Quantitative genetic modeling and inference in the presence of non-ignorable missing data

Ingelin Steinsland[1], Camilla Thorrud Larsen[2], Alexandre Roulin[3,*], and Henrik Jensen[4,*]

[1]Department of Mathematical Sciences, NTNU, 7491 Trondheim, Norway, ingelins@math.ntnu.no ,
[2]Department of Eletric Power Engineering, NTNU, 7491 Trondheim, Norway, camilla.t.larsen@elkraft.ntnu.no
[3]Department of Ecology and Evolution, University of Lausanne Biophore 1015 Lausanne, Switzerland, alexandre.roulin@unil.ch
[4]Centre for Biodiversity Dynamics, Department of Biology, NTNU, 7491 Trondheim, Norway, Henrik.Jensen@ntnu.no
[*]Both AR and HJ are senior authors of this work

January 30, 2014

Keywords: animal model, missing not at random, *tyto alba*, shared parameter model, sex-linked inheritance

Running title: Quantitative genetics & non-ignorable missing data.

Data archiving: Supplementary online material.

Word count abstract: 185

Word count paper: 7250

Number of figures: 4

Number of tables: 3

Abstract

Natural selection is typically exerted at some specific life stages. If natural selection takes place before a trait can be measured, using conventional models can cause wrong inference about population parameters. When the missing data process relates to the trait of interest a valid inference requires explicit modeling of the missing process. We propose a joint modeling approach, a shared parameter model, to account for non-random missing data. It consists of an animal model for the phenotypic data and a logistic model for the missing process, linked by the additive genetic effects. A Bayesian approach is taken and inference is made using integrated nested Laplace approximations. From a simulation study we find that wrongly assuming that missing data are missing at random, can result in severely biased estimates of additive genetic variance. Using real data from a wild population of Swiss barn owls *Tyto alba* our model indicates that the missing individuals would display large black spots, and we conclude that genes affecting this trait are under selection already before it is expressed. Our model is a tool to correctly estimate the magnitude of both natural selection and additive genetic variance.

1 Introduction

As emphasized in a recent review, a number of key issues in ecology and evolutionary biology can be only tackled using data collected in populations over many years (Clutton-Brock and Sheldon, 2010). For instance, long-term studies in the common tern (*Sterna hirundo*) have identified traits that are naturally selected in an age-specific manner (e.g. Rebke et al., 2010), which can then explain patterns of population dynamics (e.g. Ezard et al., 2007). Selection is typically exerted at different intensities throughout life and identifying the life stages when selection is maximally exerted on a given phenotype will bring essential information on its adaptive function. This is however not an easy task because gathering information at all life stages can be logistically difficult. This is a problem because failing to collect data in the life stage when selection is maximally exerted on a given phenotype may give the wrong impression that this trait is not or only weakly selected. Data can be missing either because the entire population cannot be momentarily monitored or because animals are counter-selected even before the trait of interest can be measured. A very important question to answer is whether a specific trait is under (indirect) selection even before it is expressed. We here present quantitative genetic methods that allow us to identify whether the missing individuals from a long-term dataset with a known pedigree are not a random sample of the population.

Quantitative genetic studies of wild populations and domestic breeds often suffer from a considerable amount of missing data for a multitude of reasons, including that some individuals escape capture, migrate out of the

study area or die before the trait is measured (Nakagawa and Freckleton, 2008). In animal and plant breeding it is often the case that only a subset of individuals selected for reproduction or cultivation are measured (Im et al., 1989; Piepho and Mohring, 2006). Also in quantitative genetic analysis in medical research missing data is often a challenge (e.g. Verbeke and Molenberghs, 2000; Bechger et al., 2002).

Current methods used to estimate relevant genetic parameters are mixed models often called animal models (Henderson, 1975; Lynch and Walsh, 1998). These models implicitly assume that the observed sample is a random and representative sample of the population under study. However, missing data may compromise this randomization justification, leading to biased inferences. How missing data affect statistical inference depends on why and how the data are missing, i.e. the nature of the missing data process. Little is however known about the potential effects of missing data on the bias of quantitative genetic estimates, in particular for wild populations.

According to Little and Rubin (2002) missing data theory distinguishes between *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR). When missing data are MCAR, the missing data process is independent of any observed or unobserved data, i.e. the missing observations are purely a random sample of the potential full sample. Accidentally deleting an observation is an example of the MCAR mechanism. Given that the observed data provides sufficient power to the analysis, statistical inferences should be unbiased under MCAR. However, MCAR is a strong and rarely realistic ecological assumption (Hadfield, 2008; Nakagawa and Freckleton, 2008). A less stringent assumption is that the

data are MAR, in which the missing data process may depend on observed rather than unobserved data. Any systematic pattern of missingness can be mediated by the observed data under MAR, and thus conditional on observed data the missing data process is random. Effective computational methods for handling missing data under the MAR assumption are well established, such as multiple imputations or the EM algorithm (Little and Rubin, 2002). In the Bayesian or likelihood setting, the missing data process is said to be *ignorable* under MCAR and MAR (Im et al., 1989; Little and Rubin, 2002). This means that a valid inference can be obtained based on the model for the observed data only, ignoring the missing data process. Finally, when neither the assumption of MCAR or MAR holds the missing data are MNAR, where even after accounting for all available observed information the missing data process still depends on the missing observations themselves, perhaps in addition to observed data. The missing data process is in general *non-ignorable* under MNAR and a valid inference would require the missing data process to be explicitly modeled and incorporated into the modeling procedure.

Missing data in evolutionary studies might be particularly important when the reason for missingness is that individuals die before a trait is expressed or measured. These individuals are referred to as the 'invisible fraction' (Grafen, 1988) and will often constitute a substantial amount of the missing data. When this mortality is non-random in relation to the trait of interest, e.g. lighter individuals have higher mortality than heavier individuals, missing data are MNAR in most cases (Hadfield, 2008). Because the probability of death before the trait is expressed, and hence missingness,

depends on the phenotypic value of the trait it would imply that viability selection acts directly on the focal trait or indirectly via genetically correlated traits. Moreover, the distributional properties of the observed sample will differ from those of the potential full sample, leading to biased inferences. This is illustrated in Figure 1 with an example of parent-offspring regression on height data, where the parameter of interest is the heritability given by the slope of the regression line. The left panel shows the regression on the complete data with no missing values. In the middle panel we deleted 32 observations at random. The regression line and hence the heritability is only slightly changed and asymptotically the heritability estimate is equal to that from the full data set. The right panel shows the regression on the data set after we deleted data on the 32 tallest offspring. The reason for missingness is directly linked to the response variable itself (i.e. the height of offspring), and the data are missing not at random (MNAR). It is obvious from the figure that the estimated heritability is significantly downward biased under MNAR. If data for the offspring of the 32 tallest parents were missing, but the height of the parents were known, we would have an example of missing at random (MAR).

