NTNU – Trondheim
Norwegian University of
Science and Technology

Modelling and Inference for Bayesian Bivariate Animal Models using Integrated Nested Laplace Approximations

Eirik Dybvik Bøhn

Master of Science in Physics and Mathematics
Submission date: June 2014
Supervisor: Ingelin Steinsland, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem formulation

To formulate bivariate additive genetic models such that integrated nested Laplace approximations (INLA) can be used for inference. Further the methodology should be tested through simulation studies, and a case study of partial diallel design of Scots pine.

TMA4905
Industrial Mathematics, Master's Thesis: Modelling and Inference for Bayesian Bivariate Animal Models using Integrated Nested Laplace Approximations

Eirik Dybvik Bøhn

June 15, 2014

Abstract

In this study we focus on performing inference on bivariate animal models using Integrated Nested Laplace Approximation (INLA). INLA is a methodology for making fast non-sampling based Bayesian inference for hierarchical Gaussian Markov models. Animal models are generalized mixed models (GLMM) used in evolutionary biology and animal breeding to identify the genetic part of traits. Bivariate animal models are derived and shown to fit the INLA framework. Simulation studies are conducted to evaluate the performance of the models. The models are fitted to a real data set of Scots pine to investigate correlations and dependencies.

Sammendrag

I dette studiet fokuserer vi på å utføre inferens på bivariate slektstrebaserte modeller ved å bruke Integrated Nested Laplace Approximations (INLA). INLA er en metodikk for å utføre Bayesiansk inferens på hierarkiske Gaussiske Markov modeller. Slektstrebaserte modeller er generaliserte mikset modeller som blir brukt i evolusjonsbiologi og dyreavl for å identifisere andelen av et trekk som er bestemt av gener. Bivariate slektstrebaserte modeller blir utledet og vist til å passe i INLA rammeverket. Simuleringsstudier er utført for å evaluere modellene. Modellene er også tilpasset et ekte datasett for å undersøke korrelatsjon og avhengigheter.

Contents

Glossary of notations

Variable	Meaning
$y_{i,j}$	Trait (tree height) for individual i at age level j
$a_{i,j}$	Additive genetic effect for individual i at age level j
$u_{i,j}$	Breeding value for individual i at age level j
$z_{i,j}$	Individual environmental effects for individual i at age level j
$\sigma^2_{a_j}$	Additive genetic variance for age level j
$\sigma^2_{z_j}$	Individual variance for age level j
$\mathbf{A}$	Additive relationship matrix
$\mathbf{I}$	Identity matrix
$\mathbf{\Sigma}$	Covariance matrix
$\mathbf{Q}$	Precision matrix

Table 1: Variable glossary used in this study

1 Introduction

The ability to identify if a given trait is explainable by environmental effects or genetics is of great interest in evolutionary biology, animal breeding and plant breeding. In the study of quantitative traits it is interesting to identify the cause of the given trait. Quantitative traits are the product of multiple genes acting additively. Genes are said to act additively if a set of two or more genes have the same effect as the sum of those same genes individually (Lynch & Walsh 1998).

The essence of the model I use, the animal model, is the assumption that animal i's trait, y_i, can be divided into a genetic part, a_i, and an individual part z_i. The genetic part, a_i are the additive genetic effects, also known as the breeding value. To calculate the breeding value for individual i, its pedigree and data from its relatives are required.

Calculation of breeding values have been popular in animal and plant breeding for decades. This has been successful in many disciplines, e.g. increased meat yield from beef cattle and higher milk production in dairy cattle (Simm 1998).

The modelling is performed in a Bayesian framework. All parameters are then considered random variables, an it is (in theory) straightforward to account for all uncertainty in parameter estimates. Bayesian modelling also solves many of the issues regarding analysis of breeding values discussed in (Postma 2006) and (Wilson et al. 2009) as both breeding values and functions of breeding values are considered random variables, and hence both uncertainty and dependencies are accounted for. This flexibility has made Bayesian animal models increasingly popular.

Both in plant and animal breeding and in evolutionary biology, it is often of interest to consider several traits simultaneously, e.g. amount and quality of milk. In plant and animal breeding selection of multiple traits is often desired. This requires knowledge about the additive genetic variances and correlation between traits. The additive genetic covariance matrix is also of interest to understand evolution in the wild because it contains information about evolutionary trajectories of several traits. (Lynch & Walsh 1998)

This study is based on the same dataset as (Finley et al. 2009). The data consists of height measurements from two grids with Scots pine (*Pinus sylvestris L.*)

in northern Sweden. The northern grid contains 2598 trees and the southern grid contains 2372 trees. In addition to the trait (Finley et al. 2009) examined, tree height at 26, we include tree height at age 10 as well.

To be able to look at the correlation between the two measurements, we propose a bivariate animal model. The bivariate animal model is under the same assumptions as the univariate animal model and allows us model two traits simultaneously. Modelling two traits simultaneously enables us to take both additive genetic correlation and individual environmental correlation into account. We will investigate the performance of several bivariate animal models, specified in Section 2.7.

To do inference on Bayesian animal models there are two popular methods, Markov Chain Monte Carlo, MCMC, and Integrated nested Laplace approximation, INLA. INLA being the more recent method. In (Holand et al. 2013) these two methods are compared by using them on Bayesian models that include genetic terms. Their conclusion was that MCMC is flexible, but slow and INLA is less flexible, but faster. In their example the computation times were 24 hours for MCMC and 1 hour for INLA. In this paper I will use the INLA method. A brief description of INLA is found in Section 2.5.

The goal of this paper is to demonstrate that also the bivariate animal model fits the INLA framework. When it is established that it is possible to use INLA for inference on the models, simulation studies are performed to explore identifiability issues with the bivariate animal model. The simulation studies are based on the same pedigree as the Scots pine data.

After the simulation studies are concluded, a case study is performed. The case study looks at the Scots pine data and assess the correlation between the two measurements.

This paper is organized as follows:
Section 2 gives a overview of the data, presents our model and a brief description of the methods that were used . Section 3 contains four simulation studies. Section 4 presents the results with some comments. Section 5 contains the case study. Section 6 includes discussion of the results as well as some interesting topics for further research. Appendix A gives an example of R-INLA implementation.

2 Background and models

This section presents the Scots pine study system. The simulation studies are based on this system, and it is analysed in Section 4.

2.1 Scots pine data

Our study system consists of two plantations of Scots pine in northern Sweden. It started out in 1971 by breeding 52 parent trees according to a partial diallel design.(Figure 1)

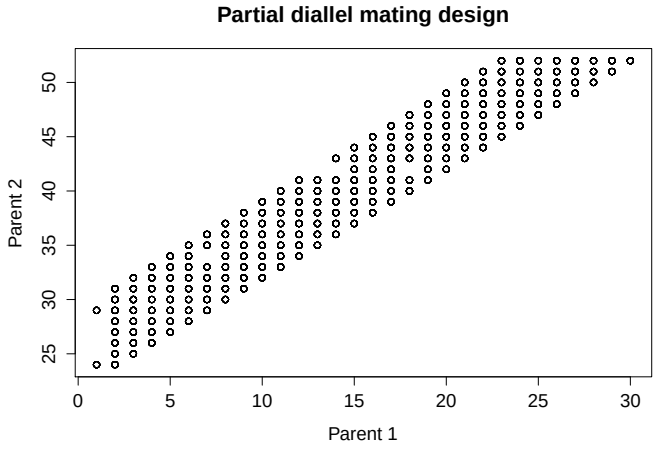


Figure 1: Mating design for the Scots pine

The parent trees were assumed unrelated and the seedlings were planted out on the grids unrestricted randomly. Each grid is divided into 8.8 x 22 m blocks each containing 40 seedlings placed on a 2.2 x 2.2 m grid. The northern grid contained 105 blocks and the southern grid contained 99 blocks.(Figure 2)
The design was done by Skogforsk (trial identification S23F7110264 Vindeln).

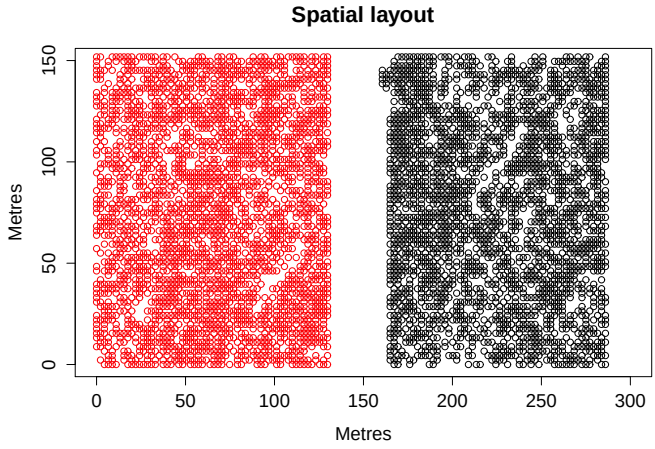
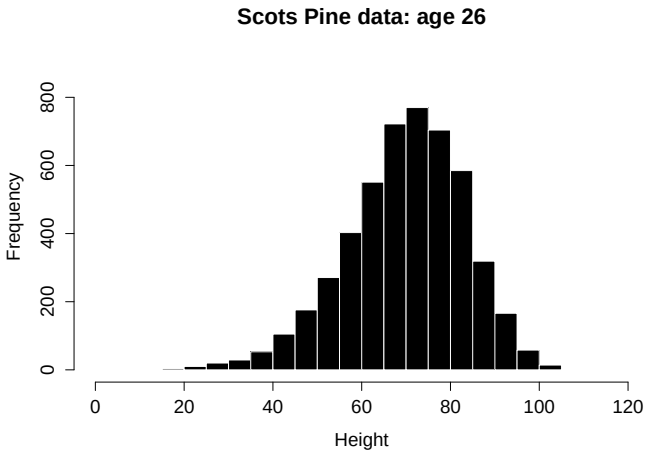
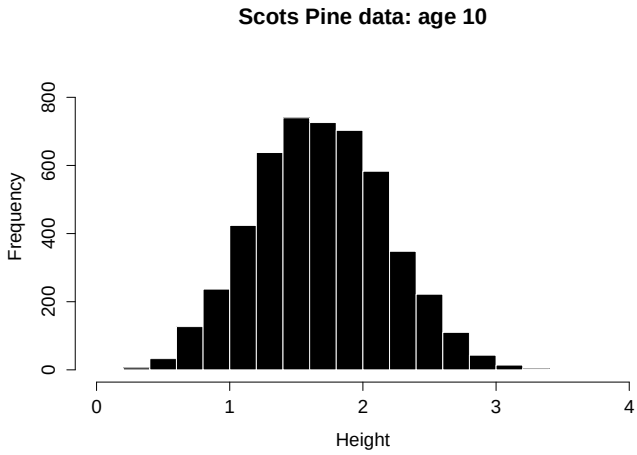


Figure 2: Spatial layout of the surviving trees. The red dots are the northern grid and the black dots are the southern grid

In 1997 there were 4970 surviving trees, 2598 on the northern grid and 2372 on the southern grid. Different kinds of measurements were done on the trees, including tree height, during their lifetime. This paper will look at the height measurements at age 10 and age 26 (Figure 3a and 3b). Ten of the trees died between age 10 and 26 and is therefore removed from the dataset.



(a) Height of the Scots Pine at age 26



(b) Height of the Scots Pine at age 10

Figure 3: Height of the Scots Pine data

2.2 Univariate animal model

The animal model is based on the assumption that the trait is determined by genetic and environmental effects, where the genetic effects is known as the breeding value.

$$\text{Trait} = \text{Genetic effects} + \text{Environmental effects}$$

To be able to identify the cause of the given trait we use a methodology that is able to utilize the pedigree of the population. This enables us to do inference and get information about how large contribution the different effects have on the trait.

A general linear mixed model (GLMM) made based on these assumptions is called the animal model and is used to make inference about the genetic parameters. The model links trait values to genetic and environmental effects using information about the relation between the individuals. This model has been used for breeding purposes in plant- and animal breeding for a long time with great success.
We present a Gaussian animal model

$$y_i = \beta_0 + a_i + z_i \tag{1}$$

where y_i is some trait, i.e height, β_0 is the intercept, a_i is the additive genetic effect

and z_i is the individual environmental effect.
Equation (1) can be written in matrix form for the whole population

$$\mathbf{y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{W}\mathbf{a} + \mathbf{z} \tag{2}$$

where $\mathbf{y}$ is a vector of traits, i.e heights, $\boldsymbol{\beta}$ is a vector with the intercepts, $\mathbf{I}$ is the identity matrix, $\mathbf{a}$ is a vector with the additive genetic effects, $\mathbf{W}$ and $\mathbf{B}$ are known incidence matrices and $\mathbf{z}$ is a vector with the individual environmental effects. The individual environmental effects are assumed independent and Gaussian distributed; $\mathbf{z} \sim \mathbf{N}(0, \mathbf{I}\sigma_z^2)$. The breeding values are assumed to follow a dependency structure given by the pedigree and Gaussian distributed; $\mathbf{a} \sim \mathbf{N}(0, \mathbf{A}\sigma_a^2)$.

