3.6), we notice that for the matrix $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ to have an inverse, it must be nonsingular. Thus, we must have more observations than the number of basis functions, $N > M$. Otherwise the matrix $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ will be singular because its size is $M+1 \times M+1$ and the matrix has at most rank N. Note also that we can apply MLE to find an estimate of σ^2.

3.3.1 Regularization

A common approach to obtain estimates for the weight parameters in Equation (3.5) when $N \leq M$ is to apply a regularization. This approach is still meaningful to use even when $N > M$ as a technique used to avoid overfitting the training data. In regularized regression, we search for the values of $\boldsymbol{w}$ that minimize

$$||\boldsymbol{y} - \boldsymbol{\Phi w}||^2 + \lambda P(\boldsymbol{w}), \tag{3.7}$$

where λ is a tuning parameter and $P(\boldsymbol{w})$ is a penalty term.

Two of the classical regularization methods are the ridge regression [55] and the lasso [56], which differs in the choice of penalty term. The estimate of the weights from ridge regression is obtained by using an l_2 penalty term $P(\boldsymbol{w}) = ||\boldsymbol{w}||_2^2$ in Equation (3.7), while the lasso method uses an l_1 penalty term $P(\boldsymbol{w}) = ||\boldsymbol{w}||_1$. An advantage of the ridge regression is that there exists an analytical expression for $\hat{\boldsymbol{w}}$ by using MLE. When $\lambda = 0$, the ridge estimate will be equal to the maximum likelihood estimate in Equation (3.6), while when λ increases it will shrink the estimates of the elements in $\boldsymbol{w}$ toward zero. However, none of the estimates will be set exactly to zero. The main advantage of lasso over ridge is that some of the estimates of $\boldsymbol{w}$ will be set exactly to zero for sufficiently large tuning parameter λ, hence the method can be used to obtain a sparse model. The cost of using the l_1 penalty term is that there does not exist a closed form expression for the weight parameters.

Another popular regularization method is the Elastic net [57], which can be seen as a combination of the ridge regression and the lasso. An advantage with these regularized methods is their ability to balance the trade-off between model complexity and validation. However, all these methods require tuning of the penalty parameter λ, which is often done by using cross-validation [58].

Chapter 4

Sparse Bayesian learning

This chapter gives an introduction to the framework of the sparse Bayesian learning models that are developed in Papers B and C. The models will mainly be presented assuming an extended linear regression framework, as described in Section 3.3.

4.1 Bayesian inference

Assume a general parametric model where $\boldsymbol{\theta}$ denotes the vector of all model parameters. Opposed to the traditional frequentist approach, discussed in Section 3.2, where the elements in $\boldsymbol{\theta}$ are assumed to be fixed but unknown quantities, the Bayesian approach considers the parameters to be random variables. The random variables are assigned a prior distribution, denoted by $p(\boldsymbol{\theta})$, that captures our prior beliefs about the parameters and can be determined before we have any data.

Using the prior and the likelihood of the observations given the parameters, $p(\boldsymbol{y}|\boldsymbol{\theta})$, a posterior distribution can be found from Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{4.1}$$

where the denominator, $p(\boldsymbol{y})$, is the marginalized probability of the data:

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}. \tag{4.2}$$

The posterior distribution in Equation (4.1) incorporates information from both the data and the prior. Thus, prior knowledge can be used to

improve the posterior. The posterior distribution can be used to obtain estimates of the parameters (using, e.g., the mean or the mode), and it is also one of the main components when obtaining new predictions.

The major challenge in the Bayesian framework is to solve the integral in Equation (4.2) as this integral is most often not possible to calculate analytically. There exist therefore a variety of different approaches to tackle this, e.g., by applying numerical solutions or different approximations.

4.1.1 Maximum a posteriori

Maximum a Posteriori (MAP) estimation is a common method to obtain point estimates in the Bayesian framework, and is closely related to the classical MLE. When applying a MAP estimation, we are searching for the parameter values that maximizes the posterior distribution, i.e., the mode of the posterior distribution. From Equation (4.1) we see that the posterior is proportional to the likelihood times the prior:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{4.3}$$

Thus, maximizing the posterior in Equation (4.1) is equivalent to maximizing Equation (4.3), as the denominator in Equation (4.1) is independent of the parameters in $\boldsymbol{\theta}$. This approach thereby avoids the problem of finding the marginal distribution $p(\boldsymbol{y})$. The cost of this shortcut is, however, that we do not obtain a complete distribution for the posterior, only the posterior mode.

When comparing Equation (4.3) with Equation (3.3), we notice that the MAP approach is identical to the MLE except for the inclusion of the prior. The effect of adding a prior is analogue to applying a regularization in the classical framework. The prior has a regularization effect, and estimates from both the ridge regression and the lasso, described in Section 3.3.1, can be interpreted in a Bayesian framework as MAP estimates. The lasso solution can be interpreted as a MAP estimate when the prior for the parameters is a Laplace distribution, while the solution from the ridge regression can be obtained when applying a Gaussian prior [56, 54].

4.2 Bayesian hierarchical models

There are two main reasons for creating hierarchical models. These types of models originate from data that has a layered structure. The second motivation is that they can be created in such a way that we obtain a sparse model. In this thesis, the main focus on the hierarchical models

has been the second point of view, and we discuss how to obtain sparse models in Section 4.5, but let us first explore the structure of the Bayesian hierarchical models.

Assume we have the model specifications from Section 3.3, where the parameters of interest are the weights in $\boldsymbol{w}$. Further, let $p(\boldsymbol{w}|\boldsymbol{\alpha})$ be the prior distribution, where $\boldsymbol{\alpha}$ contains the so called hyperparameters. This name indicates that they are parameters one level below $\boldsymbol{w}$. The hyperparameters in $\boldsymbol{\alpha}$ may again be given a prior distribution, $p(\boldsymbol{\alpha})$, which is referred to as a hyperprior. In addition, we also have to consider the noise parameter, σ^2, from Equation (3.5), which is given a prior distribution, $p(\sigma^2)$. The model can now be represented by a hierarchical representation:

$$\boldsymbol{y}|\boldsymbol{w}, \sigma^2 \sim p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) \tag{4.4}$$
$$\boldsymbol{w}|\boldsymbol{\alpha} \sim p(\boldsymbol{w}|\boldsymbol{\alpha}) \tag{4.5}$$
$$\boldsymbol{\alpha} \sim p(\boldsymbol{\alpha})$$
$$\sigma^2 \sim p(\sigma^2)$$

The equation above is one of the simpler hierarchical representations, and other models may have several layers (see, e.g., [44]). The BLS presented in Paper B adds an additional layer for the hyperprior $\boldsymbol{\alpha}$ where the hyperprior is given by $p(\boldsymbol{\alpha}|\lambda)$. The hyperparameter λ is estimated by using the empirical Bayes approach, which is described in Section 4.4. Sections 4.3 and 4.4 describe two different Bayesian procedures for hierarchical models.

4.3 Fully Bayesian approach

A fully Bayesian approach of the hierarchical model structure from Section 4.2 threats the complete set of parameters and hyperparameters as random variables, and assign prior distributions to all of them. The posterior distribution from Equation (4.1) for the hierarchical model in Section 4.2 is given by

$$p(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\sigma^2)}{p(\boldsymbol{y})}, \tag{4.6}$$

where the normalizing constant is calculated as

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\sigma^2)\,\mathrm{d}\boldsymbol{w}\,\mathrm{d}\boldsymbol{\alpha}\,\mathrm{d}\sigma^2. \tag{4.7}$$

However, even though it may be possible to integrate over either $\boldsymbol{w}$ or the hyperparameters, the complete marginalization over all of these variables in Equation (4.7) is not analytically tractable [59]. As mentioned in Section 4.1, it is only for a few simple problems where the posterior distribution in Equation (4.1) can be calculated analytically. A common approach is to use Markov Chain Monte Carlo (MCMC) methods to generate samples that come from the posterior distribution. Thus, MCMC methods avoid the problem of calculating analytical solutions. However, for certain hierarchical models, MCMC simulations are computationally inefficient and time consuming. This is especially an issue for higher-dimensional cases.

4.4 Empirical Bayes

An alternative to the fully Bayesian approach is the empirical Bayes. Within this framework one seek to find an approximation of the posterior distribution in Equation (4.1) by using the observed data to determine the hyperparameters of the prior distribution. There exist several different model specifications within this framework, but we continue with the hierarchical model structure given in Section 4.2 to introduce the approach.

The empirical Bayes approach uses the following approximation for the posterior distribution in Equation (4.6):

$$p(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2 | \boldsymbol{y}) \approx p(\boldsymbol{w} | \boldsymbol{y}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2), \tag{4.8}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ are point estimates for the elements in $\boldsymbol{\alpha}$ and σ^2. In order to evaluate the approximation in Equation (4.8), we need to find the posterior distribution for the weight parameters, $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2)$, and estimated values for $\boldsymbol{\alpha}$ and σ^2. Because of the dependence between $\boldsymbol{w}$, $\boldsymbol{\alpha}$ and σ^2, the right hand side of Equation (4.8) is usually calculated by an iterative process; see Paper C for further details.

The concept of conjugate priors is important when searching for an analytically tractable expression of $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2)$. If the prior distribution in Equation (4.5) is a conjugate to the likelihood in Equation (4.4), it implies that the posterior distribution for the weights, $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2)$, is in the same probability distribution family as the prior [60, p. 35]. The posterior distribution for the weights, $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2)$, can then be derived in closed form from Bayes' theorem:

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2)}. \tag{4.9}$$

The situations where the prior is a conjugate to the likelihood simplify the calculations not only for the posterior of the weights, but also later when estimating the hyperparameters. It is however not always possible to apply a conjugate prior. In a classification setting, the likelihood in Equation (4.4) is often a Bernoulli and a conjugate prior may not be an appropriate prior to use. Instead, the posterior for the weights can often be approximated by using the Laplace method, which is explained in Section 4.6. In the next two subsections we investigate two different variants within the empirical Bayes framework for obtaining the estimates of $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$, and we also study how these two variants relate to the true posterior distribution in Equation (4.6).

4.4.1 Type-II maximum likelihood

The type-II maximum likelihood procedure obtains estimates of $\boldsymbol{\alpha}$ and σ^2 from the marginal likelihood, which can be found from the numerator in Equation (4.9) by integrating over the weight parameters in $\boldsymbol{w}$:

$$p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2) = \int p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{w}. \tag{4.10}$$

An analytic expression for the marginal likelihood in Equation (4.10) is available when the prior distribution is a conjugate prior for the likelihood. Otherwise the Laplace approximation can often be used to approximate the marginal likelihood in Equation (4.10) by a Gaussian distribution (see Section 4.6). The estimates for $\boldsymbol{\alpha}$ and σ^2 are then calculated as

$$\hat{\boldsymbol{\alpha}}, \hat{\sigma}^2 = \underset{\boldsymbol{\alpha}, \sigma^2}{\arg\max} \; p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2).$$

To indicate why maximizing the marginal likelihood can obtain a reasonable approximation of the true posterior distribution, notice that the posterior distribution from Equation (4.6) can be decomposed as

$$p(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) = p(\boldsymbol{w}|\boldsymbol{\alpha}, \sigma^2, \boldsymbol{y}) p(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}).$$

Further, the posterior distribution for the hyperparameters is proportional to

$$p(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2) p(\boldsymbol{\alpha}) p(\sigma^2), \tag{4.11}$$

by using Bayes' theorem. If the hyperpriors $p(\boldsymbol{\alpha})$ and $p(\sigma^2)$ are relatively flat, they can be neglected such that the estimates of $\boldsymbol{\alpha}$ and σ^2, obtained

from maximizing the marginal likelihood function $p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2)$, will give the same solution as maximizing the right hand side of Equation (4.11).

The empirical Bayes procedure follows a similar approach as the ordinary MLE described in Section 3.2, except that the hyperparameters are found by maximizing the marginal likelihood. The RVM [26, 27] uses this framework, where very broad gamma distributions are assumed as hyperpriors for $\boldsymbol{\alpha}$ and σ^2. Since the parameters of the hyperprior are chosen such that the distribution is flat over a wide range of $\boldsymbol{\alpha}$, (and similarly for σ^2), the particular expression of such a function is irrelevant except for issues of computation convenience [59]. In general, estimates of the hyperparameters can be found directly from the marginal likelihood in Equation (4.10) when the hyperpriors are assumed to be flat or non-informative.

4.4.2 Type-II maximum a posteriori

When the chosen priors are not flat or non-informative then they should be accounted for in Equation (4.11). The BLS method presented in Paper B, in addition to the FLAP by Babacan et al. [45] and the PFCVM_{LP} method by Jiang et al. [29], all use an informative prior for the hyperparameters. When using informative priors for the hyperparameters, the estimates can be found by maximizing Equation (4.11) with respect to the hyperparameters, which is an empirical Bayes MAP estimate for the hyperparameters:

$$\hat{\boldsymbol{\alpha}}, \hat{\sigma}^2 = \underset{\boldsymbol{\alpha}, \sigma^2}{\arg\max}\ p(\boldsymbol{y}|\boldsymbol{\alpha}, \hat{\sigma}^2) p(\boldsymbol{\alpha}) p(\sigma^2).$$

For both the marginal likelihood estimates in Equation (4.10) and here, we need that the posterior distribution $p(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y})$ is sharply peaked around the values $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ in order to obtain a good approximation as we are relying on the posterior modes [59].

The empirical Bayes has been criticized for not including the uncertainty from the hyperparameters. However, in Paper C we discuss a possible method that extends the empirical Bayes such that the uncertainty of the hyperparameters can be included in the Laplace approximation.

