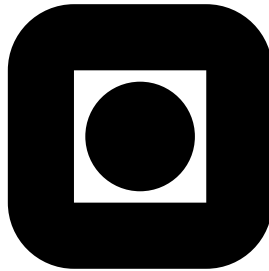NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Making Inference from Bayesian Animal Models utilising Gaussian Markov Random Field properties

by

Ingelin Steinsland and Henrik Jensen

PREPRINT
STATISTICS NO. 10/2005

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

Making Inference from Bayesian Animal Models utilising Gaussian Markov Random Field properties

Ingelin Steinsland[*] & Henrik Jensen[†]

December 5, 2005

Abstract

Numerical efficient methods for sampling and evaluation of Gaussian Markov Random Fields (GMRFs) are used for making inference from Bayesian animal models (also known as additive genetic models, that are versions of general linear models). For single-trait animal models an approximation to the posterior distribution of variance components and the heritability can be found without using Markov chain Monte Carlo (MCMC) methods. For the multiple-trait animal model a two-block Gibbs sampler can be used, also for large datasets.

The above methodology is successfully used to study the genetic architecture of morphological traits in a house sparrow meta-population. The pedigree consists of 3572 birds and there are data for seven traits, i.e. the Bayesian animal model has more that 25000 variables. The results provide strong indications of possibilities for, but also constraints on micro-evolution.

KEYWORDS: Additive genetic model, quantitative genetics, conditional auto-regression, Markov chain Monte Carlo, Gibbs sampler, blocking.

[*]Department of Mathematical Sciences, Norwegian University of Science and Technology 7491 Trondheim, Norway, e-mail: ingelins@math.ntnu.no

[†]Department of Biology, Norwegian University of Science and Technology, 7491 Trondheim, Norway, e-mail: henrik.jensen@bio.ntnu.no

1 Introduction

Quantitative genetics is the study of quantitative characters. The theory is based on the assumption that the characters are determined by a large number of genes as well as by different environmental effects. Important pioneer work done by Karl Pearson on regression and Ronald Fisher on analysis of variance (ANOVA) were motivated by the need for estimating quantitative genetic parameters. An important quantitative genetic parameter is the *heritability* of a trait; how much of the phenotypic (observed traits) variance in a population can be explained by different genes. See e.g. Lynch and Walsh (1998) for an introduction to quantitative genetics. The key assumption in the model we use, *the animal model*, is that animal i's trait, y_i (e.g. the length of an elephant's trunk) can be divided into a genetic part (or inherited part), u_i, and an environmental part ϵ_i. The value of u_i is referred to as the breeding value of individual i. To calculate individual breeding values one need data from related animals. For decades calculation of breeding values have been standard in plant and animal breeding programs, for breeding animals or plants with desired properties. For domestic populations this has been a success; beef cattle give more meat and diary cows more milk, see e.g. Simm (1998).

Until recently the required data have not been available at any large scale for wild populations. Though, with easier genotyping methods, it is today achievable to decide family structure also for natural populations, see e.g. Kruuk (2004) and references therein.

Both in plant and animal breeding and in evolutionary biology it is often of interest to consider several traits simultaneously. For example in plant and animal breeding, selection for multiple traits are often desired, e.g. yield and quality of wheat. This requires knowledge about the additive genetic variances and correlations between traits. The additive genetic covariance matrix is also of interest to understand evolution because it contains information about evolutionary trajectories of several traits; Roff (1997), Lande and Arnold (1983), and about possible bottlenecks under natural selections, Lynch and Walsh (1998). A finding of several studies of wild populations is that they do not respond to natural selection as expected, Kruuk et al. (2001). A possible reason is that only single traits were considered; Kruuk et al. (2002), Jensen et al. (2006a).

In this paper we take a Bayesian approach to analyse the additive genetic structure of a house sparrow population using the animal model. To our knowledge a Bayesian animal model has not earlier been used for a wild population, though suggested in Kruuk (2004). In animal breeding a Bayesian approach is well established, see e.g. Sorensen and Gianola (2002) or Blasco (2001). Inference has to be done by Markov Chain Monte Carlo (MCMC) methods and it is often necessary to consider tens of thousands of variables. Making inference is therefore computationally expensive. Traditionally Gibbs samplers have been used, Sorensen and Gianola (2002).

The breeding values for the individuals in a population are Gaussian with a dependency given by the family structure, i.e. by the corresponding pedigree. A pedigree can be viewed as a directed acyclic graph (DAG); each individual is a node, and there are edges from parents to offsprings. From a DAG the conditional independence graph can be found by *moralising* the DAG; 'unmarried' parents are 'married' (an edge is inserted between them), and the direction of the edges are removed, see Figure 1 for an illustration. A multivariate Gaussian model with a conditional independence structure is known as a *Gaussian Markov random field* (GMRF) model, where the *Markov property* refer to the conditional independence structure. We will see that the animal model is a GMRF model, and a multiple trait animal model is a

multivariate Gaussian Markov random field (MGMRF) model.

Computationally efficient sampling and evaluation methods are available for GMRFs and MGMRFs, Rue and Held (2005). In this paper we utilise the computational benefits of GMRFs for the animal model. For a multiple trait animal model we set up a two-block Gibbs sampler, and we are also able to make inference for models with constraints on breeding values or group level effects. For the single-trait animal models we are able to calculate the joint posterior probability density for the genetic and environmental variances on a grid, which gives us almost exact posterior distribution for the heritability without doing MCMC. Because of the computational efficiency, we are able to do these analyses for large pedigrees and many traits.

In Section 2 the data are introduced. The animal model is fully specified in Section 3. Section 4 contains a review of relevant GMRF theory. In Section 5 the inference methodology is set up. The data are analysed, and to some extent interpreted in Section 6. Section 7 ends the paper with a conclusion.

