Vebjørn Rekkebo

Extending quantitative genetic models to estimate mutational variance

Master's thesis in Industrial Mathematics
Supervisor: Stefanie Muff
June 2021

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Vebjørn Rekkebo

Extending quantitative genetic models to estimate mutational variance

Master's thesis in Industrial Mathematics
Supervisor: Stefanie Muff
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Acknowledgements

First, I want thank my supervisor, Stefanie Muff, for being extremely helpful during all stages of my Master's thesis. She has introduced me to a new, highly interesting field of applied statistics. Our regular meetings have brought many constructive discussions and helped me continuously progress towards the finish line. The last year, she has dedicated a lot of time and effort to help me, and this thesis could not have been finished without her.

Secondly, I want to thank Jane Reid for providing an amazing data set and an already working model. She also lent me her book for almost a year, more or less voluntarily. Thanks also to the rest of the people at the Centre for Biodiversity Dynamics, who gave feedback along the way. Even though we never met, I would like to thank Lukas Keller for coming up with the idea for the research in this thesis. I also want to thank Matthew Wolak for implementing some of the functions used in my work.

Finally, I want to express my appreciation for my family and friends, who have regularly brought me out of my cave so I could stay in touch with the world during the pandemic. This semester would not have been the same without their efforts.

Summary

This thesis introduces extensions to the standard animal model, a type of generalized linear mixed model in the field of quantitative genetics. The model makes use of the information in a known pedigree structure of animal or plant populations to disentangle their phenotypic variation into environmental and additive genetic effects, as well as examining the influence of other factors on the phenotypic trait. In wild study populations the animal model helps to gain specific knowledge that is particularly needed in the context of conservation, for example to quantify evolutionary responses to both natural and artificial processes.

Mutations have been suggested as one explanation for the continued response in long-term selection experiments in which it is expected that a selection plateau has been reached. As a tool to explore this theory, we suggest and investigate a method to separate mutational variance from other sources of additive genetic variance in the animal model, based on the already known pedigree structure.

As an example, we fit an animal model including mutation effects with data from a song sparrow population on the Mandarte island in Canada, using the Bayesian frameworks INLA and Stan. Moreover, a resampling method is used to look at temporal changes in random effects and the corresponding variances. As expected from previous insight from other populations, the estimated increase in variance, from one generation to the next, accounts for a minor part of the total phenotypic variance, but resampling reveals a rapid increase over time. This suggests a surprisingly large inflow of additive genetic variance from mutations, but there are signs of overestimation. Further work on the subject should include testing the model on simulated data, likely unveiling confounding between mutational variance and other additive genetic effects.

Sammendrag

Dette prosjektet introduserer utvidelser av dyremodellen, en type generalisert lineær blandet modell innen kvantitativ genetikk. Modellen utnytter informasjonen i en kjent slektskapsstruktur for dyre- eller plantepopulasjoner for å dele deres fenotypiske variasjon inn i miljøeffekter og additive genetiske effekter, samt å undersøke påvirkningen andre faktorer har på det fenotypiske trekket. I ville studiepopulasjoner kan dyremodellen hjelpe til med å skaffe spesifikk kunnskap som spesielt trengs innen konservering, for eksempel ved å kvantifisere evolusjonære responser på både naturlige og kunstige prosesser.

Mutasjoner er foreslått som en forklaring på den kontinuerlige responsen i langvarige seleksjonseksperimenter hvor det forventes at et seleksjonsplatå er nådd. Som et verktøy til å utforske denne teorien, foreslår og undersøker vi en metode for å separere mutasjonsvarians fra andre kilder til additiv genetisk varians i dyremodellen, basert på den kjente slektskapsstrukturen.

Som et eksempel, inkluderer vi mutasjonseffekter i en dyremodell med data fra en sangspurv-populasjon på øya Mandarte i Canada, implementert i de Bayesiske rammeverkene INLA og Stan. I tillegg benytter vi en resamplingsmetode for å se på årlige endringer i tilfeldige effekter og de tilsvarende variansene. Som forventet fra tidligere innsikt fra andre populasjoner, står den estimerte økningen i additiv genetisk varians, fra en generasjon til den neste, for en mindre del av den totale fenotypiske variansen, men resamplingen avdekker en rask økning over tid. Dette indikerer et overraskende stort tilskudd av additiv genetisk varians fra mutasjoner, men det finnes tegn til overestimering. Videre arbeid bør inneholde testing på simulerte data, noe som trolig vil vise at mutasjonsvariansen ikke er fullstendig separert fra andre additive genetiske effekter.

Table of Contents

Chapter 1

Introduction

Population genetics is a field within evolutionary biology where genetic differences between and within populations are studied (Conner and Hartl 2004). Genetic differences appear as results of the four evolutionary forces selection, genetic drift, gene flow and mutations. The interest lies in detecting these forces in populations, understanding their past impact and predicting how populations will be affected in the future. One approach to study population genetics is by the tools developed within the subfield of quantitative genetics (Lynch and Walsh 1998). Methods in quantitative genetics are based on predictions and summary statistics for related individuals, rather than knowledge on specific genetic material in single individuals.

Quantitative genetics is built around the assumption that many traits follow the infinitesimal model (e.g. Barton et al. 2017). That is, traits are quantitative and do not fall into distinct categories, but their values are assumed to be affected by an infinite number of genetic components, each with an infinitely small additive effect. A common goal is to disentangle the total variation in a trait into separate parts caused by either environmental or genetic components. A measure of the genetic diversity in focal phenotypic traits, makes it possible to predict a population's response to selection or potential for adaptation to new environmental factors. By Fisher's Fundamental Theorem of Natural Selection, the rate of change in fitness in a population is equal to the additive genetic variance in fitness (Fisher 1930), therefore many argue that a population at equilibrium should have no additive genetic variation in fitness (e.g. Kimura 1958). Despite this, high levels of inherited additive genetic variance are consistently found in traits under selection (Lynch and Walsh 1998). This variation is believed to be maintained partly by mutations and partly by balancing selection, but their relative importance are not known (Barton and Keightley 2002).

A central model in quantitative genetics is a type of generalized linear mixed models, the so-called *animal model* (e.g. Kruuk 2004, Wilson et al. 2010). The animal model is often used as a tool to disentangle additive genetic variance and environmental variance in both domestic and wild populations, based on the relationships between individuals. In this project we will focus on an extension to the animal model that allows for estimation of new additive genetic variance per generation due to newly emerging mutations (Wray 1990). Estimating mutational variance in wild populations can contribute to understanding the maintained additive genetic variance in fitness traits.

As an example of applying the mutational animal model, we perform a quantitative genetics analysis with data from song sparrows (Melospiza melodia) on the small Mandarte Island. The song sparrow population has been monitored since 1975 by researchers from the University of British Columbia, Canada (Smith et al. 2006). They have built an almost complete pedigree of the sparrow population over years, which is necessary to properly estimate mutation variance. Due to consistent immigration, we apply a genetic groups extension (Wolak and Reid 2017, Muff et al. 2019) to the model, separating native and immigrant individuals. Moreover, we fit the model in the Bayesian frameworks INLA (Rue et al. 2009) and Stan (Stan Development Team 2021), and utilize posterior samples to model temporal changes in the population (Sorensen et al. 2001). The main interest lies in estimating the mutational variance and investigating how including mutation effects influences other estimates.

Chapter 2

Background

In this chapter we introduce the most important statistical concepts and the relevant background in quantitative genetics that is needed to understand the methodological extensions proposed and applied in later chapters.

2.1 Generalized linear mixed models

Generalized mixed models (GLMMs) are an extension of the generalized linear models (GLMs) (Zuur et al. 2009), which allows for a mix of fixed and random effects, thereby *mixed* models. Random effects do not take determined values, but represent the deviations around the expected value determined by fixed effects. Thus, fitting a GLMM does not only involve estimating the fixed effects, but also the distributional parameters of random effects. A general GLMM can be defined using vector notation. Letting the response vector $\boldsymbol{y}$ be linked to the linear predictor $\boldsymbol{\eta}$ through some link function g, the GLMM is given as

$$g(\mathrm{E}[\boldsymbol{y}]) = \boldsymbol{\eta} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \ , \tag{2.1}$$

where $\mathrm{E}[\boldsymbol{y}]$ is the conditional expectation of $\boldsymbol{y}$, $\boldsymbol{\mu}$ denotes the intercept vector, $\boldsymbol{\beta}$ is a vector of fixed effects and $\boldsymbol{\gamma}$ is the random effect vector with some given multivariate distribution. The design matrices $\mathbf{X}$ and $\mathbf{Z}$, for fixed and random effects respectively, are built to correctly relate the effects to the response.

To show the purpose of random effects we take a look at an example. Imagine a study on an animal population where we have taken several measurements from each individual. Repeated measurements are naturally not independent from each other and thus may violate the assumption of independent residuals in the GLM (Zuur et al. 2009). One approach to avoid this violation is to only use the mean value of each individual. With this approach we lose power compared to a model where every individual is contributing several measurements. Another possibility is to include identities as a categorical fixed effect with one level per individual. However, the latter approach reduces the degrees of freedom for every level included. Moreover, the main interest in a study often lies on the overall effects in a population, rather than in specific individuals.

As a third option, we can introduce a random identity effect γ_i, for individual i. In the simplest case we assume $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$ to be independent and identically distributed between different individuals i. Now, instead of estimating a fixed effect based on means or a categorical variable for each individual, we estimate one parameter, σ_γ^2, which is the between-individual variance in the population conditioned on the fixed effects. In conclusion, random effects are suitable when different observations are not independent, that is, when we know there is some covariance structure in the data. Other types of covariance structures than those imposed by repeated measurements exist, some of which are central in the following sections.

2.2 The animal model

A specific type of linear mixed models is the animal model (e.g. Kruuk 2004, Wilson et al. 2010). A simple linear animal model for individual i's continuous phenotypic trait y_i, only including the intercept μ, the *breeding value* a_i and residual error ϵ_i, can be stated as

$$y_i = \mu + a_i + \epsilon_i \ ,$$

where we assume $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The breeding value is a random effect based on the relatedness between individuals, and it is the defining feature of the animal model. In a population of size N, the breeding value a_i for a quantitative phenotypic trait is the total additive effect of an individual's genotype on the trait expressed relative to the population mean phenotype (Wilson et al. 2010). Since close relatives are likely to share large parts of their genotype, covariance between breeding values in a population must be accounted for. With access to a pedigree, we can therefore utilize the animal model to quantify effects of genotypes without the need of genotypic data.

Applying Mendelian laws of inheritance, we can build the covariance structure of breeding values from information on relatedness between individuals. Letting $\boldsymbol{a}$ be the vector of breeding values following a multivariate normal distribution, we can write

$$\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\sigma_a^2) \ ,$$

where σ_a^2 is the population's additive genetic variance and $\mathbf{A}$ is an $N \times N$ additive genetic relatedness matrix. The elements of $\mathbf{A}$ are given by two times the coefficient of coancestry, that is $A_{ij} = 2\Theta_{ij}$ (Kruuk 2004). In other words, the ijth entry of $\mathbf{A}$ is twice the probability that an allele drawn at random from individual i is identical by descent to one drawn at random from individual j. Given no inbreeding, $2\Theta_{ij}$ equals 1 for $i = j$, $1/2$ for parent-offspring or full-sibling pairs and so on. Inbreeding refers to the situation where mating individuals are closely related, which means the total amount of unique genetic material within the pair is smaller than expected in an unrelated pair. Hence, in the case of inbreeding, the probability of alleles being identical by descent will necessarily be larger (Wilson et al. 2010). Specifically, the diagonal elements are given by $A_{ii} = 1 + F_i$, where F_i denotes individual i's *coefficient of inbreeding*, a measure of how inbred i is (Wright 1922).

The main interest from an animal model is often the estimation of the additive genetic variance σ_a^2. It serves as a scaling factor in the covariance structure of breeding values and can be interpreted as the part of the variance in an individual's phenotype caused by additive genetic effects (Kruuk 2004). Non-additive genetic effects, such as those due to dominance or epistasis, are extremely difficult to estimate in non-experimental settings and are usually neglected in wild populations (Wilson et al. 2010).

There are usually other sources of covariance that should be accounted for to obtain a valid estimate of σ_a^2. If available, these should be included in the animal model as random or fixed effects. Such sources include simple correlating effects like time of measurement and individual traits like sex, but also environmental effects that can falsely be interpreted as additive genetic effects (Kruuk and Hadfield 2007). An example of such environmental effects is *common environmental effects*. Confounding with additive genetic variance happens if individuals sharing similar environments are more related to each other than to individuals with different environments (Wilson et al. 2010). This can typically be birds bound to their nests, but is also relevant in many other species where the surroundings have an impact in early stages of life.