Unfortunately, it is not possible to tell from the data at hand whether the missing data process is ignorable or non-ignorable (Little and Rubin, 2002). A general approach to account for non-ignorable missing data is to base inferences on a joint model of both the data and the missing data process. Joint modeling approaches for handling non-ignorable missing data are frequently appearing in biostatistics (Diggle and Kenward, 1994; Little, 1995; Verbeke and Molenberghs, 2000; Bechger et al., 2002). Based on the ideas presented

in such literature, we propose a joint modeling approach for phenotypic data and the missing data process to accommodate potential non-ignorable missing data due to the invisible fraction (e.g. caused by viability selection before age of measurement). For heritable traits, at least some information on the missing observations can be recovered from observed phenotypic values of their relatives which will be reflected in the breeding values. We suggest a shared parameter model (e.g. Vonesh et al., 2006) which assumes conditional independence between the data model and the missing data process given the breeding values.

In the present study we first explore how inferences vary under assumptions of MAR (ignoring the missing data process) and MNAR (joint modeling), for various missing data processes and different heritabilities. Next we consider a phenotypic trait in the barn owl (*Tyto alba*), the diameter of black plumage spots displayed on the ventral body side, that is highly heritable ($h^2 = 0.82$) and for which the expression is only weakly sensitive to environmental factors (Roulin and Djikstra, 2003). Hence, additive genetic variance can be accurately estimated, providing a unique opportunity to evaluate whether our joint modeling approach to account for missing data is indeed efficient when applied to data collected by evolutionary biologists. The barn owl pedigree is also used for the simulation study.

The key reason for considering the barn owl is that we previously showed that the size of black spots is directionally selected, with females displaying larger black spots having a survival advantage in the first-year of life. As a consequence of this selection, we could demonstrate microevolution with the mean spot size having significantly increased in our population over the

course of 12 years (Roulin et al. 2010). We thus ask the question of whether mortality among nestling females (i.e. before these young produce feathers and hence the black spots) is random with respect to spot size. If selection is already acting before females produce the black spots, it would indicate that at least at that stage selection is acting on genetically correlated traits.

We take a Bayesian approach to inference, which can be performed efficiently and accurately without simulations using integrated nested Laplace approximations (INLA). The INLA methodology was introduced by Rue and Martino (2006) and Rue et al. (2009) and provides a fast and deterministic alternative to traditional Markov chain Monte Carlo (MCMC) methods. INLA has proven very efficient in the modeling of Gaussian traits (Steinsland and Jensen, 2010; Holand et al., 2013) and it allows us to draw inferences from the joint model in a reasonable amount of time.

The rest of the article is organized as follows. Section 2 presents the barn owl system used and introduces the Bayesian animal model framework. Furthermore, our joint model formulation is specified, the connection between our model and a bivarite model of a focal trait and missingness is established, and a joint model is set up for the barn owl system. Results from the simulation study and from applying our joint model for barn owls are presented in Section 3. In Section 4 the method and our findings are discussed. Conclusions are drawn in Section 5. Data for the barn owl system, R-code and additional tables and figures are given in the online supplementary material.

2 Methods and materials

2.1 Field data

The barn owl is a medium-sized nocturnal bird that preys upon small mammals captured in the open landscape. On the ventral body side its plumage varies both within and among populations, as well as within families, with respect to pheomelanin-based coloration (variation from white to dark reddish) and number and size of black eumelanic spots located at the tip of feathers. These traits are sexually dimorphic with females being on average darker reddish and displaying on average more and larger black spots. Variation in the size of eumelanic spots is particularly interesting as it was shown to co-vary with a number of physiological, morphological and behavioral traits (Roulin and Ducrest, 2011; Van den Brink et al., 2012). We have also recently shown in our Swiss population that females are positively selected for large spots (Roulin et al., 2010). Here we use data collected between 1996 and 2007 (a subset of the dataset in Roulin et al. (2010), as we have removed all owls that hatched before 1996) on the size of eumelanic spots of individuals breeding in 110 nest boxes put up in barns over the study area, a plain covering 190 km^2. The pedigree was constructed by assuming the social parents were the biological parents, as extra-pair paternity is rare in the barn owl (Roulin et al., 2004). Breeding females were distinguished from males by the presence of a brood patch, and sex of each nestling was determined from blood cell DNA using sex specific molecular markers. For a more thorough description of the fieldwork and methods, see e.g. Roulin et al. (2010).

The pedigree used in this work consists of N = 2999 barn owls, of which

1550 were females and 1449 were males, and where sex and hatch year were known for all individuals. Spot measurements are available for 2476 owls (1293 females and 1183 males), i.e. 17 % are missing. The barn owls fledge at an age of 55-60 days, while the plumage spots are expressed after 40-45 days. Nest boxes are visited frequently during the breeding season, and we have spot measurements for all owls that fledged, except in year 2000 when plumage spots were not measured. Hence for the 298 owls that hatched in 2000 we neither have data on spot diameter nor do we know whether they survived until they fledged. For owls that hatched in the other years than 2000, 225 out of 2701 are missing, i.e. 8 % are missing. Most of these died before they were 20 days old because of brood reduction due to food shortage. The spot diameter data were standardized to have zero mean and unit variance.

2.2 Bayesian animal models

A popular approach to quantitative genetic analysis for domestic and wild populations is the use of generalized linear mixed models (GLMM), so-called *animal models* (Henderson, 1975; Lynch and Walsh, 1998; Sorensen and Gianola, 2002; Kruuk, 2004). The animal model links phenotypic values to different genetic and environmental effects through information from large pedigrees, to estimate important quantitative genetic parameters.

The scope of this article is restricted to analysis of a single trait at a time, in which case the vector of phenotypic values $\boldsymbol{y}$ of all individuals in a

population can be written

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector containing group level effects, or 'fixed effects', and $\boldsymbol{a}$ is the vector of additive genetic effects, called breeding values. $\boldsymbol{\epsilon}$ is a vector of random individual effects, and $\boldsymbol{B}$ and $\boldsymbol{X}$ are known incidence matrices.

Bayesian inference from animal models requires the likelihood for the phenotypic values as well as prior distributions for the latent variables and hyperparameters to be defined. Continuous traits are expected to be approximately Gaussian distributed and generated from the following conditional probability distribution;

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\sigma_\epsilon^2} \sim \mathrm{N}(\boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{a}, \boldsymbol{I}\sigma_\epsilon^2), \tag{2}$$

where σ_ϵ^2 is the variance of random individual environmental effects and $\boldsymbol{I}$ denotes the identity matrix.