4.4.3 Prediction

Having found estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ by maximizing the marginal likelihood in Equation (4.10) or the posterior in Equation (4.11) we can get the approximation of the posterior distribution for $\boldsymbol{w}$ in Equation (4.8). A predicted

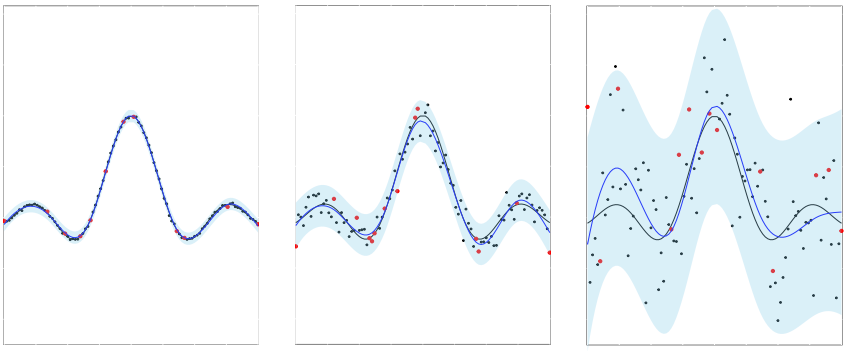

Figure 4.1: The Sinc function (black line) and its approximation by using the BLS method (blue line) from data generated for different values of σ. The red dots are the relevance vectors and the black dots are the remaining data. The blue shaded area corresponds to ± 2 predictive standard deviations.

distribution can be obtained by applying the following approximation

$$p(y^*|\boldsymbol{y}) \approx p(y^*|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2) = \int p(y^*|\boldsymbol{w}, \hat{\sigma}^2) p(\boldsymbol{w}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2)\, \mathrm{d}\boldsymbol{w}.$$

From the predictive distribution, we obtain both the mean and the variance for a new prediction, where the mean can be used as a point estimate for the new prediction.

Figure 4.1 shows the predictions obtained from the BLS method when the training data are generated from the sinc function, $f(x) = \sin(x)/x$, for different values of σ. The blue line shows the prediction by the BLS method, while the black line is the true sinc function. The training data are shown as the black dots, while the red dots represent the relevance vectors (see Section 4.5 for the definition of relevance vectors). The blue shaded area corresponds to ± 2 predictive standard deviations. The uncertainty of the predictions increases according to the increasing noise of the training data.

4.5 Sparse priors

The sparse Bayesian methods, described in Papers B and C, use a prior for the weights that includes individual hyperparameters for each weight parameter. This allows some of the weights to be estimated to zero, such

that a sparse model can be obtained. A common choice is to use a Gaussian prior for the weights with individual variance (or precision) parameters as hyperparameters:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} \mathcal{N}(0, \alpha_i), \tag{4.12}$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_N)^\top$. This form of prior that includes individual hyperparameters is known as an automatic relevance determination (ARD) prior and was first proposed by Mackay [61]. When α_i goes to zero, the probability of w_i to have a value close to zero becomes significantly high. Sparse models can therefore be obtained by setting some of the α_i's to zero (or equivalently to infinity if they are used as precision parameters, as for the RVM [26]). When the weights are set to zero, the corresponding basis functions will be pruned. The sparsity property of the BLS method, presented in Paper B, is illustrated in Figure 4.2. This Figure shows the approximation from the BLS method (the blue line) to the sinc function (the black line). The training data are shown as the black dots. The red dots are the training data that corresponds to the nonzero α_i's. Those data are called relevance vectors and it is only the relevance vectors that are used in the final model to obtain the approximated function in blue.

Priors on the form in Equation (4.12) allow sparsity by letting the w_i's have individual hyperparameters. However, sparsity requires that some of the α_i's are set to zero during estimation. Thus, in order to understand why sparse solutions arise, the estimation procedure of the α_i's must be analysed. This procedure will depend on the specific model structure, but for the RVM method Faul and Tipping [30] have proposed a detailed analysis based on the maximum of the log marginal likelihood. Faul and Tipping [30] show that the log marginal likelihood, as a function of a single α_i, has a maximum that can be obtained explicitly. The same approach is used for the BLS method in Paper B (see Section 2.3 in Paper B). Bishop [59] proposes another way of understanding sparsity for the original RVM method by Tipping [26], by its relation to a Gaussian process.

4.6 The Laplace approximation

The Laplace approximation of a general function $f(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^n$ is a Taylor expansion of $f(\boldsymbol{x})$ around its global maximum, where only the first three terms are used. Denoting the global maximum as $\boldsymbol{x}_0$, the expansion

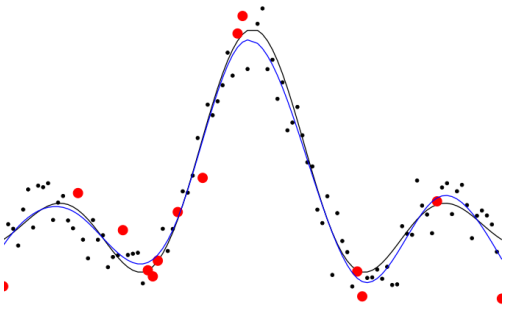

Figure 4.2: The BLS approximation is shown in blue, while the true sinc function is the black line. The training data are the black dots. The enlarged red dots are the relevance vectors.

of the logarithm of $f(\boldsymbol{x})$ is given by

$$\begin{aligned}\log f(\boldsymbol{x}) \approx \log f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^\top \frac{\partial \log f(\boldsymbol{x}_0)}{\partial \boldsymbol{x}} \\ + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top \frac{\partial^2 \log f(\boldsymbol{x}_0)}{\partial \boldsymbol{x}^2}(\boldsymbol{x}-\boldsymbol{x}_0).\end{aligned} \tag{4.13}$$

The second term of Equation (4.13) equals zero, since $\boldsymbol{x}_0$ is the global maximum of the function, and therefore also of $\log f$. Further, since the second order partial derivatives of $\log f$ evaluated at $\boldsymbol{x}_0$ are negative, Equation (4.13) simplifies to

$$\log f(\boldsymbol{x}) \approx \log f(\boldsymbol{x}_0) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top \left| \frac{\partial^2}{\partial \boldsymbol{x}^2} \log f(\boldsymbol{x}_0) \right| (\boldsymbol{x}-\boldsymbol{x}_0)^2.$$

This expression can be rewritten as the logarithm of a normal density by letting $\hat{\boldsymbol{\Sigma}}^{-1}$ denote the second order partial derivatives of $\log f$ evaluated at $\boldsymbol{x}_0$:

$$\log f(\boldsymbol{x}) \approx \log f(\boldsymbol{x}_0) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\boldsymbol{x}_0).$$

The function $f(\boldsymbol{x})$ can therefore be approximated by a Gaussian where the mean and the variance can be obtained by finding the maximum of $\log f(\boldsymbol{x})$ and calculating the second order partial derivatives at this point.

As mentioned briefly in Section 4.4, the Laplace approximation can be applied when there is no direct calculation of the posterior distribution for the weights in Equation (4.9) or the marginal likelihood in Equation (4.10). Let us see how this applies to the regression example from Section 4.4. The Laplace approximation of the posterior distribution, $p(\boldsymbol{w}|\boldsymbol{y}, \sigma^2)$, in Equation (4.9) is obtained from the joint likelihood of the weights and the data since $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2) \propto p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})$. Hence, maximizing $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2)$ is equivalent to maximizing $p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})$. The mean and the covariance matrix of the Laplace approximation of Equation (4.9) is therefore given by

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \log(p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})), \tag{4.14}$$

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{\partial^2}{\partial \boldsymbol{w}^2} \log(p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})) \right)^{-1}. \tag{4.15}$$

The marginal likelihood can be calculated from the joint likelihood of the weights and the data, as seen in the example from Equation (4.10). Given the Laplace approximation of $p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})$, the approximation of the marginal likelihood is obtained by integrating out the weights

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2) &\approx \int p(\boldsymbol{y}|\hat{\boldsymbol{w}}, \sigma^2)p(\hat{\boldsymbol{w}}|\boldsymbol{\alpha}) \exp\left\{ -\frac{1}{2}[\boldsymbol{w} - \hat{\boldsymbol{w}}]^\top \hat{\boldsymbol{\Sigma}}^{-1} [\boldsymbol{w} - \hat{\boldsymbol{w}}] \right\} \mathrm{d}\boldsymbol{w} \\ &= p(\boldsymbol{y}|\hat{\boldsymbol{w}}, \sigma^2)p(\hat{\boldsymbol{w}}|\boldsymbol{\alpha}) \sqrt{2\pi}^N \det(\hat{\boldsymbol{\Sigma}}). \end{aligned}$$

Of course, finding the maximum of the right hand side of Equation (4.14) and the second order partial derivatives of Equation (4.15) may be far from straight forward. In some cases, such as for the RVM method, there exists analytical solutions. However, even if an analytical expression may exist for a method, it may be cumbersome to derive. In Paper C we show how to use the TMB package to solve these equations.

4.6.1 The Template Model Builder

The open source `R` package TMB by Kristensen et al. [43] is applied in Paper A, as it is a flexible tool for estimating the parameters of the sibling distribution. The focus of this section, however, is to introduce the implementation of sparse Bayesian models using TMB, which is the main topic of Paper C.

TMB provides an easy setup where the user only needs to specify a joint likelihood of latent variables and the data as a `C++` function. The

TMB package can then be used to calculate the Laplace approximation of the marginal likelihood of the specified joint likelihood. The latent variables are automatically integrated out using Automatic Differentiation (AD) [62]. AD is a set of techniques that is used to evaluate the derivatives of functions automatically to machine precision in a very efficient manner. Thus, we do not have to manually calculate any derivatives of the joint likelihood, nor implement and maintain the code. The optimization of the Laplace approximation in order to estimate hyperparameters can then easily be done by, e.g., Newton's method, in `R` because the TMB package both evaluates the marginal likelihood and the gradient. In Paper C we implement and test several different sparse Bayesian methods by using TMB.

Chapter 5

The BLS for classification

This section presents an extension of the BLS method, presented in Paper B, that adapts the BLS to a two category classification setting. The sparse Bayesian methods that are presented in Paper B is the RVM [27] and the FLAP [45], in addition to the proposed BLS. Among these, it is only the RVM that is developed to solve classification problems. Both the theory and the numerical results of the BLS consider regression problems. However, as we show in this chapter, the BLS can be extended to solve classification problems by following the approach of Tipping and Faul [27]. In a classification setting the noise parameter σ^2 (from the regression setting) is omitted. The prior for the weights in the BLS reduces therefore to the FLAP prior. The theory for the BLS in a classification setting will be presented her, and a small example of a classification dataset that compares the BLS with the RVM.

The BLS for two category classification problems follows a similar framework as the regression setting, described in Paper B, but the likelihood function from the regression setting, $p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)$, must be transformed. The probabilistic outputs are continuous values in $[0, 1]$, and the sigmoid function can be used to obtain mappings from $[-\infty, \infty]$ to $[0, 1]$. When combining this transformation with a Bernoulli distribution we obtain the likelihood function:

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{N} \Big\{ S\big((\boldsymbol{\Phi w})_i\big) \Big\}^{y_i} \Big\{ 1 - S\big((\boldsymbol{\Phi w})_i\big) \Big\}^{1-y_i}, \tag{5.1}$$

where the targets y_i take values 0 or 1, $(\boldsymbol{\Phi w})_i$ is the i'th element of the vector $\boldsymbol{\Phi w}$ from Equation (3.5), and the sigmoid function is given

by $S(z) = 1/(1 + e^{-z})$.

The prior and the hyperprior for the BLS in a classification setting are given by:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} \mathcal{N}(0, \alpha_i), \qquad \alpha_i \geq 0, \tag{5.2}$$

$$p(\boldsymbol{\alpha}|\lambda) = \prod_{i=0}^{N} \frac{\lambda}{2} \mathrm{e}^{-\frac{\lambda\alpha_i}{2}}, \qquad \lambda \geq 0,$$

$$p(\lambda) = \frac{b^a}{\Gamma(a)}(\lambda)^{a-1} e^{-b\lambda}, \qquad a, b \geq 0, \tag{5.3}$$

where each α_i is an individual variance parameter for the corresponding w_i. Note that a non-informative prior is used for λ in the numerical results from Paper B by setting the parameters $a = b = 0$ in Equation (5.3).

It is however not possible to calculate analytical expressions for the posterior of the weights, $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha})$, and the marginal likelihood, $p(\boldsymbol{y}|\boldsymbol{\alpha})$, in a classification setting due to the likelihood in Equation (5.1). To overcome this problem, we can use the Laplace approximation, described in Section 4.6, to approximate the posterior, $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha})$, by a Gaussian distribution. The Laplace approximation of the posterior is found from the joint likelihood of the targets and the weights by using that $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}) \propto p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})$. The logarithm of the joint likelihood can be found from Equation (5.1) and (5.2), and is given by

$$\begin{aligned}\log[p(\mathbf{y}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})] = &\sum_{i=1}^{N} \Big\{ y_i \log S\big((\boldsymbol{\Phi}\boldsymbol{w})_i\big) + (1 - y_i) \log \Big[1 - S\big((\boldsymbol{\Phi}\boldsymbol{w})_i\big)\Big] \Big\} \\ &- \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Lambda}^{-1}\mathbf{w},\end{aligned} \tag{5.4}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with elements $\{\alpha_i\}_{i=0,\ldots,N}$. For fixed values of $\boldsymbol{\alpha}$, Equation (5.4) can be maximized with respect to $\boldsymbol{w}$ to search for

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \log[p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})]. \tag{5.5}$$

The estimate in Equation (5.5) can be found by using iteratively reweighted least squares (IRLS) [See, Ch 4.3.3] [59]. Thus both the gradient vector

and the Hessian matrix of Equation (5.4) are required:

$$\frac{\partial}{\partial \boldsymbol{w}} \log[p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})] = \boldsymbol{\Phi}^\top(\boldsymbol{y} - \boldsymbol{\psi}) - \boldsymbol{\Lambda}^{-1}\boldsymbol{w}, \tag{5.6}$$
$$\frac{\partial^2}{\partial \boldsymbol{w}^2} \log[p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})] = -(\boldsymbol{\Phi}^\top\boldsymbol{\Psi}\boldsymbol{\Phi} + \boldsymbol{\Lambda}^{-1}),$$

where $\boldsymbol{\psi} = \big[S\big((\boldsymbol{\Phi}\boldsymbol{w})_1\big), \ldots, S\big((\boldsymbol{\Phi}\boldsymbol{w})_N\big)\big]^\top$ and the matrix $\boldsymbol{\Psi}$ is diagonal with elements $\Psi_i = S\big((\boldsymbol{\Phi}\boldsymbol{w})_i\big)\big[1 - S\big((\boldsymbol{\Phi}\boldsymbol{w})_i\big)\big]$. At convergence of the IRLS algorithm, the negative Hessian represents the inverse covariance matrix for the Gaussian approximation of the posterior distribution. The mode of the Laplace approximation, which represents the mean of the Gaussian approximation, is obtained by setting Equation (5.6) to zero. The mean and the covariance of the approximated posterior distribution are therefore given by

$$\hat{\boldsymbol{w}} = \Lambda\boldsymbol{\Phi}^\top(\boldsymbol{y} - \boldsymbol{\psi}), \tag{5.7}$$
$$\hat{\boldsymbol{\Sigma}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Psi}\boldsymbol{\Phi} + \boldsymbol{\Lambda}^{-1})^{-1}. \tag{5.8}$$

The estimates of the hyperparameters are obtained by applying the empirical Bayes approach from Section 4.4. The posterior of the hyperparameters is given by

$$p(\boldsymbol{\alpha}, \lambda|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\lambda)p(\lambda). \tag{5.9}$$

The marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\alpha})$ can however not be computed analytically in the classification setting. The Laplace method is therefore used to approximate the marginal likelihood:

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{w} \approx p(\boldsymbol{y}|\hat{\boldsymbol{w}})p(\hat{\boldsymbol{w}}|\boldsymbol{\alpha})(2\pi)^{N/2}\det(\hat{\boldsymbol{\Sigma}})^{1/2},$$

where the estimates $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{\Sigma}}$ are given in Equation (5.7) and (5.8). The estimates of $\boldsymbol{\alpha}$ and λ can now be obtained from maximizing Equation (5.9), and are used in the same iterative procedure as for the regression setting described in Paper B.