Including K fixed effects and L additional random effects, a general animal model for individual i can be stated as follows. Let η_i be the linear predictor linked to the continuous phenotypic trait y_i through $\eta_i = g(\mathrm{E}[y_i])$, where g is some link function and $\mathrm{E}[y_i]$ is the conditional expectation of y_i. Let x_{ik} denote the measurement of fixed effect $k \in \{1, \ldots, K\}$ and μ be the model intercept. Let $\boldsymbol{\gamma}^{(l)}$ be a vector for random effect number l, where $l \in \{1, \ldots, L\}$, and let $\gamma_i^{(l)}$ denote said effect for the group to which individual i belongs. In the simplest case, the L random effects are assumed to be independently normally distributed between each group with zero mean, so we let $\gamma_i^{(l)} \sim \mathcal{N}(0, \sigma_l^2)$ be the additional random effects. Then we can write

$$\eta_i = \mu + \sum_{k=1}^{K} x_{ik}\beta_k + \sum_{l=1}^{L} \gamma_i^{(l)} + a_i \ .$$

The matrix form of this formula is equal to Equation 2.1 with breeding values included in the random vector. With all random effects assumed to be normally distributed with zero mean, the conditional means and variances are given by

$$\mathrm{E}\left[\eta_i | \boldsymbol{x_i}\right] = \mu + \sum_{k=1}^{K} x_{ik}\beta_k \quad \text{and} \quad \mathrm{Var}\left[\eta_i | \boldsymbol{x_i}\right] = \sum_{l=1}^{L} \sigma_l^2 + \sigma_a^2 \ .$$

Note that fixed effects are expected to explain some of the total variance, which means their inclusion changes the interpretation of this variance measure. For example a model including a fixed effect for sex will estimate the variance after accounting for a systematic sex effect, and thus giving a smaller variance estimate compared to a model without any fixed effects (Wilson 2008).

For the sake of comparability between different species, populations and traits, the additive genetic variance is often scaled to achieve the trait's heritability. In the simple case of a Gaussian trait, the heritability is defined as

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2} \ , \tag{2.2}$$

where σ_p^2 is the total phenotypic variance, that is, the sum of all variance components (e.g. Conner and Hartl 2004). Heritability is thereby the proportion of the total phenotypic variance that is due to additive genetic causes. Animal breeders are particularly interested in this measure because it is an important factor when predicting the response to selection, but it can also say something about a population's ability to adapt to rapid changes in the environment.

For a model with non-Gaussian traits, where we use a link function, the parameters are computed on a latent scale corresponding to the linear predictor $\boldsymbol{\eta}$ instead of the scale of the trait $\boldsymbol{y}$. A consequence of computing variances on a latent scale, with the inverse link function being non-linear, is generating non-additive genetic variance on the data scale. As a first step, calculating the heritability on a "liability scale" can be done by adding a so-called "link variance" to the other variance components in the denominator of Equation 2.2 (de Villemereuil 2018). For the probit link function, the link variance corresponds to the variance of a value drawn in a standard normal distribution, which is 1. Thus, with a probit animal model the heritability on a liability scale is

$$h^2_{\text{liab}} = \frac{\sigma_a^2}{\sigma_p^2 + 1} \ . \tag{2.3}$$

To transform the heritability to the observed data scale we can apply

$$h^2_{\text{obs}} = \frac{t^2}{p(1-p)} h^2_{\text{liab}} \ , \tag{2.4}$$

where p is the proportion of the focal binary phenotypic trait and t is the density of a standard normal distribution at the pth quantile (de Villemereuil 2018). Note that there are different practices for estimating heritability, which can lead to wrong comparisons between different studies. Moreover, the concept of scaling σ_a^2 by the phenotypic variance has been strongly criticized (see e.g. Hansen et al. 2011). The fixed effects included in a model will for example affect variance estimates, and thus heritability estimates are dependent on the model design.

2.3 Genetic groups extension

All relatedness measures in the animal model are computed relative to a defined base population (Lynch and Walsh 1998). The base population typically consists of imaginary "phantom parents" of all individuals whose true parents are unknown or not identified in the pedigree (Quaas 1988). Phantom parents are assumed to share the same genetic parameters, be entirely unrelated and each only having one offspring (Wolak and Reid 2017). As a consequence, σ_a^2 is the additive genetic variance of the base population, and not the population as a whole.

The base population is not exclusively parents of individuals from the earliest generation in the pedigree (founder population). Missing parents in a pedigree also arise in later generations due to failure of observation or immigration. However, depending on the study system, non-founders with unknown parents may have systematically different genotypes to the founder population and thus violate model assumptions (Wolak and Reid

2017). An example would be consistent immigration from a nearby population, adapted to a different local environment, where immigrants have lower fitness in their new habitat compared to native individuals. This would lead to estimated breeding values and additive genetic variances being biased towards values among immigrants.

It is possible to account for this systematic bias by defining different *genetic groups* for which the base population gets partitioned into (e.g. Wolak and Reid 2017, Muff et al. 2019). Instead of assuming all breeding values to have expected value zero, different genetic groups can have different means. Phantom parents are assumed unrelated and can only be members of one group each, while their descendants can partially inherit memberships from different groups as the genotypes mix through between-group mating. Define q_{ir} as individual i's proportion of membership to group r (Wolak and Reid 2017). Then phantom parents have $q_{ir} = 1$ if they belong in group r and $q_{ir} = 0$ otherwise. If individual i is not a phantom parent, q_{ir} is the mean of its two parents' membership proportions in group r. Consequently, group memberships are inherited through generations. Note that, in this thesis, genetic groups are assumed to have equal amounts of genetic variance.

Having genetic groups that allow for different means, it is useful to introduce the "total additive genetic effect" u_i (Wolak and Reid 2017). It is defined by

$$u_i = \sum_{r=1}^{R} q_{ir} g_r + a_i \ ,$$

where R is the number of genetic groups and g_r is the mean of group r, or the *genetic group effect*. The breeding value a_i can now be interpreted as i's deviation from the expected value according to its composition of group inheritance. Letting $\mathbf{Q}$ be an $N \times R$ matrix with q_{ir} as elements and $\boldsymbol{g}$ be the vector of genetic group effects, total additive genetic effects are distributed as $\boldsymbol{u} \sim \mathcal{N}(\mathbf{Q}\boldsymbol{g}, \mathbf{A}\sigma_a^2)$ (Wolak and Reid 2017). Note that we introduce in the next section an additional additive genetic effect, which means u_i will no longer be the total additive genetic effect and σ_a^2 does not capture all additive genetic variance. Nevertheless, we will stick with these terms instead of introducing new ones.

Estimating g_r is most easily done by explicitly estimating each group effect as a fixed effect in the animal model. However, we need to constrain one group's mean additive effect equal to zero, or we will have an infinite number of solutions, that is, an identifiability problem. For this purpose we have to choose a reference group with $g_r = 0$. That is, other groups' genetic effects will denote the deviation in mean total additive genetic effect from the reference.

2.4 Mutation effects

Although there exist methods for estimation of the additive genetic effect of mutational variance, additive genetic effects are only very rarely split into standing genetic variance and mutation effects (but see Wray 1990; Casellas and Medrano 2008). This may lead to upwards bias in the additive genetic variance, especially in long-term selection experiments (see e.g. Casellas et al. 2010). In addition, the inclusion of mutational effects naturally offers means to quantify mutation variance in populations.

Inclusion of mutational effects as individual random effects in the animal model is based on a set of assumptions (Wray 1990). First, individuals in the base population are assumed to have no mutational effects because any mutations in their genome contribute to the base additive genetic effects. Second, in individuals of the first and subsequent generations, new mutations are assumed to arise independently. And finally, mutational effects are assumed to be small with mean zero and contributing a new additive genetic variance of σ_m^2 per generation, if inbreeding effects are ignored. The covariance structure of mutational effects $\boldsymbol{m}$ is closely related to the one of breeding values. Let the mutational covariance matrix be $\mathbf{M}\sigma_m^2$. With t being the total number of generations, $\mathbf{M}\sigma_m^2$ can be partitioned as

$$\mathbf{M}\sigma_m^2 = \sum_{k=1}^{t} \mathbf{A}_k \sigma_m^2 ,$$

where $\mathbf{A}_k$ is the covariance matrix of additive effects attributed to mutations arising in generation k. The elements of $\mathbf{A}_k$ are the additive genetic relationships if ancestors born in generations 0 to $k-1$ are ignored (Wray 1990). This design ensures that mutational effects, independently arising in each generation, are inherited like other additive genetic effects.

A challenge with the construction of the mutational covariance matrix $\mathbf{M}$ (or its inverse, $\mathbf{M}^{-1}$) is dividing populations into t non-overlapping generations k. Due to mechanisms such as inbreeding, extra-pair paternity and full-siblings being born in different years, generations are often overlapping and difficult to separate from each other. Luckily, according to Matthew Wolak (email communication, July 7, 2020), methods based on algorithms for $\mathbf{A}^{-1}$ (Quaas 1976; Meuwissen and Luo 1992) do not need defined generations to construct $\mathbf{M}^{-1}$, but instead trace each individual one-by-one to determine each ancestor's contribution to identical-by-descent mutational effects shared between two individuals. Consequently, two full-siblings can be born in different years, but would will share the same mutational effect relatedness as full-siblings born in the same year. On the other hand, this method is more vulnerable to missing parents in the pedigree due to observation failure or genotyping uncertainty. Moreover, since this approach does not directly depend on generations, exact interpretation of σ_m^2 is complicated.

Similar to the additive genetic variance, estimating *mutational heritability* in a Gaussian trait can be done by the formula

$$h_m^2 = \sigma_m^2 / \sigma_p^2 , \tag{2.5}$$

and potentially transformed similar to h^2 in Equation 2.3 and 2.4. However, earlier estimates of mutation variance have often been reported on the form σ_m^2/σ_e^2, where σ_e^2 denotes the total environmental variance (that is the sum of non-genetic variance components). These estimates have ranged from $1 \cdot 10^{-4}$ to $5 \cdot 10^{-2}$ (Lynch 1988). Estimates for mutational variance are small because σ_m^2 only estimates the increment of the per generation variation due to new mutations. Note that mutational variance accumulates over generations and its contribution to the genetic variation in populations thus increases over time.

2.5 Bayesian inference, MCMC and INLA

In the classical, frequentist approach to statistics, the parameter(s) $\boldsymbol{\theta}$ is considered to be some unknown, but fixed, value. In that approach, $\boldsymbol{\theta}$ is thus entirely estimated by the data, that in some way is generated by $\boldsymbol{\theta}$. In the Bayesian approach however, we utilize our prior knowledge to $\boldsymbol{\theta}$ by giving it a *prior distribution* $p(\boldsymbol{\theta})$ before doing any other estimation. Giving parameters a distribution may represent the fact that they are truly varying and/or reflect that our knowledge of the parameters is imperfect. Either way, it provides an additional layer of flexibility. When observing new data $\boldsymbol{y}$, we get more information about $\boldsymbol{\theta}$ and obtain a *posterior distribution* $p(\boldsymbol{\theta}|\boldsymbol{y})$. Given a likelihood of the data $p(\boldsymbol{y}|\boldsymbol{\theta})$, the posterior distribution is given by Bayes theorem to be proportional to $p(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$, that is, the likelihood times the prior.

2.5.1 Markov chain Monte Carlo methods

Fitting Bayesian models is most commonly done with Markov chain Monte Carlo (MCMC) methods. The purpose of a general MCMC algorithm is to draw samples from some target density $p(\boldsymbol{\theta})$. The idea of an MCMC sampler is to construct a Markov chain $\{\Theta_i\}_{i=1}^{\infty}$ so that $\lim_{i\to\infty} \Pr(\Theta_i = \boldsymbol{\theta}) = p(\boldsymbol{\theta})$, that is the chain converges to the target distribution. After a sufficient amount of iterations i the transitions of the Markov chain $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}, \ldots$ essentially form a sample from $p(\boldsymbol{\theta}|\boldsymbol{y})$. In the case of a regression model, the algorithm draws samples of the different parameters $\boldsymbol{\theta}$ based on their priors $p(\boldsymbol{\theta})$ and the likelihood of the data $p(\boldsymbol{y}|\boldsymbol{\theta})$ with the aim to obtain approximate posterior distributions $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Stan is a state-of-the-art platform for statistical modeling and high-performance statistical computation (Stan Development Team 2021). One of Stan's applications is a variant of MCMC called Hamiltonian Monte Carlo (HMC). HMC uses an approximate Hamiltonian dynamics simulation based on numerical integration to generate more efficient transitions in the Markov chain (see Betancourt and Girolami 2013). Moreover, Stan uses the "no-U-turn sampling" (NUTS) algorithm for automatic parameter tuning (see Hoffman and Gelman 2014). The parameter tuning is done during "warmup iterations" (or burn-in), and provides approximately optimized transitions in the following iterations, from which the posterior samples are taken. For efficient computations, Stan is implemented with `C++`, allowing several chains to run in parallel on different cores, which also reduces autocorrelation in the resulting samples. Wrappers in `R` such as `rstanarm` and `brms`, makes Stan a user-friendly and reliable platform for Bayesian problems.

2.5.2 Latent Gaussian models and INLA

Despite great developments in the recent years, MCMC methods can still be slow and impose issues with both convergence and mixing, especially with large data sets. For a large class of models, using *integrated nested Laplace approximations* (INLAs) has in the last decade become a popular alternative (Rue et al. 2009). INLA relies on a combination of analytical approximations and efficient numerical integration schemes to achieve highly accurate deterministic approximations to posterior quantities of interest (Rue et al. 2009). Benefits of using INLA over MCMC are mainly its fast computation even for large models, but also that INLA does not suffer from slow convergence or bad mixing in generated samples.