2   Data

In this study we analyse data from an insular meta population of a small Passerine bird, the house sparrow (*Passer domesticus*). From 1993 the entire population of house sparrows has been monitored on five islands off the coast of Helgeland, Northern Norway. On these islands house sparrows live near human settlements, and mostly nest inside barns of dairy farms. During the breeding season nests are thoroughly searched for, and nestlings are ringed and a blood sample is taken before fledging.

In addition, adult and fledged juveniles (birds born the same summer) are captured with mist nets during the summer. A large porportion ($> 70\%$) of the adult birds present on each island a given year are marked. House sparrows not registered as nestlings are ringed and a blood sample is taken at first capture. When captured, several morphological traits are measured for both adult males and adult females; tarsus length ('foot length'), wing length, bill depth, bill height and body mass. The males have a black throat badge. The feathers on a region of their throat have black bases. Most of these feathers also have black tips, which makes up the visible badge. The other black based fathers have gray tips. In this study both the total badge size (area with black feather bases) and visible badge size (area with black tips) are registered. Hence, we have five morphological traits for females and seven for males. However, for some of the birds the record is not full. For all registered birds, also their sex, hatch year and hatch island are known.

The blood samples are used to determine genetic parenthood, see Jensen et al. (2003, 2004), and hence we get a pedigree for the birds on the study islands. We have used data from 1993-2002. There are 3572 birds in the pedigree. For 53% both parents are known, for 27% only one parent and 14% have neither any known parents nor children. For 824 house sparrows we have records of one or more adult traits as well as sex, hatch year and hatch island. We use adjusted one-year old traits (as described in Jensen et al. (2006a), most of the traits change during life, e.g. the badges generally get larger). The data are standardised for each trait. Most of the birds stay on the island they are born: Of the birds in our dataset about 6% migrated from one study island to another, see also Altwegg et al. (2000) and Tufto et al. (2005).

For a more thorough description of the field work and study area, see Altwegg et al. (2000),

Ringsby et al. (1999, 2002), Saether et al. (1999) and Jensen et al. (2004).

3 The Animal Model

The animal model is a general linear mixed model, and it is also known as the additive genetic model. Good references are Lynch and Walsh (1998) or for a Bayesian approach (from an animal breeding point of view) Sorensen and Gianola (2002).

Let y_i denote the trait value(s) for bird i. If only one trait is considered y_i is a scalar, while in a multiple traits situation y_i is a vector of length the number of traits, say of length m. One then assumes that the actual trait value for an animal i is the result of *group level effects* $\beta_{f,l(i)}$, *genetic effect* u_i and *environmental effect* ϵ_i;

$$y_i = \sum_{f=1}^{n_f} \beta_{f,l(i)} + u_i + \epsilon_i \tag{1}$$

Group level effects account for different levels $l = 1 \ldots L_{n_f}$ in n_f different groups, and where individual i has level $l(i)$. E.g. in the house sparrow males are for some traits generally larger than females, and sex is a natural group effect with two levels. Other group levels in natural populations can be e.g. year of birth and location. In a classical approach these effects would be treated as fixed effects. In a Bayesian setting there are no such thing as fixed effects, as all effects are treated as random effects. The genetic effect is the part determined by genes, i.e. the inherited part. Both the genetic and environmental effect are given zero mean Gaussian priors; $u_i \sim N(0, \Sigma_u)$ and $\epsilon_i \sim N(0, \Sigma_\epsilon)$, where Σ_u is the genetic covariance matrix (size $m \times m$) and Σ_ϵ is the environmental covariance matrix (size $m \times m$). We give both these matrices the conjugated inverted Wishart distribution as prior.

In the pedigree there are often many birds without any observed trait value. Let n be the number of birds in the pedigree, and denote the set of birds with (at least one) observed trait value(s) $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ and the number of birds with observed trait $|\mathcal{I}| = N$. Let $\boldsymbol{y} = (y_i)_{i \in \mathcal{I}}$, a vector of all traits of birds with trait data, $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$, a vector of all breeding values for all the birds in the pedigree, and $\boldsymbol{\epsilon} = (\epsilon_i)_{i \in \mathcal{I}}$ all the environmental effects for the birds with trait data. The full model then reads;

$$\boldsymbol{y} = X\boldsymbol{\beta} + W\boldsymbol{u} + \epsilon \tag{2}$$

where X and W are known incidence matrices. The environmental effects are assumed independent between birds; $\boldsymbol{\epsilon} \sim N(0, I \otimes \Sigma_\epsilon)$ with I the identity matrix of size $(N \times N)$ and $\otimes$ the Kronecker product. I.e. the environmental effects can be interpreted as residuals. The breeding values are assumed to have a dependency structure corresponding to how the birds are related; $\boldsymbol{u} \sim N(0, A \otimes \Sigma_u)$, where A is a $n \times n$ matrix known as the *relationship matrix*. Element (i, j) of A is two times the *coefficient of coancastry* between bird i and bird j, see e.g. Malecot (1969). In calculations only A^{-1} is needed, and its non-zero structure is found from moralising the pedigree. The non-zero values can be calculated according to Quaas (1976).

If there are more than one group level effect, there is an identification problem. We solve this problem by constraining the effect of the levels to sum to zero for all but one fixed effect; $\sum_{l=1}^{L} \beta_l = 0$. In addition, a breeding value is a property that is relative to the population, and therefore we also constrain the breeding values for each trait t to sum to zero; $\sum_{i=1}^{n} u_{it} = 0$. We write these constraints as $C_\beta \boldsymbol{\beta} = \boldsymbol{0}$ and $C_u \boldsymbol{u} = \boldsymbol{0}$ where $\boldsymbol{0}$ denotes a vector of zeros.

The variables of interest are typically the covariance matrixes Σ_u and Σ_ϵ and functions of these. In particular, the *heritability* for each trait t is of interest;

$$h_t^2 = \frac{\Sigma_{u,tt}}{\Sigma_{u,tt} + \Sigma_{\epsilon,tt}}. \tag{3}$$

Here Σ_{tt} refers to element (t,t) of matrix Σ. Also individual breeding values or trends in breeding values can be of interest.