To perform Bayesian inference on the animal model, likelihood for the observed data and prior distributions for the latent variables and hyperparameters must be defined. The traits are Gaussian distributed.

$$\mathbf{y} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\beta} + \mathbf{W}\mathbf{a}, \mathbf{I}\boldsymbol{\sigma}_{\mathbf{z}}^{\mathbf{2}}) \tag{3}$$

The animal model is a latent Gaussian model, since the latent variable($\mathbf{a}$) is assigned a Gaussian prior.

$$\mathbf{a}|\sigma_a^2 \sim \mathcal{N}(0, \mathbf{A}\sigma_a^2) \tag{4}$$
$$\tag{5}$$

Further,

$$\mathbf{z}|\sigma_z^2 \sim \mathcal{N}(0, \mathbf{I}\sigma_z^2) \tag{6}$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, \mathbf{I}10^3) \tag{7}$$
$$\tag{8}$$

where $\mathbf{A}$ is the additive relationship matrix. $\mathbf{A}$ is a symmetric matrix whose elements are twice the coefficient of co-ancestry. The coefficient of co-ancestry is the probability that two homologous genes, one from individual i and the other from j, are identical by descent, i.e. are ascended from the same ancestral gene. (Lynch & Walsh 1998)

To complete the model priors need to be assigned to the hyperparameters, σ_z^2 and σ_a^2. We have chosen to give them independent inverse gamma priors.

$$\sigma_a^2 \sim \mathbf{IG}(0.5, 0.5) \tag{9}$$
$$\sigma_z^2 \sim \mathbf{IG}(0.5, 0.5) \tag{10}$$

2.3 Bivariate animal model

The bivariate animal model is an extension of the univariate animal model, Equation (1). It is based upon the same assumptions as the univariate animal model, i.e that a trait is determined by genetic and environmental effects. The bivariate animal model applies to situations were two response variables are of interest simultaneously. This can either be two different traits or a trait measured at two different life stages. Modelling two traits simultaneously have been of interest in animal breeding for a long time. (Smith 1936) considered the problem of selecting among varieties of wheat differing in yield and quality traits and (Hazel 1943) applied some of the ideas to pig breeding schemes where body weights and scores had been collected in each of the animals. We assume the following model

$$\begin{aligned} y_{i,1} &= \beta_{i,1} + a_{i,1} + z_{i,1} \\ y_{i,2} &= \beta_{i,2} + a_{i,2} + z_{i,2} \end{aligned} \tag{11}$$

where $y_{i,j}$ is some trait, $\beta_{i,j}$ is the intercept, $a_{i,j}$ is the additive genetic effect and $z_{i,j}$ is the individual environmental effect, for $j \in \{1, 2\}$.
Equation 11 can be written in matrix form for the whole population

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 & 0 \\ 0 & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \tag{12}$$

where $\mathbf{y}_j$ is a vector containing the traits, $\boldsymbol{\beta}_j$ is a vector containing the intercepts, $\mathbf{a}_j$ is a vector containing the additive genetic effects and $\mathbf{z}_j$ is a vector with the individual environmental effects. $\mathbf{B}_j$ and $\mathbf{W}_j$ are known incidence matrices.
We now set $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]^T$, $\boldsymbol{a} = [\boldsymbol{a}_1, \boldsymbol{a}_2]^T$ and $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]^T$ with appropriate partitions for matrices $\mathbf{B}$ and $\mathbf{W}$ such that

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & 0 \\ 0 & \mathbf{B}_2 \end{bmatrix} \text{ and } \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & 0 \\ 0 & \mathbf{W}_2 \end{bmatrix}$$

To perform Bayesian inference on the bivariate animal model, likelihood for the observed data and prior distributions for the latent variables and hyperparameters must be defined. The traits are Gaussian distributed

$$\mathbf{y} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\beta} + \mathbf{W}\mathbf{a}, \mathbf{I} \otimes \boldsymbol{\Sigma}_z) \tag{13}$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{\Sigma}_z = \begin{bmatrix} \sigma^2_{z_1} & \rho_z \sigma_{z_1} \sigma_{z_2} \\ \rho_z \sigma_{z_1} \sigma_{z_2} & \sigma^2_{z_2} \end{bmatrix}$ is the individual environmental covariance matrix.

The bivariate animal model is a latent Gaussian model, since the latent variables are assigned Gaussian priors.

$$\mathbf{a}|\mathbf{G}, \mathbf{A} \sim \mathcal{N}(0, \mathbf{G} \otimes \mathbf{A}) \tag{14}$$
$$\mathbf{z}|\mathbf{\Sigma}_z \sim \mathcal{N}(0, \mathbf{I} \otimes \mathbf{\Sigma}_z) \tag{15}$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, \mathbf{I}10^3) \tag{16}$$

where $\mathbf{A}$ is the additive relationship matrix and $\mathbf{G} = \begin{bmatrix} \sigma^2_{a_1} & \rho_a \sigma_{a_1} \sigma_{a_2} \\ \rho_a \sigma_{a_1} \sigma_{a_2} & \sigma^2_{a_2} \end{bmatrix}$ is the additive genetic covariance matrix. The covariance matrices, $\mathbf{G}$ and $\mathbf{\Sigma}_z$ are assigned inverted Wishart priors. To complete the model priors need to be assigned to the hyperparameters, $\sigma^2_{z_j}$ and $\sigma^2_{a_j}$. We have chosen to give them independent inverse gamma priors.

$$\sigma^2_{a_j} \sim \mathbf{IG}(0.5, 0.5) \tag{17}$$
$$\sigma^2_{z_j} \sim \mathbf{IG}(0.5, 0.5) \tag{18}$$

For a more in depth introduction to animal models, see (Sorensen & Gianola 2002).

2.4 Gaussian Markov Random Fields

In this section we briefly review Gausian Markov random fields (GMRF), for a more thorough description see (Steinsland & Jensen 2010) and (Rue & Held 2005).
Gaussian Markov random fields models are multivariate Gaussian models with a Makrov property. The Markov property refers to conditional independence structure, often visualized with a conditional independence graph. In a conditional independence graph, each variable is a node. If two variables are conditionally dependent, conditioned on all the other variables, there is a edge between their nodes.
In our setting, we can think of each node as a tree. For breeding values, we can find the conditional independence graph from the pedigree. A pedigree is a directional acyclic graph with arrows from parents to offspring. According to graph-theory (Wermuth & Lauritzen 1982), we find the conditional independence graph by inserting edges between parents with a common offspring, and removing the direction of the parents-offspring edges, see Figure 4. Hence, each tree is conditionally dependent only on its parents, its offspring and the other parent(s) of its offspring.

The Markov structure is reflected in the nonzero pattern of the precision matrix,

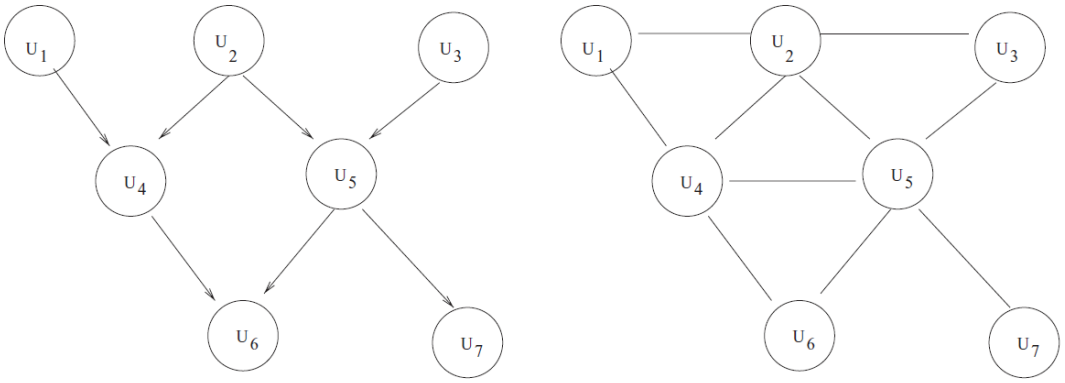


Figure 4: Left: the pedigree for the breeding values of seven trees. Tree 1 and 2 are the parents of tree 4, and tree 2 and 3 are the parents of tree 5. Right: the conditional independence graph that can be found from the pedigree. Edges are inserted between the trees with a common child, and the direction of the parent-child edges is removed.

the inverse of the covariance matrix, $\mathbf{Q} = \mathbf{\Sigma}^{-1}$: Only off-diagonal elements that correspond to conditionally dependent variables (two nodes with edges between) are nonzero. It is this sparseness of Q that imposes computational benefits for sampling and evaluation of GMRF. The computationally expensive part of both these operations is the calculation of $\mathbf{Q}$'s Cholesky factor $\mathbf{L}$, $\mathbf{Q} = \mathbf{L}^T\mathbf{L}$ ($\mathbf{L}$ is lower triangular). In most cases, a sparse precision matrix imposes a sparse Cholesky factor, fewer elements have to be calculated, and the computations are orders of magnitude cheaper than for a full $\mathbf{Q}$.

2.5 INLA and latent Gaussian models

Latent Gaussian models are hierarchical models were we assume a n_p-dimensional latent field $\mathbf{x}$ to be point-wise observed through $n_d \leq n_p$ data $\mathbf{y}$, $f(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n_d} f(\mathbf{y}_i|\mathbf{x})$. The latent field $\mathbf{x}$, includes both random and fixed effects and is assumed to have a Gaussian density conditional on hyperparameters $\phi : \mathbf{x}|\phi \sim \mathcal{N}(0, \mathbf{Q}^{-1}(\phi))$. Where $\mathbf{Q}$ is the precision matrix.

Our models are in a Bayesian framework, therefore the goal of inference is to obtain posterior distributions for the latent variables and hyperparameters. This can be

achieved by using Markov Chain Monte Carlo methods to sample from the posterior of he latent variables and hyperparameters. The drawback of using MCMC is that it is often time consuming and can suffer from slow mixing and convergence. An alternative approach is to use integrated nested Lapalce approxmiation. INLA is a fairly new methodology, introduced by (Rue et al. 2009), that provides a recipe for computing approximations to marginal posterior densities for the latent variables and hyperparameters.
INLA has the potential to be far more efficient, in terms of running times, than MCMC, if the latent Gaussian field satisfies some properties. First, the latent Gaussian field, $\mathbf{x}$, often of large dimension, admits conditional independence properties, that is it should be a GMRF with a sparse precision matrix $\mathbf{Q}$ (Rue & Held 2005). Second, because INLA needs to integrate over the hyperparamater space ϕ, the number of hyperparameters should not be too large. The models specified in Section 2.7 have sparse precision matrices and relatively few hyperparameters.
The INLA methodology explores the joint posterior of the hyperparameters, $\pi(\boldsymbol{\phi}|\mathbf{y})$, by utilizing that the identity

$$\pi(\boldsymbol{\phi}|\mathbf{y}) = \frac{\pi(\mathbf{x}_0, \boldsymbol{\phi}|\mathbf{y})}{\pi(\mathbf{x}_0|\boldsymbol{\phi}, \mathbf{y})} \tag{19}$$

is valid for any value of $\mathbf{x}_0$. It can be shown that both $\pi(\mathbf{x}_0, \boldsymbol{\phi}|\mathbf{y})$ and $\pi(\mathbf{x}_0|\boldsymbol{\phi}, \mathbf{y})$ can be evaluated efficiently up to a normalising constant, independent of $\boldsymbol{\phi}$ when the likelihood is Gaussian (Steinsland & Jensen 2010). Therefore we are able to evaluate the unormalised posterior, $\pi(\boldsymbol{\phi}|\mathbf{y})$, for every value of $\boldsymbol{\phi}$ by inserting a value of $\mathbf{x}$ in Equation (19). In order to choose good evaluation points of $\pi(\boldsymbol{\phi}|\mathbf{y})$ its mode is found by a numerical optimisation algorithm. Then he Hessian at the mode is used to distribute evaluation points.
Since all our models, Section 2.7, have Gaussian likelihoods, the accuracy of the approximations depends only on the numerical integration scheme.

For a more complete description of the INLA methodology, see (Rue et al. 2009). In this paper these procedures are done using the package R-INLA in R (`http://www.r-inla.org/`).

2.6 Bivariate Gaussian models

A vector $Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ has a bivariate Gaussian distribution, $\mathcal{N}(\mu, \Sigma)$, if its probability density function is

$$f(Y) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(Y-\mu)^T\Sigma^{-1}(Y-\mu)) \tag{20}$$

where Σ is the covariance matrix and μ is a column vector containing the means. If $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, where ρ is the correlation coefficient, the density function becomes

$$f(Y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}(\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2})) \tag{21}$$

2.6.1 Bivariate Gaussian distribution as the sum of independent Gaussian vectors

Any bivariate Gaussian distribution can be constructed as a sum of two univariate Gaussian vectors. To sample from a simple bivariate Gaussian model we can draw two independent Gaussian variables, $\dot{z}_{i,1} \sim \mathcal{N}(0, \sigma_{\dot{z}_1}^2)$ and $\dot{z}_{i,2} \sim \mathcal{N}(0, \sigma_{\dot{z}_2}^2)$. Then we define $y_{i,1} = z_{i,1}$ and $y_{i,2} = \alpha z_{i,1} + z_{i,2}$, where α is a scale parameter that defines the dependency between $y_{i,1}$ and $y_{i,2}$.

We now set $\mathbf{Y}_z = \mathbf{W}_z\mathbf{x}_z$, where $\mathbf{W}_z = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix}$ and $\mathbf{x}_z = \begin{bmatrix} \dot{z}_{i,1} \\ \dot{z}_{i,2} \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\dot{z}_1}^2 & 0 \\ 0 & \sigma_{\dot{z}_2}^2 \end{bmatrix})$, which allows us to calculate the distribution of $\mathbf{Y}_z$. Since $z_{i,1}$ and $z_{i,2}$ are independent we are able to use $\mathbf{Y}_z \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{W}_z\mathbf{\Sigma}_{\mathbf{x}_z}\mathbf{W}_z^T)$ to show that $\mathbf{Y}_z \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\dot{z}_1}^2 & \sigma_{\dot{z}_1}^2\alpha \\ \sigma_{\dot{z}_1}^2\alpha & \alpha^2\sigma_{\dot{z}_1}^2 + \sigma_{\dot{z}_2}^2 \end{bmatrix})$ The distribution of $\mathbf{Y}_z$ is the equivalent of a bivariate Gaussian distribution with $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, where $\sigma_1^2 = \sigma_{\dot{z}_1}^2$, $\sigma_2^2 = \alpha^2\sigma_{\dot{z}_1}^2 + \sigma_{\dot{z}_2}^2$ and $\rho = \frac{\alpha\sigma_{\dot{z}_1}^2}{\sqrt{\sigma_{\dot{z}_1}^2\sigma_{\dot{z}_2}^2}}$.