One condition for using INLA is that the model needs to be a latent Gaussian model (LGM). This is a wide class of models containing GLMs, GLMMs, time series, spatial models and several more (Rue et al. 2009). An LGM consists of three elements: a likelihood function

$$\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta} \sim \prod_i p(y_i|\eta_i(\boldsymbol{z}),\boldsymbol{\phi}) \ ,$$

a latent Gaussian field

$$\boldsymbol{z}|\boldsymbol{\phi} \sim p(\boldsymbol{z}|\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}(\phi)) \ ,$$

and the hyperpriors

$$\boldsymbol{\phi} \sim p(\boldsymbol{\phi}) \ .$$

Here, $\boldsymbol{z}$ is the latent Gaussian field (can be interpreted as the joint distribution of the parameters in the linear predictor) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Moreover, $\boldsymbol{\phi}$ is a vector of hyperparameters and $\eta_i(\boldsymbol{z})$ is the ith linear predictor connecting the data y_i to the latent field through some known link function.

The linear predictor with K fixed effects β_k and corresponding data x_{ik} can be expressed in a general form as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} x_{ik}\beta_k + \sum_{l=1}^{L} f_l(v_{il}) + \epsilon_i \ ,$$

where β_0 is the intercept, ϵ_i is the residual error and $\boldsymbol{f}$ is a set of functions on corresponding covariates $\boldsymbol{v}$. These functions can take many forms, for example random effects with Gaussian priors. For more information on the INLA computing scheme we refer to Rue et al. (2009).

2.5.3 Penalized complexity priors

The choice of priors in models is a highly debated topic in Bayesian statistics. A common practice in animal models is, in the absence of prior knowledge, to use for example gamma distributions with small parameters. Such priors are assumed to be uninformative, however this is not necessarily the case (Lambert et al. 2005). Moreover, many popular prior distributions have parameters that are not intuitive for the user.

A recently introduced alternative are penalized complexity (PC) priors, which are robust and intuitive in their use (Simpson et al. 2017). These priors are designed to penalize deviation from a simple defined base model, based on the Kullback-Leibler divergence (Kullback and Leibler 1951). The user needs only specify the two parameters α and U, which decide how much prior weight is assigned to certain values of the parameter θ, below the chosen threshold. With $0 < \alpha < 1$, the parameter's prior probability is given by Pr$(\theta > U) = \alpha$. As an example, let the prior be PC$(1, 0.05)$. Then the prior distribution assigns 5% of the weight on $\theta > 1$, that is Pr$(\theta \leq 1) = 0.95$. This demonstrates how PC priors are intuitive in their use, as opposed to most classical choices. In addition, the `R-INLA` package (Rue et al. 2009) provides the necessary functionality for the application of PC priors to be easy to implement, while Stan requires some more work from the user.

For this thesis, it is most interesting to look at a random effect $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{R}^{-1})$, where $\mathbf{R^{-1}}$ is some known covariance structure and $\tau^{-1} = \sigma_\gamma^2$ is the variance parameter of interest. Given the penalty parameter $\lambda > 0$, the PC prior for τ writes

$$p(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp(-\lambda\tau^{-1/2})$$

(Simpson et al. 2017). Both INLA and Stan require priors for the standard deviation σ_γ, and a change of parameter to σ_γ yields

$$\begin{aligned} p(\sigma_\gamma) &= \frac{\lambda}{2}(\sigma_\gamma^{-2})^{-3/2}\exp\left(-\lambda(\sigma_\gamma^{-2})^{-1/2}\right) \cdot \left|\frac{-2}{\sigma_\gamma^3}\right| \\ &= \lambda\exp(-\lambda\sigma_\gamma)\ , \end{aligned}$$

which is the exponential distribution with rate λ. Thus, for the standard deviation of a random effect, the PC prior is equivalent to the exponential distribution with $\lambda = -\ln(\alpha)/U$. This property makes the PC priors for specific parameters possible to implement also with Stan and other frameworks that do not inherently provide such functionality. A set of PC priors for the standard deviation σ_γ with $U = 1$ and $\alpha \in \{0.01, 0.15\}$ are shown in Figure 2.1. The figure illustrates how most weight is put on values close to 0, and how the α parameter controls how much weight is put on each side of $U = 1$. Similarly, it can be shown that the PC prior for a random effect variance σ_γ^2 is given by

$$p(\sigma_\gamma^2) = \frac{\lambda}{2}(\sigma_\gamma^2)^{-1/2}\exp\left(-\lambda(\sigma_\gamma^2)^{1/2}\right)\ ,$$

with $\lambda = -\ln(\alpha)/U^{1/2}$. Notice that $p(\sigma_\gamma^2)$ goes towards infinity when σ_γ^2 approaches 0. Figure 2.1 illustrates how $p(\sigma_\gamma^2)$ is steeper close to 0 than $p(\sigma_\gamma)$ is, penalizing values for deviating from 0. Moreover, $p(\sigma_\gamma^2)$ has a long narrow tail, potentially allowing for more extreme values. Changes to the α parameter lead to relatively small differences in $p(\sigma_\gamma^2)$ visually, but small changes may still significantly affect the posteriors.

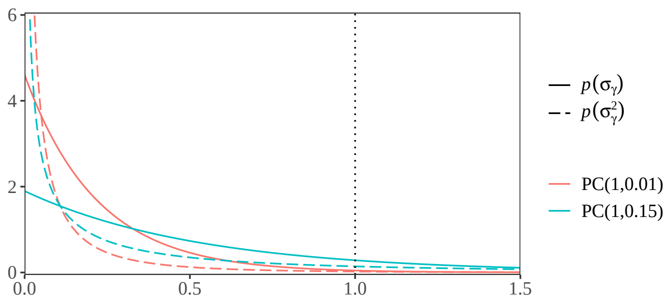

Figure 2.1: Penalized complexity priors PC$(1, \alpha)$ for standard deviation $p(\sigma_\gamma)$ (solid lines) and variance $p(\sigma_\gamma^2)$ (dashed lines) of a random effect γ. The vertical dotted line illustrates that 1 is always the $1 - \alpha$ quantile in PC$(1, \alpha)$. The y-axis is cut at $y = 6$ because the variance priors go towards infinity when σ_γ^2 is close to 0.

Chapter 3

Methods

The basis for this project are the data and animal model used by Reid et al. (2021) in their analysis of immigration effects on local evolution in fitness. They provide a thorough description of the field system, data set and modelling choices. Here we will describe the most relevant information from their work, together with own modelling choices. We will use a genetic groups animal model with mutational effects on measurements of a binary trait, survival from independence to adulthood, to investigate the effect of mutations in a small song sparrow population (Smith et al. 2006). The population inhabits Mandarte island BC, Canada, and is assumed to be descendants from two genetic groups: natives and immigrants.

3.1 Data description

The Mandarte song sparrows have been monitored continuously since 1975 and the main data set consists of 2478 observations from individuals born in the 26-year long period 1993-2018. Mandarte is a small island (approximately 6 hectares), and adult sparrows typically stay in the same breeding territories across years. This behavior makes the Mandarte song sparrows a suitable study population because it becomes relatively easy to keep track of all individuals. A pair of song sparrows can produce up to 3 broods during the season April-August, with 1 to 4 offspring in each brood. During the study period, the island was systematically surveyed to identify the chicks that survived to independence from parental care, and these constitute our data set. New surveys were done in April each year to identify whether young sparrows lived through their first winter, a measure defined as survival to adulthood (Smith et al. 2006).

Genetic parentage data was collected since 1993, and revealed some wrong parentage assignment in the pre-1993 pedigree (Sardell et al. 2010). Therefore, we only use phenotypic data from individuals born in the period 1993-2018. However, to avoid artificial zero-values from 1993-individuals, relationships from the whole study period are used to calculate inbreeding coefficients and genetic group memberships (details in Reid et al. 2021). Despite no observed emigration from Mandarte (Wilson and Arcese 2008), there was in average approximately 1 immigrant from surrounding islands per year. Molecular genetic analyses verified that immigrants were relatively unrelated to the resident population at the time of arrival (Reid et al. 2021), which justifies a division of the base population into two genetic groups: natives and immigrants. Defined phantom parents of the native population count 15 founders, while there were 33 reproducing immigrants since 1975.

3.2 Model description

Since an extensive animal model on the Mandarte song sparrows had already been constructed by Reid et al. (2021), we chose to include the same fixed and random effects as in the respective publication, but added the mutational effects. As response y_i we used the binary trait, survival from independence to adulthood, which is closely related to fitness. Because we only had two genetic groups, we defined natives as the reference group. We denoted by g the single fixed genetic group effect for immigrants, together with q_i describing individual i's proportion of genes inherited from the immigrant group. Due to known inbreeding depression (Reid et al. 2014) potentially biasing estimates of σ_a^2 (Reid and Keller 2010), we included inbreeding coefficients F_i as fixed effects. Further, since song sparrows often rear multiple broods within each summer, we defined brood date as the date on which the first egg in the individuals clutch was estimated to have been laid (where January 1st is day 1). Additionally, the model included the individual's natal year (numbered as 1 to 44) to account for linear phenotypic changes due to the environmental factors and lastly, the binary effect sex, where 1 was related to males.

Random effects of main interest were the breeding values a_i and mutational effects m_i, but we needed to account for two other effects that were not properly covered by linear fixed effects. Those were non-linear effects of the natal year and the multi-level nest effect (represented by 1109 unique nest IDs), which is a common environmental effect. All random effects were assumed normally distributed with zero mean and covariances $\mathbf{A}\sigma_a^2$, $\mathbf{M}\sigma_m^2$, $\mathbf{I}\sigma_{year}^2$ and $\mathbf{I}\sigma_{nest}^2$, where $\mathbf{A}$ and $\mathbf{M}$ are relatedness matrices defined in sections 2.2 and 2.4 and $\mathbf{I}$ is the $N \times N$ identity matrix. The full model for individual i is then given as

$$\eta_i = \mu + \sum_{k=1}^{5} x_{ik}\beta_k + \sum_{l=1}^{2} \gamma_i^{(l)} + q_i g + a_i + m_i \ ,$$

where η_i is the linear predictor, μ denotes the model intercept and $\gamma_i^{(l)}$ are the random natal year and nest effects respectively. Furthermore, x_{ik} are the measurements corresponding to the mentioned fixed effect coefficients β_k and q_i describes membership to the immigrant genetic group with coefficient g. Since survival to adulthood is a binary trait, a natural choice was to do binary regression with the probit link function. The probit link function is defined by

$$g(y_i) = \Phi^{-1}(y_i) \ ,$$

where $\Phi^{-1}(.)$ corresponds to the inverse of the cumulative distribution function of the standard normal distribution. Note that we here, due to instability of the results when using INLA, deviated from the model of Reid et al. (2021), which was a logit model.

The model was implemented using INLA and Stan, and every variance component needed a suitable prior distribution. In accordance with earlier arguments, PC priors were chosen for every parameter in the model. As a starting point, we fitted a model without mutation effects (model 0) with priors PC$(1, 0.01)$ for each variance component. The resulting posterior marginal distributions provided evidence for variances being smaller than 1. Thus, due to lack of prior knowledge and low prior sensitivity in model 0, we used PC$(1, 0.01)$ for all variance components except the mutational variance in the main model (model 1). For the mutational variance we knew that earlier estimates of σ_m^2/σ_e^2 were for sure smaller than 0.05 (Lynch 1988). Moreover, summing up the modes for environmental variance marginals in the model without mutational variance and the link variance, we could use 1.2 as a proxy for σ_e^2. This relates to an upper bound of the mutational standard deviation of $\sqrt{0.05 \cdot 1.2} = 0.24$. Hence, we chose to apply a PC$(0.25, 0.01)$ prior to σ_m (i.e. priors are set for standard deviations in INLA and Stan). In addition, a set of other PC priors were implemented for mutational variance to explore the effect of different prior assumptions.

Given the data and model formulation, INLA efficiently approximates all posterior marginal distributions. However, INLA (and Stan) has no built-in model that directly enables the use of a mutational covariance structure. Therefore, for computation of the $\mathbf{A}$- and $\mathbf{M}$-matrices, or their inverse, we used the `R`-package `nadiv` (Wolak 2012). Note that at the time of this thesis, functions for computing $\mathbf{M}$ were not yet released (current version 2.17). Thus, because Stan requires the non-inverted covariance matrices, we had to numerically invert $\mathbf{M}^{-1}$ for use in the Stan models.

The exact same model assumptions were used in both the INLA and MCMC implementations. The Stan model was implemented through the `R` package `brms`, which provides a user-friendly interface for Bayesian generalized multilevel modelling. A drawback for using `brms` is that it is not optimized for models with several random effects with a large number of levels. It requires repeatedly computing Kronecker products of the covariance matrices, which is slow for large matrices. Therefore, it was not feasible to run very long Markov chains with this implementation. As a result, we used the Stan implementation mostly as a validation model for the INLA implementation. We ran 4 chains with 10000 iterations each, of which 5000 were warmup iterations, resulting in a posterior sample of $N_s = 20000$ for each model parameter in model 1. In addition, model 0 was run with 4×1000 iterations (500 warmup iterations), generating samples of size $N_s = 2000$. It is likely that 2000 samples was not enough to generate accurate posterior distributions for model 0, but was sufficient to somewhat confirm the trends seen with the INLA implementation.

After running the INLA computation we were provided with marginal posterior distributions of the focal parameters. From the marginals it is straightforward to extract for example, point estimates and credible intervals. However, obtaining posteriors of transformed variables such as the heritability required more work. INLA provides a resampling method with which we could generate samples from the joint posterior distribution, equivalent to samples from an MCMC sampler. Given a sufficiently large sample for the different variance components, we could obtain approximate posteriors of transformed statistics, such as heritabilites. The full implementation in `R` (R Core Team 2020) is presented in Appendix A.