For the birds with observed adult traits, the record is not necessarily full. We use the notation $\boldsymbol{y} = \{\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}\}$. As the females do not have badges, we treat female badges as missing observations.

Hence, the posterior distributions we want to evaluate are; the posterior of the breeding values and genetic and environmental covariance matrices $\pi(\Sigma_u, \Sigma_\epsilon, \boldsymbol{u}, \boldsymbol{\beta}|\boldsymbol{y}_{obs})$, the posterior of the covariances matrices $\pi(\Sigma_u, \Sigma_\epsilon|\boldsymbol{y}_{obs})$ and/or the posterior of the heritabilities $\pi(h_t^2|\boldsymbol{y}_{obs})$.

4   Gaussian Markov Random Fields

The parameter estimating methodology introduced in Section 5 is made possible for problems of our size by computationally efficient sampling and evaluation algorithms for GMRFs. In this section we give a short introduction to GMRFs and their properties relevant for this work. For a thorough introduction to GMRFs and its application, see Rue and Held (2005).

GMRF models are multivariate Gaussian models with a Markov property, and are also known as conditionally autoregressive (CAR) models as introduced by Besag (1974, 1975). The Markov property refers to a conditional independence structure, often visualised with a conditional independence graph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with n nodes (or vertexes); $\mathcal{V} = \{1, 2, \ldots, n\}$ and edges $\mathcal{E} = \{(i,j)\}$ with $i, j \in \mathcal{V}$. Then $\boldsymbol{x}$ is a GMRF with respect to $\mathcal{G}$ if and only if there is no edge only between nodes corresponding to variables that are conditionally independent. Let $\boldsymbol{x}_{-ij}$ denote all variables but x_i and x_j; $\boldsymbol{x}_{-ij} = \boldsymbol{x} \setminus \{x_i, x_j\}$. If there is no edge between node i and j, then x_i and x_j are conditionally independent; $\{i, j\} \notin \mathcal{E} \Rightarrow x_i \perp x_j | \boldsymbol{x}_{-ij}$. In our setting we can think of each node as a house sparrow, and the edges are obtained by moralising their pedigree, see Figure 1. Therefore, each bird is conditionally dependent only on its parents, its offspring and the other parent(s) of its offspring. The Markov structure is reflected in the non-zero pattern of the precision matrix $Q = \Sigma^{-1}$: Only off-diagonal elements that correspond to conditionally dependent variables (two nodes with edges between) are non-zero. It is this sparseness of Q that imposes computational benefits for sampling and evaluation of GMRFs. The computationally expensive parts of both these operations are calculation of Q's Cholesky factor L, $Q = L^T L$ (L is lower triangular). In most cases a sparse precision matrix impose a sparse Cholesky factor, fewer elements have to be calculated, and the computations are orders cheaper then for a full Q. While the computational complexity of Cholesky decomposition is $\mathcal{O}(n^3)$ for a full matrix, it is $\mathcal{O}(n)$ for a GMRF on a one-dimensional graph (i.e. a times series structure) and $\mathcal{O}(n^{3/2})$ for a GMRF on a two-dimensional graph, Rue and Held (2005, Ch. 2.5.). Our pedigree-based graph is between one- and two-dimensional. A sample from $N(0, Q^{-1})$ can be obtained by solving $L^T \boldsymbol{x} = \boldsymbol{z}$ for $\boldsymbol{x}$ with $\boldsymbol{z}$ a vector of independent standard Gaussian samples and where L is Q's Cholesky factor. To evaluate the probability density function of a GMRF the determinant of Q is needed, and $\det(Q) = \prod_{i=1}^{n} l_{ii}$, where l_{ii} is element (i,i) of L. There are also other useful computationally achievable possibilities for GMRFs: Let $\boldsymbol{x} = (\boldsymbol{x}_A^T, \boldsymbol{x}_B^T)$ be

a n-dimensional GMRF with respect to a 2-dimensional graph. We are able to sample from and evaluate conditional distributions, $\pi(\boldsymbol{x}_A|\boldsymbol{x}_B)$, with computational complexity $\mathcal{O}(n_A^{3/2})$ where $n_A = \dim(\boldsymbol{x}_A)$, Rue and Held (2005, Ch. 2.3.2.). We can sample and evaluate also with deterministic constraints, i.e. from $\pi(\boldsymbol{x}|C\boldsymbol{x} = \boldsymbol{a})$, where C is a known matrixes of size $(n \times k)$, $\boldsymbol{a}$ a vector of length k and $\pi(\boldsymbol{x})$ a GMRF, Rue and Held (2005, Ch. 2.3.3.). The extra complexity due to the constraint is $\mathcal{O}(k^3)$, e.g. almost no extra cost for small k (i.e. few constraints).

In our case each node is a house sparrow, and the corresponding variable its breeding value. But each bird has seven traits, and hence seven breeding values. Marginally for each bird, its breeding values are multivariate Gaussian without a Markov structure. Therefore each node contains not only one variable, but a vector of variables. This is known as a multivariate GMRF (MGMRF) with respect to the pedigree-based graph $\mathcal{G}$. Since a MGMRF itself is a GMRF (with respect to another graph) all properties for GMRFs are also valid for MGMRFs.

For all GMRF sampling and evaluation routines, we have used the library GMRFLib, Rue and Follestad (2002).