Animal models are so-called latent Gaussian models, in which the latent variables $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\epsilon})$ are assigned Gaussian prior distributions. The random individual effects $\boldsymbol{\epsilon}$ are assumed independent between observations, with zero mean.

$$\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_\epsilon^2). \tag{3}$$

The variance of $\boldsymbol{\epsilon}$ is often a parameter of direct interest and we let σ_ϵ^2 enter the prior as an unknown hyperparameter.

The breeding values $\boldsymbol{a}$ are also assigned zero mean Gaussian prior, and

have a covariance structure corresponding to how individuals within the population are related (e.g. Lynch and Walsh, 1998)

$$\boldsymbol{a} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{A}\sigma_a^2). \tag{4}$$

Here, $\boldsymbol{A}$ is the additive genetic relationship matrix and σ_a^2 is the additive genetic variance, which is an unknown hyperparameter.

Finally, each group level effect (e.g. sex and hatch year) are assumed independent and given zero mean Gaussian prior distribution. The variance of the group level effects are often not of explicit interest, at least not in the present work, so we set the variance to a fixed value to ease computational efforts. To reflect vague prior knowledge about the group level effects, the variance is set to a high value and the prior becomes

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}10^3). \tag{5}$$

To set up the full Bayesian model, the priors for the hyperparameters (σ_ϵ^2 and σ_a^2) must be specified. We use independent inverse Gamma priors with known parameters for the variance components

$$\begin{aligned} \sigma_\epsilon^2 &\sim \mathrm{IG}(a_\epsilon, b_\epsilon) \\ \sigma_a^2 &\sim \mathrm{IG}(a_a, b_a). \end{aligned} \tag{6}$$

Bayesian animal models are considered efficient in dealing with missing data as the missing data are treated as random variables and does not require deletion or imputation of incomplete cases (O'Hara et al., 2008; Steinsland

and Jensen, 2010). However, animal models implicitly assume any missing data are MCAR or MAR. To account for potential non-random (directional) selective processes resulting in missing data, we propose to use a joint modeling approach where the missingness is considered informative (part of the data) and modeled together with the phenotypic trait.

2.3 Joint model formulation

Consider a population of N individuals ($i = 1, ..., N$) and let $\boldsymbol{y} = \{y_i\}$ be the vector of potential phenotypic measurements for the population, in the sense that some measurements may be missing. Moreover, let $\boldsymbol{m} = \{m_i\}$ denote the vector of missing data indicators, defined such that $m_i = 0$ if y_i is observed and $m_i = 1$ if y_i is missing. The vector $\boldsymbol{m}$ is fully observed and describes the distribution of missingness in the population. In the presence of non-ignorable missing data, this vector provides additional information to the analysis of the trait of interest and should be treated as part of the data (Little and Rubin, 2002).

In principle, one would like to consider the joint density $p(\boldsymbol{y}, \boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{\phi})$, where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are parameter vectors describing the measurement model and the missing data process, respectively. A major challenge for the joint modeling approach is that the correct model for the missing data process is rarely known. But as most traits of interest in quantitative genetics and/or selection studies are to some extent heritable, at least some information about the genetic component of the invisible fraction can be recovered from observed phenotypic values of recorded relatives. The genetic correlation between the

trait of interest and missingness therefore gives valuable information about the missing data process (Hadfield, 2008). By assuming that the measurement model and the missing data process are independent given the additive genetic effects $\boldsymbol{a}$ we can factorize the joint model as

$$p(\boldsymbol{y}, \boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{m}|\boldsymbol{a}, \boldsymbol{\phi}). \tag{7}$$

This model falls within the class of shared parameter models (Little, 1995; Vonesh et al., 2006), and implies that all association between the trait of interest and the missing process is induced by the additive genetic effects.

The shared parameter model (SPM) is obtained by specifying a conditional model for the phenotypic data $p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})$, for which we use the animal model (1), and a conditional model for the missing process $p(\boldsymbol{m}|\boldsymbol{a}, \boldsymbol{\phi})$. The missingness is represented by binary variables, and assumed to come from the conditional distribution $\boldsymbol{m}|\boldsymbol{\pi} \sim \text{Bin}(1, \boldsymbol{\pi})$. We model the probability of individual i being missing $\pi_i = \Pr(m_i = 1|\boldsymbol{\phi})$ using a logistic model

$$\text{logit}(\pi_i) = \eta_i = \boldsymbol{v}_i^T \kappa + \gamma a_i, \qquad i = 1, ..., N, \tag{8}$$

where κ is a vector containing group level effects relevant to the missing process and $\boldsymbol{v}_i^T$ is a design vector (row vector of indexes to assign the appropriate group level effects to π_i). a_i is the breeding value of individual i, and this parameter appears in both models and is what links the two models together. Because the two responses $\boldsymbol{y}$ and $\boldsymbol{m}$ can be related, the (scalar) hyperparameter γ describes the nature of this association. Consequently, if

$\gamma = 0$ the two models are unrelated and there is nothing to be gained by a joint analysis.

Priors on the latent variables and the hyperparameters for the missing data process must be specified to complete the modeling set-up. The prior for the additive genetic effects a_i is specified in (4). The group level effects κ are in conformity with the group level effects entering the animal model assigned independent zero mean Gaussian prior with known variance

$$\boldsymbol{\kappa} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}10^3). \tag{9}$$

Hence, also the missing process is a latent Gaussian model, with latent field $\boldsymbol{\eta}$. The association parameter γ is an unknown hyperparameter and is assigned zero mean Gausian prior with known variance

$$\gamma \sim \mathrm{N}(0, 10^3). \tag{10}$$

Even though our model has two responses $(\boldsymbol{y}, \boldsymbol{m})$, the SPM is a univariate animal model. Only the additive genetic effect of the trait $\boldsymbol{y}$ is included the model. But this does not imply that the genetic correlation between the trait and the missing process is 1. These issues are further discussed below.

2.4 Relation to bivariate animal model for focal trait and missingness

For our data on spot diameter the missing process corresponds to pre-juvenile survival. In Appendix A we set up a bivariate animal model (BAM) for our

focal trait, i.e. spot diameter, and the missing process, i.e. pre-juvenile survival. Further it is shown that the shared parameter model introduced in Section 2.3, corresponds to using the BAM. Importantly, the SPM simplifies the inference when the additive genetics of the focal trait and its association with the missing process / pre-juvenile survival is of interest, and not the additive genetic variance of the missing process. We also find that $\gamma\boldsymbol{a}$ is not the breeding value of the missing process, but the part of the additive effect the missing process share with the focal trait.