3.2.1 Cohort resampling

Mutational variance σ_m^2 is defined as the increase in additive genetic variance due to mutations from generation 0 (i.e. the base population) to the next generation. This is a measure that is difficult to make use of in practice, especially when we have overlapping generations. Therefore, it is interesting to divide posterior samples of mutational effects – and other genetic parameters – into different cohorts. Sorensen et al. (2001) propose a method for a resampling scheme that lets us generate posterior distributions of genetic parameters in arbitrary cohorts. Because the individuals are all associated with a natal year, choosing cohorts based on the year of birth makes it possible to assess temporal changes in the focal parameters.

The resampling scheme can be described with the mutational effects $\boldsymbol{m}$ as an example. Let N_s be the number of samples, either from resampling with INLA or the Stan Markov chains. For iteration n we have a sample $\boldsymbol{m}_n$, from which we get mutational values $m_{i,n}$ for each individual $i = 1, \dots, N$. Let t relate to a set C_t of individuals from a given cohort, and N_t be the number of individuals in the cohort. Then

$$\bar{m}_n(t) = \frac{1}{N_t} \sum_{i \in C_t} m_{i,n} \qquad \text{and} \qquad \hat{\sigma}^2_{m,n}(t) = \frac{1}{N_t - 1} \sum_{i \in C_t} (m_{i,n} - \bar{m}_n(t))^2$$

estimate the mean mutational value and mutational variance, respectively, in cohort C_t for iteration n in the given sampling method. Note that where Sorensen et al. (2001) would use the population variance to estimate $\hat{\sigma}_m^2$, we instead use the sample variance. This method is repeated for each sample $\boldsymbol{m}_n$ for $n = 1, \dots, N_s$, leaving us with approximate posterior distributions for mean mutation value and mutational variance for the given cohort. For comparison, cohort-wise posteriors distributions for $\bar{a}(t)$, $\bar{u}(t)$ and their respective variances were also computed for each model.

Chapter 4

Results and Discussion

In this chapter we present posterior results for fixed effect coefficients, random effect variances and heritabilites, as well as cohort-wise posterior distributions for genetic parameters from the INLA implementation. Moreover, we present some observations related to accounting for mutation effects in the animal model with INLA and Stan.

4.1 Parameter estimates

From the posterior marginal distributions of the INLA implementation of model 1, we obtained point estimates of fixed effects on the latent probit scale by taking the posterior mean (Table 4.1). Moreover, all estimates are reported with the corresponding 95% highest posterior density (HPD) credible interval (CI). The effect of the inbreeding coefficient was clearly different from zero, and estimated to be -4.29 (95% CI from -6.00 to -2.66). The negative value provided further evidence for inbreeding depression in the song sparrow population. The genetic group effect g was estimated to be -1.14 (95% CI from -2.02 to -0.24), thus, there was evidence that immigrant genome was associated with lower fitness. The posterior marginal of the brood date effect had most weight on small negative values (estimate -0.006, 95% CI from -0.009 to -0.004), suggesting an advantage in hatching early in the season, as opposed to late summer. Moreover, males were more likely to survive through their first year than female juveniles (estimate: 0.28, 95% CI from 0.16 to 0.39). On the other hand, there was no evidence for a linearly decreasing trend of the natal year effect (95% CI from -0.063 to 0.008), even though mean juvenile survival had a clearly decreasing trend over the study period.

As recommended by He and Hodges (2008), we mainly looked at posterior modes for variances, but also reported the posterior mean values in Table 4.1. For mutational variance we obtained $8\text{e}{-4}$ (95% CI from $7\text{e}{-5}$ to 0.0063), which is a very small estimate relative to the largest variance components, but remember that σ_m^2 only relates to the increase in additive genetic variance in one generation and should accumulate over time. The posterior marginal for σ_a^2 spanned larger values, with the 95% CI covering 0.01 to 0.16 (posterior mode 0.04). The posterior mode for nest variance was 0.001, which sug-

Table 4.1: Posterior mean (for fixed effects), posterior mode; mean (for random effect variances) and 95% HPD CI for the animal model on juvenile survival that accounted for mutational variance (model 1) and the model not accounting for mutational variance (model 0) generated with INLA.

Summary of posterior distributions for model parameters

		Model 1		Model 0	
	Parameter	Estimate	95% CI	Estimate	95% CI
Fixed effects					
	F	-4.29	$(-6.00, -2.66)$	-4.24	$(-5.93, -2.62)$
	g	-1.14	$(-2.02, -0.24)$	-1.16	$(-2.05, -0.25)$
	Natal year	-0.026	$(-0.063, 0.008)$	-0.025	$(0.061, 0.009)$
	Brood date	-0.006	$(-0.009, -0.004)$	-0.006	$(-0.009, -0.004)$
	sex	0.28	$(0.16, 0.39)$	0.28	$(0.16, 0.39)$
Variances					
	σ^2_{nest}	$0.001; 0.018$	$(2e{-}5, 0.051)$	$8e{-}4; 0.019$	$(2e{-}5, 0.054)$
	σ^2_{year}	$0.20; 0.23$	$(0.11, 0.38)$	$0.19; 0.23$	$(0.10, 0.38)$
	σ^2_a	$0.04; 0.07$	$(0.01, 0.16)$	$0.04; 0.06$	$(0.01, 0.13)$
	σ^2_m	$8e{-}4; 0.0024$	$(7e{-}5, 0.0063)$		
	h^2	$0.05; 0.09$	$(0.01, 0.23)$	$0.07; 0.09$	$(0.02, 0.18)$
	h^2_m	$0.0010; 0.0036$	$(1e{-}4, 0.0097)$		

gests small variance due to common environmental effects, although the 95% CI spanned substantially larger (and smaller) values (2e−5 to 0.051), reflecting the relatively large uncertainty in the marginal posterior. Unlike for the corresponding fixed effect, the natal year variance was substantial, with a posterior mode of 0.20 (95% CI from 0.11 to 0.38), and was thus the largest variance component, explaining the majority of environmental variance after adjusting for fixed effects.

Estimates of heritabilities were found using the resampling scheme from INLA with 10^5 samples from the joint distribution of each variance parameter. Using the heritability formulas introduced above, we obtained the posterior heritability distributions on the data scale. The posterior mode for h^2 was 0.05 (95% CI from 0.01 to 0.23). This point estimate of h^2 was in the smaller range compared to earlier estimates (e.g. Kruuk 2004). The small value may be an indication of relatively large environmental variation, a consequence of a genetically homogeneous population or the fact that σ^2_a is generally expected to be low in fitness traits when the population is under strong selection. The posterior mode of h^2_m was 0.0010 (95% CI from 1e−4 to 0.0097). Moreover, using the same procedure for σ^2_m/σ^2_e, we obtained a posterior mode of 0.0010 (95% CI from 0.0001 to 0.0101). This estimate falls perfectly inside the range of $1 \cdot 10^{-4}$ to $5 \cdot 10^{-2}$ reported by Lynch (1988). Again, large environmental variation leads to smaller values. Note that it is somewhat unclear whether the transformation to data scale holds also for σ^2_m/σ^2_e. Moreover, scaling variance components by other variance components is often criticized, and thus, these results should not be overinterpreted.

To examine the outcome of including mutational effects, we could compare estimates from model 1 (main model) with estimates from the already fitted model 0 (no mutational effects). Posterior statistics from model 0 are presented on the right side in Table 4.1. The table reveals only minor changes to point estimates for all the modeling components that were included in both models. Furthermore, a comparison of posterior marginal distributions for variance components in the two models is presented in Figure 4.1. The figure shows that posterior marginal distributions for σ^2_{nest} and σ^2_{year} were essentially unaltered between the models. Moreover, the estimate of mutational variance σ^2_m seems to be so small that the distribution of standing additive genetic variance σ^2_a was only marginally altered.

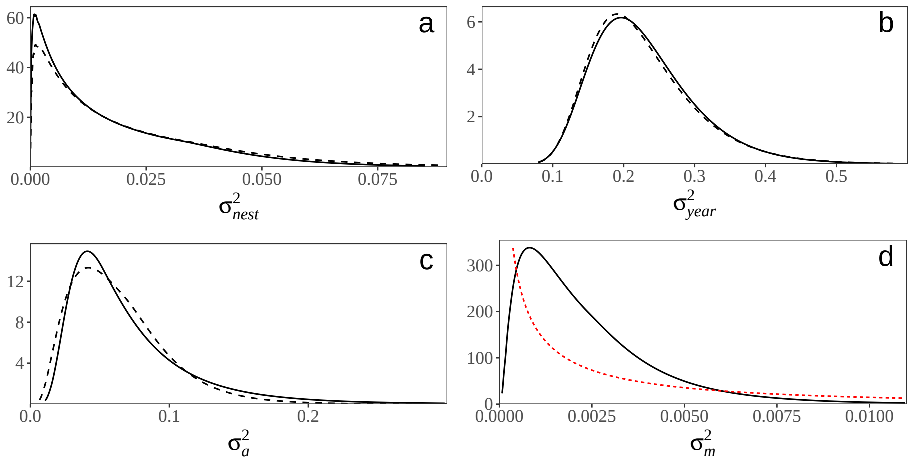

Figure 4.1: Marginal posterior distributions of the variances for the animal model on juvenile survival that accounted for mutational variance generated with INLA. The dashed lines for a,b and c denote distributions for the model without mutational variance, and the dashed red line in d denotes the prior distribution of σ^2_m.

Posteriors for σ^2_m and σ^2_a from the INLA implementation turned out to be highly sensitive to the prior of the former. Relatively small changes in the given prior distribution for σ^2_m led to notable changes in the two marginal distributions (see Appendix B). Priors with more weight on larger values led to right-shifted posteriors for σ^2_m, as one would expect. Still, estimated distributions consistently differed from the prior distribution (see Figure B.1d). The PC-priors penalize values that deviate from zero, and still the posterior marginal distribution for σ^2_m clearly differed from zero. This suggested that the data and covariance structure were quite informative, despite the observed sensitivity to the prior. Moreover, the additive genetic variance σ^2_a appears to be confounded with the mutational variance, that is, a right shift of the posterior marginal for σ^2_a comes together with a left shift of σ^2_m, and vice versa.

4.1.1 Stan implementation

Implementing model 1 in Stan generated very similar posterior marginal distributions to the results from the INLA implementation for most parameters, but there were some deviations (summary of posterior distributions for model parameters are presented in Appendix C). Estimates for the five fixed effects seem reliable since they had close to identical distributions in the two implementations. Approximate posterior marginal distributions for the random effect variances are displayed in Figure 4.2. The most interesting changes, compared to the posterior marginals from INLA in Figure 4.1, were the changes in shape for σ_a^2 and σ_m^2, which seemed to be heavily influenced by the shape of their priors. In Figure 4.2d the prior for σ_m^2 is included, illustrating how its shape is similar to the posterior marginal of σ_m^2. However, the posterior clearly differs from the prior for σ_m^2. Unlike INLA, Stan seemed to have trouble with very small values, arguably putting too much weight close to 0 for all variances, except σ_{year}^2. Hence, the posterior modes for these parameters do not seem reliable. On the other hand, differences in posterior means between the two implementations were not too large, implying some learning from the data also with Stan. Moreover, since we worked with means in the cohort-wise results, both implementations were somewhat useful.

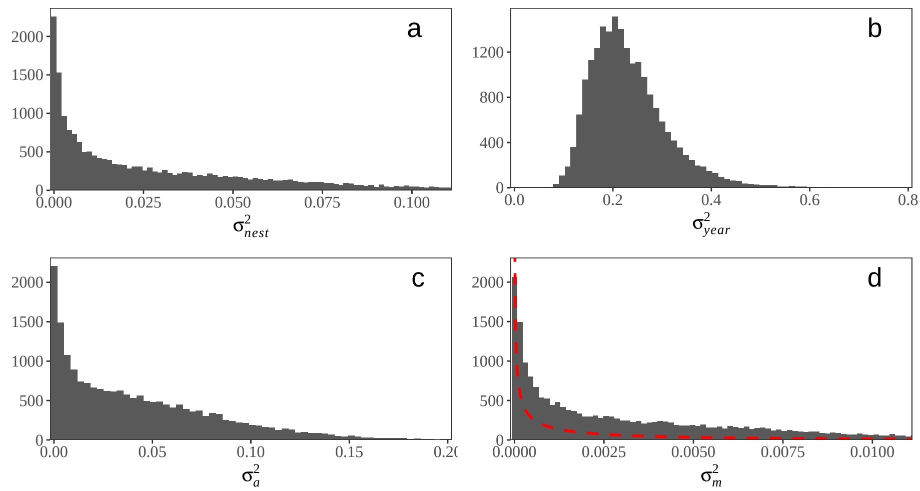

Figure 4.2: Approximate marginal posterior distributions of the random effect variances for the animal model on juvenile survival that accounted for mutational variance implemented with Stan. In d) the prior for mutational variance $p(\sigma_m^2)$ is included as a dashed red line.