5.1 Ripley's synthetic data

The following example considers a binary classification problem, where the data are from Ripley's synthetic data [63]. Each class is generated by

using a mixture of two Gaussian distributions. The data set consists of 125 training observations for each class. The performance of a model can then be evaluated based on a test dataset of size 1000. The class distributions were designed to allow a best possible error rate of about 8% [63]. The training error rate is given by

$$\frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i), \tag{5.10}$$

where $\hat{y}_i$ is the predicted class label for the ith observation and N is the size of the training data. The test error rate is found similarly by calculating Equation (5.10) when using a test dataset as input and average over the K test observations. The aim is to obtain the lowest possible test error rate.

We compare the fast RVM by Tipping and Faul [27] with the BLS from Paper B, where the BLS has been adjusted for classification as described in the above section.

The result for the 250 Ripley's training data is shown in the top row of Figure 5.1. The corresponding training error rate is 10.4% for the RVM model and 11.6% for the BLS model. The number of relevance vectors is 7 for the RVM and 13 for the BLS. The model obtained from the RVM method is therefore the most sparse model on this dataset. The result for the associated 1000 example test set is shown in the bottom row of Figure 5.1. The test error rate is 9.6% for the RVM model and 9.0% for the BLS model. The BLS model provides the best prediction result, but is at the same time not as sparse as the RVM. The overall performance of these two methods on this dataset is very similar.

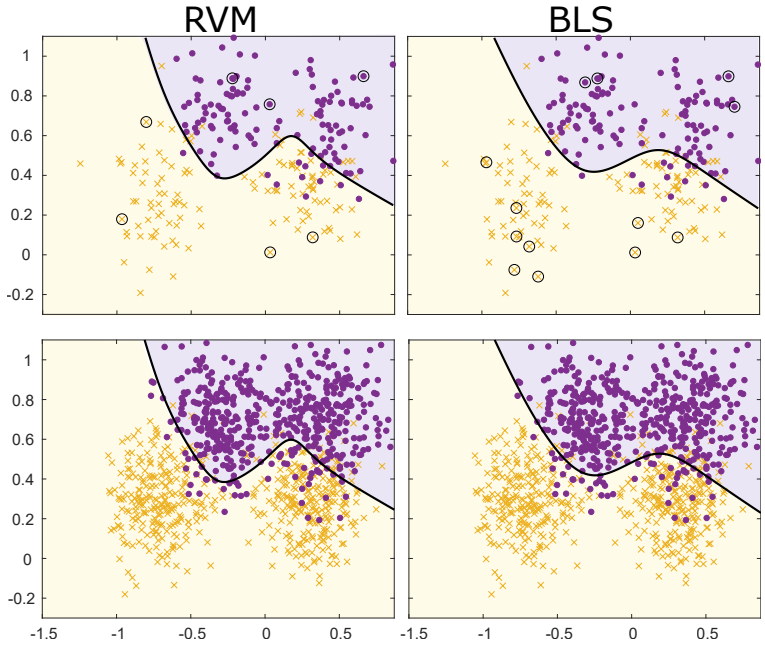

Figure 5.1: Ripley's Gaussian mixture data set. The top row shows the result on the 250 training data where the relevance vectors are shown circled. The bottom row shows the classifiers on the 1000 example test set.

Chapter 6

Summary of papers

Paper A

Title:	The sibling distribution for multivariate life time data
Authors:	Ingvild M. Helgøy, Hans J. Skaug
Journal:	Sankhya B
DOI:	10.1007/s13571-021-00259-w

In Paper A, we introduce a multivariate sibling distribution for continuous lifetimes. The distribution is defined in terms of the age-at-death of m siblings. The framework from the stable population theory is used to find an expression for the joint distribution. The sibling distribution is constructed by defining a reference time point where we set the time $t = 0$ and the assumption that the mother was alive at $t = 0$. There is no information about the mother, except that she had m offspring and was alive at this reference time point. The latent variables in this distribution are the birth and death time-points of the mother. The positive dependence between the lifetimes of the siblings is due to the shared dependence of their mother's life span, in addition to conditioning on their death time points. The death time points are the parameters of this distribution, in addition to the age dependent birth rate, $\beta(a)$, and death rate, $\phi(a)$. The sibling distribution is defined for general functions $\beta(a)$ and $\phi(a)$, but the explicit expression for the bivariate density is derived by assuming constant birth and death rates, i.e., $\beta(a) = \beta$ and $\phi(a) = \phi$. We prove that the bivariate constant-rate siblings distribution is MPT_2, which is a strong dependence property [64]. The constant rate sibling distribution is also related to the

Block-Basu distribution [42] which gives additional insight to the sibling distribution (see Section 4 of Paper A).

Paper B

Title: A Bayesian Lasso Based Sparse Learning Model
Authors: Ingvild M. Helgøy, Yushu Li
Journal: Submitted

Paper B considers sparse Bayesian learning methods. These methods have shown great success in several different fields and the recent years, particularity within the field of compressive sensing. We refer the reader to the introduction of Paper B and Paper C for an overview. The sparse Bayesian methods produce models that are sparse and as a consequence robust to noisy training data and therefore generalize well when applied to new unseen data. The main contribution in Paper B is the development of a new sparse Bayesian learning model, called BLS. The new model applies the hierarchical structure from the Bayesian lasso [44] in a kernel based scheme, such that general nonlinear regression problems can be solved. Furthermore, the Bayesian lasso is not sparse, while the BLS can provide sparse solutions based on a fast learning algorithm. The framework from the empirical Bayes (see Section 4.4) is applied in order to obtain a sparse model. BLS is tested on both simulated and real data, and the results show that the BLS model is comparable with existing models. In addition, the results show that the BLS method works particular well for irregular datasets that have a high degree of noise.

Paper C

Title: Sparse Bayesian Learning using TMB (Template Model Builder)
Authors: Ingvild M. Helgøy, Hans J. Skaug, Yushu Li
Journal: Submitted

One of the disadvantages of the sparse Bayesian models presented in Paper B, is that sparsity is only obtained with respect to the weight parameters when applied to general nonlinear regression problems. Because many dataset may include features that are not significant, a favourable model should also include sparsity with respect to features, i.e., it should perform feature selection (or dimension reduction). Sparse Bayesian models that also perform feature selection are studied in Paper C, where they are implemented using the TMB package [43]. The implementation using TMB,

simplifies the process of estimating the parameters since the algorithm can be constructed by only defining the likelihood and the prior density. The presented framework, makes it easy to adjust and modify the models. We use this flexible framework to extend some of the existing sparse Bayesian learning models to also include dimension reduction. In Paper B we only considered sparse Bayesian methods for solving regression problems. In Paper C we consider both regression and classification problems.

Bibliography

[1] A. J. Lotka. "Relation between birth rates and death rates". In: *Science* 26 (1907), pp. 21–22.

[2] A. J. Lotka. "The stability of the normal age distribution". In: *Proceedings of the National Academy of Sciences* 8 (1922), pp. 339–345.

[3] J. Lotka Alfred and F. Sharpe. "A problem in age-distribution". In: *Philosophical Magazine, Series* 21.6 (1911), pp. 435–438.

[4] H. Inaba. *Age-Structured Population Dynamics in Demography and Epidemiology*. Springer, 2017.

[5] L. Lefkovitch. "The study of population growth in organisms grouped by stages". In: *Biometrics* 21.1 (1965), pp. 1–18.

[6] M. E. Gurtin and R. C. MacCamy. "Non-linear age-dependent population dynamics". In: *Archive for Rational Mechanics and Analysis* 54 (1974), pp. 281–300.

[7] P. A. Werner and H. Caswell. "Population growth rates and age versus stage-distribution models for teasel (Dipsacus sylvestris Huds.)" In: *Ecology* 58.5 (1977), pp. 1103–1111.

[8] H. Caswell. "A general formula for the sensitivity of population growth rate to changes in life history parameters". In: *Theoretical Population Biology* 14.2 (1978), pp. 215–230.

[9] P. Carr. "Introduction to Multiregional Mathematical Demography". In: *Journal of the Operational Research Society* 27.2 (1976), pp. 405–406.

[10] D. Ediev. "On the existence and uniqueness of the remaining life expectancy in the model of a stable population". In: *Mathematical Models and Computer Simulations* 13.6 (2021), pp. 964–970.

[11] J. W. Vaupel and F. Villavicencio. "Life lived and left: Estimating age-specific survival in stable populations with unknown ages". In: *Demographic Research* 39 (2018), pp. 991–1008.

[12] X. Song and R. D. Mare. "Shared lifetimes, multigenerational exposure, and educational mobility". In: *Demography* 56.3 (2019), pp. 891–916.

[13] H. Caswell, C. de Vries, N. Hartemink, G. Roth, and S. F. van Daalen. "Age× Stage-classified Demographic Analysis: A Comprehensive Approach". In: *Ecological Monographs* 88.4 (2018), pp. 560–584.

[14] A. S. Srinivasa Rao and J. R. Carey. "Generalization of Carey's equality and a theorem on stationary population". In: *Journal of mathematical biology* 71.3 (2015), pp. 583–594.

[15] T. Kallonen, H. J. Brodrick, S. R. Harris, J. Corander, N. M. Brown, V. Martin, S. J. Peacock, and J. Parkhill. "Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131". In: *Genome Research* 27.8 (2017), pp. 1437–1449.

[16] S. M. Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[17] A. I. Dale. *A history of inverse probability: From Thomas Bayes to Karl Pearson*. Springer Science & Business Media, 2012.

[18] M. Stephens and D. J. Balding. "Bayesian statistical methods for genetic association studies". In: *Nature Reviews Genetics* 10.10 (2009), pp. 681–690.

[19] D. J. Schaid, W. Chen, and N. B. Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504.

[20] J. Marchini and B. Howie. "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7 (2010), pp. 499–511.

[21] T. Auld, A. W. Moore, and S. F. Gull. "Bayesian neural networks for internet traffic classification". In: *IEEE Transactions on Neural Networks* 18.1 (2007), pp. 223–239.

[22] J. M. Hernández-Lobato and R. Adams. "Probabilistic backpropagation for scalable learning of bayesian neural networks". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015, pp. 1861–1869.

[23] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. "Hands-on Bayesian neural networks—A tutorial for deep learning users". In: *IEEE Computational Intelligence Magazine* 17.2 (2022), pp. 29–48.

[24] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[25] D. P. Wipf and B. D. Rao. "Sparse Bayesian learning for basis selection". In: *IEEE Transactions on Signal processing* 52.8 (2004), pp. 2153–2164.

[26] M. E. Tipping. "Sparse Bayesian learning and the Relevance Vector Machine". In: *Journal of machine learning research* 1 (2001), pp. 211–244.

[27] M. E. Tipping and A. C. Faul. "Fast Marginal Likelihood Maximisation for Sparse Bayesian Models". In: *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*. 2003, pp. 276–283.

[28] Y. Mohsenzadeh, H. Sheikhzadeh, and S. Nazari. "Incremental Relevance Sample-feature Machine: A fast marginal likelihood maximization approach for joint feature selection and classification". In: *Pattern Recognition* 60 (2016), pp. 835–848.

[29] B. Jiang, C. Li, M. D. Rijke, X. Yao, and H. Chen. "Probabilistic Feature selection and Classification Vector Machine". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.2 (2019), pp. 1–27.

[30] A. C. Faul and M. E. Tipping. "Analysis of Sparse Bayesian Learning". In: *NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. 2002, pp. 383–389.

[31] B. Demir and S. Erturk. "Hyperspectral Image Classification using Relevance Vector Machines". In: *IEEE Geoscience and Remote Sensing Letters* 4.4 (2007), pp. 586–590.

[32] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki. "Sparse Bayesian classification of EEG for brain–computer interface". In: *IEEE Transactions on Neural Networks and Learning Systems* 27.11 (2015), pp. 2256–2267.

[33] S. Liu, J. Jia, Y. D. Zhang, and Y. Yang. "Image reconstruction in electrical impedance tomography based on structure-aware sparse Bayesian learning". In: *IEEE Transactions on Medical Imaging* 37.9 (2018), pp. 2090–2102.

[34] D. Li, M. Han, and J. Wang. "Chaotic time series prediction based on a novel robust echo state network". In: *IEEE Transactions on Neural Networks and Learning Systems* 23.5 (2012), pp. 787–799.