2.6.2 Bivariate animal model as the sum of independent Gaussian vectors

The same procedure can be used to construct a simple bivariate animal model. We start by defining two independent additive genetic effects, $\dot{a}_{i,1} \sim \mathcal{N}(0, A)$ and $\dot{a}_{i,2} \sim \mathcal{N}(0, A)$, where A is the additive genetic relationship matrix. Then set $y_{i,1} = \sigma_{\dot{a}_1}\dot{a}_{i,1}$, $y_{i,2} = \kappa\dot{a}_{i,1} + \sigma_{\dot{a}_2}\dot{a}_{i,2}$, where κ is a scale parameter that defines the dependency between $\dot{a}_{i,1}$ and $\dot{a}_{i,2}$.Then define $\mathbf{Y}_a = \mathbf{W}_a\mathbf{x}_a$,
where $\mathbf{W}_a = \begin{bmatrix} \sigma_{\dot{a}_1} & 0 \\ \rho & \sigma_{\dot{a}_2} \end{bmatrix}$ and $\mathbf{x}_a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix})$.
Since $a_{i,1}$ and $a_{i,2}$ are independent we can utilize that $\mathbf{Y}_a \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{W}_a\mathbf{\Sigma}_{\mathbf{X}_a}\mathbf{W}_a^T)$
to show that $\mathbf{Y}_a \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\dot{a}_1}A & \kappa\sigma_{\dot{a}_1}A \\ \kappa\sigma_{\dot{a}_1}A & \kappa^2A + \sigma^2_{\dot{a}_2}A \end{bmatrix}) = \mathbf{Y}_a \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} \otimes \mathbf{A})$,
where $\otimes$ denotes the Kronecker product.
We recognize $\mathbf{Y}_z$ as the genetic part of the bivariate animal model, Section 2.3, where $\sigma^2_{a_1} = \sigma^2_{\dot{a}_1}$, $\sigma^2_{a_2} = \kappa^2 + \sigma^2_{\dot{a}_2}$ and $\rho_a = \frac{\kappa\sigma_{\dot{a}_1}}{\sqrt{\sigma^2_{\dot{a}_1}(\kappa^2+\sigma^2_{\dot{a}_2})}}$.

The sum of $\mathbf{Y}_a$ and $\mathbf{Y}_z$ equals the bivariate animal model specified in Section 2.3 with no intercept, $\mathbf{\Sigma}_z = \begin{bmatrix} \sigma^2_{\dot{z}_1} & \sigma^2_{\dot{z}_1}\alpha \\ \sigma^2_{\dot{z}_1}\alpha & \alpha^2\sigma^2_{\dot{z}_1} + \sigma^2_{\dot{z}_2} \end{bmatrix}$ and $\mathbf{G} = \begin{bmatrix} \sigma^2_{\dot{a}_1} & \kappa\sigma_{\dot{a}_1} \\ \kappa\sigma_{\dot{a}_1} & \kappa^2 + \sigma^2_{\dot{a}_2} \end{bmatrix}$.
$\sigma^2_{\dot{a}_1}$ can be interpreted as the genes that influence height at age 10, $\kappa^2 + \sigma^2_{\dot{a}_2}$ as the genes that influence height at age 26 and $\kappa\sigma_{\dot{a}_1}$ as the genetic covariance. $\sigma^2_{\dot{z}_1}$ can be interpreted as the individual environmental effects that influence height at age 10, $\alpha^2\sigma^2_{\dot{z}_1} + \sigma^2_{\dot{z}_2}$ as the individual environmental effects that influence height at age 26 and $\sigma^2_{\dot{z}_1}\alpha$ as the individual environmental covariance.
This model, and variations of it, will be used in both the simulation studies, Section 3, and the case study, Section 5. The variations are specified in Section 2.7.

2.7 Bivariate Animal model specification

This section presents the models and priors used in the simulation studies, Section 3, and the case study, Section 5.

2.7.1 The basic bivariate model

The first model is without additive genetic effects, but with correlated individual environmental effects.

$$\begin{aligned} y_{i,1} &= z_{i,1} \\ y_{i,2} &= \alpha z_{i,1} + z_{i,2} \end{aligned} \tag{22}$$

The latent variables, $z_{i,1}$ and $z_{i,2}$, are assigned zero mean Gaussian prior distributions,

$$z_{i,j} \sim \mathcal{N}(0, \mathbf{I}\sigma^2_{z_j}) \tag{23}$$

where $\sigma^2_{z_j}$ is an unknown hyperparameter. To complete the full Bayesian model the hyperparameters, $\sigma^2_{z_j}$ and α, are assigned priors, we use inverse gamma and zero mean Gaussian respectively.

$$\sigma^2_{z_j} \sim \mathbf{IG}(0.5, 0.5) \tag{24}$$

$$\alpha \sim \mathcal{N}(0, 10) \tag{25}$$

This model, Equation (22), is the equivalent of a bivariate Gaussian distribution with covariance matrix $\Sigma = \begin{bmatrix} \sigma^2_{z_1} & \alpha\sigma^2_{z_1} \\ \alpha\sigma^2_{z_1} & \alpha^2\sigma^2_{z_1} + \sigma^2_{z_2} \end{bmatrix}$ and will later be referred to as the basic bivariate model.

2.7.2 The additive bivariate model

The second model assumes that the traits are determined by independent individual environmental effects and dependent additive genetic effects.

$$\begin{aligned} y_{i,1} &= a_{i,1} + z_{i,1} \\ y_{i,2} &= \kappa a_{i,1} + a_{i,2} + z_{i,2} \end{aligned} \tag{26}$$

The latent variables, $z_{i,j}$ and $a_{i,j}$, are assigned zero mean Gaussian prior distributions,

$$z_{i,j} \sim \mathcal{N}(0, \mathbf{I}\sigma^2_{z_j}) \tag{27}$$

$$a_{i,j} \sim \mathcal{N}(0, \mathbf{A}\sigma^2_{a_j}) \tag{28}$$

where $\sigma^2_{z_j}$ and $\sigma^2_{a_j}$ are unknown hyperparametrs. To complete the full Bayesian model the hyperparameters, $\sigma^2_{z_j}$, $\sigma^2_{a_j}$ and κ, are assigned priors. We use inverse gamma priors for the variances and zero mean Gaussian for κ.

$$\sigma^2_{z_j} \sim \mathbf{IG}(0.5, 0.5) \tag{29}$$
$$\sigma^2_{a_j} \sim \mathbf{IG}(0.5, 0.5) \tag{30}$$
$$\kappa \sim \mathcal{N}(0, 10) \tag{31}$$

This model, Equation (26), is the equivalent of a bivariate Gaussian distribution with covariance matrix $\Sigma = \begin{bmatrix} \sigma^2_{a_1} A + \sigma^2_{z_1} & \kappa\sigma^2_{a_1} A \\ \kappa\sigma^2_{a_1} A & \kappa^2\sigma^2_{a_1} A + \sigma^2_{a_2} A + \sigma^2_{z_2} \end{bmatrix}$ and will later be referred to as the additive bivariate model.

2.7.3 The environmental bivariate model

The third model assumes that the traits are determined by dependent individual environmental effects and independent additive genetic effects.

$$\begin{aligned} y_{i,1} &= a_{i,1} + z_{i,1} \\ y_{i,2} &= a_{i,2} + \alpha z_{i,1} + z_{i,2} \end{aligned} \tag{32}$$

The latent variables, $z_{i,j}$ and $a_{i,j}$, are assigned zero mean Gaussian prior distributions,

$$z_{i,j} \sim \mathcal{N}(0, \mathbf{I}\sigma^2_{z_j}) \tag{33}$$
$$a_{i,j} \sim \mathcal{N}(0, \mathbf{A}\sigma^2_{a_j}) \tag{34}$$

where $\sigma^2_{z_j}$ and $\sigma^2_{a_j}$ are unknown hyperparametrs. To complete the full Bayesian model the hyperparameters, $\sigma^2_{z_j}$, $\sigma^2_{a_j}$ and α, are assigned priors. We use inverse gamma priors for the variances and zero mean Gaussian for α.

$$\sigma^2_{z_j} \sim \mathbf{IG}(0.5, 0.5) \tag{35}$$
$$\sigma^2_{a_j} \sim \mathbf{IG}(0.5, 0.5) \tag{36}$$
$$\alpha \sim \mathcal{N}(0, 10) \tag{37}$$

This model, Equation (32), is the equivalent of a bivariate Gaussian distribution with covariance matrix $\Sigma = \begin{bmatrix} \sigma^2_{z_1} + \sigma^2_{a_1} A & \alpha\sigma^2_{z_1} \\ \alpha\sigma^2_{z_1} & \alpha^2\sigma^2_{z_1} + \sigma^2_{z_2} + \sigma^2_{a_2} A \end{bmatrix}$ and will later be referred to as the environmental bivariate model.

2.7.4 The full bivariate model

The fourth model assumes that the traits are determined by dependent individual environmental effects and dependent additive genetic effects.

$$\begin{aligned} y_{i,1} &= a_{i,1} + z_{i,1} \\ y_{i,2} &= \kappa a_{i,1} + a_{i,2} + \alpha z_{i,1} + z_{i,2} \end{aligned} \tag{38}$$

The latent variables, $z_{i,j}$ and $a_{i,j}$, are assigned zero mean Gaussian prior distributions,

$$z_{i,j} \sim \mathcal{N}(0, \mathbf{I}\sigma^2_{z_j}) \tag{39}$$

$$a_{i,j} \sim \mathcal{N}(0, \mathbf{A}\sigma^2_{a_j}) \tag{40}$$

where $\sigma^2_{z_j}$ and $\sigma^2_{a_j}$ are unknown hyperparametrs. To complete the full Bayesian model the hyperparameters, $\sigma^2_{z_j}$, $\sigma^2_{a_j}$, κ and α are assigned priors. We use inverse gamma priors for the variances and zero mean Gaussian priors for κ and α.

$$\sigma^2_{z_j} \sim \mathbf{IG}(0.5, 0.5) \tag{41}$$

$$\sigma^2_{a_j} \sim \mathbf{IG}(0.5, 0.5) \tag{42}$$

$$\kappa \sim \mathcal{N}(0, 10) \tag{43}$$

$$\alpha \sim \mathcal{N}(0, 10) \tag{44}$$

This model, Equation (38), is the equivalent of a bivariate Gaussian distribution with covariance matrix $\Sigma = \begin{bmatrix} \sigma^2_{z_1} + \sigma^2_{a_1} A & \kappa\sigma^2_{a_1} aA + \alpha\sigma^2_{z_1} \\ \kappa\sigma^2_{a_1} A + \alpha\sigma^2_{z_1} & \alpha^2\sigma^2_{z_1} + \sigma^2_{z_2} + \kappa^2\sigma^2_{a_1} A + \sigma^2_{a_2} A \end{bmatrix}$ and will later be referred to as the full bivariate model.

2.8 Our models as Latent GMRF models in INLA framework

The bivariate animal models, specified in Section 2.7, are latent GMRF, since the pedigree imposes a Markov structure. Individual i's breeding value is only dependent on it's parents, offspring and the other parents of its offspring. Information about the rest of the population will not affect that particular individual's breeding value. Since the breeding values form a GMRF, the inverse of the relationship marix, $\mathbf{A}^{-1}$, is a sparse matrix (Steinsland & Jensen 2010).

The models are latent GMRF models with latent field $\mathbf{x} = (\mathbf{z_i}, \mathbf{a_i})$ and hyperparameter vector θ includes the variances $(\sigma^2_{a_i}, \sigma^2_{z_i})$ and the parameters in the likelihood function. Since the inverse of $\mathbf{A}$ is sparse, the precision matrix for the latent field $\mathbf{x}$ is sparse.

Those properties makes our models suitable for fast and efficient computation of inference with INLA. We have implemented our models in INLA using multiple likelihoods and the copy function (Martins et al. 2013). The multiple likelihood feature allows us to fit our bivariate models with independent likelihoods. To achieve independent likelihoods we have specified the variance of the likelihoods, σ^2_L, to be very small, which can be interpreted as measurement errors. The disadvantage is that we are unable to use deviance information criterion (Spiegelhalter et al. 2002) to compare models. The copy feature allows us to estimate our dependence parameters.
The full bivariate animal model, Equation (38), can be formulated in the INLA framework with likelihoods

$$y_{i,1} \sim \mathcal{N}(\eta_{i,1}, \sigma^2_L) \tag{45}$$

$$y_{i,2} \sim \mathcal{N}(\eta_{i,2}, \sigma^2_L) \tag{46}$$

where $\eta_{i,1} = a_{i,1} + z_{i,1}$ and $\eta_{i,2} = \kappa a_{i,1} + a_{i,2} + \alpha z_{i,1} + z_{i,2}$. An example of implementation of the additive bivariate animal model, Equation (26), using R-INLA is found in Appendix A.

3 Simulation studies

Simulation studies were conducted to evaluate the performance of the the proposed models. Datasets were simulated based on the same pedigree and structure as the

Scots pine dataset. The simulated datasets were then fitted to the simulation model.

3.1 Simulation study 1

In this study phenotypic values were simulated according to the basic bivariate model, Equation (22), which states that the phenotypic values are subject to individual environmental effects with dependence.

$$\begin{aligned} y_{i,1} &= z_{i,1} \\ y_{i,2} &= \alpha z_{i,1} + \gamma z_{i,2} \end{aligned} \tag{47}$$

The individual environmental effects were generated from a zero mean univariate Gaussian distribution, $z_{i,j} \sim \mathcal{N}(0,1)$, α varied according to Table 2 and $\gamma = \sqrt{1-\alpha^2}$ to simulate standardized values. 100 datasets were simulated for each value of α.

α	0	0.2	0.4	0.6	0.8
$\text{corr}(y_{i,1}, y_{i,2})$	0	0.2	0.4	0.6	0.8

Table 2: Parameters used in the simulation of Equation (47)

For each simulation posterior mean, standard deviation and 95% credible intervals were calculated for $\sigma^2_{z_{i,1}}$, $\sigma^2_{z_{i,2}}$ and α. Results follows in Section 4.