4.2 Cohortwise results

Using the posterior samples generated from model 1 with INLA, we found approximate posterior distributions for the genetic parameters in each defined cohort. Figure 4.3a-c illustrates how the distributions of $\sigma_a^2(t)$, $\sigma_u^2(t)$ and $\sigma_m^2(t)$ changed over different cohorts C_t. The additive genetic variance $\sigma_a^2(t)$ was very stable through the years 1993-2018. As seen in Figure 4.3d, the mean of $\sigma_a^2(t)$ had some variation from year to year, but the overall mean $\sigma_a^2(t)$ during the study period was 0.044, which is in accord with the general parameter estimate. The corresponding $\sigma_u^2(t)$ was overall larger than $\sigma_a^2(t)$ and also had larger fluctuations between each cohort. This behaviour is natural because the samples of u_i is the sum of two sampled values, $q_i g$ and a_i, instead of being directly sampled from a set of different σ_u^2. The mean value of $\sigma_u^2(t)$ appeared to vary between approximately 0.05

and 0.08 (Figure 4.3d), and there were weak indications of a slightly decreasing trend in $\sigma_u^2(t)$.

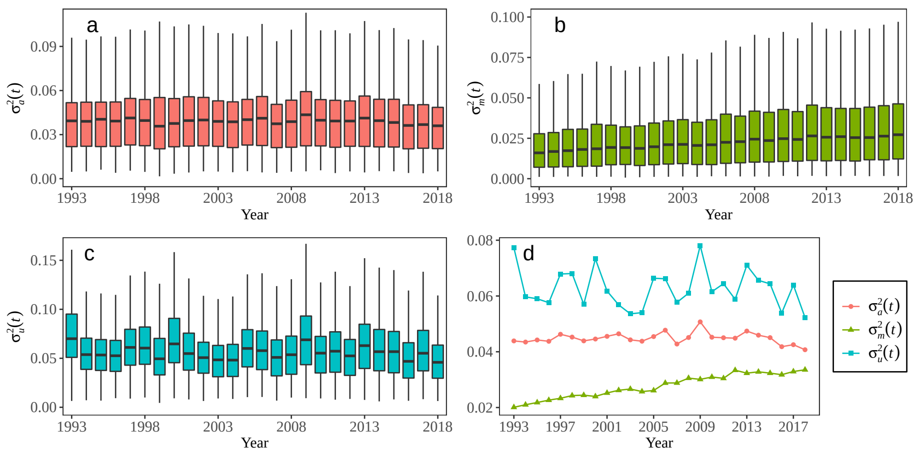

Figure 4.3: Boxplots of cohort-wise distributions for a) additive genetic variance $\sigma_a^2(t)$, b) total additive genetic variance $\sigma_u^2(t)$ and c) mutational variance $\sigma_m^2(t)$ from the animal model on juvenile survival accounting for mutational variance implemented with INLA. Horizontal lines denote medians, boxes denote first and third quartiles and whiskers denote the most extreme value within 1.5 times the inter quartile range. Outliers are not included in the figures. Figure d) displays the posterior mean variances for each cohort.

The posterior cohort-wise mutational variance $\sigma_m^2(t)$ told a more surprising story. The yearly mean values of $\sigma_m^2(t)$ during the period 1993-2018 were much larger than the first posterior estimate. The mean $\sigma_m^2(t)$ grew steadily up to 2012, starting at 0.020 in 1993 and reaching 0.033 in 2012. However, the growth seemed to stop after 2012. Fitting a simple linear model of $\sigma_m^2(t)$ on natal year, for the period 1993-2018, gave a slope of 5.3e$-$4, which was somewhat smaller, but of the same order of magnitude as the estimate of overall σ_m^2 (8.1e$-$4). However, the linear model predicts that $\sigma_m^2(1975) = 0.011$, which on this scale is much larger than the expected 0 for the base population. This indicated that the growth in $\sigma_m^2(t)$ was faster in the earlier years (i.e. before 1993), and that the value reached in 2012 might have been close to a maximum.

An interesting finding in the work of Reid et al. (2021) was how immigration counteracts the expected increase in additive genetic effects in the population. Figure 4.4 illustrates how this phenomenon is present in our model 1 as well. As expected from the non-zero additive genetic variance in survival and consistent directional selection, there was a clear increase in cohort-wise mean breeding value $\bar{a}(t)$ in the period 1993-2018. An increase in $\bar{a}(t)$ would normally imply a higher value on the phenotypic trait (i.e. higher survival rate), but there are more effects to consider. The genetic group effects $q_i g$ are also additive genetic effects. Cohort-wise mean immigrant group membership $\bar{q}(t)$ in the population increased through the observed period, and because g was negative, $\bar{q}(t)g$ decreased accordingly. Thus, since the increase in $\bar{a}(t)$ and the decrease in $\bar{q}(t)g$ were of similar magnitude, the total additive genetic effects $\bar{u}(t) = \bar{a}(t) + \bar{q}(t)$ were quite stable, with no clear increasing or decreasing trend. In model 0 however, $\bar{u}(t)$ had a substantial increase (see Figure 4.5).

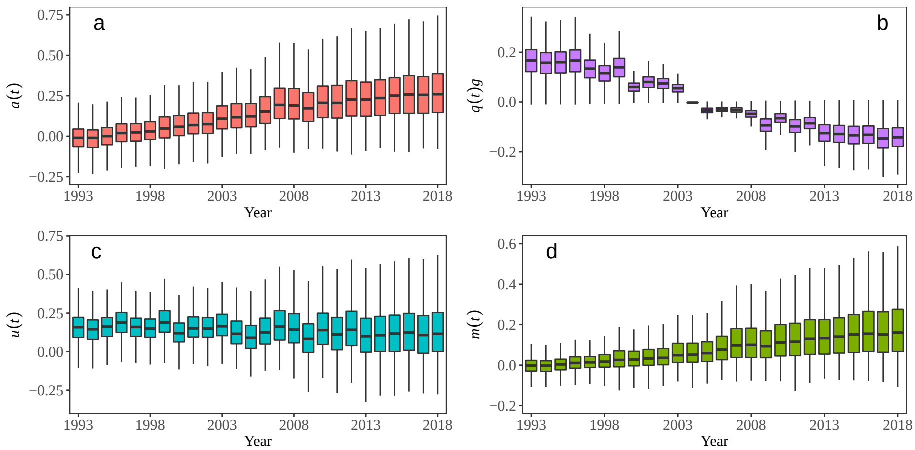

Figure 4.4: Boxplots of cohort-wise distributions for a) breeding values $a(t)$, b) genetic group effects $q(t)g$, c) total additive genetic values $u(t)$ and d) mutational effects $m(t)$ from the animal model on juvenile survival accounting for mutational variance implemented with INLA. Horizontal lines denote medians, boxes denote first and third quartiles and whiskers denote the most extreme value within 1.5 times the inter quartile range. Outliers are not included in the figures.

The discrepancy in $\bar{u}(t)$ between model 1 and model 0 can be explained by the addition of mutational effects. Figure 4.5 illustrates how the mean mutational effects increased over generations compared to other additive genetic effects in model 1 and model 0. The increase in $\bar{m}(t)$ can be interpreted as the net effect of mutations on survival to adulthood in the population being positive, indicating substantial directional selection on new mutations. If we compare mean mutational effects to breeding values, $\bar{m}(t)$ was typically $50 - 60\%$ of $\bar{a}(t)$. Mutational effects are additive genetic values and would be included in the breeding values in a model without mutational effects, explaining why $\bar{a}(t)$ in model 0 are larger than in model 1. If we instead look at the sum of $\bar{m}(t)$ and $\bar{u}(t)$ in model 1 (Figure 4.5), we are close to resembling $\bar{u}(t)$ from model 0, especially in the first half of the study period, showing that the sum of all additive genetic effects was almost constant between model 0 and model 1.

Since the covariance structures of σ_a^2 and σ_m^2 are based on the same relatedness measures, some confounding between the two parameters was expected. Moreover, mutations are assumed to generate additive genetic variance. Thus, we expected σ_m^2 to be absorbed in σ_a^2 in models that do not consider mutational effects. Figure 4.6 shows mean $\sigma_u^2(t)$ for model 0 and model 1. From the figure, it is clear that mean $\sigma_u^2(t)$ behaved similarly in the two models, and that estimating mutational variance in model 1 led to a shift in the value of $\sigma_u^2(t)$. If we let $\Delta\sigma_u^2(t)$ be the difference between $\sigma_u^2(t)$ in model 0 and model 1 respectively, we expected $\Delta\sigma_u^2(t)$ to have similar values to $\sigma_m^2(t)$. In Figure 4.7, $\Delta\sigma_u^2(t)$ and $\Delta\sigma_a^2(t)$ are plotted together with $\sigma_m^2(t)$ for the focal period. Although $\Delta\sigma_u^2(t)$ had a spiky profile like $\sigma_u^2(t)$, it started at a similar value to $\sigma_m^2(t)$ in 1993 and somewhat followed

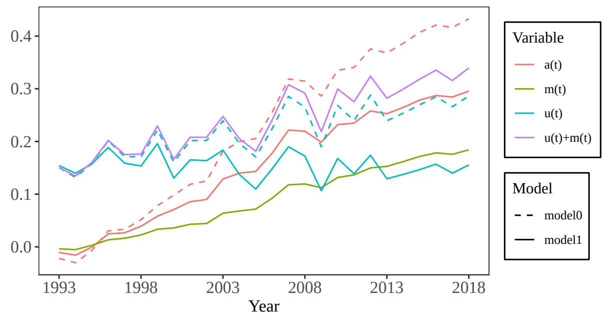

Figure 4.5: Comparison of cohort-wise means of different additive genetic effects from the animal model on juvenile survival accounting for mutational variance (model 1) and the model not accounting for mutational variance (model 0).

the increasing trend up until the last five years, where both $\Delta\sigma_u^2(t)$ and $\Delta\sigma_a^2(t)$ started to decrease, while $\sigma_m^2(t)$ flattened out. The correspondence between $\sigma_a^2(t)$ and $\Delta\sigma_u^2(t)$ was not perfect, but deviations were small enough to meet our expectations.

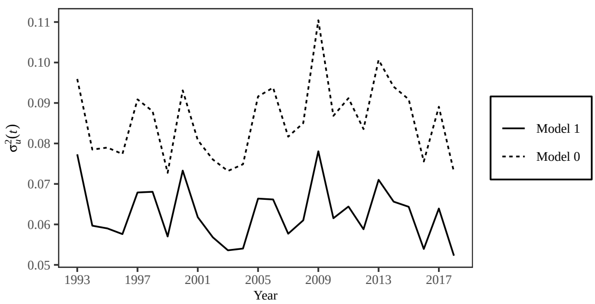

Figure 4.6: Posterior means of cohort-wise distributions for total additive genetic variance $\sigma_u^2(t)$ from the animal model on juvenile survival accounting for mutational variance (model 1) and the model not accounting for mutational variance (model 0).

The posterior distributions of $\sigma_a^2(t)$, $\sigma_m^2(t)$ and $\sigma_u^2(t)$ for individuals born in $t = 1993$ are displayed in Figure 4.8. Both $\sigma_a^2(1993)$ and $\sigma_m^2(1993)$ had two clear modes in their posterior distributions, whereas $\sigma_u^2(1993)$ had a regular distribution with one mode. At first glance this looked like a genetic group problem, where one group had larger values of genetic variance than the other, and the group difference was corrected by adding the genetic group effects $q_i g$ to the breeding values a_i. However, the posterior marginals reported in Section 4.1 showed no sign of such problems. The real reason for bimodal posteriors was the sampling algorithm from INLA, where all 10000 samples of random effects were generated from only 25 different configurations of hyperparameters (i.e. variances), which was not enough to precisely represent the actual distribution (personal communication with Håvard Rue, February 26, 2021). Therefore, the shapes of posterior marginal distributions of σ_a^2 and σ_m^2 were not precisely conveyed into the cohort-wise posteriors.

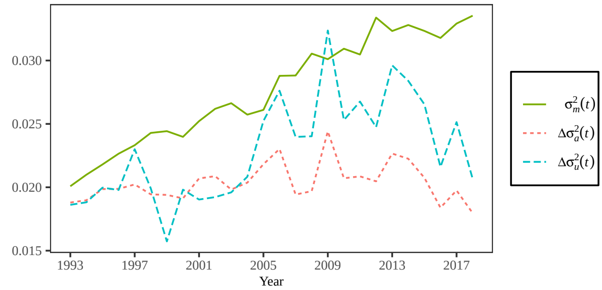

Figure 4.7: Posterior mean of cohort-wise mutational variance $\sigma_m^2(t)$ compared to the difference in additive genetic variance $\Delta\sigma_a^2(t)$ and difference to total additive genetic variance $\Delta\sigma_a^2(t)$ between the animal model on juvenile survival accounting for mutational variance and the model not accounting for mutational variance.

The reason why $\sigma_u^2(1993)$ had a simpler distribution was the, wide and almost symmetric, posterior of the genetic group coefficient g, which had greater impact to the shape of $\sigma_u^2(1993)$ than the variance in breeding values. Still, these cohort-wise posteriors captured most of the features we were interested in, and could be used to model temporal changes in the genetic parameters.