[35] R. K. Agrawal, F. Muchahary, and M. M. Tripathi. "Ensemble of relevance vector machines and boosted trees for electricity price forecasting". In: *Applied Energy* 250 (2019), pp. 540–548.

[36] D. L. Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[37] S. Foucart, H. Rauhut, S. Foucart, and H. Rauhut. *An invitation to compressive sensing*. Springer, 2013.

[38] S. Liu, Y. D. Zhang, T. Shan, and R. Tao. "Structure-aware Bayesian compressive sensing for frequency-hopping spectrum estimation with missing observations". In: *IEEE Transactions on Signal Processing* 66.8 (2018), pp. 2153–2166.

[39] M. Shekaramiz, T. K. Moon, and J. H. Gunther. "Bayesian compressive sensing of sparse signals with unknown clustering patterns". In: *Entropy* 21.3 (2019), p. 247.

[40] D. Calvetti, E. Somersalo, and A. Strang. "Hierachical Bayesian models and sparsity: l2-magic". In: *Inverse Problems* 35.3 (2019), p. 035003.

[41] H. Djelouat, M. Leinonen, and M. Juntti. "Spatial correlation aware compressed sensing for user activity detection and channel estimation in massive MTC". In: *IEEE Transactions on Wireless Communications* 21.8 (2022), pp. 6402–6416.

[42] H. W. Block and A. Basu. "A continuous, bivariate exponential extension". In: *Journal of the American Statistical Association* 69.348 (1974), pp. 1031–1037.

[43] K. Kristensen, A. Nielsen, C. Berg, H. Skaug, and B. Bell. "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software, Articles* 70.5 (2016), pp. 1–21. DOI: 10.18637/jss.v070.i05.

[44] T. Park and G. Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686.

[45] S. D. Babacan, R. Molina, and A. K. Katsaggelos. "Bayesian compressive sensing using Laplace priors". In: *IEEE Transactions on Image Processing* 19.1 (2010), pp. 53–63.

[46] N. Keyfitz and H. Caswell. *Applied Mathematical Demography*. Vol. 47. Springer, 2005.

[47] T. R. Malthus. "An Essay on the Principle of Population as It Affects the Future Improvements of Society". In: *J. Johnson, London* (1798).

[48] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.

[49] H. Caswell. *Matrix population models*. Vol. 1. Sinauer Sunderland, 2000.

[50] R. F. Gunst and R. L. Mason. *Regression analysis and its application: A data-oriented approach*. CRC Press, 2018.

[51] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, et al. "Supervised machine learning: A review of classification techniques". In: *Emerging artificial intelligence applications in computer engineering* 160.1 (2007), pp. 3–24.

[52] H. Abdi and L. J. Williams. "Principal Component Analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[53] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. "A review of robust clustering methods". In: *Advances in Data Analysis and Classification* 4.2 (2010), pp. 89–109.

[54] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Vol. 2. Springer, 2009.

[55] A. E. Hoerl and R. W. Kennard. "Ridge regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[56] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[57] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical methodology)* 67.2 (2005), pp. 301–320.

[58] M. W. Browne. "Cross-validation methods". In: *Journal of Mathematical Psychology* 44.1 (2000), pp. 108–132.

[59] C. M. Bishop. *Pattern Recognition and Machine Learning*. springer, 2006.

[60] A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Vol. 3. Chapman and Hall/CRC, 2013.

[61] D. Mackay. "A new method for estimating dissociation constants of competitive and non-competitive antagonists with no prior knowledge of agonist concentrations." In: *British journal of pharmacology* 111.1 (1994), p. 219.

[62] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics (SIAM), 2008.

[63] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 2007.

[64] S. Karlin and Y. Rinott. "Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions". In: *Journal of Multivariate Analysis* 10.4 (1980), pp. 467–498.

Part II

Scientific results

A

Paper A

The sibling distribution for multivariate life time data

Ingvild M. Helgøy, Hans J. Skaug

Published in Sankhya B
DOI: 10.1007/s13571-021-00259-w

Sankhyā : The Indian Journal of Statistics
2022, Volume 84-B, Part 1, pp. 340–363

The Sibling Distribution for Multivariate Life Time Data

Ingvild M. Helgøy and Hans J. Skaug
University of Bergen, Bergen, Norway

Abstract

A flexible class of multivariate distributions for continuous lifetimes is proposed. The distribution is defined in terms of the age-at-death of m siblings. The expression for the joint density is derived using classical results from mathematical demography. The parameters of the distribution are the age-specific birth and death rates, in addition to a vector of relative death times for the m siblings. For the case of constant birth and death rates we are able to derive an explicit expression for the bivariate sibling density, which is proven to be MTP_2, and hence has positive dependence. Further, we show that a special case of the sibling distribution belongs to the Block-Basu class of multivariate distribution. In the general case, with age-dependent birth and death rates, evaluation of the density involves numerical integration, but is still feasible.

AMS (2000) *subject classification.* Primary 62N99; Secondary 60E05.
Keywords and phrases. Copula, Frailty, Life time distribution, Mathematical demography, Close-Kin Mark-Recapture.

1 Introduction

Classes of multivariate densities for multivariate life time and survival data are well studied in the statistical and demographic literature (Hougaard, 2001; Barreto-Souza and Mayrink, 2019). A common approach for making survival times positively dependent goes via "shared frailties" (Hougaard, 2001, Chpt. 7). A frailty is a latent random variable that proportionally scales the hazard rate in a group of individuals, hence inducing dependence between otherwise independent lifetimes. In the present paper we introduce a new class of latent variable models, named the "sibling distribution", which is defined in terms of the age-at-death for each of m half siblings. There is no information available about their common mother, except that she had m offspring in total, and was alive at a specified point in time, taken to be $t = 0$ for convenience. The two latent variables of the model are the mother's birth and death times. The building blocks of the sibling distribution are

the individual birth rates $\beta(a)$ and death rates $\phi(a)$, where a is the age of an individual. We define the distribution for general functions, $\beta(a)$ and $\phi(a)$, but for most part we shall assume that $\beta(a)$ and $\phi(a)$ are constants as functions of a.

The positive dependence between the sibling's life times comes from conditioning on their (absolute) times of death, in combination with a shared dependence on the mother's life span. The times-of-death become parameters of the distribution. This somewhat implicit construction will be seen to be a mixture distribution, and can be studied using general theory for multivariate dependence (Shaked and Spizzichino, 1998; Khaledi and Kochar, 2001). We prove that the bivariate constant-rate sibling distribution is multivariate totally positive of order 2 (Karlin and Rinott, 1980), which for instance imply that the correlation is positive.

The constant-rate sibling distribution turns out to have marginal distributions that are perturbated exponential distributions. The exponential distribution plays a special role for univariate life times and several multivariate extensions can be found in the literature (Marshall and Olkin, 1967; Freund, 1961; Block and Basu, 1974; Arnold and Strauss, 1988; Gumbel, 1960; Hougaard, 1986; Sarkar, 1987). One of these extensions was introduced by Block and Basu (Block and Basu, 1974). The Block-Basu bivariate lifetime distribution can be derived by omitting the singular part of a bivariate exponential distribution as outlined by Marshall and Olkin (Marshall and Olkin, 1967), but can also be viewed as a reparametrization of Freund's distribution (Freund, 1961). We will see that the constant-rate (birth and death) sibling distribution reduces to a Block-Basu distribution, which will be used to shed light on the sibling distribution.

An alternative route to construction of multivariate life time distributions goes via copulae (Andersen, 2005). The implication also goes in the other direction; our sibling distribution induces a novel symmetric two-parameter copula.

The remaining part of the paper is organized as follows. In Section 2 we introduce the general sibling distribution. Explicit expressions in the bivariate case, along with positive dependence property, are derived in Section 3. In Section 4 we discuss the relationship to the Block-Basu distribution and in Section 5 we address simulation and parameter estimation. Finally, we provide a discussion in Section 6.

2 The Sibling Age Distribution

Consider a female who over her lifespan is known to have had m offspring. Denote by t_j and x_j the time of death and age-at-death, respectively, of the

j'th offspring. The offspring are arbitrarily ordered, not according to the time of birth. We shall view t and x as random variables taking values on the real line. For notational simplicity we let $j = 0$ refer to the mother, and we condition on the fact that the mother was alive at time $t = 0$, i.e. on the event that $t_0 - x_0 \leq 0 \leq t_0$. This assumption should be kept in mind at all times when reading this paper. We define $a_0 = x_0 - t_0$ as the age of the mother at $t = 0$, and we denote the joint density of (a_0, t_0) by $g(a_0, t_0)$. Let $y_j = t_j - x_j$, be the birth time of the *j'th* offspring. Please refer to Fig. 1 for an illustration of key quantities.

We denote random variables by capital letters. Conditionally on $(A_0, T_0) = (a_0, t_0)$, and hence on $X_0 = x_0 = a_0 + t_0$, the density of Y_j is

$$f_Y(y_j|a_0, t_0) = \frac{\beta(y_j - y_0)}{\int_0^{x_0} \beta(a)da}, \qquad y_j \in (y_0, t_0), \tag{2.1}$$

where $\beta(a)$ is the age-specific rate at which the mother produces offspring. We shorten our notation for conditional densities, e.g. we write f_Y rather than the full $f_{Y|A_0,T_0}$. The joint conditional density of (X_j, T_j) is

$$f_{X,T}(x_j, t_j|a_0, t_0) = f_Y(y_j|a_0, t_0)f_X(x_j) = f_Y(t_j - x_j|a_0, t_0)f_X(x_j), \tag{2.2}$$

on the support

$$\{(x_j, t_j) : (t_j - t_0)_+ \leq x_j \leq t_j - y_0\}, \tag{2.3}$$

where $z_+ = \max(z, 0)$. The constraints on x_j express the fact that $x_j \geq 0$ and $y_0 \leq y_j \leq t_0$, i.e. the offspring must be born in the time window when the mother is alive (see Fig. 1). The latter is related to Eq. 2.3 via the algebraic equivalence $y_0 \leq y_j \leq t_0 \Leftrightarrow t_j - t_0 \leq t_j - y_j \leq t_j - y_0$. Further, the marginal density $f_X(x_j)$ in Eq. 2.2 is defined in terms of the survival function $l(x_j) = \Pr(X > x_j)$ via $f_X(x_j) = -l'(x_j)$. Finally, the marginal density of T_j is obtained from Eq. 2.2 and Eq. 2.3,

$$f_T(t_j|a_0, t_0) = \int_{(t_j - t_0)_+}^{t_j - y_0} f_Y(t_j - x|a_0, t_0)f_X(x)dx, \qquad t_j \geq y_0. \tag{2.4}$$

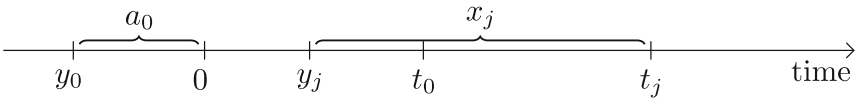

Figure 1: Birth (y_0) and death (t_0) times of mother, and corresponding times (y_j and t_j) for the j'th offspring. Further, a_0 is the age of the mother at the reference point $t = 0$

The births and deaths of the m siblings are assumed to be conditionally independent, given $(A_0, T_0) = (a_0, t_0)$, so the joint density of $X_{1:m} = (X_1, \ldots, X_m)$ and $T_{1:m} = (T_1, \ldots, T_m)$ is $\prod_{j=1}^m f_{X,T}(x_j, t_j|a_0, t_0)$, where the density $f_{X,T}(x_j, t_j|a_0, t_0)$ is given by Eq. 2.2. We are now in position to define the sibling distribution as the conditional distribution of $X_{1:m}$, given $T_{1:m} = t_{1:m}$.

REMARK 1. The sibling distribution of $X_{1:m}$ has density

$$f(x_{1:m}|t_{1:m}) = \frac{\int_0^\infty \int_0^\infty \prod_{j=1}^m f_{X,T}(x_j, t_j|a_0, t_0) g(a_0, t_0) da_0 dt_0}{\int_0^\infty \int_0^\infty \prod_{j=1}^m f_T(t_j|a_0, t_0) g(a_0, t_0) da_0 dt_0}, \quad x_{1:m} \in R_+^m, \tag{2.5}$$

where $f_{X,T}$ and f_T are given by Eqs. 2.2 and 2.4, respectively, and g is the joint density of (A_0, T_0) for which we will derive the density (2.7) below.

The parameters of the sibling distribution are $t_{1:m} \in R^m$, in addition to whatever parameters are hidden in the functional forms of the functions $\beta(a)$ and $l(a)$. Note that the t_j are not restricted to be positive, i.e. the offspring may have died before $t = 0$. Mostly, we shall parameterize $l(a)$ in terms of the age-specific death rate $\phi(a)$, which is related to the survival function through the well known relationship $l(x) = \exp\left(-\int_0^x \phi(a) da\right)$.

In order to derive an expression for the density $g(a_0, t_0)$, occurring in Eq. 2.5, we use the theory for stable age distributions from mathematical demography (Caswell and Keyfitz, 2005), which we now briefly review. A population in which the age-specific rates $\phi(a)$ and $\beta(a)$ do not change with time will settle into a stable age distribution. Further, the population will grow at a rate r given as the solution to the "characteristic equation"

$$\int_0^\infty \beta(a) l(a) e^{-ra} da = 1.$$

The stable age distribution has density $f_A(a) = l(a)e^{-ra} / \int_0^a l(u)e^{-ru} du$, for $a \geq 0$. Our point of view is that the mother is randomly selected among all females alive at $t = 0$, so that the density of A_0 is given by f_A. This is yet not taking into account the fact that she has m offspring over her life time. The joint density of A_0 and T_0 is

$$f_{A_0,T_0}(a_0, t_0) = f_{T_0|A_0}(t_0|a_0) f_{A_0}(a_0), \quad a_0 \geq 0, t_0 \geq 0, \tag{2.6}$$

where $f_{T_0|A_0}(t_0|a_0) = -l'(a_0 + t_0)/l(a_0)$. Conditionally on A_0 and T_0, and hence on the length of the time period $X_0 = A_0 + T_0$ that she is alive, her total number of offspring M is Poisson distributed with mean $B(x_0) = \int_0^{x_0} \beta(u) du$.