5 Parameter estimation

5.1 A two-blocks Gibbs sampler

To make inference from the multiple-trait animal model we use a Gibbs sampler to obtain samples from $\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}_{miss}, \Sigma_u, \Sigma_\epsilon|\boldsymbol{y}_{obs})$. Traditionally single-site (or nearly single-site) Gibbs samplers have been used, which can give serious mixing problems, see Sorensen and Gianola (2002, Ch. 13.5). Well known remedies for mixing problems are blocking; Liu et al. (1994), Liu (1994) and reparametrisation; Gelfand et al. (1995). We will use a Gibbs sampler with only two blocks; Σ_u and Σ_ϵ in one block and all remaining variables (breeding values $\boldsymbol{u}$, level variables $\boldsymbol{\beta}$ and missing data $\boldsymbol{y}_{miss}$) in the other block. A similar approach is empirically explored and found more successful than reparametrisation for a single-trait animal model with a small synthetic dataset in Garcia-Cortes and Sorensen (1996).

Let $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}_{miss})$ and write the constraints as $C\boldsymbol{x} = \boldsymbol{0}$. Our Gibbs sampler is found in Algorithm 1.

Algorithm 1 *Two blocks Gibbs sampler*

- Initialise $\boldsymbol{x} = \boldsymbol{x}^0$.

- for $i = 1 : \text{niter}$ (*niter* is the number of iterations)

  - $(\Sigma_u^i, \Sigma_\epsilon^i) \sim \pi(\Sigma_u, \Sigma_\epsilon|\boldsymbol{x}^i, \boldsymbol{y}_{obs})$
  - $\boldsymbol{x}^i \sim \pi(\boldsymbol{x}|\Sigma_u^i, \Sigma_\epsilon^i, \boldsymbol{y}_{obs}, C\boldsymbol{x} = \boldsymbol{0})$

- Return $(\Sigma_u^1, \Sigma_u^2, \dots, \Sigma_u^{niter})$, $(\Sigma_\epsilon^1, \Sigma_\epsilon^2, \dots, \Sigma_\epsilon^{niter})$ and $(\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^{niter})$.

In Appendix 1 we show that $\pi(\boldsymbol{x}|\Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs}, C\boldsymbol{x} = \boldsymbol{0})$ is a MGMRF with constraints, and can be sampled from and evaluated using efficient algorithms for GMRFs. Without constraints on $\boldsymbol{x}$, $\pi(\Sigma_u, \Sigma_\epsilon|\boldsymbol{x}, \boldsymbol{y}_{obs}) = \pi(\Sigma_u|\boldsymbol{x}, \boldsymbol{y}_{obs})\pi(\Sigma_\epsilon|\boldsymbol{x}, \boldsymbol{y}_{obs})$ where both factors are known inverted Wishart distributions (see e.g. Schafer (1997)). The constraints disturb this conjungacy

slightly. Therefore, we do a Metropolis-Hastings (M-H) step when updating $(\Sigma_u, \Sigma_\epsilon)$. As proposal distribution we use the full conditional distribution for $(\Sigma_u, \Sigma_\epsilon)$ without constraints on $\boldsymbol{x}$. The M-H step requires evaluation of $\pi(\Sigma_u, \Sigma_\epsilon | \boldsymbol{x}, \boldsymbol{y}_{obs})$ up to a normalisation constant. We are able to meet this requirement, see Appendix 2.

5.2 Exact posterior for the single-trait model

In Appendix 1 and 2 we show that it is possible to evaluate both $\pi(\boldsymbol{x}, \Sigma_u, \Sigma_\epsilon | \boldsymbol{y}_{obs})$ and $\pi(\boldsymbol{x} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs})$ up to a normalisation constant that do not depend on $(\Sigma_u, \Sigma_\epsilon)$. Hence we are able to evaluate the unnormalised posterior $\pi(\Sigma_u, \Sigma_\epsilon | \boldsymbol{y}_{obs})$ using

$$\pi(\Sigma_u, \Sigma_\epsilon | \boldsymbol{y}_{obs}) = \frac{\pi(\Sigma_u, \Sigma_\epsilon, \boldsymbol{x}^* | \boldsymbol{y}_{obs})}{\pi(\boldsymbol{x}^* | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs})} \tag{4}$$

for any value of $\boldsymbol{x}^*$ (e.g. a sample from $\pi(\boldsymbol{x} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs})$).

In the animal model with one trait the covariance matrices Σ_u and Σ_ϵ reduce to two scalars; the variances σ_u^2 and σ_ϵ^2, respectively. The posterior $\pi(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$ is only two-dimensional, and it is feasible to evaluate $\pi(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$ on a grid of $(\sigma_u^2, \sigma_\epsilon^2)$ using (4). This gives us a unnormalised discretisation of $\pi(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$ on a lattice which is straightforward to normalise. We denote this discretisation $\pi_d(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$.

Furthermore, we can use $\pi_d(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$ to calculate an approximation to the posterior of the heritability (3); $\pi_a(h^2 | \boldsymbol{y}_{obs}) \approx \pi(h^2 | \boldsymbol{y}_{obs})$. The details of these calculations are given in Appendix 3. How crude the approximation $\pi_a(h^2 | \boldsymbol{y}_{obs})$ is, depends on the discretisation of $(\sigma_u^2, \sigma_\epsilon^2)$.

6 Results

We have used the methodology from Section 5 to analyse the data introduced in Section 2 with the animal model described in Section 3. As prior parameters for Σ_u and Σ_ϵ we have used inverted Wishart distributions $\Sigma \sim W(V, \nu)$ with parameters $\nu = (\text{number of traits} + 1)$ and V the identity matrix. This corresponds to flat priors on all heritabilities and genetic and environmental correlations. Hatch year and hatch island are used as group levels for all traits, and sex for all traits except the badge measures. For our standardised data we have made inference from both the multivariate animal model using the two-blocks Gibbs sampler, and from the seven single-trait animal models using our ability to calculate posterior distributions without doing MCMC.