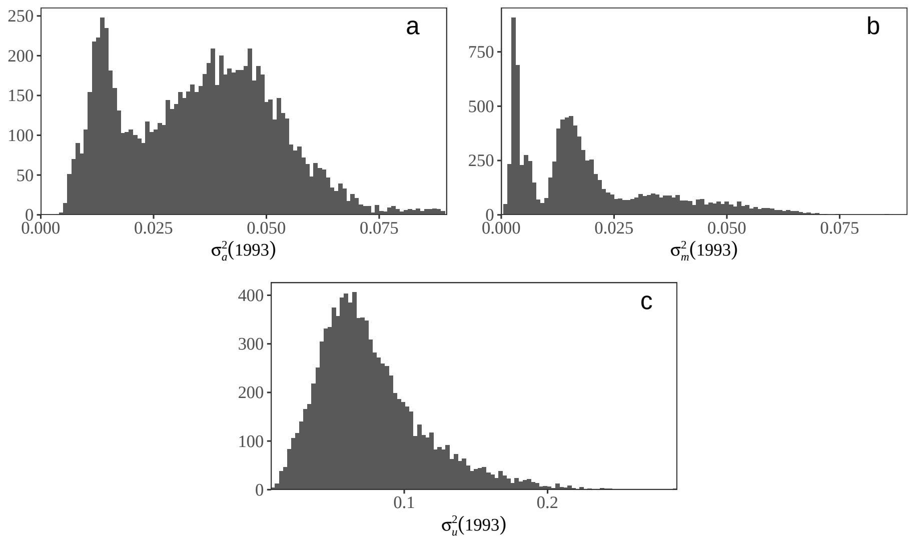

Figure 4.8: Histograms for a) additive genetic variance σ_a^2, b) mutational variance σ_m^2 and c) total additive genetic variance for the 1993 cohort generated with INLA.

4.2.1 Results from Stan implementation

Although posterior marginal distributions of σ_a^2 and σ_m^2 were quite different between the results from the INLA and Stan implementations, using the resampling method with samples from Stan led to similar conclusions as before. The new posterior distributions for $\sigma_a^2(1993)$, $\sigma_m^2(1993)$ and $\sigma_u^2(1993)$, displayed in Figure 4.9, had shapes very similar to the respective posterior marginals in Figure 4.2. This check gave further evidence for the complicated distributions of $\sigma_a^2(1993)$ and $\sigma_m^2(1993)$ from the INLA samples being caused by the sampling algorithm. Moreover, all analyses on cohort-wise results from the INLA implementation were repeated for the Stan implementation (see Appendix C). Apart from some minor differences in size and slope of the parameters, all conclusions drawn from the INLA implementation could be drawn from the Stan implementation as well. Note that the sample size for model 0 in Stan was only 2000, so that comparisons between model 0 and model 1 could not very meaningfully be done.

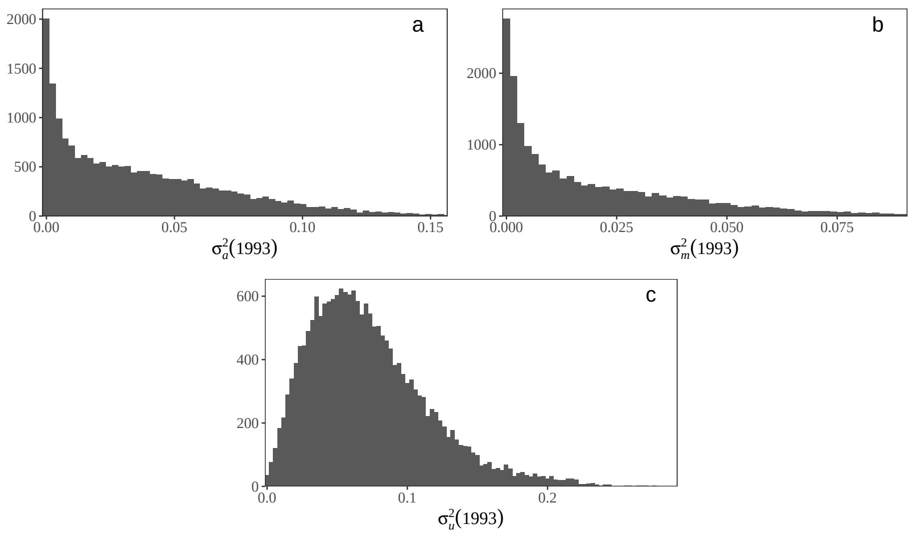

Figure 4.9: Histograms for a) additive genetic variance σ_a^2, b) mutational variance σ_m^2 and c) total additive genetic variance for the 1993 cohort generated with Stan.

Chapter 5

Discussion and conclusion

By extending a Bayesian genetic groups animal model with mutational effects, we have obtained posterior distributions of genetic parameters related to fitness in a song sparrow population. We have applied a resampling scheme, with samples from the Bayesian frameworks INLA and Stan to estimate temporal changes in these genetic parameters. The posterior marginal distribution of mutational variance had most weight on relatively small values, but point estimates lay within the expected range and were in agreement with earlier studies.

A major challenge regarding mutational variance was how it should be interpreted. The definition states that σ_m^2 measures how much variance is generated from mutations from the base population to the next generation (Wray 1990). However, the base population consisted of a group of 15 founders primarily present in 1975 and 33 immigrants sporadically arriving between 1975 and 2018. Hence, it was difficult to place the base population on a timeline, especially when we do not have data between 1975 and 1993. Since the estimated mean $\sigma_m^2(1993)$ was as high as 0.020, the increase in additive genetic variance due to mutations must have been faster between the pre-1993 generations than post-1993. According to Reid et al. (2019), the length of a song sparrow generation should be approximately 2.5 years, implying the number of generations between 1975 and 2018 would be around 17.5. Treating the posterior mean of σ_m^2, instead of the posterior mode, as a linear per generation increase since 1975 would correspond to a value of $0.0024 \cdot 17.5 = 0.042$ for $\sigma_m^2(2018)$, which is not extremely far from 0.033 (the estimated mean for $\sigma_m^2(2018)$). Remembering that 1975 is not necessarily a completely accurate placement for the base population, interpreting the posterior mean σ_m^2 as the per generation increase in additive genetic variance due to mutations might be reasonable. On the other hand, the estimated mean σ_m^2 for the Stan implementation was 0.0035, and would correspond to a value of 0.061 for $\sigma_m^2(2018)$, almost twice as large as the estimated mean $\sigma_m^2(2018)$. This showed that the relationship between posterior estimates of σ_m^2 and $\sigma_m^2(t)$ is complicated and needs further inspection.

Results from model 1 suggest that mutational effects and mutational variance increased rapidly in the study period. The increase was larger than one would expect from other studies, which raised the concern that σ_m^2 might have been overestimated. According to Henrik Jensen (personal communication, May 7, 2020), the majority of mutations have neutral or negative effects on fitness, and even strongly beneficial mutations are very likely to get lost because of genetic drift. Thus, the modelled increase in mutational effects and mutational variance in fitness are likely to not only stem from mutations. Seeing how sensitive σ_m^2 was to the choice of prior distribution, we cannot exclude the possibility that the chosen prior assigned too much weight on larger values. Another problematic point is the strong temporal increase of the sum $\bar{u}(t) + \bar{m}(t)$ (see Figure 4.5), which includes all additive genetic effects. This increase would mean increased fitness over time in the population, as opposed to an approximate selection-migration balance in the results of Reid et al. (2021). Note that this increase in total additive genetic value was present also in model 0, possibly indicating confounding with other parameters than the mutational variance.

Implementing the models in Stan gave some validation to the results from INLA, but did also reveal some difficulties. Firstly, posteriors marginals for σ_a^2 and σ_m^2 differed in shape, but the means were somewhat in agreement. Judging from the shapes of the posterior marginals in Figure 4.2, it seems like Stan has some problems with distinguishing complicated random effects from 0. We recognize that running more MCMC iterations could have possibly helped on this issue, but it seems unlikely that doing more iterations would drastically change the posteriors. Since we used a pedigree including 244 individuals without any measurements, both the breeding values and mutational effects had more levels than data points. The nest effect also had more than 1000 levels, partly explaining why σ_a^2, σ_m^2 and σ_{nest}^2 were hard to estimate, while σ_{year}^2 (26 levels) was consistent between all models. Disregarding differences in posterior shapes, INLA and Stan gave very similar cohort-wise results with only minor differences. This confirmed that the animal model accounting for mutational variance is evaluated sufficiently similar in different frameworks.

Comparing model 1 to model 0 (Figure 4.6 and 4.7) showed that $\Delta\sigma_u^2(t)$ increased similarly to $\sigma_m^2(t)$ up to 2013. This similarity indicates that $\sigma_u^2(t)$ includes most of $\sigma_m^2(t)$ in a model without mutational effects, while other effects are essentially unaffected. This behavior is exactly what was expected, since mutational variance is defined as an additive genetic effect itself. As described earlier, $\Delta\sigma_a^2(t)$ did not show the same growth as $\Delta\sigma_u^2(t)$, meaning the genetic group effect contains part of the mutational variance. On the other hand, these patterns were less clear for 2014-2018, where the growth of $\sigma_m^2(t)$ seemed to halt, while $\Delta\sigma_u^2(t)$ and $\Delta\sigma_a^2(t)$ were decreasing. This is believed to be a consequence of the latest cohorts having a very limited sample size, and thus we were lacking power to accurately estimate both $\sigma_a^2(t)$ and $\sigma_m^2(t)$ in these cohorts. The results for later cohorts should therefore not be overinterpreted.

Survival from independence to adulthood is undoubtedly closely connected to the fitness of an individual. However, using a binary response may negatively affect the power we have to reliably estimate all modeling parameters. In contrast to continuous traits, survival only has two possible values: survival or not survival. This affects the latent information in each individual as it for example does not distinguish between a sparrow barely surviving for a year and a sparrow with extremely high fitness. Moreover, we may be close to the border of how many parameters we can reliably estimate in this model. This may be part of the reason why disentangling σ_m^2 and σ_a^2 was difficult, especially in the Stan implementation. Using a continuous fitness trait would likely allow for more informative posterior estimation.

Future work on the mutational animal model could include some specific improvements when it comes to modeling decisions. Applying the model to simulated data seems necessary to tune model parameters and check whether the current model properly separates mutational variance from other model parameters. Further investigations of prior sensitivity is important and should be done in upcoming work. Moreover, by replacing the binary survival trait by a continuous trait and fitting the linear animal model with mutational effects, we could drastically increase power to disentangle mutational variance from other sources of additive genetic variance. The Mandarte song sparrow data luckily contain suitable continuous traits and thus provide opportunities to fit such models in the future.

Bibliography

Barton, N.H. and P.D. Keightley (2002). “Understanding Quantitative Genetic Variation”. In: *Nature Reviews. Genetics* 3.1, pp. 11–21.

Barton, N.H. et al. (2017). “The Infinitesimal Model: Definition, Derivation, and Implications”. In: *Theoretical Population Biology* 118, pp. 50–73.

Betancourt, M. J. and Mark Girolami (2013). *Hamiltonian Monte Carlo for Hierarchical Models*. arXiv: `1312.0906`.

Casellas, J. and J. F. Medrano (2008). “Within-Generation Mutation Variance for Litter Size in Inbred Mice”. In: *Genetics* 179.4, pp. 2147–2155.

Casellas, J. et al. (2010). “Accounting for Additive Genetic Mutations on Litter Size in Ripollesa Sheep”. In: *Journal of Animal Science* 88.4, p. 1248.

Conner, Jeffrey K. and Daniel L. Hartl (2004). *A Primer of Ecological Genetics.* Sunderland: Sinauer.

de Villemereuil, Pierre (2018). “Quantitative Genetic Methods Depending on the Nature of the Phenotypic Trait: Quantitative Genetics for Non-Classical Traits”. In: *Annals of the New York Academy of Sciences* 1422.1, pp. 29–47.

Fisher Ronald Aylmer, Sir (1930). *The Genetical Theory of Natural Selection*. Clarendon Press Oxford.

Hansen, Thomas F et al. (2011). “Heritability is not Evolvability”. In: *Evolutionary Biology* 38.3, pp. 258–277.

He, Yi and James S Hodges (2008). “Point Estimates for Variance-Structure Parameters in Bayesian Analysis of Hierarchical Models”. In: *Computational Statistics & Data Analysis*. Computational Statistics & Data Analysis 52.5, pp. 2560–2577.

Hoffman, Matthew D. and Andrew Gelman (2014). “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. In: *Journal of Machine Learning Research* 15.47, pp. 1593–1623.

Kimura, Motoo (1958). “On the Change of Population Fitness by Natural Selection”. In: *Heredity* 12.2, pp. 145–167.

Kruuk, Loeske E. B. (2004). “Estimating Genetic Parameters in Natural Populations Using the ‘Animal Model’”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359.1446, pp. 873–890.

Kruuk, Loeske E. B. and Jarrod D. Hadfield (2007). “How to Separate Genetic and Environmental Causes of Similarity Between Relatives”. In: *Journal of Evolutionary Biology* 20.5, pp. 1890–1903.

Kullback, S. and R. A. Leibler (Mar. 1951). “On Information and Sufficiency”. In: *Annals of Mathematical Statistics* 22.1, pp. 79–86.

Lambert, Paul C et al. (2005). “How Vague is Vague? A Simulation Study of the Impact of the use of Vague Prior Distributions in MCMC Using WinBUGS”. In: *Statistics in medicine* 24.15, pp. 2401–2428.

Lynch, Michael (1988). “The Rate of Polygenic Mutation”. In: *Genetical Research* 51.2, pp. 137–148.

Lynch, Michael and Bruce Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Vol. 1. Sunderland: Sinauer.

Meuwissen, T. and Z. Luo (1992). “Computing Inbreeding Coefficients in Large Populations”. In: *Genetics Selection Evolution* 24, pp. 305–313.

Muff, Stefanie et al. (2019). “Animal Models With Group-Specific Additive Genetic Variances: Extending Genetic Group Models”. In: *Genetics Selection Evolution* 51.1, p. 7.