3.2 Simulation study 2

In this study phenotypic values were simulated according to the additive bivariate animal model, Equation (26), which states that the phenotypic values are subject to both additive genetic effects and individual environmental effects, with dependence in the additive genetic effects.

$$\begin{aligned} y_{i,1} &= a_{i,1} + z_{i,1} \\ y_{i,2} &= \kappa a_{i,1} + \gamma a_{i,2} + z_{i,2} \end{aligned} \tag{48}$$

The individual environmental effects were generated from a zero mean univariate Guassian distribution, $z_{i,j} \sim \mathcal{N}(0, 0.5)$, the additive genetic effects, $a_{i,j}$, were generated from a zero mean multivariate Gaussian distribution with dependency structure

κ	0	0.2	0.4	0.6	0.8
$\text{corr}(y_{i,1}, y_{i,2})$	0	0.1	0.2	0.3	0.4

Table 3: Parameters used in the simulation of Equation (48)

determined by the additive relationship matrix $\mathbf{A}$, $a_{i,j} \sim \mathcal{N}(0, 0.5\mathbf{A})$, κ varied according to Table 3 and $\gamma = \sqrt{1-\kappa^2}$ to simulate standardized phenotypic values. 100 datasets were simulated for each value of κ.
For each simulation posterior mean, standard deviation and 95% credible intervals were calculated for $\sigma^2_{z_{i,1}}$, $\sigma^2_{z_{i,2}}$, $\sigma^2_{a_{i,1}}$, $\sigma^2_{a_{i,2}}$ and κ. Results follows in Section 4.

3.3 Simulation study 3

In this study phenotypic values were simulated according to the environmental bivariate animal model, Equation (32), which states that the phenotypic values are subject to both additive genetic effects and individual environmental effects, with dependence in the individual environmental effects.

$$\begin{aligned} y_{i,1} &= a_{i,1} + z_{i,1} \\ y_{i,2} &= a_{i,2} + \alpha z_{i,1} + \gamma z_{i,2} \end{aligned} \tag{49}$$

The individual environmental effects were generated from a zero mean univariate Guassian distribution, $z_{i,j} \sim \mathcal{N}(0, 0.5)$, the additive breeding values, $a_{i,j}$, were generated from a zero mean multivariate Gaussian distribution with dependency structure determined by the additive relationship matrix $\mathbf{A}$, $a_{i,j} \sim \mathcal{N}(0, 0.5\mathbf{A})$, α varied according to Table 4 and $\gamma = \sqrt{1-\alpha^2}$ to simulate standardized phenotypic values. 100 datasets were simulated for each value of α.

α	0	0.2	0.4	0.6	0.8
$\text{corr}(y_{i,1}, y_{i,2})$	0	0.1	0.2	0.3	0.4

Table 4: Parameters used in the simulation of Equation (49)

For every simulation posterior mean, standard deviation and 95% credible intervals were calculated for $\sigma^2_{z_{i,1}}$, $\sigma^2_{z_{i,2}}$, $\sigma^2_{a_{i,1}}$, $\sigma^2_{a_{i,2}}$ and α. Results follows in Section 4.

3.4 Simulation study 4

In this study the phenotypic values were simulated according to the full bivariate animal model, Equation (38), which states that the phenotypic values are subject to both additive genetic effects and individual environmental effects, with dependence in both the additive genetic effects and the individual environmental effects.

$$\begin{aligned} y_{i,1} &= a_{i,1} + z_{i,1} \\ y_{i,2} &= \kappa a_{i,1} + \gamma_\kappa a_{i,2} + \alpha z_{i,1} + \gamma_\alpha z_{i,2} \end{aligned} \tag{50}$$

The individual environmental effects were generated from a zero mean univariate Guassian distribution, $z_{i,j} \sim \mathcal{N}(0, 0.5)$, the additive breeding values, $a_{i,j}$, were generated from a zero mean multivariate Gaussian distribution with dependency structure determined by the additive relationship matrix $\mathbf{A}$, $a_{i,j} \sim \mathcal{N}(0, 0.5\mathbf{A})$, α and κ varied according to Table 5 and $\gamma_\kappa = \sqrt{1-\kappa^2}$ and $\gamma_\alpha = \sqrt{1-\alpha^2}$ to simulate standardized phenotypic values. 100 datasets were simulated for each combination of κ and α, resulting in 2500 datasets.

κ	0	0.2	0.4	0.6	0.8
α	0	0.2	0.4	0.6	0.8

Table 5: Parameters used in the simulation of Equation (50)

$$corr(y_{i,1}, y_{i,2}) = \frac{\kappa + \alpha}{2} \tag{51}$$

For every simulation posterior mean, standard deviation and 95% mean credible intervals were calculated for $\sigma^2_{z_{i,1}}$, $\sigma^2_{z_{i,2}}$, $\sigma^2_{a_{i,1}}$, $\sigma^2_{a_{i,2}}$, κ and α. Results follows in Section 4.

4 Results simulation studies

In this section the results from the simulation studies, described in Section 3, are presented and briefly discussed. Further discussion is found in Section 6.
All the models are fitted using integrated nested Laplace integration (Section 2.5). 100 datasets are simulated for each value of the dependency parameters, κ and α.

4.1 Simulation study 1

The results from simulation study 1 are summarized in Tables 6, 7 and 8.
The basic bivariate model, Equation (22), performs well for all values of α. The bias, Table 6, is relatively small for all the estimated parameters. The coverage intervals, Table 7, are all above 90%, with the exception of $\sigma^2_{a_1}$ when $\alpha = 0.6$, which is 89%. The mean credible intervals, Table 8, all cover the true value and are relatively narrow.

α	$\hat{\alpha} - \alpha$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1} - \sigma^2_{z_1}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2} - \sigma^2_{z_2}$
0	0.000	1	-0.001	1	0.002
0.2	-0.002	1	-0.006	0.96	-0.002
0.4	-0.001	1	-0.009	0.84	0.005
0.6	-0.001	1	-0.002	0.64	-0.002
0.8	0.001	1	0.001	0.36	-0.001

Table 6: Bias ($\hat{\sigma}^2_x - \sigma^2_x$) from simulation study 1. Each value is the mean of the 100 simulations per value of α

α	Coverage	$\sigma^2_{z_1}$	Coverage	$\sigma^2_{z_2}$	Coverage
0	0.94	1	0.95	1	0.95
0.2	0.97	1	0.94	0.96	0.95
0.4	0.96	1	0.97	0.84	0.95
0.6	0.98	1	0.89	0.64	0.94
0.8	0.98	1	0.93	0.36	0.91

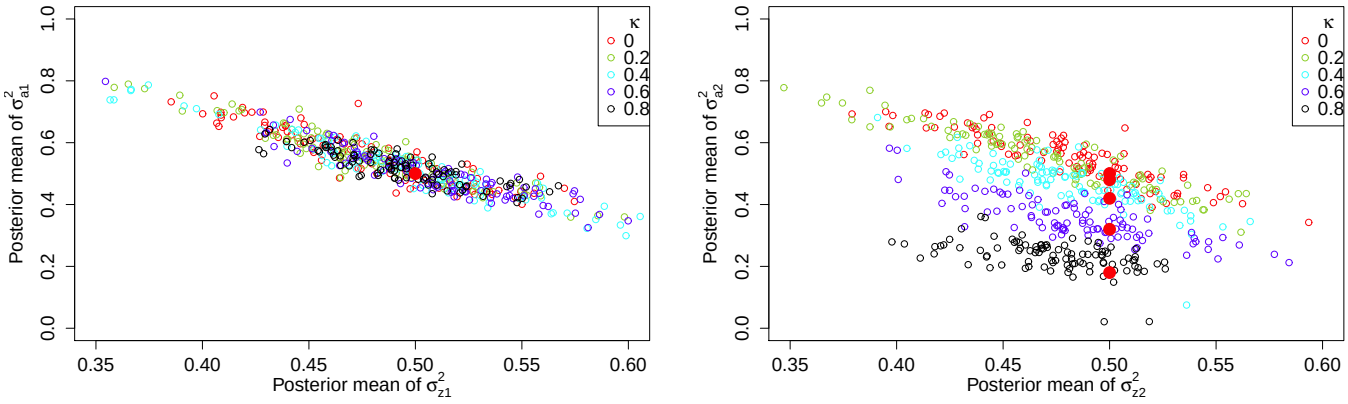
Table 7: Coverage intervals from simulation study 1. Percentages of estimates whose 95% credible intervals cover the true value.

α	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{z_2}$	CI
0	(-0.03,0.03)	1	(0.96,1.04)	1	(0.96,1.04)
0.2	(0.17,0.23)	1	(0.96,1.04)	0.96	(0.92,0.99)
0.4	(0.37,0.43)	1	(0.96,1.04)	0.84	(0.81,0.88)
0.6	(0.58,0.62)	1	(0.96,1.04)	0.64	(0.61,0.66)
0.8	(0.78,0.82)	1	(0.96,1.04)	0.36	(0.35,0.37)

Table 8: Mean credible intervals from simulation study 1. Each value is the mean upper and lower limit of the 95% CI. Intervals are in bold if the mean CI does not contain the true value.

4.2 Simulation study 2

The results from simulation study 2 are summarized in Tables 9, 10 and 11, as well as Figures 5a and 5b.
The additive bivariate model, Equation (26), performs well in regards of coverage intervals, Table 10, and mean credible intervals, Table 11. The coverage intervals are relatively large and all the mean credible intervals covers the true values. The bias, Table 9, is quite large for several of the parameters, e.g. for $\sigma^2_{a_{i,1}}$ and $\sigma^2_{z_{i,1}}$. The additive variances, $\sigma^2_{a_{i,1}}$ and $\sigma^2_{a_{i,2}}$, are clearly overestimated and the individual environmental variances, $\sigma^2_{z_{i,1}}$ and $\sigma^2_{z_{i,2}}$, clearly underestimated. Large credibility intervals together with baises might be symptoms of identification issues. Figures 5a and 5b shows the posterior mean of the additive genetic effects plotted against the individual environmental effects for $y_{i,1}$ and $y_{i,2}$ respectively. They show that the posterior means are definitely negatively correlated, which supports our suspicion regarding identification issues. The model does well overall, but seems to have some minor problems with allocating the variance between the additive genetic effects and the individual environmental effects.



(a) (b)

Figure 5: (a) Posterior mean of $\sigma^2_{a_{i,1}}$ plotted against the posterior mean of $\sigma^2_{z_{i,1}}$ from simulation study 2. The true values are $\sigma^2_{a_{i,1}} = \sigma^2_{z_{i,1}} = 0.5$, illustrated with a filled red circle. (b) Posterior mean of $\sigma^2_{a_{i,2}}$ plotted against the posterior mean of $\sigma^2_{z_{i,2}}$ from simulation study 2. The true values are $\sigma^2_{z_{i,2}} = 0.5$ and $\sigma^2_{a_{i,2}} = 0.5(1-\kappa^2)$, illustrated with filled red circles.

κ	$\hat{\kappa} - \kappa$	$\sigma^2_{a_1}$	$\hat{\sigma}^2_{a_1} - \sigma^2_{a_1}$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1} - \sigma^2_{z_1}$	$\sigma^2_{a_2}$	$\hat{\sigma}^2_{a_2} - \sigma^2_{a_2}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2} - \sigma^2_{z_2}$
0	-0.002	0.5	0.059	0.5	-0.022	0.5	0.053	0.5	-0.019
0.2	-0.012	0.5	0.056	0.5	-0.020	0.48	0.074	0.5	-0.029
0.4	-0.008	0.5	0.023	0.5	-0.006	0.42	0.049	0.5	-0.019
0.6	0.006	0.5	0.012	0.5	0.005	0.32	0.037	0.5	0.003
0.8	-0.028	0.5	0.019	0.5	-0.006	0.18	0.053	0.5	-0.026

Table 9: Bias $(\hat{\sigma}^2_x - \sigma^2_x)$ from simulation study 2. Each value is the mean of the 100 simulations per value of κ

κ	Coverage	$\sigma^2_{a_1}$	Coverage	$\sigma^2_{z_1}$	Coverage	$\sigma^2_{a_2}$	Coverage	$\sigma^2_{z_2}$	Coverage
0	0.96	0.5	0.80	0.5	0.92	0.5	0.82	0.5	0.89
0.2	0.88	0.5	0.83	0.5	0.90	0.48	0.73	0.5	0.88
0.4	0.85	0.5	0.83	0.5	0.82	0.42	0.88	0.5	0.95
0.6	0.88	0.5	0.86	0.5	0.82	0.32	0.85	0.5	0.90
0.8	0.86	0.5	0.95	0.5	0.87	0.18	0.84	0.5	0.93

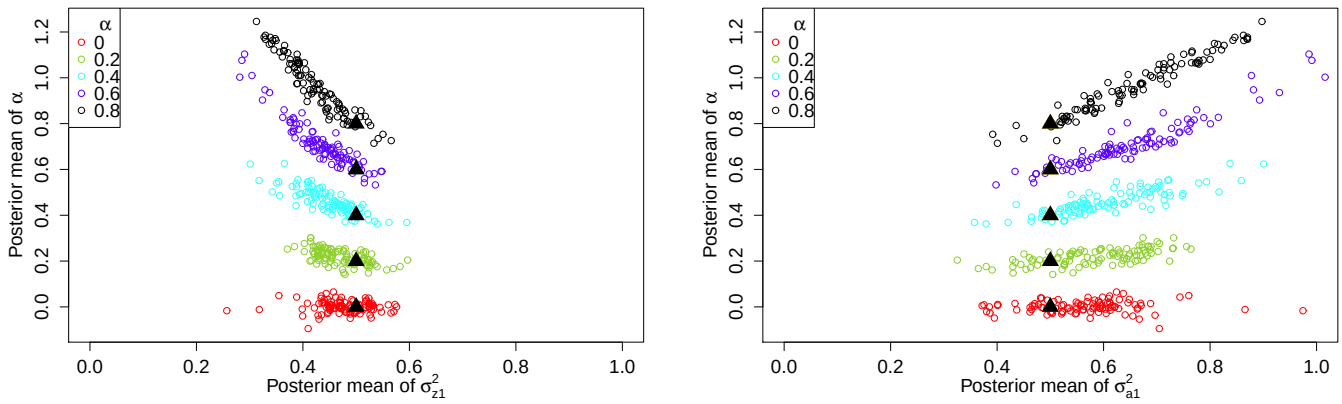
Table 10: Coverage intervals from simulation study 2. Percentages of estimates whose 95% credible intervals cover the true value.