Knowing that the mother had m offspring over her lifetime perturbs the distribution (2.6) as follows

$$g(a_0, t_0) \propto f_{A_0,T_0}(a_0, t_0) B(a_0 + t_0)^m e^{-B(a_0+t_0)}. \tag{2.7}$$

We have now completely specified the sibling distribution via its density (2.5). Instead of providing results about its properties in the general case, we turn to a special case in which explicit results can be found. In Section 6 we briefly return with some discussion of the general case.

3 Constant Birth and Death Rates

We shall refer to the situation

$$\beta(a) = \beta \quad \text{and} \quad \phi(a) = \phi \quad \text{for all } a, \tag{3.1}$$

as the constant-rate sibling distribution. Under this assumption it follows that the conditional densities (2.1) and (2.2) reduce respectively to

$$f_Y(y_j|a_0, t_0) = \frac{1}{a_0 + t_0}, \qquad y_j \in (-a_0, t_0), \tag{3.2}$$

and

$$f_{X,T}(x_j, t_j|a_0, t_0) = \frac{1}{a_0 + t_0} \phi e^{-\phi x_j}, \qquad (t_j - t_0)_+ \le x_j \le t_j + a_0. \tag{3.3}$$

Further, the marginal density (2.4) becomes

$$f_T(t_j|a_0, t_0) = \frac{e^{-\phi t_j}}{a_0 + t_0} \begin{cases} e^{\phi t_j} - e^{-\phi a_0}, & -a_0 \le t_j \le t_0, \\ e^{\phi t_0} - e^{-\phi a_0}, & t_j > t_0, \end{cases} \tag{3.4}$$

and the joint density (2.7) becomes

$$g(a_0, t_0) \propto (a_0 + t_0)^m e^{-(\phi+\beta)t_0} e^{-2\beta a_0}, \qquad a_0, t_0 \ge 0. \tag{3.5}$$

Using these expressions we are able to find an explicit expression for the sibling distribution (2.5) of order $m = 2$. We shall first assume that $t_1 = t_2$ which simplifies expressions somewhat. Derivations for the case $t_1 \neq t_2$ are very similar.

Consider the distribution of the life times X_1 and X_2, given that $T_1 = T_2 = t$, where t is the common time of death of the two siblings. We have the following expression for the sibling density, which due to symmetry is presented only for $x_1 \le x_2$.

THEOREM 1. *With constant-rates* (3.1)*, the sibling density* (2.5) *with* $m = 2$ *becomes (for* $t > 0$*):*

$$f(x_1, x_2) = C_1^{-1} \begin{cases} e^{-(\beta+\phi)(t-x_1)-\phi(x_1+x_2)}, & 0 \le x_1 \le x_2 \le t, \\ e^{2\beta(t-x_2)-(\beta+\phi)(t-x_1)-\phi(x_1+x_2)}, & 0 \le x_1 \le t \le x_2, \\ e^{2\beta(t-x_2)-\phi(x_1+x_2)}, & 0 \le t \le x_1 \le x_2, \end{cases} \tag{3.6}$$

and for $t < 0$*:*

$$f(x_1, x_2) = C_2^{-1} e^{-(2\beta+\phi)x_2-\phi x_1}, \quad t < 0 \le x_1 \le x_2, \tag{3.7}$$

where C_1 *and* C_2 *are normalizing constants.*

PROOF. See Appendix A.

By integrating over three branches in Eq. 3.6 we obtain the following expression for the normalizing constant:

$$C_1 = \begin{cases} \frac{e^{-2\phi t}(7\beta+5\phi)}{(2\beta+\phi)(\beta^2-\phi^2)} + \frac{4e^{-(\beta+2\phi)t}}{\phi(2\beta+\phi)} - \frac{2e^{-(\beta+\phi)t}}{\phi(\beta-\phi)}, & \beta \neq \phi, \\ \frac{12\phi e^{\phi t}t-7e^{\phi t}+8}{6\phi^2 e^{3\phi t}}, & \beta = \phi, \end{cases} \tag{3.8}$$

and similarly integration of Eq. 3.7 yields

$$C_2 = [(2\beta + \phi)(\beta + \phi)]^{-1}. \tag{3.9}$$

Note that $f(x_1, x_2)$ does not depend on t when $t < 0$. When we in addition set $\beta = 0$ (interpreted as a limit), X_1 and X_2 become independent, exponentially distributed. Further interpretation of the case that $t < 0$ is given in Section 4 below.

Like the exponential distribution, the constant-rate sibling distribution is closed under change of scale. If we define $X_j' = cX_j$ for $c > 0$, the parameters of the resulting sibling distribution are $\phi' = c^{-1}\phi$, $\beta' = c^{-1}\beta$ and $t' = ct$. Hence, we may set $\phi = 1$ and reparameterize the distribution in terms of (c, β, t), which for some purposes is useful.

As seen from Eq. 3.6, the density has a piecewise definition. When using symmetry to include also the case $x_1 > x_2$, the definition of the sibling density splits the first quadrant, $x_1, x_2 \ge 0$, into six regions (Fig. 2 with $t_1 = t_2$). We see that $\log\{f(x_1, x_2)\}$ is piecewise linear over these regions, and is continuous (but not differentiable) across the boundaries of the regions. The density is unimodal, with the mode at $(x_1, x_2) = (0, 0)$ when $\beta < \phi$, and while $\beta > \phi$ the mode is at $(x_1, x_2) = (t, t)$. Figure 3 shows $f(x_1, x_2)$ for three different parameter.

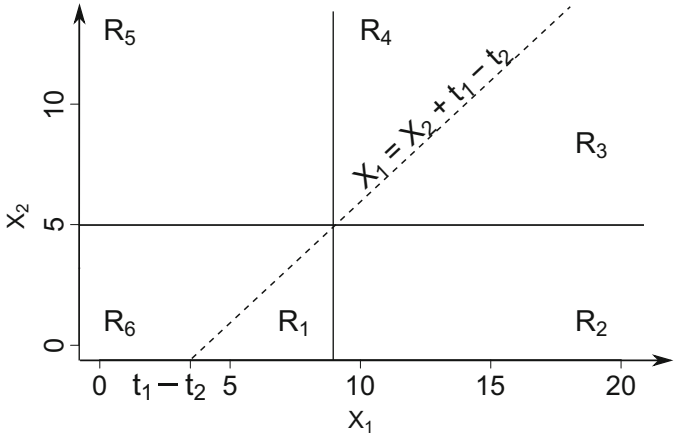

Figure 2: The six different regions of the sibling density when $t_1 = 9$ and $t_2 = 5$. The dashed line indicates ages where the siblings are born at the same time ($y_1 = y_2$)

In order to present the sibling distribution for the case $t_1 \neq t_2$, it is advantageous to introduce a general piecewise log-linear density f over the regions $R_1, \ldots, R_6$ in Fig. 2:

$$f(x_1, x_2) = C^{-1} e^{b_k + c_k x_1 + d_k x_2}, \qquad (x_1, x_2) \in R_k, \quad k = 1, \ldots, 6. \tag{3.10}$$

Here, $b_{1:6} = (b_1, \ldots, b_6)$, $c_{1:6} = (c_1, \ldots, c_6)$ and $d_{1:6} = (d_1, \ldots, d_6)$ are constants satisfying the constraints $c_2, c_3, c_4, d_3, d_4, d_5 < 0$, which are needed for

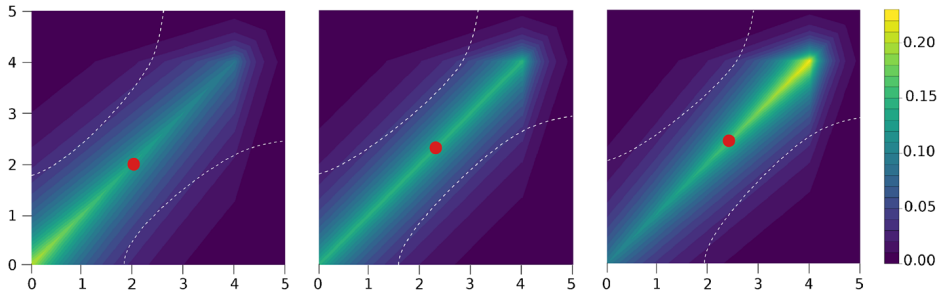

Figure 3: Bivariate sibling density $f(x_1, x_2)$ with parameters $\phi = 1$, $\beta = 0.8, 1.0, 1.2$ (left to right) and $(t_1, t_2) = (4, 4)$. The red dots show expected value (μ, μ). The dashed white curve is the contour $c = 1$ of $c(x_1, x_2)$ given by Eq. 3.15

f to be a proper density. Straightforward, but tedious, integration yields the normalization constant:

$$\begin{aligned} C = &-\frac{\mathrm{e}^{b_1}\left(\mathrm{e}^{d_1 t_2} c_1 \left(\mathrm{e}^{-d_1 t_1} - \mathrm{e}^{c_1 t_1}\right) + \mathrm{e}^{c_1 t_1}\left(c_1 + d_1\right) - c_1 - d_1\right)}{c_1 d_1 \left(c_1 + d_1\right)} \\ &+ \frac{\mathrm{e}^{c_2 t_1 + b_2}\left(1 - \mathrm{e}^{d_2 t_2}\right)}{c_2 d_2} + \frac{\mathrm{e}^{c_3 t_1 + d_3 t_2 + b_3}}{c_3\left(c_3 + d_3\right)} + \frac{\mathrm{e}^{c_4 t_1 + d_4 t_2 + b_4}}{d_4\left(c_4 + d_4\right)} \\ &- \frac{\mathrm{e}^{d_5 t_2 + b_5}\left(\mathrm{e}^{c_5 t_1} - 1\right)}{c_5 d_5} + \frac{\mathrm{e}^{d_6 t_2 + b_6}\left(c_6\left(\mathrm{e}^{-d_6 t_1} - 1\right) + d_6\left(\mathrm{e}^{c_6 t_1} - 1\right)\right)}{c_6 d_6\left(c_6 + d_6\right)}, \end{aligned} \tag{3.11}$$

where $c_1 + d_1 \neq 0$ and $c_6 + d_6 \neq 0$ in addition to the constraint $c_2, c_3, c_4, d_3, d_4, d_5 < 0$. The constants C_1 and C_2 in Theorem 1 are special cases of this.

It is easy to derive the moment generating function of Eq. 3.10:

$$M(s_1, s_2) = E\left(e^{s_1 X_1 + s_2, X_2}\right) = \frac{C(b_{1:6}, c_{1:6} + s_1, d_{1:6} + s_2, t_1, t_2)}{C(b_{1:6}, c_{1:6}, d_{1:6}, t_1, t_2)}, \tag{3.12}$$

where $c_{1:6} + s_1 = (c_1 + s_1, \ldots, c_6 + s_1)$ and $d_{1:6} + s_2 = (d_1 + s_2, \ldots, d_6 + s_2)$. Moments of X_1 and X_2 of various orders can be obtained by repeated differentiation of $M(s_1, s_2)$ at $s_1 = s_2 = 0$. The resulting expressions are complex, and not well suited for interpretation, but are nevertheless useful for numerical evaluation.

THEOREM 2. *With constant rates* (3.1)*, the sibling density* (2.5) *with* $m = 2$ *and* $t_1 \neq t_2$ *has density given by* Eq. 3.10 *with coefficients as specified in* Table 1.

PROOF. The proof is very similar to that of Theorem 1 and is omitted.

3.1. Positive Association Intuitively, X_1 and X_2 are positively associated under the sibling distribution, due to their dependence of the lifespan

Table 1: Choice of coefficients in Eq. 3.10 yielding the order 2 sibling density when $t_1 \neq t_2$

j	b_j	c_j	d_j
1	$-(\beta+\phi)t_2$	$-\phi$	β
2	$2\beta t_1 - (\beta+\phi)t_2$	$-(2\beta+\phi)$	β
3	$2\beta t_1$	$-(2\beta+\phi)$	$-\phi$
4	$2\beta t_2$	$-\phi$	$-(2\beta+\phi)$
5	$2\beta t_2 - (\beta+\phi)t_1$	β	$-(2\beta+\phi)$
6	$-(\beta+\phi)t_1$	β	$-\phi$

$(-A_0, T_0)$ of their shared mother. Further, for each marginal ($j = 1, 2$) we must have that X_j is stochastically increasing in the parameter t_j. The informal argument for the latter is that since the mother is known to have been alive at $t = 0$, the larger the death time t_j the older (x_j) the individual is likely to be. We now set out to prove these claims rigorously.

We start out by proving the positive association between X_1 and X_2. An appropriate notion of positive association is the so-called Multivariate Total Positivity of order 2 (MTP_2). A bivariate density $f(x)$, $x \in R^2$, is is said to be MTP_2 if

$$f(x \vee z)f(x \wedge z) \geq f(x)f(z), \tag{3.13}$$

for any $x, z \in R^2$, where $x \vee z = (\max(x_1, z_1), \max(x_2, z_2))$ and $x \wedge z = (\min(x_1, z_1), \min(x_2, z_2))$. See Karlin and Rinott (1980) for a comprehensive overview of properties of MTP_2 distributions.

THEOREM 3. *The sibling densities* (3.6) *and* (3.7) *are MTP*$_2$.

This is proved in Appendix C using the definition Eq. 3.13 of MTP_2 directly. We believe that an alternative proof may be based on the fact that mixture distributions, of which the numerator of Eq. 2.5 is an example, under certain conditions are MTP_2 (Khaledi and Kochar, 2001; Shaked and Spizzichino, 1998). Using this approach it may be possible to prove that more general sibling distributions than (3.6) are MTP_2.

MTP_2 is a strong positive dependence property, which among other things imply that $\text{cov}(X_1, X_2) \geq 0$. Although covariance is (arguably) not the most relevant dependence measure for life times, it nevertheless the most common dependence measure in general, and it is therefore useful to have establish this result.