Quaas, R. L. (1976). “Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix”. In: *Biometrics* 32.4, pp. 949–953.

— (1988). “Additive Genetic Model with Groups and Relationships”. In: *Journal of Dairy Science* 71.5, pp. 91–98.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Reid, Jane M et al. (2019). “Individuals’ Expected Genetic Contributions to Future Generations, Reproductive Value, and Short-term Metrics of Fitness in Free-living Song Sparrows (Melospiza Melodia)”. In: *Evolution letters* 3.3, pp. 271–285.

Reid, Jane M. and Lukas F. Keller (2010). “Correlated Inbreeding Among Relatives: Occurence, Magnitude, and Implications”. In: *Evolution* 64.4, pp. 973–985.

Reid, Jane M. et al. (2014). “Pedigree Error due to Extra-Pair Reproduction Substantially Biases Estimates of Inbreeding Depression”. In: *Evolution* 68.3, pp. 802–815.

Reid, Jane M. et al. (2021). “Immigration Counter-Acts Local Micro-Evolution of a Major Fitness Component: Migration-Selection Balance in Free-Living Song Sparrows”. In: *Evolution Letters* 5.1, pp. 48–60.

Rue, Håvard et al. (2009). “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.

Sardell, Rebecca J. et al. (2010). “Comprehensive Paternity Assignment: Genotype, Spatial Location and Social Status in Song Sparrows, Melospiza Melodia”. In: *Molecular Ecology* 19.19, pp. 4352–4364.

Simpson, Daniel et al. (2017). “Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors”. In: *Statistical science* 32.1, pp. 1–28.

Smith, James N.M. et al. (2006). *Conservation and Biology of Small Populations : The Song Sparrows of Mandarte Island*. Oxford: Oxford University Press.

Sorensen, Daniel et al. (2001). “Inferring the Trajectory of Genetic Variance in the Course of Artificial Selection”. In: *Genetics Research* 77.1, pp. 83–94.

Stan Development Team (2021). *Stan Modeling Language Users Guide and Reference Manual, 2.26*. URL: https://mc-stan.org.
Wilson, A. J (2008). "Why h2 Does Not Always Equal VA/VP?" In: *Journal of Evolutionary Biology* 21.3, pp. 647–650.
Wilson, Alastair J. et al. (2010). "An Ecologist's Guide to the Animal Model". In: *Journal of Animal Ecology* 79.1, pp. 13–26.
Wilson, Amy G. and Peter Arcese (2008). "Influential Factors for Natal Dispersal in an Avian Island Metapopulation". In: *Journal of Avian Biology* 39.3, pp. 341–347.
Wolak, Matthew E. (2012). "nadiv: An R Package to Create Relatedness Matrices for Estimating Non-Additive Genetic Variances in Animal Models". In: *Methods in Ecology and Evolution* 3.5, pp. 792–796.
Wolak, Matthew E. and Jane M. Reid (2017). "Accounting for Genetic Differences Among Unknown Parents in Microevolutionary Studies: How to Include Genetic Groups in Quantitative Genetic Animal Models". In: *Journal of Animal Ecology* 86.1, pp. 7–20.
Wray, Naomi R. (1990). "Accounting for Mutation Effects in the Additive Genetic Variance-Covariance Matrix and Its Inverse". In: *Biometrics* 46.1, pp. 177–186.
Wright, Sewall (1922). "Coefficients of Inbreeding and Relationship". In: *The American Naturalist* 56.645, pp. 330–338.
Zuur, Alain et al. (2009). *Mixed Effects Models and Extensions in Ecology With R*. Springer Science & Business Media New York.

Appendix A

Code

This appendix contains the R code used for the project.

A.1 Data preparation

We load data files with data on each sparrow, the pruned pedigree and individual genetic group memberships q_i for all sparrows. The continuous traits are then centered and sex data are transformed from 1 and 2 to 0 and 1, for females and males respectively.

```
#' Load data
qg.data.gg.inds <- read.table(
  "../data/qg.data.gg.inds.txt", header=T)
d.ped <- ped.prune.inds <- read.table(
  "../data/ped.prune.inds.txt", header=T)
d.Q <- read.table(
  "../data/Q.data.txt", header=T)

#' Center continuous covariates:
# Inbreeding coefficient:
qg.data.gg.inds$f.coef.sc <- scale(
  qg.data.gg.inds$f.coef,scale=FALSE)
# Immigrant group coefficient:
qg.data.gg.inds$g1.sc <- scale(
  qg.data.gg.inds$g1,scale=FALSE)
# Natal year:
qg.data.gg.inds$natalyr.no.sc <- scale(
  qg.data.gg.inds$natalyr.no,scale=FALSE)
# Brood date:
qg.data.gg.inds$brood.date.sc <- scale(
  qg.data.gg.inds$brood.date,scale=FALSE)
```

```r
#' Make sex binary:
qg.data.gg.inds$sex <- qg.data.gg.inds$sex.use.x1 - 1
```

A.2 Covariance matrices

Since INLA works with precisions we need to compute the inverse of matrices **A** and **M**. This is done with the `nadiv` package (Wolak 2012). For `nadiv`, INLA and Stan to work however, the data needs some restructuring.

```r
#' First load the nadiv package.
library(nadiv)

#'Get pedigree on nadiv's format
d.ped <- nadiv::prepPed(d.ped)

#' For INLA we need ids that run from 1 to the number of
#' individuals
d.ped$id <- 1:(nrow(d.ped))

#' Need a map file to keep track of the ids and data from
#' the Q-matrix
d.map <- d.ped[,c("ninecode","id")]
d.map$g1 <- d.Q[match(d.map$ninecode,d.Q$ninecode),"g1"]

#' give mother and father the id
d.ped$mother.id <- d.map[match(
  d.ped$gendam, d.map$ninecode),"id"]
d.ped$father.id <- d.map[match(
  d.ped$gensire, d.map$ninecode),"id"]

#' Make the inverse A and M matrices using the nadiv
#' package:
Ainvmatrix <- nadiv::makeAinv(
  d.ped[,c("id","mother.id","father.id")])$Ainv
Minvmatrix <- nadiv::makeMinv(
  d.ped[,c("id","mother.id","father.id")])$Minv

#' Store the id twice: Once for the breeding value and once
#' for the mutation effects
qg.data.gg.inds$id <- d.map[match(
  qg.data.gg.inds$ninecode, d.map$ninecode), "id"]
qg.data.gg.inds$idm <- qg.data.gg.inds$id
```

A.3 INLA model

Now we have all data on the correct format and are ready to define the INLA formula. Modelling choices are explained in Section 3.2.

```
#' f() encode the random effects and
#' v denoting an initial guess for the variance component
#' corresponds to initial=log(1/v))
formula.surv.ind.to.ad <- surv.ind.to.ad ~
  f.coef.sc + g1.sc + natalyr.no.sc + brood.date.sc + sex +
  #Nest variance:
  f(nestrec,
    model="iid",
    hyper=list(prec=list(
      initial=log(1/0.003),
      prior="pc.prec",
      param=c(1,0.01)))) +
  #Natal year variance:
  f(natalyr.no,
    model="iid",
    hyper=list(
    prec=list(
      initial=log(1/0.22),
      prior="pc.prec",
      param=c(1,0.01)))) +
  #Additive genetic variance
  f(id,model="generic0",
    Cmatrix=Ainvmatrix,
    hyper=list(prec=list(
      initial=log(1/.05),
      prior="pc.prec",
      param=c(1,0.01)))) +
  #Mutational variance
  f(idm, model="generic0",
    Cmatrix=Minvmatrix,
    hyper=list(prec=list(
      initial=log(1/.001),
      prior="pc.prec",
      param=c(.25,0.01))))

#Load INLA package
if(!require(INLA)){
  install.packages("INLA", repos=c(getOption("repos"),
    INLA="https://inla.r-inla-download.org/R/stable"),
    dep=TRUE)}
library(INLA)
```

```
#' Call INLA:
r.inla.surv.ind.to.ad <- inla(
  formula=formula.surv.ind.to.ad,
  family="binomial",
  data=qg.data.gg.inds,
  control.family = list(
    control.link = list(model="probit"))
)
#' Rerun INLA with the mode of last run as initial value
#' to avoid possible bias due to a bad initial guess:
r.inla.surv.ind.to.ad <- inla.rerun(r.inla.surv.ind.to.ad)
```

After running the INLA computation, most interesting estimates are found using INLA-specific functionality.

A.4 Stan model

The exact same model is implemented with the `brms` package in Stan. The only differences to the INLA call are that Stan works with covariance instead of precision and we need to define priors as exponential distributions because `brms` does not have functionality for PC priors.

```
#' brms needs non-inverted covariance matrices
Amatrix <- makeA(d.ped[,c("id","mother.id","father.id")])
#Functions for Mmatrix are not yet implemented
Mmatrix <- solve(Minvmatrix)

#' Random effects coded as (1|gr(data),cov=Matrix)
#' Formula and priors identical to INLA model
#' Four chains with 1000 iterations -> n_sample=2000
model_brms <-
  brm(surv.ind.to.ad ~ f.coef.sc + g1.sc + natalyr.no.sc +
        brood.date.sc + sex + (1|gr(id, cov=Amatrix)) +
         (1|gr(idm, cov=Mmatrix)) + (1|gr(nestrec)) +
         (1|gr(natalyr.no)),
      data = qg.data.gg.inds,
      family = bernoulli(link="probit"),
      data2 = list(Amatrix=Amatrix, Mmatrix=Mmatrix),
      prior = c(prior(exponential(-log(0.01)/1),
                     class=sd, group=nestrec),
               prior(exponential(-log(0.01)/1),
                     class=sd, group=natalyr.no),
               prior(exponential(-log(0.01)/1),
                     class=sd, group=animal),
```

```
                prior(exponential(-log(0.01)/0.25),
                       class=sd, group=mutation)),
      inits = "0",
      chains = 4,
      iter = 1000)
```

A.5 Resampling

We then utilize INLA's resampling function to obtain posterior distributions of h^2 and h_m^2 on the data scale from posterior distributions of precisions.

```
#' To obtain the posterior marginal of heritability 'h2'
#' and mutational heritability 'hm2', we need to resample
#' from the posterior of the hyperparameters:
nsamples <- 10^5
sample.posterior <- inla.hyperpar.sample(
  n=nsamples, r.inla.surv.ind.to.ad)

# Compute scaling factor to get h2 on data scale
p <- mean(qg.data.gg.inds$surv.ind.to.ad)
t <- qnorm(p, lower.tail = FALSE)
h2_scale <- p*(1-p)/t^2

#' INLA works with precisions, and variance is 1/precision
h2.inla <- 1/sample.posterior[,"Precision for id"] /
  ((1/sample.posterior[,"Precision for id"]) +
   (1/sample.posterior[,"Precision for natalyr.no"]) +
   (1/sample.posterior[,"Precision for nestrec"]) +
   (1/sample.posterior[,"Precision for idm"]) + 1) /
   h2_scale

hm2.inla <- 1/sample.posterior[,"Precision for idm"] /
  ((1/sample.posterior[,"Precision for id"]) +
   (1/sample.posterior[,"Precision for natalyr.no"]) +
   (1/sample.posterior[,"Precision for nestrec"]) +
   (1/sample.posterior[,"Precision for idm"]) + 1) /
   h2_scale
```

For the cohort resampling we need posterior samples of $\boldsymbol{a}$, $\boldsymbol{m}$ and g.

```
#' Sample a and m from all individuals in pedigree
n_ind <- length(d.ped)

#' Extract samples of a, m and g from INLA
```

```r
n_sample <- 2000
posterior.sample <-
  inla.posterior.sample(n=n_sample,
                        r.inla.surv.ind.to.ad,
                        selection =
                          list("id"=0, "idm"=0, "g1.sc"=0))
breedingvalues_INLA <- matrix(NA, n_sample, n_ind)
mutationvalues_INLA <- matrix(NA, n_sample, n_ind)
groupvalues_INLA <- numeric(n_sample)

for (it in 1:n_sample){
  breedingvalues_INLA[it,] <-
    posterior.sample[[it]]$latent[1:n_ind]
  mutationvalues_INLA[it,] <-
    posterior.sample[[it]]$latent[(n_ind+1):(2*n_ind)]
  group_coefficients_INLA[it] <-
    posterior.sample[[it]]$latent[2*n_ind+1]
}

#' Extract samples of a, m and g from Stan
breedingvalues_brms <- as.matrix(
  model_brms, pars=c("r_id"))
mutationvalues_brms <- as.matrix(
  model_brms, pars=c("r_idm"))
groupvalues_brms <- posterior_samples(
  model_brms, pars="b_g1.sc")
```