The approximated posterior $\pi_d(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$ for the single-trait models for bill length and visible badge size are given in Figure 2 (based on a regular 200×200 grid covering variances from 0.005 to 1.0). Not surprisingly, we find that there is a negative correlation between the posterior estimates of σ_u^2 and σ_ϵ^2: The phenotypic variance (the variance in the trait data) not explained by level effects has to be explained either by the environmental or by the genetic variance. The posterior heritabilities for the seven single-trait models are plotted in Figure 3 together with the histograms visualising the posterior heritabilities from the multiple-trait model. The posterior distributions of the heritabilities for tarsus length, wing length, bill depth and bill length all have mean around 0.5 and are quite symmetric. Body mass, total badge and visible badge have a mean around 0.3 and left-skewed posteriors (the modes are left of 0.3 and the right tails are heavy). At the end of this section we discuss the differences in posterior distributions between single and multiple-trait models.

In Figure 4 marginal posterior densities for genetic and environmental correlations are plotted. We see that most of the distributions are wide, especially the genetic correlations involving the badge measures. The badges have lower heritability than the other traits (except body-mass), and there are many missing observations (as females do not have a badge). Hence, there is less information about these genetic correlations in the data, as the posterior distributions reflect. All environmental correlations are either close to zero or positive. The environmental correlations between tarsus length and all other traits but the badge measures are clearly positive (95% of the posteriors are above 0), as well as the environmental correlation between the badges. None of the genetic correlations are clearly negative, but the genetic correlations between tarsus length and body mass and visible badge size , between wing length and bill length and body mass, and between bill depth and bill length are clearly positive.

From the evolutionary biologists point of view, it is of interest to know if there has been any evolutionary changes and to identify possible causes for these changes. The posterior mean of the breeding values for bill depth for birds in the cohorts (with hatch year) 1993-2002 are plotted in Figure 5. There is an increasing trend in the average breeding value of cohorts with time. This trend indicates that *micro-evolution* is occurring, and is further examined in a separate paper; Jensen et al. (2006b).

The two-block Gibbs sampler ran for 200000 iterations after a burn-in of 30000 iterations, and we have tested the convergence with trace plots and estimated auto-correlation functions (ACFs) for some selected variables, see Figure 6. We find that the auto-correlation is large for all hyper-parameters (parameters in Σ_u and Σ_ϵ), especially those with a wide posterior distribution. Variables in $\boldsymbol{u}$ and $\boldsymbol{\beta}$ on the other side have very small auto-correlation (note the different scales for ACFs in Figure 6). Hence the hyper-parameters (Σ_u and Σ_ϵ) and functions of them mix slower than the field variables (here $\boldsymbol{\beta}$ and $\boldsymbol{u}$), a result also reported by others; Diggle et al. (2003), Steinsland and Rue (2005).

Comparing the posterior distributions from the multiple-trait model with those from the single trait models in Figure 4, we see that the distributions roughly agrees. Interestingly, going from the single to the multiple-trait model, the posterior for tarsus length gets a bit wider (i.e. more uncertain), for the bill depth quite a bit wider, for bill length the mode shifts slightly to the right, and the distribution gets narrower. For body mass and the two badge measures the multiple-trait model posteriors are more skewed than the single-trait posteriors. An exploration can be found in Figure 7, which shows plots of posteriors for the genetic variance of the badge measures and the genetic correlation between the badges. From the left-hand plot we find that larger genetic variance can be justified for the visible badge if there are strong genetic correlation between the badges (genetic correlation 0 corresponds to the single trait model). Hence, also the marginal posterior distribution for the genetic variance is wider than for the single-trait model, as seen in right-hand plot in Figure 7.

7 Discussion

In this paper we have combined the animal model with recent developments in efficient sampling and evaluation of GMRFs. This results in a robust two-block Gibbs sampler for the multiple-trait animal model. For the single-trait animal model, we are able to calculate posteriors without doing MCMC. We have analysed a dataset with 3572 individuals with seven traits, i.e. a model with more than 25000 variables. With todays computers we are able to deal with larger models. There are also good opportunities to use parallel computers or

PC-clusters to sample and evaluate GMRFs, see Steinsland (2003). Parallel computing can be used to decrease the computation time or to increase the possible problem size. For a single-trait model we will be able to handle a pedigree with more than 1000000 animals. This makes our method very useful also for animal breeders that may have large datasets.

Bayesian animal models have, as far as we know, never earlier been used for natural populations. Evolutionary biologists have other goals for their analysis than animal breeders. They want to observe, explain and predict natural selection, while animal breeders want to impose selection to obtain evolutionary changes of commercial interest. In addition, important system conditions (e.g. weather conditions and number of predators) can not be controlled in natural systems, and this gives extra uncertainty to the models of evolutionary biologists. Furthermore, uncertainty about parameter estimates has in general not been considered in further analysis. For example, it is today common to estimate response to selection (which is a function of Σ_u and Σ_ϵ) without giving any uncertainty for this estimate. Using a Bayesian approach it is straightforward to assess this uncertainty.

Based on this paper there are several interesting directions for further research. The animal model can be extended to also include dominance effects, genotype×environment effects and repeated records. We have reported a few biologically interesting results here, these and others are now being further examine from a evolutionary biologists point of view. Next, we want to extend the Bayesian approach to other analyses connected with the animal model (e.g. response to selection). Further, in our dataset there are trait data on birds that are left out because we do not know hatch island and hatch year (especially from the first years of the study). By modelling also the covariates (as done in King et al. (2005)) we would be able to utilise this information. To identify environmental factors that can explain the large spatial and temporal changes in phenotype and/or genotype are also biologically very interesting. This might be obtained by combining the methodology from this paper with Bayesian model choice and reversible jump MCMC.