3.2. Marginal Distribution and Copula The marginal densities in Eq. 3.6 are both given as

$$f(x) = \frac{e^{-\phi(t+x)}}{C_1} \begin{cases} e^{-\beta(t-x)} \left[\frac{\beta+\phi}{\beta\phi} - \frac{e^{-\beta x}}{\beta} - \frac{2\beta e^{\phi(x-t)}}{\phi(2\beta+\phi)} \right], & x \in (0, t], \\ e^{2\beta(t-x)} \left[\frac{\beta+\phi}{\beta\phi} - \frac{e^{-\beta t}}{\beta} - \frac{2\beta e^{\phi(t-x)}}{\phi(2\beta+\phi)} \right], & x \in (t, \infty), \end{cases} \tag{3.14}$$

where C_1 is given by Eq. 3.8. As a local measure of dependency between X_1 and X_2 we introduce

$$c(x_1, x_2) = \frac{f(x_1, x_2)}{f(x_1)f(x_2)}. \tag{3.15}$$

The $c = 1$ contour of $c(x_1, x_2)$ is displayed in Fig. 3. The region in which $c(x_1, x_2) > 1$ is located around the diagonal $x_1 = x_2$. This reflects the positive dependence in the sibling distribution.

We can obtain an analytical expression for the cumulative joint distribution $F(x_1, x_2)$ by integrating Eq. 3.6. Similarly, we get an expression for the cumulative marginal distribution function $G(x)$ by integrating Eq. 3.14. Then we can define a copula (Nelsen, 2007), $F\left(G^{-1}(x_1), G^{-1}(x_2)\right)$, based on the sibling distribution, where G^{-1} denotes the inverse of G. Because the sibling distribution is closed under change of scale, we set $\phi = 1$, and the copula thus has β and t as free parameters. We do not explore this copula further in this paper.

We next prove that X (X_1 or X_2) is stochastically increasing in t, in the sense of the following theorem.

THEOREM 4. *For any $t' > t > 0$ we have*

$$P(X > x|T = t) \leq P(X > x|T = t'). \tag{3.16}$$

It turns out to be easier to prove the more general statement that (X_1, X_2) is multivariate stochastically increasing (Shaked and Shanthikumar, 2007, p. 265), which imply Eq. 3.16. The reason this is simpler is that the key quantity, the ratio $f(x; t')/f(x; t)$, which is involved in Theorem 6.B.8. of Shaked and Shanthikumar (2007, p. 265), takes on a simpler form for the bivariate density (3.6) than for the univariate density (3.14). The details of the proof are given in Appendix D.

Stochastic monotonicity of a random variable X implies that $E(X)$ is an increasing function of t (Shaked and Shanthikumar, 2007, p. 4). This means that, for given ϕ and β, there is a one-to-one correspondence between t and $\mu = E(X)$. This fact will play a crucial role when we later devise an estimator for the parameters ϕ, β and t.

3.3. The Role of β and t Because the family of constant-rate sibling distributions is closed under change of scale we set $\phi = 1$. In this section we will study the effect of varying β and t on two characteristics: the correlation (COR) between X_1 and X_2 and $\mathrm{CV}(X_j) = \sqrt{\mathrm{Var}(X_j)}/\mathrm{E}(X_j)$. Without conditioning on $T_j = t$, we have that X_j is exponentially distributed with rate $\phi = 1$. The process of conditioning on T_j can be expected to deduce $\mathrm{CV}(X_j)$. Rather than trying to prove this formally, we provide numerical evidence.

Figure 4 shows correlation and CV as functions of β (top) and t (bottom), and indeed we see that $\mathrm{CV} \leq 1$ for all β and t. When β and t are both close to zero we have $\mathrm{CV} \approx 1$ and $\mathrm{COR} \approx 0$, which reflects the fact that X_1 and X_2 are then approximately independent and exponentially distributed. For increasing β the correlation increases, but not necessarily monotonically, and approaches an asymptotic level (top-left). From the corresponding plot of

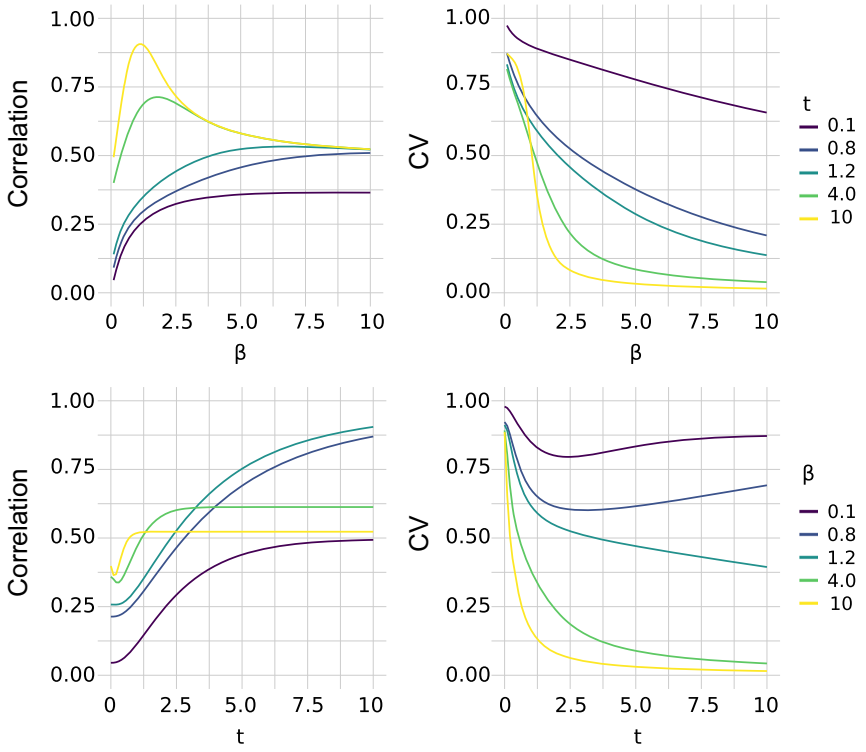

Figure 4: Correlation (left) and CV (right) of the sibling distribution ($\phi = 1$) as a function of its parameters. The parameter ϕ is set to 1 and the plots on the top row are plotted as a function of β for five different values of t. The plots on the bottom are plotted as a function of t for five different values of β

the CV (top-right) we notice that the overall trend is that the CV decreases as the value of β increases. The decrease is steepest for the highest values of t.

Further, we see from the bottom row of Fig. 4 that the correlation increases as a function of t, except for very small t. It can be shown that when $\beta = \phi$ the correlation approaches 1 as $t \to \infty$. When $\beta \neq \phi$, the correlation does not approach 1 as $t \to \infty$, but flattens out at a lower value which depends on the value of β. The CV decrease quickly as a function of t, especially for the higher values of β. When $\beta < \phi$ we see that the CV first decrease, then starts to increase for higher values of t.

4 Relationship to the Block-Basu Distribution

In this section we clarify the relationship between the constant-rate sibling distribution and the Block-Basu distribution (Block and Basu, 1974),

and we shall use this relationship to interpret the sibling distribution. One way of deriving the Block-Basu distribution goes via (Freund, 1961), and we will refer to this as the "Freund interpretation". Let now X_1 and X_2 be the lifetimes of two components assumed to be independently exponentially distributed with rate parameters α_1 and α_2, respectively. When one of the component fails, the rate for the remaining component changes from α_1 to α_1^* or from α_2 to α_2^*, depending on which component fails first. The resulting density (see Freund (1961)) is

$$f(x_1, x_2) = \begin{cases} \alpha_1 e^{-x_1(\alpha_1+\alpha_2)} \alpha_2^* e^{-\alpha_2^*(x_2-x_1)}, & 0 \le x_1 \le x_2, \\ \alpha_2 e^{-x_2(\alpha_1+\alpha_2)} \alpha_1^* e^{-\alpha_1^*(x_1-x_2)}. & 0 \le x_2 \le x_1. \end{cases} \tag{4.1}$$

When setting

$$\alpha_j = \beta + \phi, \qquad \alpha_j^* = 2\beta + \phi, \qquad j = 1, 2, \tag{4.2}$$

we see that Eq. 4.1 reduces to the symmetric sibling density (3.7) with $t \le 0$. Since $t = 0$ is the time point at which the mother is known to have been alive, we are effectively considering a sibling distribution where both offspring are dying before their mother.

The Freund interpretation yields that $X^{(1)} = \min(X_1, X_2)$, i.e. the age of the youngest sibling, has an exponential distribution with rate parameter $\alpha_1 + \alpha_2 = 2(\beta + \phi)$. Further, the age difference between the oldest and the youngest, $X^{(2)} = \max(X_1, X_2) - \min(X_1, X_2)$, has an exponential distribution with rate parameter $\alpha^* = 2\beta + \phi$. In the following paragraphs we interpret the rates (4.2) in the context of the sibling distribution.

If we consider the case with $t = 0$, we know that the mother and her two offspring were all alive at (or just prior to) $t = 0$. The Freund interpretation requires us to look backwards in time, starting from $t = 0$. The "failure" of a component corresponds to an offspring being born. We first look at the event $X^{(1)} > x$, which can be broken into three sub events:

(i) The mother was born prior to $t = -x$, i.e. $a_0 > x$. Because the stable age distribution of A_0 is exponential with rate β, we have $P(A_0 > x) = \exp(-\beta x)$.

(ii) Both offspring were born prior to $-x$, and because we are conditioning on there being $m = 2$ siblings in total, this implies that there were no additional births in $(-x, 0)$. The latter has probability $\exp(-\beta x)$.

(iii) Both offspring survived the interval $(-x, 0)$, which has probability $\exp(-2\phi x)$.

When combining the independent events i)–iii) we get the Freund interpretation $P\left(X^{(1)} > x\right) = \exp\left[-2(\beta+\phi)x\right]$.

The event $X^{(2)} > x$ can be interpreted similarly, but we must shift our point of view backwards in time to $t = -x^{(1)}$ when the youngest sibling was born. The mother would have to be born prior to $t = -(x + x^{(1)})$. Using the stable age distribution of A_0, this has conditional probability

$$P\left(A_0 > x + x^{(1)} | A_0 > x^{(1)}\right) = \exp\left(-\beta x\right).$$

Secondly, there couldn't have been any births between $t = -\left(x + x^{(1)}\right)$ and $t = -x^{(1)}$, which has probability $\exp\left(-\beta x\right)$. Finally, the offspring that was born at $t = -x^{(2)}$ survived from $t = -\left(x + x^{(1)}\right)$ until $t = -x^{(1)}$, which has probability $\exp\left(-\phi x\right)$. In total we get $P\left(X^{(2)} > x\right) = \exp\left[-(2\beta+\phi)x\right]$, which is the Freund interpretation of the sibling distribution. Similar arguments applies to the situation $t < 0$.

Finally, we discuss a few additional insights gained from the Freund interpretation (4.2). First, the age difference between the siblings, $X_1 - X_2$ follows a Laplace distribution with rate $2\beta + \phi$. Further, note that $\alpha_j^* \to \alpha_j = \phi$ as $\beta \to 0$. Hence, X_1 and X_2 are independent in the limit $\beta \to 0$, each having an exponential distribution with rate ϕ.

5 Simulation, Estimation and Application to Real Data

We first devise an algorithm for sampling (x_1, x_2) from the density (3.6). Rather than sampling directly from Eq. 3.6, which would be feasible albeit a bit technical, we choose to go back to the definition of the sibling distribution. This involves explicitly sampling (a_0, t_0) for the mother. As a byproduct, our algorithm sheds light on the sibling distribution, through an expression for the conditional density of (A_0, T_0), given (T_1, T_2).

We also construct a hybrid moment/maximum likelihood estimator for the parameter vector $\theta = (\beta, \phi, t)$. The statistical properties of this estimator are investigated on simulated data.

5.1. Simulation The joint density of (A_0, T_0), (X_1, T_1) and (X_2, T_2) is given by

$$g(a_0, t_0) \prod_{j=1}^{2} f(x_j, t_j | a_0, t_0) = f(t_1, t_2) f(a_0, t_0 | t_1, t_2) \prod_{j=1}^{2} f(x_j | t_j, a_0, t_0), \quad (5.1)$$

where for notational simplicity we suppress subscripts on densities and range of variables in this section. The quantities on the left-hand side are given

by Eqs. 3.3 and 3.5, while the right-hand side is a generic refactoring of the joint density in terms of conditional densities. Dividing through by $f(t_1, t_2)$ in Eq. 5.1 we obtain the target distribution, and the right-hand side (5.1) suggests the following algorithm for sampling (X_1, X_2) conditionally on (T_1, T_2):

(i) Sample (A_0, T_0) from $f(a_0, t_0|t_1, t_2)$,

(ii) Using (a_0, t_0) from (i), draw X_j from $f(x_j|t_j, a_0, t_0)$, independently for $j = 1, 2$.

Step (ii) is straight forward, and is seen from Eq. 3.3 to amount to sampling from an exponential distribution with x_j constrained to a certain interval. Step (i) requires more careful consideration. We have

$$f(a_0, t_0|t_1, t_2) \propto g(a_0, t_0) \prod_{j=1}^{2} f(t_j|a_0, t_0),$$

where $f(t_j|a_0, t_0)$ is given by Eq. 3.4 and $g(a_0, t_0)$ by Eq. 3.5. We have found experimentally for $t_1 = t_2$ that the following two-step procedure works well. We start by independently drawing A_0 and T_0 from exponential distributions with rates 2β and $\min(\phi + \beta, t_1^{-1}, t_2^{-1})$, respectively. This is repeated K times to get a pre-sample $\{(A_{0k}, T_{0k}), k = 1 \ldots, K\}$, from which a single pair (A_0, T_0) is drawn with probabilities

$$p_k \propto f(a_{0k}, t_{0k}|t_1, t_2) \exp\left[2\beta a_{0k} + \min(\phi + \beta, t_1^{-1}, t_2^{-1}) t_{0k}\right].$$

5.2. Estimation Consider n observation pairs $\{(x_{1i}, x_{2i}); i = 1, \ldots, n\}$ from Eq. 3.6. While $f(x_1, x_2)$ is continuous as a function of $\theta = (\beta, \phi, t)$, it is not differentiable in t at $t = x_1$ and $t = x_2$. This implies that the log-likelihood

$$l(\beta, \phi, t) = n^{-1} \sum_{i=1}^{n} \log\left(f(x_{1i}, x_{2i}; \beta, \phi, t)\right)$$

has $2n$ points where the derivative is not differentiable. Standard numerical optimization algorithms typically either do not use derivative information at all, or requires the objective function to be continuously differentiable in all variables. The former types of algorithms are slow and unstable, and the latter type are not directly applicable to our setting. We thus devise a special two-stage estimation algorithm.