Samples of mean and variance for each cohort can be extracted with the following functions.

```r
# Get id of all individuals born in one year
get_cohort <- function(year){
  with(qg.data.gg.inds, unique(id[natalyr == year]))
}

#' Function to extract samples of mean and variance
#' in a given cohort
find_cohort_stats <- function(cohort,
                              breedingvalues,
                              groupvalues,
                              mutationvalues) {
  #Declare vectors to save cohort results
  Va_samples <- numeric(n_sample)
  meana_samples <- numeric(n_sample)
  meanu_samples <- numeric(n_sample)
```

```
  Vm_samples <- numeric(n_sample)
  meanm_samples <- numeric(n_sample)
  Vu_samples <- numeric(n_sample)
  meanqg_samples <- numeric(n_sample)
  for (it in 1:n_sample) {
    #Declare vectors for each sample
    n_t <- length(cohort)
    a_t <- numeric(n_t)
    u_t <- numeric(n_t)
    m_t <- numeric(n_t)
    qg_t <- numeric(n_t)
    #Only one g per sample
    g_t <- groupvalues[it]
    i <- 1
    #Get samples for each individual in cohort
    #NB! Indexing is slightly different for Stan samples
    for (ind in cohort) {
      a_t[i] <- breedingvalues[it,ind]
      qg_t[i] <- qg.data.gg.inds$g1.sc
                      [qg.data.gg.inds$id==ind] * g_t
      u_t[i] <- a_t[i] + qg_t[i]
      m_t[i] <- mutationvalues[it,ind]
      i <- i + 1
    }
    #Save mean and variance of cohort for each sample
    Va_samples[it] <- var(a_t)
    meana_samples[it] <- mean(a_t)
    meanu_samples[it] <- mean(u_t)
    Vm_samples[it] <- var(m_t)
    meanm_samples[it] <- mean(m_t)
    Vu_samples[it] <- var(u_t)
    meanqg_samples[it] <- mean(qg_t)
  }
  return(list(Va = Va_samples, meana = meana_samples,
              meanu = meanu_samples, Vm = Vm_samples,
              meanm = meanm_samples, Vu = Vu_samples,
              gq = meanqg_samples))
}

#' Function to extract samples of mean and variance
#' in multiple cohorts - each year in years
get_yearly_cohort_stats <- function(years,
                                    breedingvalues,
                                    groupvalues,
                                    mutationvalues){
```

```
  n_years <- length(years)
  #Declare matrices to store results
  yearly_Va <- matrix(NA, n_sample, n_years)
  yearly_Vm <- matrix(NA, n_sample, n_years)
  yearly_Vu <- matrix(NA, n_sample, n_years)
  yearly_a <- matrix(NA, n_sample, n_years)
  yearly_m <- matrix(NA, n_sample, n_years)
  yearly_u <- matrix(NA, n_sample, n_years)
  yearly_qg <- matrix(NA, n_sample, n_years)
  i <- 1
  #Get stats for each year
  for (year in years){
    cohort <- get_cohort(year)
    n_cohort <- length(cohort)
    cohort_stats <-
      find_cohort_stats(cohort, breedingvalues,
                        groupvalues, mutationvalues)
    yearly_Va[,i] <- cohort_stats$Va
    yearly_Vm[,i] <- cohort_stats$Vm
    yearly_Vu[,i] <- cohort_stats$Vu
    yearly_a[,i] <- cohort_stats$meana
    yearly_m[,i] <- cohort_stats$meanm
    yearly_u[,i] <- cohort_stats$meanu
    yearly_qg[,i] <- cohort_stats$meanqg
    i <- i + 1
  }
  return(list(Va=yearly_Va, Vm=yearly_Vm, Vu=yearly_Vu,
              a=yearly_a, m=yearly_m, u=yearly_u,
              qg=yearly_qg))
}
```

Lists of cohort-wise samples can be generated by simple function calls.

```
#Get individuals born in 1998
cohort <- get_cohort(1998)
#Get statistics for the 1998 cohort
cohort_stats <-
  find_cohort_stats(cohort, breedingvalues,
                    groupvalues, mutationvalues)
#Get statistics for all cohorts
years <- 1993:2018
all_cohort_stats <-
  get_yearly_cohort_stats(years, breedingvalues,
                          groupvalues, mutationvalues)
```

Appendix B

Prior sensitivity with INLA

In order to test how the prior distribution of σ_m^2 affects posterior marginal distribution, we defined models with the following priors for σ_m^2: PC$(0.20, 0.01)$ and PC$(0.30, 0.01)$. The posterior marginal distributions of these two models, together with the main model (with prior PC$(0.25, 0.01)$) and the model without mutational effects, are shown in Figure B.1.

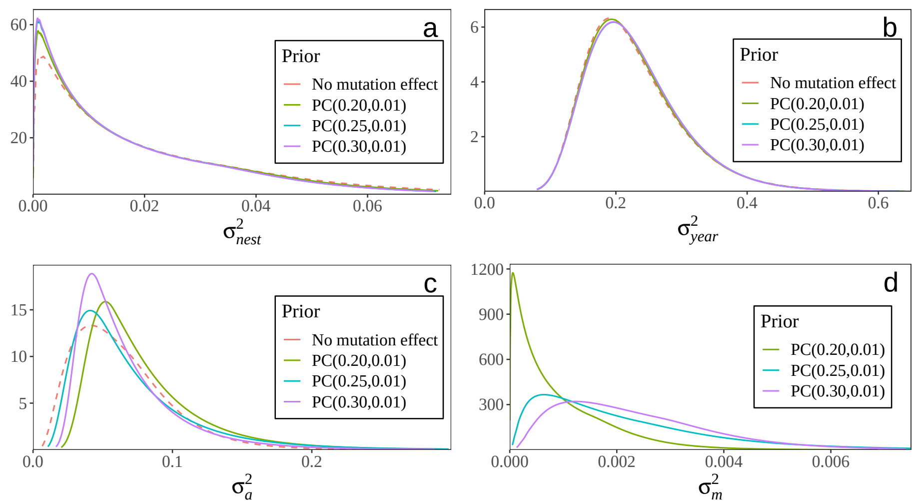

Figure B.1: Marginal posterior distributions of the variances for the animal model on juvenile survival that accounted for mutational variance with different prior distributions for σ_m^2, compared to a model not accounting for mutational variance.

From Figure B.1 it is clear that relatively small changes to the prior of σ_m^2 greatly affects the posterior marginal distribution of σ_m^2. As expected, environmental variances σ_{nest}^2 and σ_{year}^2 are essentially unaffected by changes in σ_m^2. On the other hand, σ_a^2 seems to be confounded with σ_a^2 as a right shift of the posterior marginal for σ_a^2 comes together with a left shift of σ_m^2, and vice versa. Interestingly, very small estimates of σ_m^2 leads to a right shift in the marginal of σ_a^2 relative to the non-mutational model. These results show that correlation between the additive genetic variance and mutation effects needs to be investigated in upcoming work.

Appendix C

Stan results

In the analysis of model 1, mainly results from the INLA implementation were used. Because running the Markov chains was very slow with multiple random effects with large covariance matrices in Stan, it was not feasible to generate large samples with this method. Therefore, the results from the Stan implementation were mainly used as validation for the INLA results, and are presented here. Overall, results between the two implementations were very similar, and main differences are discussed in Chapter 4. Keep in mind that the number of sampling iterations for model 0 in Stan was only 2000, so that comparisons between model 0 and model 1 should not receive too much attention.

Table C.1: Posterior mean (for fixed effects), posterior mode; mean (for random effect variances) and 95% HPD CI for the animal model on juvenile survival that accounted for mutational variance (model 1) generated with Stan.

Summary of posterior distributions for model parameters in Stan model

		Posterior statistics	
	Parameter	Estimate	95% CI
Fixed effects			
	F	-4.27	$(-5.96, -2.61)$
	g	-1.13	$(-2.05, -0.20)$
	Natal year	-0.03	$(-0.06, 0.01)$
	Brood date	-0.006	$(-0.009, -0.003)$
	sex	0.28	$(0.16, 0.40)$
Variances			
	σ^2_{year}	$5e-04; 0.03$	$(8e-10, 0.10)$
	σ^2_{year}	$0.21; 0.23$	$(0.11, 0.39)$
	σ^2_a	$3e-04; 0.05$	$(9e-09, 0.13)$
	σ^2_m	$6e-05; 0.0035$	$(1e-13, 0.0113)$
	h^2	$9e-04; 0.07$	$(1e-08, 0.18)$
	h^2_m	$8e-05; 0.005$	$(2e-13, 0.017)$

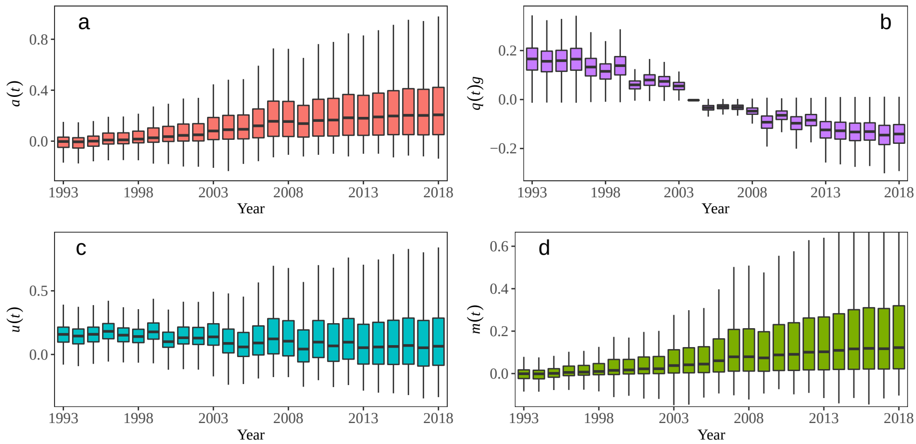

Figure C.1: Boxplots of cohort-wise distributions for a) breeding values $a(t)$, b) genetic group effects $q(t)g$, c) total additive genetic values $u(t)$ and d) mutational effects $m(t)$ created by samples from the animal model on juvenile survival accounting for mutational variance generated with Stan. Horizontal lines denote medians, boxes denote first and third quartiles and whiskers denote the most extreme value within 1.5 times the inter quartile range. Outliers are not included in the figures.

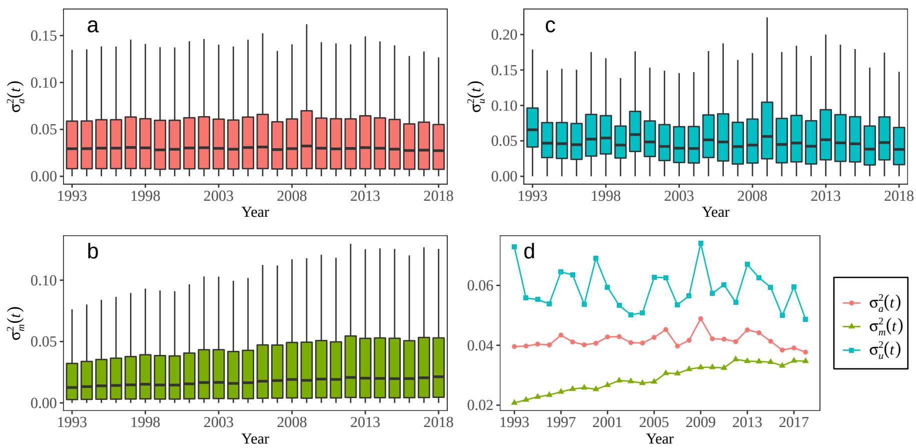

Figure C.2: Boxplots of cohort-wise distributions for a) additive genetic variance $\sigma_a^2(t)$, b) total additive genetic variance $\sigma_u^2(t)$ and c) mutational variance $\sigma_m^2(t)$ created by samples from the animal model on juvenile survival accounting for mutational variance generated with Stan. Horizontal lines denote medians, boxes denote first and third quartiles and whiskers denote the most extreme value within 1.5 times the inter quartile range. Outliers are not included in the figures. Figure d) displays the posterior mean variances for each cohort.

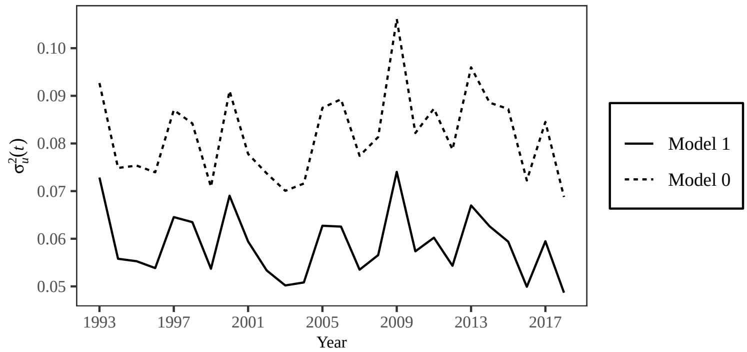

Figure C.3: Posterior means of cohort-wise distributions for total additive genetic variance from the animal model on juvenile survival accounting for mutational variance and the model not accounting for mutational variance (dotted line) generated with Stan.

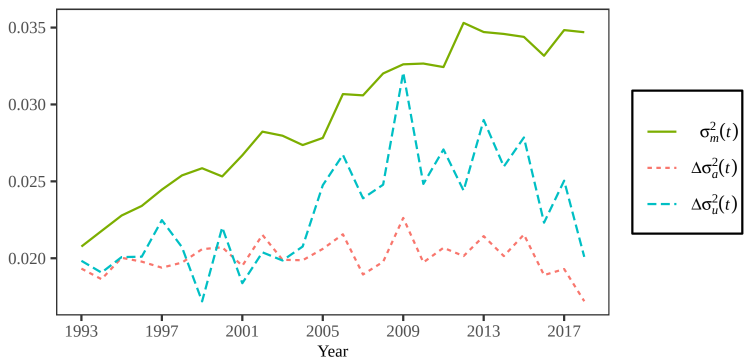

Figure C.4: Posterior mean of cohort-wise mutational variance compared to the difference in additive genetic variance and total additive genetic variance between the animal model on juvenile survival accounting for mutational variance and the model not accounting for mutational variance generated with Stan.

NTNU

Norwegian University of
Science and Technology