Acknowledgement

A Appendix

A.1 Derivation of $\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}_{miss} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs}, C_u \boldsymbol{u} = \boldsymbol{0}, C_\beta \boldsymbol{\beta} = \boldsymbol{0})$

We start considering the unconstrained case and by finding the conditional distribution $\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y} | \Sigma_u, \Sigma_\epsilon)$ by

$$
\begin{aligned}
\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y} | \Sigma_u, \Sigma_\epsilon) \quad &\propto \quad \pi(\boldsymbol{y} | \boldsymbol{u}, \boldsymbol{\beta}, \Sigma_u, \Sigma_\epsilon) \pi(\boldsymbol{u} | \Sigma_u) \pi(\boldsymbol{\beta}) \\
&\propto \quad \exp(-\frac{1}{2}(\boldsymbol{y} - (X\boldsymbol{\beta} + W\boldsymbol{u}))^T (I \otimes \Sigma_\epsilon)^{-1}(\boldsymbol{y} - (X\boldsymbol{\beta} + W\boldsymbol{u}))) \cdot \\
&\qquad \exp(-\frac{1}{2}\boldsymbol{u}^T (A \otimes \Sigma_u)^{-1}\boldsymbol{u}) \cdot \exp(\boldsymbol{\beta}^T \Sigma_\beta^{-1} \boldsymbol{\beta})
\end{aligned}
$$

where I is a generic notation for an identity matrix with appropriate dimension. From here it is easy to show that $(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}) | \Sigma_u, \Sigma_\epsilon \sim N(0, Q^{-1})$ with precision matrix (only lower half given)

$$
Q = \begin{bmatrix} A^{-1} \otimes \Sigma_u^{-1} + W^T(I \otimes \Sigma_\epsilon^{-1})W & & \\ -X^T(I \otimes \Sigma_\epsilon^{-1})W & \Sigma_\beta^{-1} + X^T(I \otimes \Sigma_\epsilon^{-1})X & \\ -(I \otimes \Sigma_\epsilon^{-1})W & -(I \otimes \Sigma_\epsilon^{-1})X & I \otimes \Sigma_\epsilon^{-1} \end{bmatrix}.
$$

Q is a sparse matrix which is block diagonal with some other non-zero blocks. The off-diagonal non-zero blocks are given by A^{-1} as well as non-zero blocks for corresponding elements in $\boldsymbol{y}$, $\boldsymbol{u}$ and $\boldsymbol{\beta}$.

Further, let $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}_{miss})$, and partition Q between $\boldsymbol{x}$ and $\boldsymbol{y}_{obs}$;

$$
\boldsymbol{x}, \boldsymbol{y}_{obs} | \Sigma_u, \Sigma_\epsilon \sim N(0, \begin{bmatrix} Q_{xx} & Q_{xo} \\ Q_{ox} & Q_{oo} \end{bmatrix}^{-1}) \tag{5}
$$

The conditional distribution of our interest, $\pi(\boldsymbol{x} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs})$, is then known to be Gaussian with expected value $Q_{xx}^{-1} Q_{xo}$ and precision matrix Q_{xx}; Rue and Held (2005, Ch. 2.2.3.). Hence, $\pi(\boldsymbol{x} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs})$ is also a MGMRF. The constraints can be written as $C\boldsymbol{x} = \boldsymbol{0}$, and the distribution of our interest is $\pi(\boldsymbol{x} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs}, C\boldsymbol{x} = \boldsymbol{0})$. This distribution can be sampled from and evaluated as described in Section 4.

A.2 Evaluation of $\pi(\Sigma_u, \Sigma_\epsilon | \boldsymbol{x}, \boldsymbol{y}_{obs})$

We know that

$$
\pi(\Sigma_u, \Sigma_\epsilon | \boldsymbol{x}, \boldsymbol{y}_{obs}) \propto \pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}_{miss} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs}) \pi(\Sigma_u) \pi(\Sigma_\epsilon)
$$

The priors of Σ_u and Σ_ϵ are straightforward to evaluate. According to Rue and Held (2005, Ch. 2.3.), also $\pi(\boldsymbol{x} | \Sigma_u, \Sigma_\epsilon, \boldsymbol{y}_{obs}, C\boldsymbol{x} = \boldsymbol{0})$ can be calculated up to a normalisation constant not dependent on Σ_u and Σ_ϵ. The computational demanding part of this evaluation is Cholesky decomposition of Q_{xx} in (5), which is already done while sampling $\boldsymbol{x}$.

A.3 Calculation of $\pi_d(\sigma_u^2, \sigma_\epsilon^2|\boldsymbol{y}_{obs})$ and $\pi_a(h^2|\boldsymbol{y}_{obs})$

The approximations are based on a lattice on the domain of $(\sigma_u^2, \sigma_\epsilon^2)$: The posterior of $(\sigma_u^2, \sigma_\epsilon^2)$ is only considered for a set of equal spaced values of $\sigma_u^2 \in \{\sigma_{u1}^2, \sigma_{u2}^2, \ldots, \sigma_{uN}^2\}$ and $\sigma_\epsilon^2 \in \{\sigma_{\epsilon 1}^2, \sigma_{\epsilon 2}^2, \ldots, \sigma_{\epsilon M}^2\}$. For these values the approximated posterior is given as

$$\pi_d(\sigma_{ui}^2, \sigma_{\epsilon j}^2|\boldsymbol{y}_{obs}) = \frac{\pi(\sigma_{ui}^2, \sigma_{\epsilon j}^2|\boldsymbol{y}_{obs})}{\sum_{k=1}^{M}\sum_{l=1}^{N} \pi(\sigma_{uk}^2, \sigma_{\epsilon l}^2|\boldsymbol{y}_{obs})}$$

where $\pi(\sigma_u^2, \sigma_\epsilon^2|\boldsymbol{y}_{obs})$ is the posterior probability density function which can be calculated up to a normalisation factor using (4). Hence, the ratio between two probability masses in π_d equals the corresponding ratio of the exact probability densities;

$$\frac{\pi_d(\sigma_{ui}^2, \sigma_{\epsilon j}^2|\boldsymbol{y}_{obs})}{\pi_d(\sigma_{uk}^2, \sigma_{\epsilon k}^2|\boldsymbol{y}_{obs})} = \frac{\pi(\sigma_{ui}^2, \sigma_{\epsilon j}^2|\boldsymbol{y}_{obs})}{\pi(\sigma_{uk}^2, \sigma_{\epsilon k}^2|\boldsymbol{y}_{obs})}.$$