2.5 Simulation study 1, SPM

We conducted a simulation study to evaluate the performance of the shared parameter model (SPM) in comparison to a naive modeling approach where the animal model was used without considering the missing data process. Data sets were simulated to represent phenotypic data with various levels of additive genetic basis, and with missing data caused by various missing-data processes. Phenotypic values were simulated across the known pedigree of the barn owl population by sampling from the following simple animal model

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{\epsilon}, \tag{11}$$

where $\boldsymbol{a} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_\epsilon^2)$, A and I as in Section 2.2. To simulate approximately standardized data we chose σ_a^2 and σ_ϵ^2 such that $\sigma_a^2+\sigma_\epsilon^2 = 1$. Thus, the models used to simulate $\boldsymbol{y}$ are determined by the values chosen for the additive genetic variance, σ_a^2, and we used $\sigma_a^2 = (0.2, 0.5, 0.8)$.

Further, individual i's trait record is missing (i.e. deleted) with probabil-

ity

$$\text{logit}(\pi_i) = \alpha + \gamma a_i. \tag{12}$$

where a_i is the additive genetic effect of individual i. In (12) α sets the average level of missing values and γ is the association parameter, which determines the strength of dependency between missing process and phenotypic values. In the simulation study we used nine sets of parameters for the missing data process: all combinations of i) three different values for the level ($\alpha = -2, -1, 0$) and ii) three different values for the association ($\gamma = -2, -1, 0$). Under model (12), the missing process is unrelated to the data model if $\gamma = 0$. Then each individual have the same probability of being missing, and hence, the data are missing at random. Because the probability density function of $\boldsymbol{a}$ is symmetric, positive values of γ will yield on average the same results with respect to the estimated additive genetic variance as their negative counterparts and only negative values for γ are therefore chosen. Negative association ($\gamma < 0$) implies that individuals with smaller genetic values are more likely to be missing than individuals with larger values.

For each combination of $(\sigma_a^2, \alpha, \gamma)$ we generated 100 complete data sets from model (11). Further, based on each complete data set, the missing data pattern was simulated by model (12) and observations deleted accordingly. Thus, we have 100 synthetic data sets with missing values for each set of model parameters. For each of these data sets parameters were estimated under both the shared parameter model (SPM) and the missing at random (MAR) model.

The parameter α set the general level of the proportion of missing data which is increasing with in α. For non-negative association ($\gamma \neq 0$) and $\alpha \neq 0.5$ the proportion of missing data also depends on γ and the additive genetic variance σ_a^2, and is increasing with stronger association and heritability. This can be understood from the unsymmetrical relation between these parameters and the *logit* link function. The proportion of missing data depends on all three parameters, and are given for our parameter sets in Table S1.

For a real data set at hand we know the proportion of missing data, but neither the additive genetic variance nor the association between the breeding values and the missing process. To get an impression of how large a bias we might get, we have also preformed a simulation study where we set the proportion of missing data to $m \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$. We use an extreme association between between breeding values and the missing process; In each simulation the individuals with the largest breeding values are set to be missing. For each proportion of missing data m we simulate 100 data sets with heritability 0.8 ($\sigma_a^2 = 0.8$ and $\sigma_\epsilon^2 = 0.2$). For these data sets MAR-models are fitted, and biases ($\hat{\sigma}_a^2 - \sigma_a^2$) calculated.

2.6 Simulation study 2, BAM

We have also preformed a simulation study where we generate data from the bivariate animal model (BAM) and make inference using the SPM. The purpose of this study was to demonstrate that the SPM gives valid inference of the genetic parameters also when the true data generating model is the BAM.

We simulated data sets using the barn owl pedigree for different G-matrices. We set $\sigma_\epsilon = 0.5$, $\sigma_a = 0.5$ and $\alpha = -1$, and chose a set of parameter values of the association parameters; $\gamma = \{0, -1, -2\}$ and for the extra additive genetic variance of the missing process $\sigma_2^2 = \{0, 0.2, 0.5, 0.8\}$. The interpretation of γ is as for the SPM. The corresponding G-matrices are given by (16), and the additive genetic correlation between the focal trait and the missing process is given by (17).

For each of these 12 parameter sets we generated 100 data sets with missing values based on the BAM presented in Section 2.4. Further, inference were made using both the MAR-model and the SPM.

2.7 Joint model for field data

We now set up a shared parameter model for the barn owls system presented in Section 2.1. The trait of interest is the diameter of black plumage spots expressed on the ventral body side. This trait has previously been shown to be under strong genetic control and not significantly sensitive to rearing environment or body condition (Roulin and Djikstra, 2003). More recent studies have also revealed that spot diameter has a partially sex-linked inheritance (Roulin et al., 2010). As spot diameter is highly heritable, the shared parameter model should be able to account for potential non-random missingness in relation to this trait, making it suitable for demonstration of the proposed methodology.

Roulin et al. (2010) found that selection exerted on spot size directly, or on unmeasured traits highly genetically correlated with spot size favored

females with large spots and weakly favored males with smaller spots. Strong directional selection on females caused an increase in spot diameter in the population over the study period from 1996 to 2007. The results indicate that spot diameter is under viability selection and thus a modeling approach assuming MAR, like the animal model, might not be valid.

We follow the modeling framework presented in Section 2.4, but to account for sex-linked inheritance we used an extended animal model as presented in Roulin et al. (2010). Further, as selection seems to favor opposite characteristics of spot diameter in the two sexes, we allow the parameters in the missing data process to be sex specific. Hence, we have two sets of parameters for the missing data process, one for males and one for females. The full model then reads:

$$y_i = \boldsymbol{\beta}_{sex(i)} + \boldsymbol{\beta}_{year(i)} + a_i + z_i + \epsilon_i \tag{13a}$$

$$\text{logit}(\pi_i^m) = \alpha + \boldsymbol{\kappa}_{year(i)} + \gamma_a^m a_i + \gamma_z^m z_i \tag{13b}$$

$$\text{logit}(\pi_i^f) = \alpha + \boldsymbol{\kappa}_{year(i)} + \gamma_a^f a_i + \gamma_z^f z_i \tag{13c}$$

where a superscript of m and f indicates parameters corresponding to males and females, respectively. The other parameters have an interpretation equivalent to that in model (1) and (8). Since the owls that hatched in 2000 are missing because they hatched in 2000, they have another missing process that the others. Therefore, these birds are not included in the model with their trait status nor their missing status. They only contribute to the model through the connections they provide in the pedigree.

When comparing our results in this study with the results in Roulin et al.

(2010), note that there are slight differences in the data, pedigrees and models used and that the sex-linked variance is for the heterogametic sex in Roulin et al. (2010), while it is for the homogametic sex here (which is twice as large).