κ	CI	$\sigma^2_{a_1}$	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{a_2}$	CI	$\sigma^2_{z_2}$	CI
0	(-0.07,0.07)	0.5	(0.42,0.71)	0.5	(0.41,0.57)	0.5	(0.42,0.70)	0.5	(0.41,0.57)
0.2	(0.10,0.28)	0.5	(0.42,0.71)	0.5	(0.41,0.56)	0.48	(0.41,0.70)	0.5	(0.40,0.56)
0.4	(0.26,0.52)	0.5	(0.40,0.70)	0.5	(0.43,0.56)	0.42	(0.34,0.61)	0.5	(0.41,0.57)
0.6	(0.45,0.74)	0.5	(0.40,0.64)	0.5	(0.45,0.56)	0.32	(0.25,0.49)	0.5	(0.42,0.57)
0.8	(0.62,0.92)	0.5	(0.43,0.64)	0.5	(0.44,0.55)	0.18	(0.14,0.34)	0.5	(0.42,0.55)

Table 11: Mean credible intervals from simulation study 2. Each value is the mean upper and lower limit of the 95% CI. Intervals are in bold if the mean CI does not contain the true value.

4.3 Simulation study 3

The results from simulation study 3 are summarized in Tables 12, 13 and 14, as well as Figures 6a and 6b.
The environmental bivariate model, Equation (32), performs adequate for low values of α. We see that the bias, Table 12, is relatively small, the coverage intervals, Table 13, are relatively large and the mean credible intervals, Table 14, covers the true values. When α increases problems arises. As α increases we see that the estimations are more biased, the coverage intervals decreasing and some of the mean credible intervals does not contain the true value. This is illustrated in Figure 6, where the posterior mean of $\sigma^2_{z_1}$ and $\sigma^2_{a_1}$ are plotted against the posterior mean of α. We see clearly from the figure that as α increases the estimates for $\sigma^2_{z_1}$ and $\sigma^2_{a_1}$ becomes less accurate. We suspect this may be related to prior sensitivity, which we will take a closer look at in Section 4.5.



(a) (b)

Figure 6: (a) Posterior mean of α plotted against the posterior mean of $\sigma^2_{z_{i,1}}$. The true value is $\sigma^2_{z_{i,1}} = 0.5$, illustrated with filled black triangles. (b) Posterior mean of α plotted against the posterior mean of $\sigma^2_{a_{i,1}}$. The true value is $\sigma^2_{a_{i,1}} = 0.5$, illustrated with filled black triangles.

α	$\hat{\alpha}-\alpha$	$\sigma^2_{a_1}$	$\hat{\sigma}^2_{a_1}-\sigma^2_{a_1}$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1}-\sigma^2_{z_1}$	$\sigma^2_{a_2}$	$\hat{\sigma}^2_{a_2}-\sigma^2_{a_2}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2}-\sigma^2_{z_2}$
0	0.001	0.5	0.064	0.5	-0.023	0.5	0.039	0.5	-0.008
0.2	0.019	0.5	0.083	0.5	-0.031	0.5	0.058	0.48	-0.027
0.4	0.060	0.5	0.109	0.5	-0.051	0.5	0.011	0.42	-0.015
0.6	0.117	0.5	0.154	0.5	-0.066	0.5	-0.029	0.32	-0.018
0.8	0.165	0.5	0.157	0.5	-0.076	0.5	-0.074	0.18	-0.020

Table 12: Bias ($\hat{\sigma}^2_x - \sigma^2_x$) from simulation study 3. Each value is the mean of the 100 simulations per value of α.

α	Coverage	$\sigma^2_{a_1}$	Coverage	$\sigma^2_{z_1}$	Coverage	$\sigma^2_{a_2}$	Coverage	$\sigma^2_{z_2}$	Coverage
0	0.92	0.5	0.79	0.5	0.89	0.5	0.82	0.5	0.84
0.2	0.89	0.5	0.66	0.5	0.91	0.5	0.84	0.48	0.89
0.4	0.75	0.5	0.62	0.5	0.72	0.5	0.83	0.42	0.82
0.6	0.59	0.5	0.45	0.5	0.53	0.5	0.83	0.32	0.76
0.8	0.40	0.5	0.35	0.5	0.35	0.5	0.55	0.18	0.64

Table 13: Coverage intervals from simulation study 3. Percentages of estimates whose 95% credible intervals cover the true value.

α	CI	$\sigma^2_{a_1}$	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{a_2}$	CI	$\sigma^2_{z_2}$	CI
0	(-0.05,0.05)	0.5	(0.42,0.71)	0.5	(0.41,0.57)	0.5	(0.40,0.70)	0.5	(0.42,0.58)
0.2	(0.16,0.27)	0.5	(0.44,0.73)	0.5	(0.40,0.57)	0.5	(0.42,0.70)	0.48	(0.39,0.54)
0.4	(0.37,0.68)	0.5	(0.48,0.73)	0.5	(0.39,0.53)	0.5	(0.40,0.66)	0.42	(0.35,0.47)
0.6	**(0.61,0.81)**	0.5	**(0.54,0.76)**	0.5	(0.38,0.50)	0.5	(0.37,0.62)	0.32	(0.25,0.36)
0.8	**(0.86,1.05)**	0.5	**(0.57,0.73)**	0.5	**(0.39,0.48)**	0.5	(0.35,0.55)	0.18	(0.13,0.20)

Table 14: Mean credible intervals from simulation study 3. Each value is the mean upper and lower limit of the 95% CI. Intervals are in bold if the mean CI does not contain the true value.

4.4 Simulation study 4

The results from simulation study 4 are summarized in Tables 15 and 16.
The full bivariate model, Equation (38), suffers from the same problems as the environmental bivariate model, Equation (32), and the additive bivariate model, Equation (26), combined. As κ increases we see that the estimations are more biased and the coverage intervals decreasing. The mean credible intervals are not included in this study, since the bias and coverage intervals already indicate poor inference. We also recognize that the bias for $\sigma^2_{a_j}$ and $\sigma^2_{z_j}$ have opposite signs as symptoms of identification issues. These overall poor inference results needs to be investigated further. We suspect they are related to prior sensitivity. In Section 4.5 we will run some diagnostic tests on the worst case scenarios of this simulation study. The parameter pairs we have chosen to investigate further are marked with * in the tables.

κ	$\hat{\kappa}-\kappa$	α	$\hat{\alpha}-\alpha$	$\sigma^2_{a_1}$	$\hat{\sigma}^2_{a_1}-\sigma^2_{a_1}$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1}-\sigma^2_{z_1}$	$\sigma^2_{a_2}$	$\hat{\sigma}^2_{a_2}-\sigma^2_{a_2}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2}-\sigma^2_{z_2}$
0	-0.042	0	0.021	0.5	0.006	0.5	0.001	0.5	-0.040	0.5	0.020
0	-0.032	0.2	0.015	0.5	0.044	0.5	-0.018	0.5	-0.015	0.48	0.004
0	0.071	0.4	0.049	0.5	0.174	0.5	-0.086	0.5	-0.004	0.42	0.002
0	0.325	0.6	-0.121	0.5	0.395	0.5	-0.185	0.5	0.049	0.32	0.023
0	0.514	0.8	-0.178	0.5	0.482	0.5	-0.259	0.5	0.099	0.18	0.051
0.2	-0.122	0	0.051	0.5	0.020	0.5	-0.005	0.48	-0.028	0.5	0.019
0.2	-0.050	0.2	0.028	0.5	0.093	0.5	-0.042	0.48	-0.010	0.48	0.001
0.2	-0.028	0.4	0.070	0.5	0.144	0.5	-0.064	0.48	-0.020	0.42	-0.009
0.2(*)	0.252	0.6	-0.099	0.5	0.522	0.5	-0.251	0.48	0.033	0.32	0.013
0.2	0.355	0.8	-0.077	0.5	0.451	0.5	-0.226	0.48	0.024	0.18	0.037
0.4	-0.123	0	0.049	0.5	0.049	0.5	-0.020	0.42	0.008	0.5	0.008
0.4	-0.069	0.2	0.008	0.5	0.144	0.5	-0.069	0.42	0.013	0.48	-0.006
0.4	-0.005	0.4	0.014	0.5	0.412	0.5	-0.188	0.42	0.008	0.42	-0.002
0.4	0.126	0.6	-0.102	0.5	0.530	0.5	-0.276	0.42	-0.032	0.32	0.024
0.4	0.128	0.8	-0.031	0.5	0.200	0.5	-0.095	0.42	0.008	0.18	0.010
0.6	-0.158	0	0.034	0.5	0.108	0.5	-0.047	0.32	0.043	0.5	0.008
0.6	-0.165	0.2	0.055	0.5	0.162	0.5	-0.076	0.32	0.015	0.48	0.001
0.6(*)	-0.136	0.4	-0.025	0.5	0.550	0.5	-0.260	0.32	0.058	0.42	-0.017
0.6	-0.001	0.6	0.006	0.5	0.420	0.5	-0.211	0.32	-0.042	0.32	0.021
0.6	0.074	0.8	-0.032	0.5	0.172	0.5	-0.080	0.32	-0.004	0.18	0.003
0.8	-0.243	0	-0.026	0.5	0.254	0.5	-0.128	0.18	0.092	0.5	-0.011
0.8	-0.304	0.2	0.045	0.5	0.345	0.5	-0.166	0.18	0.130	0.48	-0.029
0.8	-0.259	0.4	0.070	0.5	0.516	0.5	-0.256	0.18	0.085	0.42	-0.015
0.8	-0.068	0.6	-0.034	0.5	0.271	0.5	-0.131	0.18	0.017	0.32	-0.005
0.8	-0.031	0.8	0.017	0.5	0.326	0.5	-0.148	0.18	-0.007	0.18	0.001

Table 15: Bias $(\hat{\sigma}^2_x - \sigma^2_x)$ from simulation study 4. Each value is the mean of the 100 simulations per combination of κ and α. The rows marked with * is selected for further investigation in Section 4.5

κ	Cover	α	Cover	$\sigma^2_{a_1}$	Cover	$\sigma^2_{z_1}$	Cover	$\sigma^2_{a_2}$	Cover	$\sigma^2_{z_2}$	Cover
0	0.82	0	0.83	0.5	0.87	0.5	0.82	0.5	0.75	0.5	0.71
0	0.83	0.2	0.89	0.5	0.86	0.5	0.93	0.5	0.77	0.48	0.77
0	0.68	0.4	0.77	0.5	0.45	0.5	0.48	0.5	0.74	0.42	0.76
0	0.13	0.6	0.54	0.5	0.16	0.5	0.21	0.5	0.58	0.32	0.55
0	0.01	0.8	0.60	0.5	0.01	0.5	0.02	0.5	0.09	0.18	0.58
0.2	0.65	0	0.69	0.5	0.86	0.5	0.88	0.48	0.86	0.5	0.73
0.2	0.78	0.2	0.78	0.5	0.53	0.5	0.69	0.48	0.80	0.48	0.80
0.2	0.66	0.4	0.70	0.5	0.47	0.5	0.60	0.48	0.72	0.42	0.58
0.2(*)	0.02	0.6	0.37	0.5	0.03	0.5	0.04	0.48	0.62	0.32	0.40
0.2	0.07	0.8	0.53	0.5	0.08	0.5	0.11	0.48	0.38	0.18	0.45
0.4	0.69	0	0.75	0.5	0.78	0.5	0.90	0.42	0.92	0.5	0.87
0.4	0.74	0.2	0.78	0.5	0.42	0.5	0.46	0.42	0.83	0.48	0.78
0.4	0.89	0.4	0.79	0.5	0.19	0.5	0.27	0.42	0.76	0.42	0.76
0.4	0.11	0.6	0.26	0.5	0.00	0.5	0.01	0.42	0.61	0.32	0.42
0.4	0.47	0.8	0.75	0.5	0.40	0.5	0.48	0.42	0.64	0.18	0.65
0.6	0.57	0	0.82	0.5	0.59	0.5	0.74	0.32	0.90	0.5	0.88
0.6	0.46	0.2	0.69	0.5	0.40	0.5	0.53	0.32	0.79	0.48	0.69
0.6(*)	0.05	0.4	0.55	0.5	0.03	0.5	0.03	0.32	0.73	0.42	0.74
0.6	0.95	0.6	0.35	0.5	0.11	0.5	0.17	0.32	0.73	0.32	0.66
0.6	0.58	0.8	0.76	0.5	0.34	0.5	0.49	0.32	0.57	0.18	0.59
0.8	0.25	0	0.76	0.5	0.15	0.5	0.20	0.18	0.38	0.5	0.83
0.8	0.21	0.2	0.53	0.5	0.19	0.5	0.17	0.18	0.32	0.48	0.68
0.8	0.02	0.4	0.44	0.5	0.03	0.5	0.03	0.18	0.28	0.42	0.96
0.8	0.64	0.6	0.81	0.5	0.27	0.5	0.29	0.18	0.90	0.32	0.90
0.8	0.78	0.8	0.68	0.5	0.29	0.5	0.37	0.18	0.68	0.18	0.58

Table 16: Coverage intervals from simulation study 4. Percentages of estimates whose 95% credible intervals cover the true value. The rows marked with * is selected for further investigation in Section 4.5

4.5 Diagnostic simulations

In this section we investigate the poor results from simulation study 3, Section 4.3, and simulation study 4, Section 4.4. We suspect the poor results to be due to prior sensitivity. This will be investigated by refitting the models with varying variance

for the priors of κ and α, while keeping the prior mean equal to the true value. Another possible explanation is the size of the dataset. To investigate this we will refit the bivariate environmental model, Equation (32), to a dataset twice the size of the one used in the simulation studies. All the datasets were simulated 100 times for each value of κ and α.