The approximation $\pi_d(\sigma_u^2, \sigma_\epsilon^2|\boldsymbol{y}_{obs})$ can be viewed as a lattice of probability masses. Element (i, j) corresponds to a specific set of parameter values, $(\sigma_{ui}^2, \sigma_{\epsilon j}^2)$, and hence to a heritability $h^2(i, j) = \sigma_{ui}^2/(\sigma_{ui}^2 + \sigma_{\epsilon j}^2)$. To find an approximation to $\pi(h^2|\boldsymbol{y}_{obs})$ we consider a discretisation $h_0^2 < h_1^2 < \ldots h_{nh}^2$ with $h_0^2 = 0$ and $h_{nh}^2 = 1$. We construct a discretised approximation $\pi_d(h^2|y_{obs}) \approx \pi(h^2|\boldsymbol{y}_{obs})$ based on $\pi_d(\sigma_{ui}^2, \sigma_{\epsilon j}^2|\boldsymbol{y}_{obs})$: For each element (i, j) the probability mass $\pi_d(\sigma_{ui}^2, \sigma_{\epsilon^2 j}|\boldsymbol{y}_{obs})$ is divided between the discretised heritabilities closest to $h^2(i, j)$, i.e. between h_k^2 and h_{k+1}^2 with k such that $h_k^2 \leq h^2(i, j) \leq h_{k+1}^2$. A fraction $(1 - \frac{h^2(i,j)-h_k^2}{h_{k+1}^2-h_k^2}) \cdot \pi_d(\sigma_{ui}^2, \sigma_{\epsilon^2 j}|\boldsymbol{y}_{obs})$ is dedicated to $\pi_d(h_k^2|\boldsymbol{y}_{obs})$ the remaining to $\pi_d(h_{k+1}^2|\boldsymbol{y}_{obs})$. A continuous approximation $\pi_a(h^2|\boldsymbol{y}_{obs}) \approx \pi(h^2|\boldsymbol{y}_{obs})$ is made by a linear interpolation of the probability masses of $\pi_d(h^2|\boldsymbol{y}_{obs})$ followed by a normalisation.

References

Altwegg, R., Ringsby, T., and Saether, B.-E. (2000). Phenotypic correlates of dispersal in a metapopulation of house sparrows *Passer domesticus*. *Journal of Animal Ecology*, 69:762–770.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36(2):192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195.

Blasco, A. (2001). The Bayesian controversy in animal breeding. *Journal of Animal Science*, 79(8):2023–2046.

Diggle, P. J., Ribeiro Jr., P. J., and Christensen, O. F. (2003). An introduction to model-based Geostatistics. In Møller, J., editor, *Spatial Statistics and Computational Methods*, Lecture Notes in Statistics; 173, pages 43–86. Springer-Verlag, Berlin.

Garcia-Cortes, L. and Sorensen, D. (1996). On a multivariate implementation of the Gibbs sampler. *Genetics, Selection, Evolution*, 28(1):121–126.

Gelfand, A., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear meixed models. *Biometrika*, 82(3):479–488.

Jensen, H., Saether, B.-E., Ringsby, T. H., Tufto, J., Griffith, S., and Ellegren, H. (2003). Sexual variation in heritability and genetic correlations of morphological traits in house sparrow (*Passer domesticus*). *Journal of Evolutionary Biology*, 16(6):1296–1307.

Jensen, H., Saether, B-E. Ringsby, T. H., Tufto, J., Griffith, S., and Ellegren, H. (2004). Lifetime reproductive success in relation to morphology in the house sparrow *passer domesticus*. *Journal of Animal Ecology*, 73:599–611.

Jensen, H., Steinsland, I., Ringsby, T. H., and Saether, B. E. (2006a). Effects of indirect selection on the evolution of a sexual measurment in the house sparrow (*passer domesticus*). Manuscipt.

Jensen, H., Steinsland, I., Ringsby, T. H., and Saether, B. E. (2006b). Micro-evolution in a house-sparrow popultion. In process.

King, R., Brooks, S. P., Morgan, B. J. T., and Coulson, T. (2005?). Factors influencing soay sheep survival: A Bayesian analysis. *Biometrics*. in press.

Kruuk, L. (2004). Estimating genetic parameters in natural populations using the 'animal model'. *Phil. Trans. of the Royal Society of London, series B-biological sciences*, 359:874–890.

Kruuk, L., Merila, J., and Sheldon, B. C. (2001). Phenotypic selection on a heritable size trait revisited. *American Naturalist*, 158(6):557–571.

Kruuk, L., Slate, J., Pemberton, J., Brotherstone, S., Guinness, F., and Clutton-Brock, T. (2002). Antler size in red deer: Heritability and selection but no evolution. *Evolution*, 56(8):1683–1695.

Lande, R. and Arnold, S. J. (1983). The measurment of selection on correlated charecters. *Evolution*, 37:1210–1226.

Liu, J. S. (1994). The collapsed Gibbs sampler with application to a gene regulation problem. *Journal of the American Statistical Association*, 89:958–966.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrica*, 81(1):27–40.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quatitative Traits*. Sinauer Associates, 5 edition.

Malecot, G. (1969). *The mathematics of heredity*. W.H. Freemann. Revised, edited and translated by Demetrios M. Yermanos.

Quaas, R. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, 32(4):949–953.

Ringsby, T. H., Saether, B.-E., Altwegg, R., and Solberg, E. (1999). Temporal and spatial variation in survival rates of a house sparrow, *passer domesticus*, metapopulation. *Oikos*, 85:419–425.

Ringsby, T. H., Saether, B.-E., Tufto, J., Jensen, H., and Solberg, E. (2002). Asynchronous spatiotemporal demography of a house sparrow metapopulation in a correlated environment. *Ecology*, 83:561–569.