3 Results

All models are fitted using integrated nested Laplace integration (INLA), and model choice are done using DIC. A brief description is given in Appendix B

3.1 Results, Simulation studies

The objective of simulation study 1 was to investigate model performance for different values of additive genetic variance and for the parameters governing the missing data process, i.e. the level of missingness (α) and the association between the missing process and the focal trait (γ). We calculated the bias of estimated additive genetic variance obtained by the MAR-model and the SPM in each simulation. The results are summarized in Table 1 and S2.

MAR performs well in terms of both bias and coverage under MCAR, i.e. when $\gamma = 0$, for any given values of σ_a^2 and α. This is in accordance with missing data theory, which states that a MAR-model will be valid under MCAR. Results are not (at least not systematically) sensitive to changes in σ_a^2 and α under MCAR. Further, it is clear from our study that estimated additive genetic variance obtained using the MAR-model is downward biased under a MNAR process ($\gamma = -1, -2$). To which extent, depends highly on the value for σ_a^2 and γ. Low values of these parameters ($\sigma_a^2 = 0.2$ and $\gamma = -1$)

yields relatively accurate estimates of σ_a^2, while high values ($\sigma_a^2 = 0.8$ and $\gamma = -2$) result in severe bias and poor coverage. This is a reasonable result, as the dependency between data and missingness decreases as γ approaches zero. Also, when σ_a^2 is low, there is less dependency between phenotypic and additive genetic values, and hence less dependency between the phenotypic values and missingness. Traits with low additive genetic variance are much influenced by other factors than genes. Due to the nature of the missing data process, the missingness is more random for low values of σ_a^2 and γ.

The results of the study with extreme association between breeding values and the missing process are given in Figure 3. We find that with a high heritability ($\sigma_a^2 = 0.8$ and $\sigma_\epsilon^2 = 0.2$) even a relatively low proportion of missing data (5%), can give a substantial downward bias ($\hat{\sigma}_a^2 - \sigma_a^2 = -0.18$) when assuming MAR. This bias gets more severe with larger proportion of missing data.

The results from simulation study 2 are summarized in Table 2 and Table S3. Comparing these results with the corresponding results for simulation study 1, i.e. with $\alpha = -1$ and $\sigma_a^2 = 0.5$, we find, as expected from theory, that the results are almost identical and independent of the value of σ_2^2. Thus, our SPM gives correct estimates for the additive genetic variance σ_a^2 also for data simulated from a BAM.

3.2 Results, Field data

Several models were fitted to determine the appropriate factors (sex and hatch year) to include in the data model in equation 13a, and also to decide

whether the autosomal- and/or Z-linked additive genetic component should comprise the shared parameter in the missing process models in equation 13b and 13c. The models were compared using the deviance information criterion (DIC), where the model with the lowest value of DIC is considered the best model (Spiegelhalter et al., 2002). The DIC strongly suggested that only sex should be included in the data model while only hatch year was to be included in the missing data processes. Further, the difference in DIC suggested a model with autosomal effect as shared parameter for males and that Z-linked effect as shared parameter for both males and females. Thus, we specified the shared parameter model as:

$$y_i = \boldsymbol{\beta}_{sex(i)} + a_i + z_i + \epsilon_i \tag{14a}$$

$$\text{logit}(\pi_i^m) = \alpha + \boldsymbol{\kappa}_{year(i)} + \gamma_a^m a_i + \gamma_z^m z_i \tag{14b}$$

$$\text{logit}(\pi_i^f) = \alpha + \boldsymbol{\kappa}_{year(i)} + \gamma_z^f z_i \tag{14c}$$

The corresponding MAR-model, used for comparison, is solely the extended animal model (14a).

Parameter estimates for both the SPM model and the MAR model are given in Table 3, and the marginal posterior distributions of autosomal- and Z-linked additive genetic variance are shown in Figure S1. Posterior mean of σ_a^2 is 0.46 (95% CI: 0.38 to 0.56) from SPM and 0.43 (95% CI: 0.36 to 0.54) from the MAR-model, and the posterior mean of σ_z^2 is 0.25 (95% CI: 0.17 to 0.36) from SPM and 0.27 (95% CI: 0.18 to 0.39) from the MAR-model.

Hence, phenotypic variation in spot diameter is dominated by additive genetic variance and a substantial part of this variation is attributed to Z-linked genes. The variance estimates differ only slightly between SPM and MAR, but a comparison of DIC values showed a difference of 24 in favor of the SPM, which implies SPM provides a substantially better fit than the MAR-model; the missing process is better explained by including the breeding values in the model.

According to SPM, there is a significant positive association between the missing process and the sex-linked breeding values of spot diameter for both males and females. The posterior mean of γ_z^f is 1.23 (95% CI: 0.70 to 1.78) and of γ_z^m is 0.77 (95% CI: 0.15 to 1.39), which indicates that individuals with larger z-linked spot breeding values are more likely to be missing than those with smaller breeding values, and more so for females than for males. The association between the missing process and autosomal spot diameter breeding values in males γ_a^m, is only slightly negative, with a posterior mean of -0.06 (95% CI: -0.47 to 0.37). Zero is well within the credible interval.

We have also calculated the posterior distribution of mean autosomal and z-linked breeding values for each hatch year for both the naive and the joint model, see Figure 4. Further, the posterior distributions for the difference in mean breeding values for the first (1996) and last (2007) year have been calculated for both autosomal and z-linked breeding values. The MAR model gives difference in autosomal breeding values 0.20 (95% CI: 0.10 to 0.31) and for sex-linked breeding values 0.05 (95% CI: -0.05 to 0.15), and the SPM gives differences in autosomal breeding values of 0.17 (95% CI: 0.06 to 0.28) and for sex-linked breeding values 0.11 (95% CI: 0.01 to 0.20). The MAR and

SPM gives similar differences, but the SPM model gives significant differences in both autosomal and z-linked breeding values, while the MAR only gives significant differences for the autosomal breeding values.

4 Discussion

Although some of the potential problems caused by ignoring the *missing fraction* of a population in evolutionary analysis were pointed out more than two decades ago (Grafen, 1988), this issue subsequently received little attention. Only recently has it again been brought to the attention of scientists studying natural populations (Hadfield, 2008; Nakagawa and Freckleton, 2008). Hadfield (2008) used missing data theory (Rubin, 1976) to show that the presence of non-ignorable missing data may lead to estimates of selection that are downward biased or even in the wrong direction. He also raised a call to extend existing quantitative genetic techniques to account for missing data and use such techniques in evolutionary biology studies.

We have introduced a framework that simultaneously model the missing process and the quantitative genetics of the trait of interest. A simulation study was carried out to explore when the (indirect) assumption of missing at random is critical, and how our joint models deals with this. Further, the methodology was used to reanalyse additive genetic parameters of a directionally selected melanin-based trait (in the form of black feather spots) in a Swiss barn owl population.