4.5.1 Size of the dataset

The dataset was doubled by cloning each tree except the parents. This results in a dataset where each parent has twice the amount of children. Then we fitted the environmental bivariate model, Equation (32), to the new dataset with dependency parameter, α, equal to 0.8. We chose this parameter to investigate one of the worst case scenario from simulation study 3. We see from Tables 17 and 18 that the results does not imply any significant improvement, or decline, compared to simulation study 3.

	α	$\hat{\alpha}-\alpha$	$\sigma^2_{a_1}$	$\hat{\sigma}^2_{a_1}-\sigma^2_{a_1}$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1}-\sigma^2_{z_1}$	$\sigma^2_{a_2}$	$\hat{\sigma}^2_{a_2}-\sigma^2_{a_2}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2}-\sigma^2_{z_2}$
Sim stud 3	0.8	0.165	0.5	0.157	0.5	-0.076	0.5	-0.074	0.18	-0.020
Double data	0.8	0.173	0.5	0.145	0.6	-0.087	0.5	-0.069	0.18	-0.032

Table 17: Bias ($\hat{\sigma}^2_x-\sigma^2_x$) from diagnostic study with doubled dataset compared with the bias from simulation study 3. Each value is the mean of the 100 simulations per value of α

	α	Cover	$\sigma^2_{a_1}$	Cover	$\sigma^2_{z_1}$	Cover	$\sigma^2_{a_2}$	Cover	$\sigma^2_{z_2}$	Cover
Sim stud 3	0.8	0.40	0.5	0.35	0.5	0.35	0.5	0.55	0.18	0.64
Double data	0.8	0.42	0.5	0.37	0.5	0.33	0.5	0.51	0.18	0.67

Table 18: Coverage intervals from diagnostic study with doubled dataset compared with the coverage intervals from simulation study 3. Percentages of estimates whose 95% credible intervals cover the true value.

4.5.2 Prior sensitivity

We investigate prior sensitivity in the environmental bivariate model, Equation (32), by constructing three different prior distributions for the dependency parameter, α. The mean of the priors is set equal to the true value, which is $\alpha = 0.8$ to investigate the worst case scenario from simulation study 3. The variance of the priors is set to be 1, 0.1 and 0.02, which is referred to as Prior #1, Prior #2 and Prior #3 respectively.

$$\alpha \sim \mathcal{N}(0.8, 1) \tag{52}$$
$$\alpha \sim \mathcal{N}(0.8, 0.1) \tag{53}$$
$$\alpha \sim \mathcal{N}(0.8, 0.02) \tag{54}$$

The results are summarized in Tables 19 and 20. We can clearly see an improvement of the inference as the variance of the dependency paramater prior is lowered. Both the bias, Table 19, and the coverage intervals, Table 20, are improved for all the parameters. These results implies that the environmental bivariate model, Equation (32), is prior sensitive.

	α	$\hat{\alpha} - \alpha$	$\sigma^2_{a_1}$	$\hat{\sigma}^2_{a_1} - \sigma^2_{a_1}$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1} - \sigma^2_{z_1}$	$\sigma^2_{a_2}$	$\hat{\sigma}^2_{a_2} - \sigma^2_{a_2}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2} - \sigma^2_{z_2}$
Sim stud 3	0.8	0.165	0.5	0.157	0.5	-0.076	0.5	-0.074	0.18	-0.020
Prior #1	0.8	0.155	0.5	0.142	0.5	-0.070	0.5	-0.067	0.18	-0.016
Prior #2	0.8	0.125	0.5	0.120	0.5	-0.053	0.5	-0.051	0.18	-0.013
Prior #3	0.8	0.035	0.5	0.040	0.5	-0.013	0.5	-0.006	0.18	-0.007

Table 19: Bias ($\hat{\sigma}^2_x - \sigma^2_x$) from diagnostic study with varying priors compared with the bias from simulation study 3. Each value is the mean of the 100 simulations per value of α

	α	Cover	$\sigma^2_{a_1}$	Cover	$\sigma^2_{z_1}$	Cover	$\sigma^2_{a_2}$	Cover	$\sigma^2_{z_2}$	Cover
Sim stud 3	0.8	0.40	0.5	0.35	0.5	0.35	0.5	0.55	0.18	0.64
Prior #1	0.8	0.48	0.5	0.46	0.5	0.42	0.5	0.62	0.18	0.69
Prior #2	0.8	0.55	0.5	0.62	0.5	0.65	0.5	0.70	0.18	0.72
Prior #3	0.8	0.73	0.5	0.70	0.5	0.76	0.5	0.90	0.18	0.86

Table 20: Coverage intervals from diagnostic study with varying priors compared with the coverage intervals from simulation study 3. Percentages of estimates whose 95% credible intervals cover the true value.

We investigate prior sensitivity in the full bivariate model, Equation (38), by constructing three different prior distributions for each of the dependency parameters, α and κ. The mean of the priors is set equal to the true value and the variance of the priors is set to be 1, 0.1 and 0.02 which is referred to as Prior #1, Prior #2 and Prior #3 respectively. We have chosen to investigate the full bivariate animal model with dependency parameters , (α, κ), (0.4, 0.6) and (0.6, 0.2).
The results are summarized in Tables 21 and 22. We can clearly see an improvement of the inference as the variance of the dependency paramaters prior is lowered for both our parameter pairs. Both the bias, Table 21, and the coverage intervals, Table 22, are improved for all the parameters. These results implies that the full bivariate model, Equation (38), is prior sensitive.

	κ	$\hat{\kappa}-\kappa$	α	$\hat{\alpha}-\alpha$	$\sigma^2_{a_1}$	$\hat{\sigma}^2_{a_1}-\sigma^2_{a_1}$	$\sigma^2_{z_1}$	$\hat{\sigma}^2_{z_1}-\sigma^2_{z_1}$	$\sigma^2_{a_2}$	$\hat{\sigma}^2_{a_2}-\sigma^2_{a_2}$	$\sigma^2_{z_2}$	$\hat{\sigma}^2_{z_2}-\sigma^2_{z_2}$
Sim stud 4	0.2	0.252	0.6	-0.099	0.5	0.522	0.5	-0.251	0.48	0.033	0.32	0.013
Prior #1	0.2	0.138	0.6	-0.073	0.5	0.315	0.5	-0.132	0.48	0.032	0.32	0.010
Prior #2	0.2	0.044	0.6	-0.016	0.5	0.167	0.5	-0.093	0.48	0.029	0.32	0.008
Prior #3	0.2	0.026	0.6	-0.005	0.5	0.021	0.5	-0.017	0.48	0.013	0.32	0.003
Sim stud 4	0.6	-0.136	0.4	-0.025	0.5	0.550	0.5	-0.260	0.32	0.058	0.42	-0.017
Prior #1	0.6	-0.093	0.4	-0.019	0.5	0.214	0.5	-0.163	0.32	0.037	0.42	-0.016
Prior #2	0.6	-0.062	0.4	-0.016	0.5	0.091	0.5	-0.087	0.32	0.025	0.42	-0.012
Prior #3	0.6	-0.022	0.4	-0.009	0.5	0.037	0.5	-0.047	0.32	0.009	0.42	-0.008

Table 21: Bias ($\hat{\sigma}^2_x-\sigma^2_x$) from the diagnostic studies compared with simulation study 4. Each value is the mean of the 100 simulations per combination of κ and α.

	κ	Cover	α	Cover	$\sigma^2_{a_1}$	Cover	$\sigma^2_{z_1}$	Cover	$\sigma^2_{a_2}$	Cover	$\sigma^2_{z_2}$	Cover
Sim stud 4	0.2	0.02	0.6	0.37	0.5	0.03	0.5	0.04	0.48	0.62	0.32	0.40
Prior #1	0.2	0.35	0.6	0.49	0.5	0.28	0.5	0.33	0.48	0.69	0.32	0.53
Prior #2	0.2	0.53	0.6	0.64	0.5	0.52	0.5	0.56	0.48	0.75	0.32	0.76
Prior #3	0.2	0.75	0.6	0.82	0.5	0.78	0.5	0.83	0.48	0.89	0.32	0.89
Sim stud 4	0.6	0.05	0.4	0.55	0.5	0.03	0.5	0.03	0.32	0.73	0.42	0.74
Prior #1	0.6	0.29	0.4	0.69	0.5	0.23	0.5	0.29	0.32	0.80	0.42	0.78
Prior #2	0.6	0.49	0.4	0.78	0.5	0.51	0.5	0.58	0.32	0.83	0.42	0.87
Prior #3	0.6	0.65	0.4	0.92	0.5	0.69	0.5	0.73	0.32	0.91	0.42	0.95

Table 22: Coverage intervals from the diagnostic studies compared with simulation study 4. Percentages of estimates whose 95% credible intervals cover the true value.

The inference for both the environmental bivariate animal model, Equation (32), and the full bivariate model, Equation (38), improved significantly when we varied the priors. The biases became smaller and the coverage intervals became larger. These results indicate that both models are sensitive to priors.

5 Case study

The proposed models, Section 2.7, are now fitted to the standardized Scots pine data, Section 2.1. The objective of the analysis is to estimate the dependency between the additive genetic variances and the individual environmental variances, as well as the variances themselves. This information can give us indications to similarities between growth rates and to what degree the different parameters affects height. For instance a high dependency between additive genetic variances indicates that many of the same genes control the growth rate at both age 10 and age 26, whereas a lower dependency would indicate that different genes operate at different ages.

The poor performance of the models in the simulation studies, Section 4, gives us reason to doubt our inference. We saw that as α and κ increased the results became less accurate. In this section we will treat the results of the real data as if the simulation studies were adequate. The impact of the simulation studies with respect to the case study will be discussed in Section 6.

The first model that was fitted to the standardized Scots pine data was the bivariate animal model without dependency, which assumes that the phenotypic values are the sum of additive genetic effects and individual environmental effects.

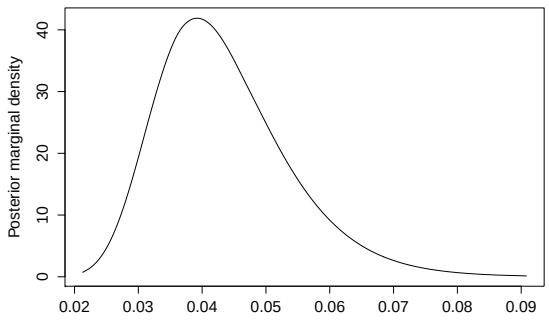
$$y_1 = a_1 + z_1 \tag{55}$$

$$y_2 = a_2 + z_2 \tag{56}$$

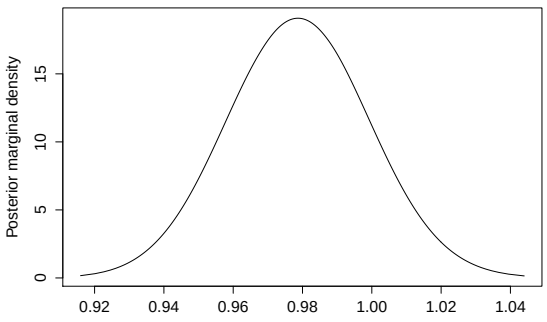
The posterior marginal densities are plotted in Figure 7 and the posterior mean with credible intervals are found in Table 23.
According to this model both traits, tree height at age 10 and age 26, have close to none trait specific additive effects, $\sigma^2_{a_1} = \sigma^2_{a_2} = 0.04$. The individual environmental effects are clearly dominating, $\sigma^2_{z_1} = 0.97$ and $\sigma^2_{z_2} = 0.98$, indicating that the tree height is mostly affected by individual environmental effects.

$\sigma^2_{a_1}$	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{a_2}$	CI	$\sigma^2_{z_2}$	CI
0.04	(0.03,0.07)	0.97	(0.94,1.02)	0.04	(0.03,0.06)	0.98	(0.93,1.02)

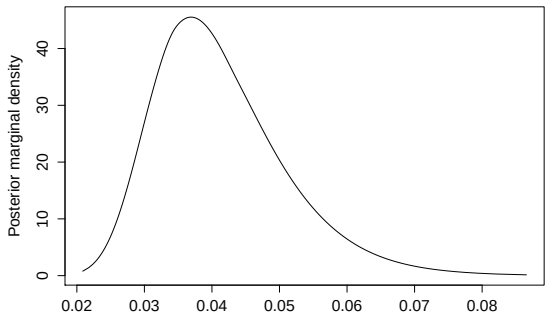
Table 23: Posterior mean with credible intervals for the bivariate animal model fitted to the Scots pine data.