Because X_1 and X_2 have the same marginal distribution we define the overall sample mean $\bar{x} = (2n)^{-1} \sum_{i=1}^{n} (x_{1i} + x_{2i})$. We denote by $\mu(\beta, \phi, t)$ the expectation of X_1 and X_2, and impose the constraint $\mu(\beta, \phi, t) = \bar{x}$ on

the parameter estimation problem. An analytical expression for $\mu(\beta, \phi, t)$ is given in Appendix B. The expression is complicated, but nevertheless well suited for numerical evaluation. Recall that we have proven earlier that $\mu(\beta, \phi, t)$ is increasing as a function of t.

Our estimation algorithm iterates between the following two steps:

1. For given $\hat{t}$, let $\hat{\beta}$ and $\hat{\phi}$ be the maximizer of $l(\beta, \phi, \hat{t})$.
2. For given $\hat{\beta}$ and $\hat{\phi}$, let $\hat{t}$ be the solution of the equation $\mu(\hat{\beta}, \hat{\phi}, t) = \bar{x}$.

Both 1) and 2) are solved numerically using the software TMB (Kristensen et al., 2016).

5.3. Simulation Experiment Using the algorithm of Section 5.1, we sampled $n = 1000$ observation pairs (x_1, x_2) to which the estimator of Section 5.2 was applied. This process was repeated 1000 times to assess the statistical properties of the estimator. Table 2 shows the results for 20 different parameter combinations (one row per combination). Moments characterizing the distribution of (X_1, X_2), obtained from the moment generating function (3.12), are also given in the table.

From Table 2 we see that the estimator for the parameter vector $\theta = (\beta, \phi, t)$ is overall stable with respect to the different combinations of the input parameters. More specifically, for the parameter t in the first column, we see that the mean values of the estimates are all very close to the true value of t, but they are slightly worse in the case when $\beta > \phi$. The same trend can be seen in the standard deviations, which are higher in these situations. For the parameter β we see that the mean of the estimates are more accurate for higher values of t, but we also see the trend with better predictions when $\beta > \phi$. The standard deviations are however quite stable for all combinations of the input parameters. The mean values of the estimates for the parameter ϕ are all very similar and they do not seem to be affected by the different combinations of the input parameters. The standard deviations are higher when $t = 2$, but otherwise quite similar. Figure 5 shows the marginal density (3.14) for some of the parameter combinations used in Table 2.

5.4. Application to Real Data We have presented the sibling distribution as a distribution for life times, but it may in fact be applied to any set of non-negative quantities with positive dependence. The constant-rate case is applicable only when the CV is less than one, and when $t_1 = t_2$ the marginals must be the same. We use the "twinData" dataset found in the R-package "OpenMx" (Neale et al., 2016) as an example. These are BMI measurements on twins (around age 18), but nevertheless satisfy the above mentioned restrictions (Table 3).

Table 2: Performance of the estimator $(\hat{\beta}, \hat{\phi}, \hat{t})$ with $n = 1000$ observations drawn using the algorithm in Section 5.2

$\hat{t}$				$\hat{\beta}$			$\hat{\phi}$			Moments		
	True	Mean	SD	True	Mean	SD	True	Mean	SD	E(X)	CV(X)	COR
1	2.00	2.05	0.21	0.60	0.63	0.05	1.00	1.01	0.15	1.20	0.65	0.37
2	2.00	2.03	0.15	0.80	0.83	0.05	1.00	1.01	0.15	1.21	0.61	0.40
3	2.00	2.02	0.08	1.20	1.23	0.06	1.00	1.01	0.13	1.27	0.54	0.46
4	2.00	2.00	0.07	1.40	1.43	0.06	1.00	1.00	0.13	1.30	0.51	0.48
5	4.00	4.06	0.15	0.60	0.64	0.04	1.00	1.01	0.05	1.84	0.66	0.57
6	4.00	4.03	0.10	0.80	0.83	0.04	1.00	1.00	0.05	2.02	0.61	0.62
7	4.00	4.01	0.06	1.20	1.21	0.05	1.00	1.01	0.06	2.43	0.49	0.68
8	4.00	4.00	0.05	1.40	1.41	0.05	1.00	1.00	0.07	2.62	0.43	0.70
9	6.00	6.05	0.18	0.60	0.64	0.04	1.00	1.00	0.04	2.30	0.71	0.69
10	6.00	6.03	0.12	0.80	0.83	0.04	1.00	1.00	0.04	2.73	0.63	0.75
11	6.00	6.01	0.06	1.20	1.21	0.04	1.00	1.01	0.05	3.71	0.45	0.80
12	6.00	6.00	0.05	1.40	1.40	0.05	1.00	1.00	0.06	4.16	0.37	0.81
13	8.00	8.03	0.25	0.60	0.65	0.04	1.00	1.00	0.03	2.59	0.75	0.76
14	8.00	8.02	0.14	0.80	0.83	0.04	1.00	1.00	0.04	3.32	0.66	0.82
15	8.00	8.01	0.06	1.20	1.20	0.04	1.00	1.01	0.05	5.12	0.42	0.87
16	8.00	8.00	0.05	1.40	1.40	0.04	1.00	1.01	0.05	5.87	0.32	0.86
17	10.00	9.98	0.33	0.60	0.65	0.03	1.00	1.00	0.03	2.77	0.79	0.80
18	10.00	10.00	0.15	0.80	0.83	0.03	1.00	1.00	0.03	3.81	0.69	0.87
19	10.00	10.00	0.06	1.20	1.20	0.04	1.00	1.01	0.04	6.63	0.39	0.90
20	10.00	9.99	0.05	1.40	1.39	0.05	1.00	1.00	0.05	7.69	0.28	0.89

The column "True" shows the values used in the simulations, while "Mean" and "SD" are, respectively, the average and standard deviation of $(\hat{\beta}, \hat{\phi}, \hat{t})$ across 1000 repetitions. The three rightmost columns show respectively $E(X)$, $CV(X) = \sqrt{\mathrm{Var}(X)}/E(X)$, and $\mathrm{COV}(X_1, X_2)$, all calculated using the moment generating function for the true parameter values

The dataset consists of BMI measurements for 3569 (male/female, monozygotic/dizygotic) twin pairs. The fitted sibling distribution is unable to accomodate the light left-hand tail of data (Fig. 6). The fitted density is huge perturbation of the unconditional (on T) distribution of X, which for the constant-rate case is an exponential distribution. This illustrates the flexibility of the distribution. Table 3 shows the parameters of the fitted distribution. The lack of fit is reflected in estimated moments not fitting the empirical moments very well. If we look at the estimated parameter values in Fig. 6 we see that these are "outside in the normal range", in the following sense. The expected life length of an individual is $1/\hat{\phi} = 1/0.94 = 1.12$, but

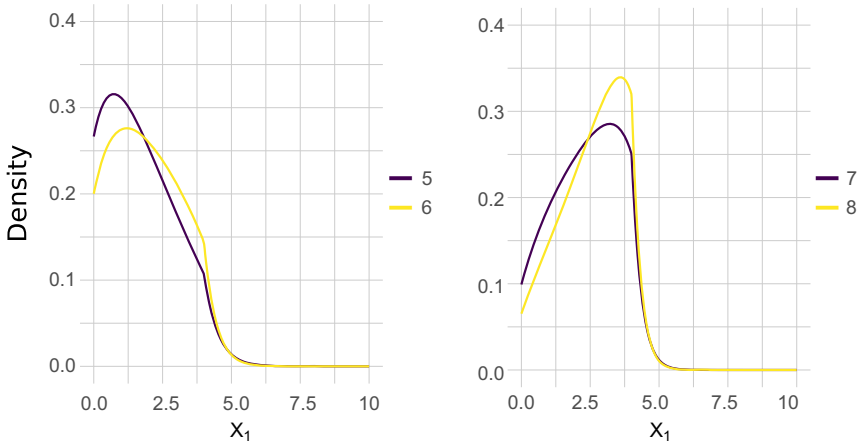

Figure 5: Marginal density (3.14) when $\beta < \phi$ (left) and when $\beta > \phi$ (right). The legend refers to the leftmost column of Table 2, which gives the parameter setting used for the different density curves

the mother-offspring duo spanned (mother's birth to offspring's death) at least $\hat{t} = 21.77$ time units.

6 Discussion

The idea of a sibling distribution was conceived during our work with the recently invented Close-Kin Mark-Recapture method (Bravington et al., 2016), in which the joint age distribution of half-siblings plays a crucial role. Its usefulness as a distribution for multivariate life time data in general remains to be explored. The fact that it is a mixture (over A_0 and T_0) of independent life times makes it amenable to analysis in the framework of Shaked and Spizzichino (1998). However, due to the conditioning on

Table 3: Fitted sibling distribution to BMI data

	Estimated	Empirical
Mean	21.79	21.77
SD	1.82	0.94
CV	0.08	0.04
COR	0.87	0.53
β	1.52	-
ϕ	0.94	-
t	23.51	-

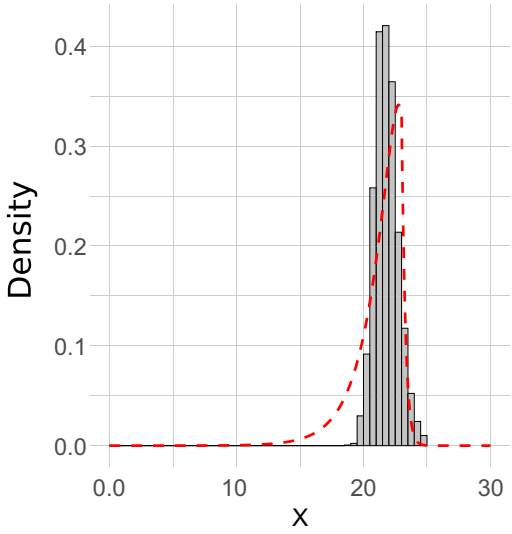

Figure 6: Marginal distribution of BMI data, with fitted sibling distribution. The dashed red curve shows the sibling density (3.6)

T_j its structure, and in particular the moments, is complicated. Moreover, the non-differentiability of the likelihood with respect to the parameter t prevents straight forward application of maximum likelihood estimation.

Our numerical experiments indicate that the constant-rate (3.1) distribution has CV ≤ 1. This is clearly limiting for a general purpose life time distribution, but this restriction can be removed by choosing a non-constant $\phi(a)$. We have not been able to obtain explicit expressions for the sibling density under more general conditions. In general, the sibling density (2.5) can be evaluated numerically. Both the numerator and denominator involves two-dimensional integrals (with respect to a_0 and t_0). The integrand of the denominator is a product of m functions on form (2.4), which each involves a one dimensional integral. This is by no means computationally prohibitive, but specially tailored numerical integration schemes would have to be devised in order for the general distribution to be practically useful.

We have shown that the constant-rate sibling distribution with $t_1 = t_2 = 0$ coincides with the exchangeable Block-Basu distribution ($\alpha_1 = \alpha_2$ and $\alpha_1^* = \alpha_2^*$). The non-exchangeable case is not a sibling distribution, as the sibling framework requires that ϕ and β are the same for both siblings. Conversely, the sibling distribution with $t_1 \neq t_2$, or $t_1 = t_2 > 0$, is not a Block-Basu distribution. Also, when allowing age-specific rates, $\phi(a)$ and $\beta(a)$, the sibling distribution is no longer a Block-Basu distribution. Kundu

and Gupta (2010) extended the Block-Basu distribution by deriving it from components that were Weibull distributed instead of exponential. The additional shape parameter makes the extended Block-Basu distribution more flexible at the cost of being less computational tractable.

Finally, we have proven that the constant-rate sibling distribution is MTP_2 and stochastically increasing in t. We conjecture, based on literature for mixture distributions (Shaked and Spizzichino, 1998, p. 273; Shaked and Shanthikumar, 2007) that these properties hold for a wider class of sibling distributions, but possibly not for all.

Acknowledgments. Parts of this work have been done in the context of CEDAS (Center for Data Science, University of Bergen, Norway).

Funding. Open access funding provided by University of Bergen (incl Haukeland University Hospital).

References

ANDERSEN, EW (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Anal.* **11**, 333–350.

ARNOLD, BC and STRAUSS, D (1988). Bivariate distributions with exponential conditionals. *J. Am. Stat. Assoc.* **83**, 522–527.

BARRETO-SOUZA, W and MAYRINK, VD (2019). Semiparametric generalized exponential frailty model for clustered survival data. *Ann. Inst. Stat. Math.* **71**, 679–701.

BLOCK, HW and BASU, A (1974). A continuous, bivariate exponential extension. *J. Am. Stat. Assoc.* **69**, 1031–1037.

BRAVINGTON, MV, SKAUG, HJ, ANDERSON, EC et al. (2016). Close-kin mark-recapture. *Stat. Sci.* **31**, 259–274.

CASWELL, H and KEYFITZ, N (2005). *Applied mathematical demography*, 3rd edn. Springer, New York.

FREUND, JE (1961). A bivariate extension of the exponential distribution. *J. Am. Stat. Assoc.* **56**, 971–977.

GUMBEL, EJ (1960). Bivariate exponential distributions. *J. Am. Stat. Assoc.* **55**, 698–707.

HOUGAARD, P (1986). A class of multivariateate failure time distributions. *Biometrika* **73**, 671–678.

HOUGAARD, P (2001). *Analysis of multivariate survival data*. Springer, New York.

KARLIN, S and RINOTT, Y (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivar. Anal.* **10**, 467–498.