The simulation study showed that especially when the heritability is high and the association between the breeding values and the missing process is

strong, we get severely biased estimates for the genetic variance, and also very low coverage. This agrees with both Hadfield (2008) and other studies including theoretical work, simulation studies and case studies. For example, Blomquist (2010) examined three proxies for individual fitness in a natural population of rhesus macaques (*Macaca mulatta*) and found that estimates of the heritability of these traits were reduced by 35-60% when non-reproductive individuals were excluded from the analysis. Mojica and Kelly (2010) showed that there was strong selection for small flowers in the yellow monkey flower *Mimulus guttatus* when viability selection prior to trait expression was taken into account in addition to fecundity. Previous studies on *Mimulus*, that did not include survival to flowering in their analyses, gave the opposite conclusion, showing positive selection on flower size (see references in Mojica and Kelly (2010)).

Our model, the shared parameters model (SPM), only requires the pedigree and the trait measurements, which implicitly gives the missing data structure. The dependency between the missing process and the trait is modeled through a linear dependency between missingness and breeding values. As demonstrated in the barn owl study, other explanatory variables, for example hatch year, can be included in the analysis, both in the model for the trait and for the missing process. The modeling framework also allows non-genetic random effects such as maternal effects or nest effects to be included in the models. From the simulation study we have seen that in the presence of NMAR, the SPM model gives unbiased estimates and good coverage. If the SPM is used when there is no association between the missing process and the breeding values, we still get unbiased estimates, but with slightly

larger credible intervals than the MAR model, see Table 2 for $\gamma = 0$).

A quantity of key interest for evolutionary biologists is the rate and direction of adaptive evolution. To precisely predict the rate and direction of the adaptive evolution of a trait both the strength and selection of selection (i.e. relationship between phenotype and fitness) and the adaptive potential (i.e. the additive genetic variance) of the trait needs to be accurately estimated (Lynch and Walsh, 1998). In our simulation study we show that ignoring missing data in quantitative genetic analyses might lead to biased estimates of additive genetic variance. As a consequence, predictions of the potential rate of evolutionary change might be wrong. Our simulation study shows that this potential problem is particularly important for traits that have high heritability, and when the relationship between trait and missing process is strong (Table 1). However, we also find that estimates of additive genetic variance may be reduced by more than 30% in the presence of non-ignorable missing data even when the heritability of the trait is only 0.2 (Table 1) if the association between the breeding value of the trait and the missing process is strong.

Based on the previous finding that females displaying larger black spots are positively selected in the barn owl, a key aim of the present study was to determine whether this pattern of selection takes place even before female nestlings produce their feathers where spots are located (i.e. since we measured this plumage trait in nestling birds, a number of individuals that die prematurely are missing from our data set). In the study of barn owl spot size we found that accounting for the missing data process improved the model fit, but that the estimated additive genetic variances of spot size

did not change much (Figure 3, Table 5). In this case we have moderate heritability, and association, but relatively few individuals are missing (about 6 %). The closest parameter set in the simulation study is ($\sigma_a^2 = 0.5$, $\alpha = -2$ and $\gamma = -1$), and the lack of any effect on the quantitative genetic estimates when including the missing data process is in accordance with the simulation study. In an analysis of the same population, Roulin et al. (2010) modeled late survival (from fledging to recruitment) and quantitative genetics separately, and survival was modeled based on trait observations. In this paper the analysis of the trait and the missing process (nestling mortality) are done jointly and the missing model is based on breeding values.

Roulin et al. (2010) found that there was a strong negative relationship between spot size and the probability of becoming missing between fledging and recruitment in females (i.e. a positive relationship between spot size and survival), and a weak positive relationship in males (Table 2 in Roulin et al. (2010)). Our results (on nestling mortality) shows that there is a positive relationship between Z-linked additive effects and missingness for both males and females, i.e. that smaller breeding values give a higher probability of surviving until fledging. Hence, there seems to be opposite selection processes for different life phases for female barn owls. Brood reduction is the dominating cause of death for nestlings, while traffic accidents is the dominating cause of death between fledging and recruitment (Baudvin, 1986). Therefore, it is not unreasonable that the selection is in opposite directions before and after fledging. Compared to Roulin et al. (2010) our study indicates a weaker positive selection on spot size in females and a stronger negative selection in males, which is in accordance with results in Roulin et al. (2011).

The results indicate that selection is taking place even before the black spots are produced. This emphasizes that spot size is genetically correlated with other traits that are under selection. Thus, owls displaying large spots were missing from our study not because they displayed large spots, but because this trait is associated with a number of phenotypes that are under selection. Indeed, it is shown that spot size displayed by mothers is correlated with offspring quality measures including parasite resistance, resistance to oxidative stress and an increase in corticosterone levels, appetite and the ability to withstand lack of food (Roulin and Ducrest, 2011). That nestling stage selection acts on genetically correlated traits, does of course not exclude the possibility that large spots are themselves under direct selection at a later stage as previous experimental studies suggested (Roulin, 1999; Roulin and Altwegg, 2007). That we estimate a positive trend in autosomal breeding values support this possibility. Our findings of negative associations between missingness and Z-linked additive genetic effects together with a positive trend in autosomal breeding values, calls for putting effort into finding the autosomal genes on which selection is positive between fledging and recruitment and the Z-linked genes on which selection is negative at the nestling stage. This study demonstrates that our SPM can be used not only to to account for missing data in quantitative genetic analyses, but also to explore evolutionary processes that can not be explored directly: the association parameter γ gives us information about the selection process.

5 Conclusion

Missing data in quantitative genetic studies can cause severely biased estimates of additive genetic variance and an underestimation of natural selection if the data are missing not at random (MNAR), seen from the trait of our interest, but the model used assumes missing at random (MAR). In such cases a joint model of the missing process and the quantitative genetics is needed. We have proposed the shared parameter model (SPM), which has proven to be successful through a simulation study. Whether a SPM is needed or not is hard to judge from a model assuming MAR: even with relatively few missing data (15%) the additive genetic variance estimate can be severely biased if the heritability is moderate to high and the association between the trait and missing process is strong. In any case, a MAR model does not give any information about the association, which might give important information on the selection process that causes the missing process. Hence, we recommend that a SPM is always fitted to check whether MAR can be assumed.

Acknowledgments

We thank fieldworkers and laboratory technicians for assistance, and S. Martino and H. Rue for technical INLA help. We are also thankful for comments and suggestions from Associate Editor Anne Charmantier, reviewer Jarrod Hadfield and two anonymous reviewers that improved the manuscript.

The study was supported by grants from the Swiss National Science Foundation (31003A-120517 to AR) and the Norwegian Research Council (grants

no. 191847 and 221956 to HJ).

References

Baudvin, H., 1986. La reproduction de la chouette effraie, *Tyto alba*. Le Jean le-Blanc 25:1–125.