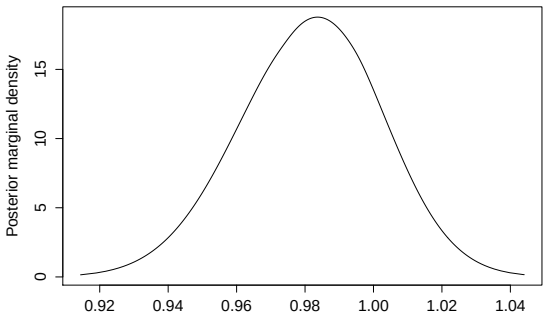
(a) Posterior marginal density for $\sigma^2_{a_1}$



(b) Posterior marginal density for $\sigma^2_{z_1}$



(c) Posterior marginal density for $\sigma^2_{a_2}$



(d) Posterior marginal density for $\sigma^2_{z_2}$

Figure 7: Posterior marginal densities for the bivariate animal model

The second model that was fitted to the standardized Scots pine data was the basic bivariate model, Equation (22). It assumes that the phenotypic values are subject to individual environmental effects with dependence.

$$y_1 = z_1 \tag{57}$$

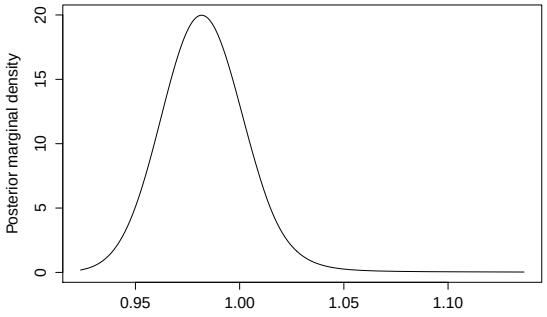
$$y_2 = \alpha z_1 + z_2 \tag{58}$$

The posterior marginal densities are plotted in Figure 8 and posterior means with credible intervals are found in Table 24.
These results suggests a fairly high dependency between the individual environmen-

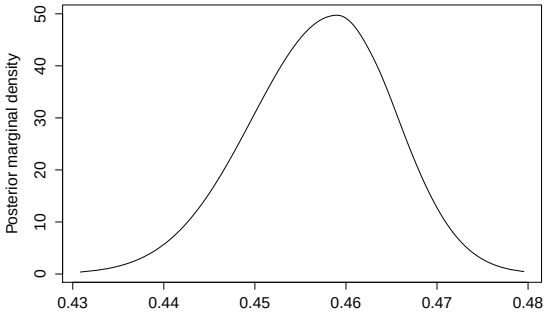
tal variances, $\alpha = 0.74$,

$$corr(y_1, y_2) = \frac{\alpha\sigma^2_{z1}}{\sqrt{\sigma^2_{z_1}(\alpha^2\sigma^2_{z_1} + \sigma^2_{z_2})}} \tag{59}$$

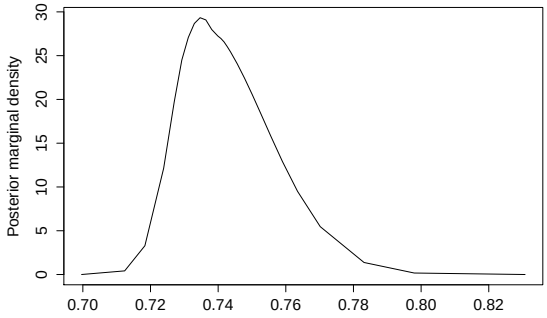
. which equals a correlation of 0.73. This means that the tree height at age 10 and at age 26 are affected by quite similar individual environmental effects, or there are dependencies due to similar genetic effects.



(a) Posterior marginal density for $\sigma^2_{z_1}$



(b) Posterior marginal density for $\sigma^2_{z_2}$



(c) Posterior marginal density for α

Figure 8: Posterior mean density for the basic bivariate model

α	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{z_2}$	CI
0.74	(0.72,0.78)	0.98	(0.94,1.03)	0.46	(0.44,0.47)

Table 24: Posterior mean with credible intervals for the basic bivariate model fitted to the Scots pine data.

The third model that was fitted to the standardized Scots pine data was the additive bivariate model, Equation (26). It assumes that the phenotypic values are subject to both additive genetic effects values and individual environmental effects, with dependence in the additive genetic effects, but not in the individual effects.

$$y_1 = a_1 + z_1 \tag{60}$$

$$y_2 = \kappa a_1 + a_2 + z_2 \tag{61}$$

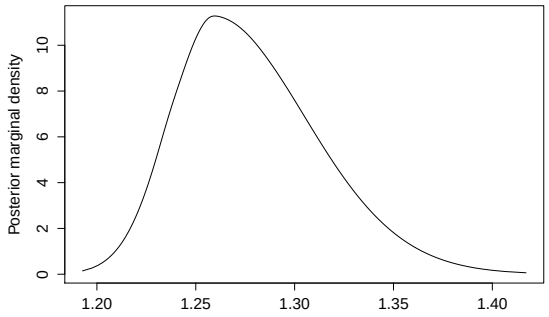
The posterior marginal densities are plotted in Figure 9 and posterior means with credible intervals are found in Table 25.
According to this model the height at age 10 is clearly dominated by the additive genetic effects, $\sigma^2_{a_1} = 1.28$, compared to the individual environmental effects, $\sigma^2_{z_1} = 0.34$. The large dependency, $\kappa = 1.13$, leads to the same tendencies for the trees at age 26. These values result in a additive genetic variance of 1.67 for y_2 and the individual environmental effects, $\sigma^2_{z_2}$, equal to 0.19.

$$corr(y_1, y_2) = \frac{\kappa\sigma^2_{a_1}}{\sqrt{(\sigma^2_{a_1} + \sigma^2_{z_1})(\kappa^2\sigma^2_{a_1} + \sigma^2_{a_2} + \sigma^2_{z_2})}} \tag{62}$$

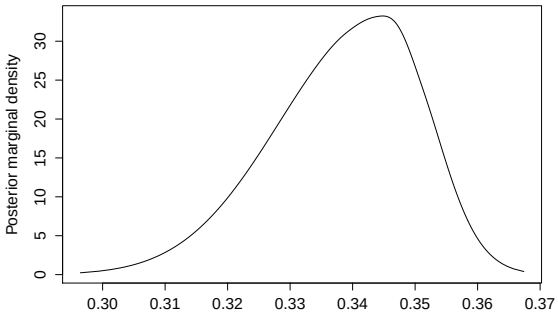
The correlation between the two traits is 0.83. The results indicate that the additive genetic effects influence tree height most at both ages. The low trait specific additive variance at age 26 and large dependency implies that many of the same genes affects height for trees at age 10 and 26.

κ	CI	$\sigma^2_{a_1}$	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{a_2}$	CI	$\sigma^2_{z_2}$	CI
1.13	(1.08,1.16)	1.28	(1.22,1.36)	0.34	(0.31,0.36)	0.04	(0.02,0.07)	0.19	(0.17,0.22)

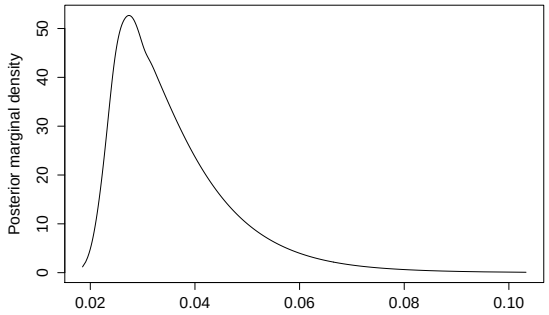
Table 25: Posterior mean with credible intervals for the additive bivariate model fitted to the scots pine data.



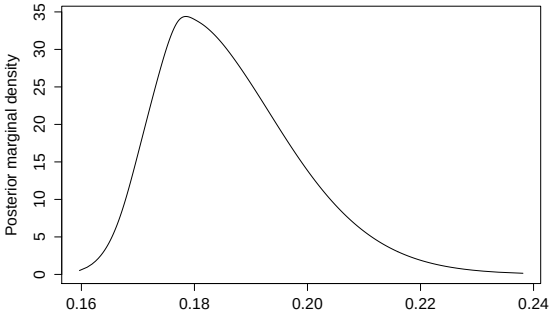
(a) Posterior marginal density for $\sigma^2_{a_1}$



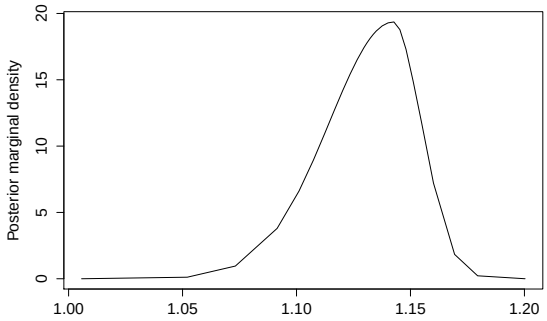
(b) Posterior marginal density for $\sigma^2_{z_1}$



(c) Posterior marginal density for $\sigma^2_{a_2}$



(d) Posterior marginal density for $\sigma^2_{z_2}$



(e) Posterior marginal density for κ

Figure 9: Posterior marginal densities for the additive bivariate model

The fourth model that was fitted to the standardized Scots pine data was the environmental bivariate model, Equation (32). It assumes that phenotypic values are determined by the sum of the individual environmental effects and additive genetic effects, with dependence in the individual effects, but not in the additive genetic effects.

$$y_1 = a_1 + z_1 \tag{63}$$
$$y_2 = a_2 + \alpha z_1 + z_2 \tag{64}$$

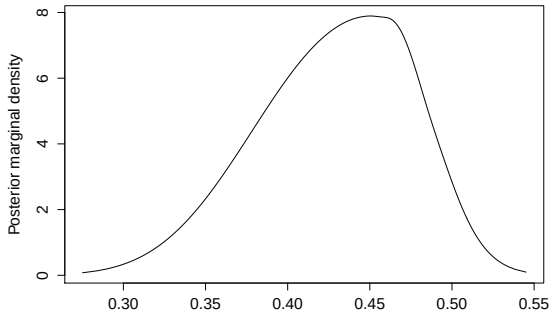
The posterior marginal densities are plotted in Figure 10 and posterior means with credible intervals are found in Table 26.
According to this model height at age 10 is affected by both additive genetic effects, $\sigma^2_{a_1} = 0.42$, and individual environmental effects, $\sigma^2_{z_1} = 0.76$. The height at age 26 is dominated by individual environmental effects, since the dependency is large, $\alpha = 0.98$. The trait specific additive genetic variance for height at age 26 is 0.04 and the individual environmental variance is 0.99.

$$corr(y_1, y_2) = \frac{\alpha\sigma^2_{z_1}}{\sqrt{(\sigma^2_{z_1} + \sigma^2_{a_1})(\alpha^2\sigma^2_{z_1} + \sigma^2_{z_2} + \sigma^2_{a_2})}} \tag{65}$$

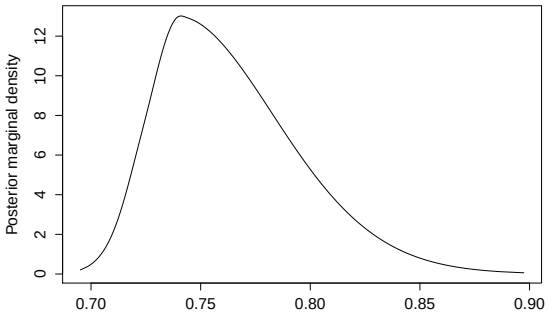
The correlation between the two traits is 0.67. These results contradict the results from the additive bivariate model, which indicated that the additive genetic effects had the most influence. This may be a symptom of identification issues. To investigate this matter further, the next model we fit to the standardized Scots pine data is the full bivariate model, which includes both dependencies.

α	CI	$\sigma^2_{a_1}$	CI	$\sigma^2_{z_1}$	CI	$\sigma^2_{a_2}$	CI	$\sigma^2_{z_2}$	CI
0.98	(0.89,1.03)	0.42	(0.33,0.51)	0.76	(0.71,0.84)	0.04	(0.02,0.06)	0.27	(0.24,0.34)

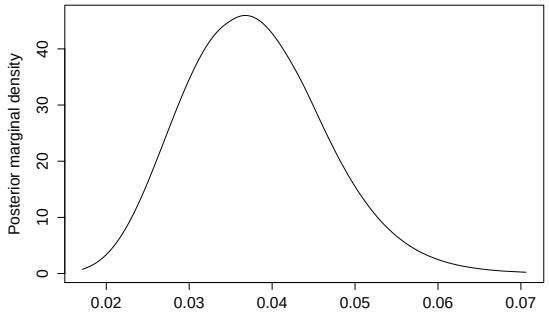
Table 26: Posterior mean with credible intervals for the additive bivariate model fitted to the scots pine data.