KHALEDI, BE and KOCHAR, S (2001). Dependence properties of multivariate mixture distributions and their applications. *Ann. Inst. Stat. Math.* **53**, 620–630.

KRISTENSEN, K, NIELSEN, A, BERG, CW, SKAUG, HJ and BELL, B (2016). TMB: Automatic differentiation and Laplace approximation. *J. Stat. Softw.* **70**, 1–21.

KUNDU, D and GUPTA, RD (2010). A class of absolutely continuous bivariate distributions. *Statistical Methodology* **7**, 464–477.

MARSHALL, AW and OLKIN, I (1967). A multivariate exponential distribution. *J. Am. Stat. Assoc.* **62**, 30–44.

NEALE, M, HUNTER, M, PRITIKIN, J, ZAHERY, M, BRICK, T, KIRKPATRICK, R, ESTABROOK, R, BATES, T, MAES, H and BOKER, S (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika* **81**, 535–549.

NELSEN, RB (2007). An introduction to copulas. Springer Science & Business Media.

SARKAR, SK (1987). A continuous bivariate exponential distribution. *J. Am. Stat. Assoc.* **82**, 667–675.

SHAKED, M and SHANTHIKUMAR, JG (2007). Stochastic orders. Springer Science & Business Media.

SHAKED, M and SPIZZICHINO, F (1998). Positive dependence properties of conditionally independent random lifetimes. *Math. Oper. Res.* **23**, 944–959.

Appendix A. Proof of Theorem 1

The quantities involved in the expression (2.5) for the sibling density are given by Eqs. 3.3, 3.4 and 3.5. The evaluation of the integrals over (a_0, t_0) in the numerator and denominator of Eq. 2.5 is made difficult by the constraints (2.3). Below we show how these constraints split the first quadrant of the (x_1, x_2) plane into six disjoint regions $R_1, \ldots, R_6$ (see Fig. 2). In each of these the integrand is just a simple exponential function. Because the density is exchangeable when $t_1 = t_2$ it is sufficient to specify the expression only over the regions R_1, R_2 and R_3.

Recall that $y_j = t_j - x_j$ denotes the birth time ($j = 0, 1, 2$). The integrand of the numerator in Eq. 2.5 is

$$\begin{aligned} & f_{X,T}(x_1, t|a_0, t_0) f_{X,T}(x_2, t|a_0, t_0) g(a_0, t_0) \\ & \propto e^{-2\beta a_0} e^{-(\beta+\phi)t_0} \times \exp\{-\phi(x_1 + x_2)\}, \end{aligned} \tag{A.1}$$

for values of a_0 and t_0 such that the constraints (2.3) are satisfied for $j = 1, 2$, and zero otherwise. The term after $\times$ does not depend on a_0 and t_0, but is included for later reference. Because we assume $x_1, x_2 \geq 0$, we can replace the inequality $(t_j - t_0)_+ \leq x_j$ in Eq. 2.3 by $t - t_0 \leq x_j$, which again is equivalent to $t_0 \geq y_j$. Similarly, the last inequality $x_j \leq t_j - y_0$ in Eq. 2.3 can be re-expressed as $a_0 \geq -y_j$. Together with the basic constraints $a_0, t_0 \geq 0$ we get $a_0 \geq \max(0, -y_1, -y_2)$ and $t_0 \geq \max(0, y_1, y_2)$. Depending on the relative values of y_1 and y_2, the lower bounds of the integrals over a_0 and t_0 will be qualitatively different. There are 6 different cases, corresponding to the partition $R_1, \ldots, R_6$ of the (x_1, x_2) plane shown in Fig. 2. When integrating (A.1) with respect to a_0 and t_0, and replacing y_j by $t - x_j$ $(j = 1, 2)$, we get

$$f_{X,T}(x_1, t) f_{X,T}(x_2, t) \propto e^{-\phi(x_1+x_2)} \times \begin{cases} e^{-(\beta+\phi)(t-x_1)}, & (x_1, x_2) \in R_1, \\ e^{2\beta(t-x_2)} e^{-(\beta+\phi)(t-x_1)}, & (x_1, x_2) \in R_2, \\ e^{2\beta(t-x_2)}, & (x_1, x_2) \in R_3, \end{cases}$$

which is Eq. 3.6.

The proof when $t < 0$ follows in the same vein, where we start out with the integrand of the numerator in Eq. 2.5 given by Eq. A.1. We must find values of a_0 and t_0 such that the constraints (2.3) still hold. The inequality $(t_j - t_0)_+ \leq x_j$ in Eq. 2.3 which is equivalent to $t_0 \geq y_j$ can be replaced with $t_0 \geq 0$ since we have $y_1, y_2 < 0$ in combination with the constraint $t_0 \geq 0$. For the last inequality in Eq. 2.3, the above arguments still apply such that this term is replace, with by $a_0 \geq -y_j$. In total we get $a_0 \geq \max(0, -y_1, -y_2)$ and $t_0 \geq 0$.

Appendix B. Analytical expressions

The expectation under the density (3.14) is

$$\begin{aligned} \mu(\beta, \phi, t) &= \mathrm{E}(X) \\ &= 6\beta\,(\phi - \beta/3)\,(\beta + \phi)^2\,(2\beta + \phi)^2\,\mathrm{e}^{(-\beta+\phi)t} \\ &\quad +4\left(\phi^2\,(t\beta - 1) + 2\beta^2\phi t + 2\beta^2\right)(-\beta + \phi)^2\,(\beta + \phi)^2\,\mathrm{e}^{-t\beta} \\ &\quad +(4 - 10t\beta)\,\phi^6 - \left(34\beta^2 t + 4\beta\right)\phi^5 - \left(18\beta^3 t + 51\beta^2\right)\phi^4 + \left(34\beta^4 t - 7\beta^3\right)\phi^3 \\ &\quad +\left(28\beta^5 t - 23\beta^4\right)\phi^2 / 8\beta\,(2\beta + \phi)\,(-\beta + \phi)\,\phi\,(\beta + \phi) \\ &\quad \left(1/2\,(2\beta + \phi)\,(\beta + \phi)\,\mathrm{e}^{(-\beta+\phi)t} + \left(-\beta^2 + \phi^2\right)\mathrm{e}^{-t\beta} - 5/4\phi\,(\phi + 7/5\beta)\right). \end{aligned}$$

This expression is obtained using computer algebra system Maple.

Appendix C. Proof of Theorem 3

We shall work with $g(x) = \log f(x)$, $x \in R^2$, which for the sibling density that Eq. 3.6 is a piecewise linear function. The MTP_2 property of f is equivalent supermodularity of g, i.e.

$$g(x \vee z) + g(x \wedge z) \geq g(x) + g(z). \tag{C.1}$$

It is trivial to check that a linear function g is supermodular, so it follows directly that Eq. 3.7 is MTP_2.

The density (3.6) requires more careful attention due to its piecewise definition. It suffices to consider x and z such that the four points x, z, $(x \vee z)$ and $(x \wedge z)$ form a non-degenerate rectangle with sides parallel to the axes. This happens when either $x_1 < z_1$ and $x_2 > z_2$ or when $x_1 > z_1$ and $x_2 < z_2$. We will consider only the former, where x is the upper-left corner (and z the lower-right corner) of the rectangle. The other case is handled in the same way. When the rectangle is degenerate (a line or a point) it can be checked that Eq. C.1 holds for any function g. Under these constraints, if x and z lie in the same region $(R_1, \ldots, R_6)$ of Fig. 7 all four corners of the rectangle lies lies in same region, and by linearity of g within each region we have that Eq. C.1 is satisfied.

The sum of two supermodular functions is again supermodular, so we may add $\phi(x_1 + x_2)$ to all three branches of the logarithm of Eq. 3.6, so that we may work with

$$g(x_1, x_2) = \begin{cases} -(\beta + \phi)(t - x_1), & 0 \leq x_1 \leq x_2 \leq t, \\ 2\beta(t - x_2) - (\beta + \phi)(t - x_1), & 0 \leq x_1 \leq t \leq x_2, \\ 2\beta(t - x_2), & 0 \leq t \leq x_1 \leq x_2. \end{cases} \tag{C.2}$$

The extension of g to all six regions of Fig. 7 (with $t_1 = t_2 = t$) is $g(x_1, x_2) = g(x_2, x_2)$ when $x_1 > x_2$. The fact that $g(x_1, x_2)$ does not depend on x_2 in regions R_1 and R_4, and not on x_2 in R_3 and R_6 are visualized via the a level curve (green dashed line) of g in Fig. 7. It is seen that g is unimodal, with the mode at $(x_1, x_2) = (t, t)$.

We start out by restricting ourselves to the case $x_2 < t$. Under the facts established above, and the assumed restrictions on x and z, there are only three qualitatively different cases that must be considered. Using the red part of Fig. 7 as a reference, these are: i) $x = A$ and $z = C'$, ii) $x = A$ and $z = D'$ and iii) $x = C$ and $z = D'$. With our geometric approach, supermodularity is something that is checked for rectangles. It has the property that we can split a rectangle (A, C, C', A') in two, (A, B, B', A')

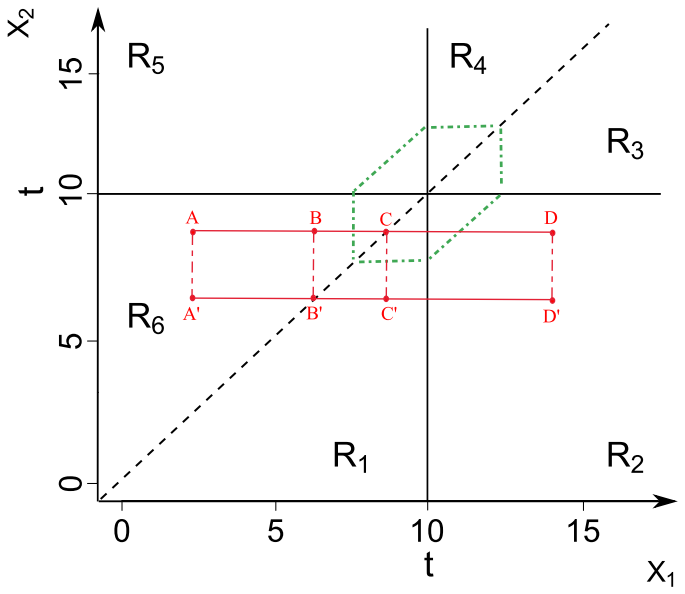

Figure 7: Part of the proof of Theorem C. The green dashed line is an example of a level curve (C.2). The red rectangles are used to prove supermodularity

and (B, C, C', B'), and it is sufficient to check Eq. C.1 for the two parts. Hence, to check all of i)–iii) it suffices to check all the red sub-rectangles in Fig. 7, which we will now do. First, (A, B, B', A') lies in a single region (R_6) so Eq. C.1 holds. For (B, C, C', B') we find by looking at the level curves of g that $g(C) > g(B)$ and $g(C') = g(B')$, which imply that Eq. C.1 holds. Finally, the (solid black) vertical line $x_1 = t$ splits (C, D, D', C') in two parts which each line entirely in R_1 and R_2, respectively. This completes the proof for $x_2 < t$.

The situation that $z_2 > t$, i.e. the red part of the figure is moved above the (solid black) horizontal line $x_2 = t$, follows by symmetry. The remaining case, $x_2 > t > z_1$, can be handled by splitting in two the rectangle horizontally at the x-axis, for each of which we know Eq. C.1 holds. Since supermodularity is also additive when splitting a rectangle horizontally, we have completed the proof.

Appendix D. Proof of Theorem 4

We prove that $f(x_1, x_2; t)$ is multivariate stochastically increasing (Shaked and Shanthikumar, 2007, Definition (6.B.1)) in the parameter t, where

$f(x_1, x_2; t)$ is given by Eq. 3.6. This implies Eq. 3.16, i.e. the marginals are also stochastically increasing (Shaked and Shanthikumar, 2007, Theorem 6.B.16 (c)).

We prove that the conditions in Theorem 6.B.8 in (Shaked and Shanthikumar, 2007) are satisfied. First, it follows from the MTP_2 property that (X_1, X_2) is "associated" in the sense of the theorem (Karlin and Rinott, 1980, Eq. (1.7)). Define the function $h(x_1, x_2) = \log[f(x_1, x_2; t')/f(x_1, x_2; t)]$. The main condition of Theorem 6.B.8 is that $h(x_1, x_2)$ is increasing in (x_1, x_2) for $t' > t$. To verify this we will check that

$$\frac{\partial h}{\partial x_1} \geq 0 \quad \text{and} \quad \frac{\partial h}{\partial x_2} \geq 0, \tag{D.1}$$

which together with the continuity of h is sufficient.

We build on the proof of Theorem 3, and denote the six regions of Fig. 7 associated with t' by $R_1', \ldots, R_6'$. We need to verify Eq. D.1 when $(x_1, x_2) \in R_j \cap R_k'$ for different values of j and k. When $j = k$ it follows that $h(x_1, x_2) = 0$ which implies Eq. D.1. Next, due to the fact that $t \leq t'$ many combinations of j and k cannot occur, and we are left with the following list to check:

(x_1, x_2)	$h(x_1, x_2)$
$R_4 \cap R_5'$	$(\phi + \beta)x_1$
$R_4 \cap R_6'$	$(\phi + \beta)x_1 + 2\beta x_2$
$R_5 \cap R_6'$	$2\beta x_2$
$R_3 \cap R_2'$	$(\phi + \beta)x_2$
$R_3 \cap R_1'$	$(\phi + \beta)x_2 + 2\beta x_1$
$R_2 \cap R_1'$	$2\beta x_1$

For all of these combinations (D.1) holds. Note that we have skipped additive terms in h that does not depend on x_1 or x_2. This completes the proof.

INGVILD M. HELGØY
HANS J. SKAUG
DEPARTMENT OF MATHEMATICS,
UNIVERSITY OF BERGEN, P.O. BOX 7803,
N-5020 BERGEN, NORWAY
E-mail: Ingvild.Helgoy@uib.no
Hans.Skaug@uib.no

Paper received: 17 September 2020; accepted 28 May 2021.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230867471 (print)
9788230842065 (PDF)