Bechger, T., D. Boomsma, and H. Koning, 2002. A limited dependent variable model for heritability estimation with non-random ascertained samples. Behavior Genetics 32:145–151.

Blomquist, G. E., 2010. Heritability of individual fitness in female macaques. Evolutionary Ecology 24:657–669.

Clutton-Brock, T. and B. C. Sheldon, 2010. Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. Trends in Ecology & Evolution 25:562 – 573.

Cox, D. and E. Snell, 1989. Analysis of Binary Data. 2 ed. : Chapman and Hall/CRC.

Diggle, P. J. and M. G. Kenward, 1994. Informative drop-out in longitudinal data analysis (with discussion). Applied Statistics 43:49–93.

Ezard, T., P. Becker, and T. Coulson, 2007. Correlations between age, phenotype, and individual contribution to population growth in common terns. Ecology 88:2496–2504.

Grafen, A., 1988. Reproductive success, chap. On the uses of data on lifetime reproductive success. University of Chicago Press, Chicago, IL.

Hadfield, J. D., 2008. Estimating evolutionary parameters when viability selection is operating. Proc. R. Soc. B 275:723–734.

Henderson, C., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Holand, A., I. Steinsland, H. Jensen, and S. Martino, 2013. Animal model and integrated nested Laplace approximations. G3, Genes, Genomes, Genetics 3:1241–1251.

Im, S., R. Fernando, and D. Gianola, 1989. Likelihood inferences in animal breeding under selection - a missing-data theory view point. Genetics Selection Evolution 21:399–414.

Kruuk, L. E. B., 2004. Estimating genetic parameters in natural populations using the 'animal model'. Philos. Trans. R. Soc. London Ser. B 359:873–890.

Little, R. J. A., 1995. Modelling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 90:1112–1121.

Little, R. J. A. and D. B. Rubin, 2002. Statistical Analysis with Missing Data. 2 ed. Wiley, New York.

Lynch, M. and J. B. Walsh, 1998. Genetics and Analysis of Quantitative Traits. Sinauer Assocs., Inc.

Mojica, J. P. and J. K. Kelly, 2010. Viability selection prior to trait expression is an essential component of natural selection. Proceedings of the Royal Society B-Biological Sciences 277:2945–2950.

Nakagawa, S. and R. P. Freckleton, 2008. Missing in action: the dangers of ignoring missing data. Trends in Ecology & Evolution 23:592–596.

O'Hara, R. B., J. Cano, O. Ovaskinen, C. Teplitsky, and J. S. Alho, 2008. Bayesian Approaches in Evolutionary Quantitative Genetics. Journal of Evolutionary Biology 21:949–957.

Piepho, H. and J. Mohring, 2006. Selection in cultivar trials - Is it ignorable? Crop Science 46:192–201.

Rebke, M., T. Coulson, P. H. Becker, and J. W. Vaupel, 2010. Reproductive improvement and senescence in a long-lived bird. Proceedings of the National Academy of Sciences 107:7841–7846.

Roulin, A., 1999. Nonrandom pairing by male barn owls (Tyto alba) with respect to a female plumage trait. Behavioral Ecology 10:688–695.

Roulin, A. and R. Altwegg, 2007. Breeding rate is associated with pheomelanism in male and with eumelanism in female barn owls. Behavioral Ecology 18:563–570.

Roulin, A., R. Altwegg, H. Jensen, I. Steinsland, and M. Schaub, 2010. Sex-dependent selection on an autosomal melanic female ornament promotes the evolution of sex ratio bias. Ecology Letters 13:616–626.

Roulin, A., S. Antoniazza, and R. Burri, 2011. Spatial variation in the temporal change of male and female melanic ornamentation in the barn owl. Journal of Evolutionary Biology 24:1403–1409.

Roulin, A. and C. Djikstra, 2003. Genetic and environmental components of variation in eumelanin and phaeomelanin sex-traits in the barn owl. Heredity 90:359–364.

Roulin, A. and A.-L. Ducrest, 2011. Association between melanism, physiology and behaviour: A role for the melanocortin system. European Journal of Pharmacology 660:226–233.

Roulin, A., W. Möller, L. Sasvari, C. Djikstra, A. L. Ducrest, and C. Riols, 2004. Extra-pair paternity, testes size and testosterone level in relation to colour polymorphism in the barn owl Tyto alba. J. Avian Biol 35:492–500.

Rubin, D., 1976. Inference and missing data. Biometrika 63:581–590.

Rue, H. and S. Martino, 2006. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. Journal of Statistical Planning and Inference 137:3177–3192.

Rue, H., S. Martino, and N. Chopin, 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Statist. Soc. B 71:319–392.

Sorensen, D. and D. Gianola, 2002. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer-Verlag, New York.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde, 2002. Bayesian measures of model complexity and fit. Journal Of The Royal Statistical Society Series B 64:583–639.

Steinsland, I. and H. Jensen, 2010. Utilizing Gaussian Markov Random Field Properties of Bayesian Animal Models. Biometrics 66:763–771.

Van den Brink, V., V. Dolivo, X. Falourd, A. Dreiss, and A. Roulin, 2012. Melanic color-dependent anti-predator behavior strategies in barn owl nestlings. Behavioral Ecology 23:473–480.

Verbeke, G. and G. Molenberghs, 2000. Linear mixed models for longitudinal data. Springer, New York.

Vonesh, E. F., T. Greene, and M. D. Schluchter, 2006. Shared parameter models for the joint analysis of longitudinal data and event times. Statistics in Medicine 25:143–163.

A Derivation of relation between shared parameter model and bivariate model

To set up the bivariate model we first define two independent genetic fields; $\boldsymbol{c_1} \sim N(0, A)$ and $\boldsymbol{c_2} \sim N(0, A)$), where A is the additive genetic relationship matrix. Next we define $\boldsymbol{a} = \sigma_a \boldsymbol{c_1}$ with $\sigma_a > 0$ and $\boldsymbol{u} = \rho \boldsymbol{c_1} + \sigma_2 \boldsymbol{c_2}$ with $\sigma_2 > 0$. Using the fact that $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are independent, it is straightforward to show that

$$\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{u} \end{bmatrix} \sim \mathrm{N}(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, G_0 \otimes A) \tag{15}$$

where $\otimes$ denotes a Kronecker product and

$$G_0 = \begin{bmatrix} \sigma_a^2 & \rho\sigma_a \\ \rho\sigma_a & \rho^2 + \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_a^2 & \gamma\sigma_a^2 \\ \gamma\sigma_a^2 & \gamma^2\sigma_a^2 + \sigma_2^2 \end{bmatrix} \tag{16}$$

for $\gamma = \frac{\rho}{\sigma_a}$. We recognize this as the genetic part of a BAM (e.g. Sorensen and Gianola, 2002, pg. 578), where G_0 is known as the additive genetic covariance matrix. The additive genetic correlation is given as

$$\mathrm{corr}(u_i, a_i) = \frac{\gamma\sigma_a^2}{\sqrt{\sigma_a^2(\gamma^2\sigma_a^2 + \sigma_2^2)}} \tag{17}$$

$\boldsymbol{c_1}$ can be interpreted as the genes that influence our focal trait, with σ_a as scaling factor for the focal trait and $\rho = \gamma\sigma_a$ as scaling factor for the genetic part of the missing process for these genes. The missing process might have an additive genetic part not shared, $\boldsymbol{c_2}$, scaled with σ_2. We see from equation (16) that an appropriate choice of $(\sigma_a^2, \sigma_u^2, \rho)$ can give us any G-matrix, and

hence this specification using $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ is an alternative to specifying the G_0-matrix directly.