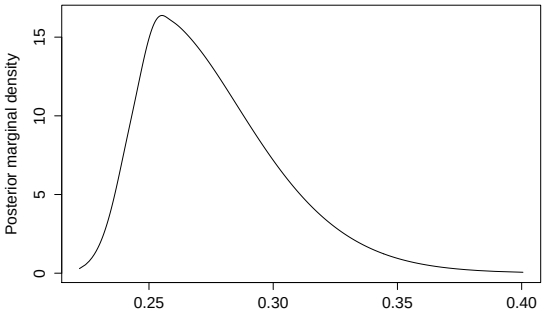
(a) Posterior marginal density for $\sigma^2_{a_1}$



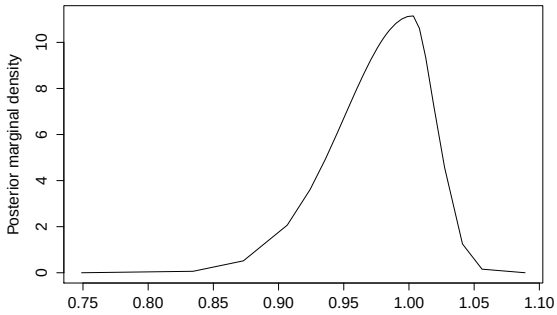
(b) Posterior marginal density for $\sigma^2_{z_1}$



(c) Posterior marginal density for $\sigma^2_{a_2}$



(d) Posterior marginal density for $\sigma^2_{z_2}$



(e) Posterior marginal density for α

Figure 10: Posterior marginal densities for the environmental bivariate model

The last model that was fitted to the standardized Scots pine data was the full bivariate model, Equation (38). It assumes that the phenotypic values are subject to both additive genetic effects and individual environmental effects, with dependence in both the additive genetic effects and the individual environmental effects.

$$y_1 = a_1 + z_1 \tag{66}$$

$$y_2 = \kappa a_1 + a_2 + \alpha z_1 + z_2 \tag{67}$$

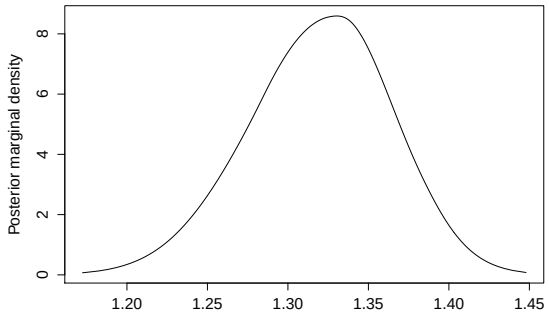
The posterior marginal densities are plotted in Figure 11 and posterior means with credible intervals are found in Table 27.
The height at age 10 is clearly dominated by the additive genetic effects, $\sigma^2_{a_1} = 1.13$, compared to the individual environmental effects, $\sigma^2_{z_1} = 0.31$. The total additive genetic variance at age 26 is 0.89 and the total individual environmental variance is 0.48, which indicates that tree height at age 26 is also dominated by additive genetic effects. The strong dependence in this model , $\kappa = 0.81$ and $\alpha = 0.53$, indicates that the influence of both additive genetic effects and individual environmental effects is quite similar between the traits.

$$corr(y_1, y_2) = \frac{\kappa\sigma^2_{a_1} + \alpha\sigma^2_{z_1}}{(\sigma^2_{z_1} + \sigma^2_{a_1})(\alpha^2\sigma^2_{z_1} + \sigma^2_{z_2} + \kappa^2\sigma^2_{a_1} + \sigma^2_{a_2})} \tag{68}$$

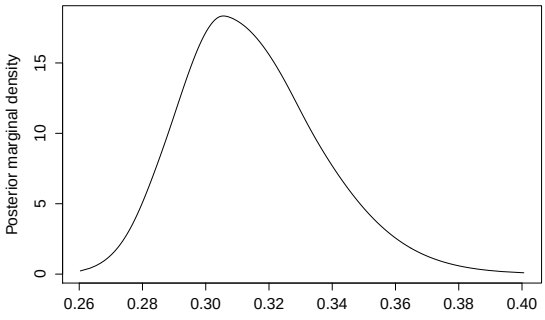
The correlation between the two traits is 0.82. These results indicate that tree height is mostly influenced by additive genetic effects at both age 10 and age 26.

κ	CI	α	CI	$\sigma^2_{a_1}$	CI	$\sigma^2_{z_1}$	CI
0.81	(0.74,0.87)	0.53	(0.40,0.67)	1.31	(1.23,1.40)	0.31	(0.28,0.36)
$\sigma^2_{a_2}$	CI	$\sigma^2_{z_2}$	CI				
0.04	(0.02,0.06)	0.40	(0.37,0.42)				

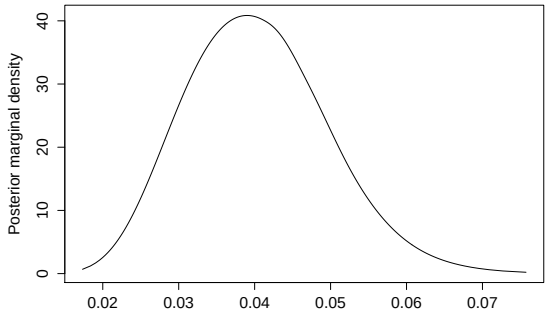
Table 27: Posterior mean with credible intervals for the full bivariate model fitted to the scots pine data.



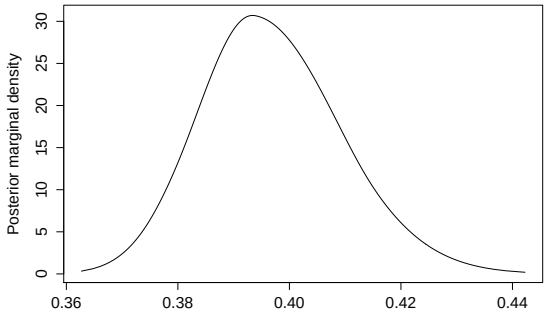
(a) Posterior marginal density for $\sigma^2_{a_1}$



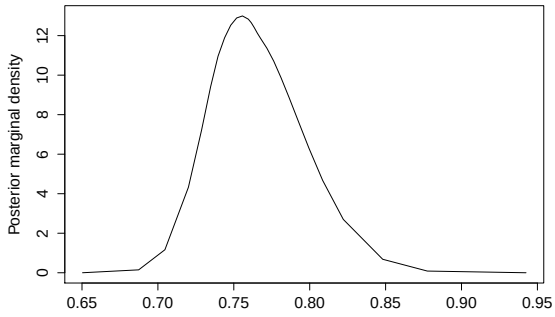
(b) Posterior marginal density for $\sigma^2_{z_1}$



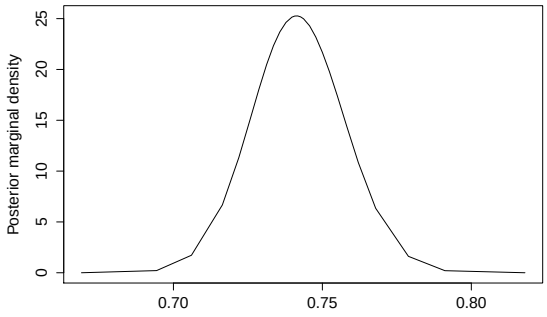
(c) Posterior marginal density for $\sigma^2_{a_2}$



(d) Posterior marginal density for $\sigma^2_{z_2}$



(e) Posterior marginal density for κ



(f) Posterior marginal density for α

Figure 11: Posterior marginal densities for the full bivariate model

All the models indicate a fairly large dependence between the two traits, both for the individual environmental effects and additive genetic effects. The results for the additive genetic effects are in particular interesting. In all the models the trait specific additive genetic variance for trees age 26 very small, $\sigma^2_{a_2} < 0.05$, while the dependency is quite large. This strongly indicate that many of the same genes influences tree height at age 10 and 26.

6 Discussion and further work

We have showed that the bivariate animal models specified in Section 2.7 are suitable for the INLA framework. This enables us to do inference on the models with the R-INLA library. R-INLA is very beneficial because it is relatively easy to specify models and inference is fast.

The results from simulation study 1, Section (4.1), indicates that the basic bivariate model, Equation 22, performs well in terms of bias, coverage intervals and mean credible intervals.

Simulation study 2 shows that the additive bivariate animal model, Equation (26), performs well in terms of coverage intervals and mean credible intervals. It seems to have some issues allocating the trait specific variances. The trait specific additive genetic effects and individual environmental effects have opposite signed bias, indicating identification issues.

Simulation study 3 shows that the environmental bivariate animal model, Equation (32), performs well when the dependency parameter, α, is low. As α increases the bias increases and the coverage intervals decreases. We suspect this behaviour to be the result of prior sensitivity and investigate the matter further in Section 4.5. Setting the mean of the prior for the dependency parameter to the true value and lowering the variance resulted in a significant improvement of the inference.

The results from simulation study 4 indicates that the full bivariate model, Equation (38), suffers from the same problems as the environmental bivariate model, Equation (32). As the dependency parameters increases the overall performance of the inference decreases. The bias increases and the coverage intervals decreases. We suspect this behaviour occurs for the same reason as for the environmental bivariate model, which is prior sensitivity. We have selected two of the parameter pairs, (κ, α), that

gives the worst outcome and investigated them further in Section 4.5. Setting the mean of the prior for the dependency parameters to the true value and lowering the variance resulted in significant improvement of the inference.

The case study indicates a large correlation between tree height at age 10 and age 26. The dependency parameters for both additive genetic effects and individual environmental effects where estimated to be quite large. These results indicate that many of the same genes influence tree height at both age 10 and age 26. Unfortunately there are several reasons indicating that we should not trust these results blindly. Two of the models estimated the trait specific additive genetic effect for trees age 10 to be larger than 1, when the data was standardized. This may be due to prior sensitivity, since the estimated trait specific additive variances are estimated to be close to zero, when we fitted an bivariate animal model without dependence. The simulation studies showed poor inference for large dependency parameters, which indicates uncertainties regarding the case study as well. The simulation study showed that the models performed well for low dependency parameters, which in the context of the case study indicates that there is in fact a quite large dependency, even though the inference may be poor.

Several aspects of this study are suited for further research. It would be very interesting to conduct a more comprehensive study regarding prior sensitivity. The results from Section 4.5 showed us that varying the prior of the dependence parameter significantly improved the inference. A study where the prior distributions of several hyperparameters are varied, could lead to some interesting results.

Another approach would be to include other effects to the models, such as dominative genetic effects and spatial effects. It is done for univariate animal models with the Scots pine data in both (Finley et al. 2009) and (Bøhn 2013), but not for bivariate animal models. It would be interesting to see the effect of dominative genetic effects compared with the additive genetic effects.

7 Acknowledgements

I would like to thank my supervisor Ingelin Steinsland for her inputs, ideas and helpful comments. She has been of immense help throughout the writing of this thesis.

References

Bøhn, E. D. (2013), 'Modeling and inference for Bayesian pedigree and spatial modeling using INLA'.

Finley, A. O., Banerjee, S., Waldmann, P. & Ericsson, T. (2009), 'Hierarchial Spatial Modeling of Additive and Dominance Genetic Variance for Large Spatial Trial Datasets'.

Hazel, L. (1943), 'The genetic bases for constructing selection indexes', *Genetics* **28**, 476–490.

Holand, A. M., Steinsland, I., Martino, S. & Jensen, H. (2013), 'Animal models and Integrated Nested Laplace Approximations'.

Lynch, M. & Walsh, B. (1998), *Genetics and Analysis of Quantitative Traits*.

Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. (2013), 'Bayesian computing with INLA: new features'.

Postma, E. (2006), 'Implications of the difference between true and predicted breeding values for the study of natural selection and micro-evolution', *Journal of Evolutionary Biology* **19**, 309–320.

Rue, H. & Held, L. (2005), 'Gaussian Markov random fields: Theory and Applications', *Monographs on Statistics and Applied Probability* **104**.

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations'.

Simm, G. (1998), *Genetic Improvement of Cattle and Sheep*.

Smith, H. F. (1936), 'A discriminant function for plant selection', *Annals of Eugenics* **7**, 240–250.

Sorensen, D. & Gianola, D. (2002), *Likelihood, Bayesian and MCMC Methods in Genetics*, Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. V. D. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Roayal Statistical Society: Series B (Statistical Methodology* **64**(4), 583–639.

Steinsland, I. & Jensen, H. (2010), 'Utilizing Gaussian Markov Random Field Properties of Bayesian Animal Models'.

Wermuth, N. & Lauritzen, S. L. (1982), *Graphical and Recursive Models for Contigency Tables*, Biometrika Trust.

Wilson, A. J., Reale, D., Clements, M. N., Morrissey, M. M., Postma, E., Walling, C. A., Kruuk, L. E. B. & Nussey, D. H. (2009), 'An ecologist's guide to the animal model', *Journal of Animal Ecology* **79**, 13–26.

A Implementation of bivaraiate Gaussian models using R-INLA

The following R-code shows how to simulate and fit bivariate Gaussian models. This example shows the procedure used in Simulation study 2, Section 3.2

$$y_1 = a_1 + z_1$$
$$y_2 = \rho a_1 + \sqrt{1-\rho^2} a_2 + z_2$$

```
#Sample the environmental and additive effects.
#simulate.breeding is a function from the Animal INLA package and
#ainv is the inverse of the additive relationship matrix
RandomDraw = function (var.e,var.a) {
  n=5012
  z1=rnorm(n,sd=sqrt(var.e))
  z2=rnorm(n,sd=sqrt(var.e))
  a1 = simulate.breeding(ainv,var.a)
  a1 = unlist(a1)
  a2 = simulate.breeding(ainv,var.a)
  a2 = unlist(a2)

  drawlist = list("z1"=z1,"z2"=z2,"a1"=a1,"a2"=a2,"n"=n)
  return(drawlist)
}

#Preparing the simulated data matrix
SimG = function(randomdraw,rho,rho.2) {
  scale = sqrt(1-rho^2)
  scale.2 = sqrt(1-rho.2^2)
  n=randomdraw$n
  y1 = randomdraw$a1 + randomdraw$z1
  y2 = scale*randomdraw$a2 + rho*randomdraw$a1 +
  scale.2*randomdraw$z2 + rho.2*randomdraw$z1

  Y=matrix(NA,2*n,2)
```

```
  Y[1:n,1]=y1
  Y[(n+1):(2*n),2]=y2
  return(Y)

}

#Prepare covariates
    y1a1=c(1:n, rep(NA, n))
    y1z1=c(1:n, rep(NA, n))
    y2a2=c(rep(NA, n), 1:n)
    y2z2=c(rep(NA, n), 1:n)
    y2rho=c(rep(NA, n), 1:n)
    intcpt = c(rep(1, n),rep(2, n))
#Specify the the formula to be used in inla()
formula=data ~
f(y1a1, model="generic0",hyper = list(theta = list(param = c(0.5, 0.5), fixed=F)),
 constr=FALSE, Cmatrix=NewSimCmatrix) +
f(y2a2, model="generic0",hyper = list(theta = list(param = c(0.5, 0.5), fixed=F)),
 constr=FALSE, Cmatrix=NewSimCmatrix) +
f(y1z1,model="iid",hyper = list(theta = list(fixed=F))) +
f(y2z2,model="iid",hyper = list(theta = list(fixed=F))) +
f(y2rho,copy="y1a1",hyper = list(theta = list(fixed=FALSE,
param=c(0,0.1),initial=1))) +
f(intcpt,model="iid",hyper = list(theta = list(fixed=TRUE))) - 1

#Fit model using inla()
  res = inla(
    formula,
    data = data.frame(simulated.data),
    family = c("gaussian", "gaussian"),
    control.family=list(list(hyper=list(prec=list(initial=10, fixed=TRUE))),
               list(hyper=list(prec=list(initial=10, fixed=TRUE)))))
```