To complete the BAM we give the focal trait likelihood as in Section 2.2, Equation (1). For the pre-juvenile survival / missing process the likelihood model is similar to the one specified in Section 2.3; $m_i \sim Bin(1, \pi_i)$ with

$$\text{logit}(\pi_i) = \boldsymbol{v}_i^T \kappa + u_i = \boldsymbol{v}_i^T \kappa + \gamma a_i + \sigma_2 c_{2i}, \quad i = 1, ..., N. \tag{18}$$

If we compare (18) with the model for the SPM in (8) we find that the only difference is that the term $\sigma_2 c_{2i}$ is added. As independent random effects are confounded with the link for binary likelihoods (Cox and Snell, 1989). This implies that it is not possible to estimate (independent) environmental effects for binary traits. Because of this we do not include an environmental effect ϵ_i for the missing process.

To finalize the Bayesian model priors have to be assigned to hyper-parameters. It is common to give the G_0-matrix the conjugate inverse-Wishart distribution, which corresponds to a inverse-gamma prior for the variance σ_a^2 and a Gaussian prior for γ.

The model can be illustrated using a graph, see figure 2, panel A. From Figure 2 B, it seems as $\boldsymbol{y}$ and $\boldsymbol{m}$ relates to c_1 in a symmetric way. But they do not as we require that the additive genetic variance $\sigma_a^2 > 0$, while γ can take any real number, including 0.

We now assume that properties of the additive genetic effects of the focal trait and their association with the missing process / pre-juvenile survival is of interest, and not the additive genetic variance of the missing process itself.

According to missing data theory (Little and Rubin, 2002), only variables that the focal trait and the missing process have in common have to be included in the model. In our case only $\boldsymbol{c_1}$ is common, and hence $\boldsymbol{c_2}$ can be omitted from the analysis. In our setting we can explain this as only the effects of genes that influence both the focal trait and the missing process has to be included in the analysis, while the effect of genes that do influence the missing process, but not the focal trait can be omitted. This model can be illustrated with the graph in Figure 2 panel B, and coincides with the SPM. The parameters and variables that are denoted similar in the corresponding BAM and SPM should be interpreted similarly, and estimates will also be similar. It is important to note that in a SPM we do not calculate the genetic variance of the missing process, or the breeding values of the missing process. We can see this in our set up since σ_2^2 is not calculated, and hence we do not have an estimate of the additive genetic variance of the missing process nor the breeding values $\boldsymbol{u}$ of the missing process. The quantity $\gamma\boldsymbol{a}$ is not the breeding value of the missing process, but the part of the additive effect the missing process share with the focal trait.

B Parameter estimation and model choice

We use integrated nested Laplace approximations (INLA) (Rue and Martino, 2006; Rue et al., 2009) to estimate relevant parameters from our models. INLA is a new non-sampling based approach to Bayesian inference available for latent Gaussian Markov random field (GMRF) models. Markov chain Monte Carlo (MCMC) is currently the standard tool for Bayesian inference

for such models. MCMC methods are however computationally very expensive and might suffer from poor mixing and convergence properties. INLA provides a fast deterministic alternative to MCMC, to accurately approximate the posterior marginals of interest.

INLA utilize two basic properties that many latent Gaussian models satisfies. The first is that the latent field $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\epsilon}, \boldsymbol{\kappa})$ admits conditional independence properties, such that the latent field is a GMRF with a sparse precision matrix (inverse covariance matrix). This enables the use of fast numerical methods for sparse matrices, which INLA benefits from in calculations. The second property is that the number of non-Gaussian hyperparameters must be small, to allow for fast numerical integration. Currently, INLA can handle models with up to 10-15 non-Gaussian hyperparameters.

It has be shown that animal models falls within the class of latent GMRF models (Steinsland and Jensen, 2010; Holand et al., 2013), and the INLA methodology is established and tested for animal models in Holand et al. (2013) including a comparing MCMC and INLA.

A further benefit the sparse structure the GMRF property gives is that also simulations from the models are fast, and this combined with fast inference enable simulation studies (Holand et al., 2013).

Model comparison in the analysis of field data are carried out using the deviance information criterion (DIC) (Spiegelhalter et al.,2002). The model with the smallest DIC is considered the best model, i.e. the model that would best predict a replicate data set which has the same structure as that currently observed. According to Spiegelhalter et al. (2002), differences in DIC of more than 10 should definitely rule out the model with the higher

DIC. In Holand et al. (2013) simulation studies showed that difference in DIC is an appropriate measure for identifying models with/with out additive genetic effects.

Caption Table 1

Bias $(\widehat{\sigma}_a^2 - \sigma_a^2)$ from simulation study 1. Each presented quantity is the mean of the bias for the 100 data sets in the simulation study with the corresponding parameters $(\alpha, \gamma, \sigma_a^2)$ using the shared parameter model (SPM) and the missing at random (MAR) model. Numbers in bold correspond to parameter sets for which the mean credible interval (see Table S2) does not cover the true additive genetic variance σ_a^2.

Caption Table 2

Bias $(\widehat{\sigma}_a^2 - \sigma_a^2)$ from simulation study 2 which has $\sigma_a^2 = 0.5$ and $\alpha = -1$ for all data sets. Each presented quantity is the mean of the bias for the 100 data sets in the simulation study with the corresponding parameters $(\alpha, \gamma, \sigma_a^2, \sigma_2^2)$ using the shared parameter model (SPM) and the missing at random (MAR) model. Numbers in bold correspond to parameter sets for which the mean credible interval (see Table S3) does not cover the true additive genetic variance σ_a^2

Caption Figure 3

The proportion of missing data against bias of additive genetic variance $(\widehat{\sigma}_a^2 - \sigma_a^2)$ using the MAR model with maximum association between breeding values and missing data (individuals with largest breeding values are missing).