Roff, D. E. (1997). *Evolutionary Quantitative Genetics*. Chapman and Hall.

Rue, H. and Follestad, T. (2002). GMRFLib: A C-library for fast and exact simulation of Gaussian Markov random fields. Statistics Report No. 1/2002, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Saether, B.-E., Ringsby, T. H., Bakke, O., and Solberg, E. (1999). Spatial and temporal variation in demography of a house sparrow metapopulation. *Journal of Animal Ecology*, 68:628–637.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

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

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

Steinsland, I. (2003). Parallel sampling of GMRFs and geostatistical GMRF models. Preprint Statistics 7/2003, Norwegian University of Science and Technology.

Steinsland, I. and Rue, H. (2005). Collapsed blocks approximations. Statistics Report No. 9/2005, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Tufto, J., Ringsby, T. H., Dhondt, A. A., Adriaensen, F., and Matthysen, E. (2005). A parametric model for estimation of dispersal patterns applied to five passerine spatially structured populations. *American Naturalist*, 165(1):E13–E26.
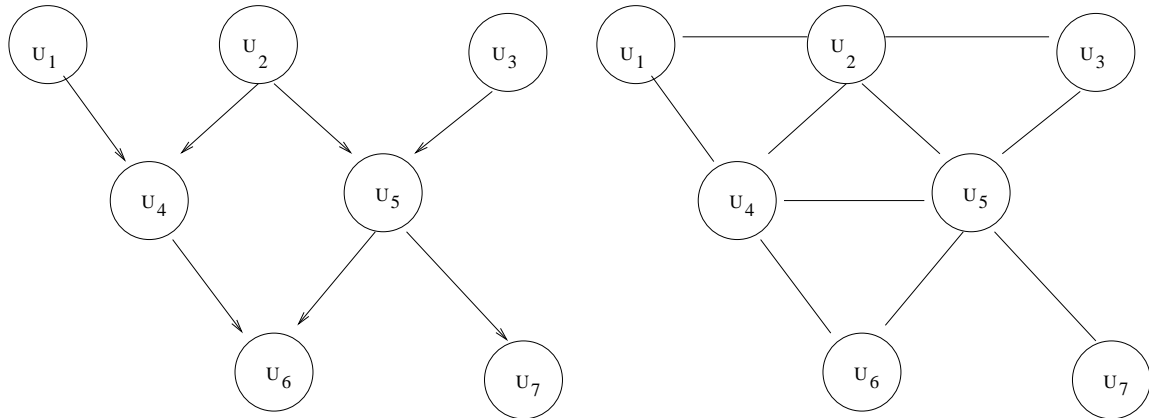
Figures

Figure 1: *Left: The pedigree / DAG for the breeding values of seven birds. Bird 1 and 2 are the parents of bird 4, and bird 2 and 3 the parents of bird 5. Bird 7 has only one known parent; bird 5. Right: The conditional independence graph found by moralising the pedigree on the left.*
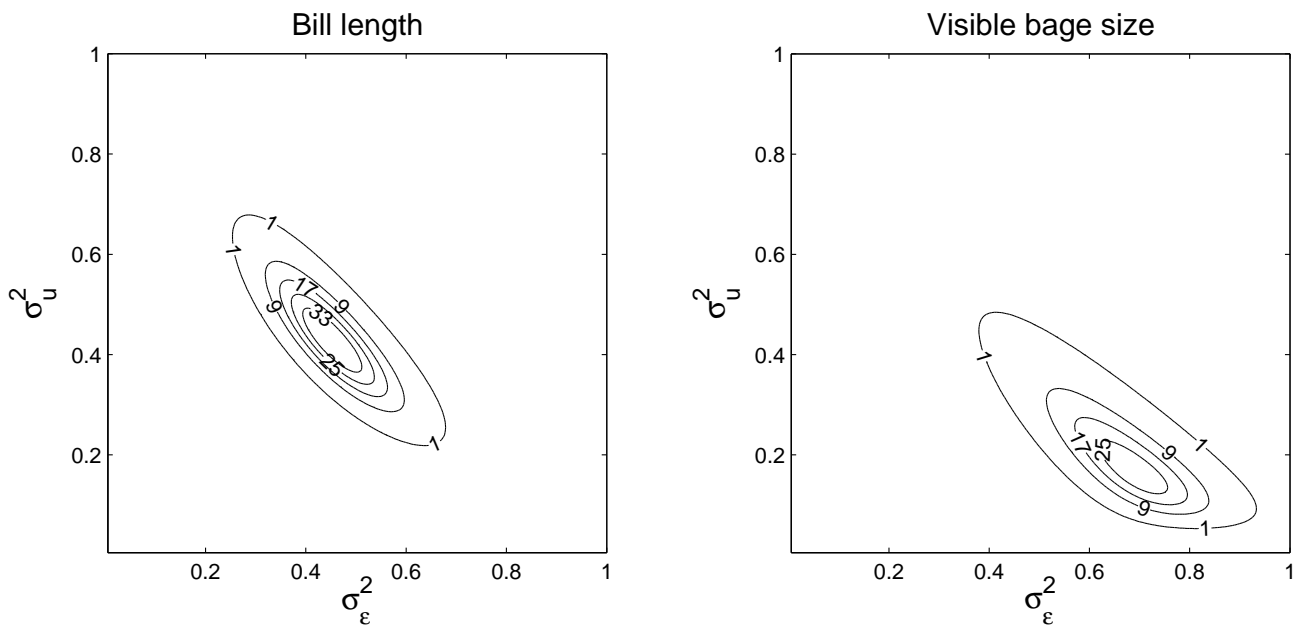
Figure 2: *Joint posterior distribution for the variance components from the single-trait models,* $\pi_d(\sigma_u^2, \sigma_\epsilon^2 | \boldsymbol{y}_{obs})$, *for bill depth (left) and visible badge size (right).*
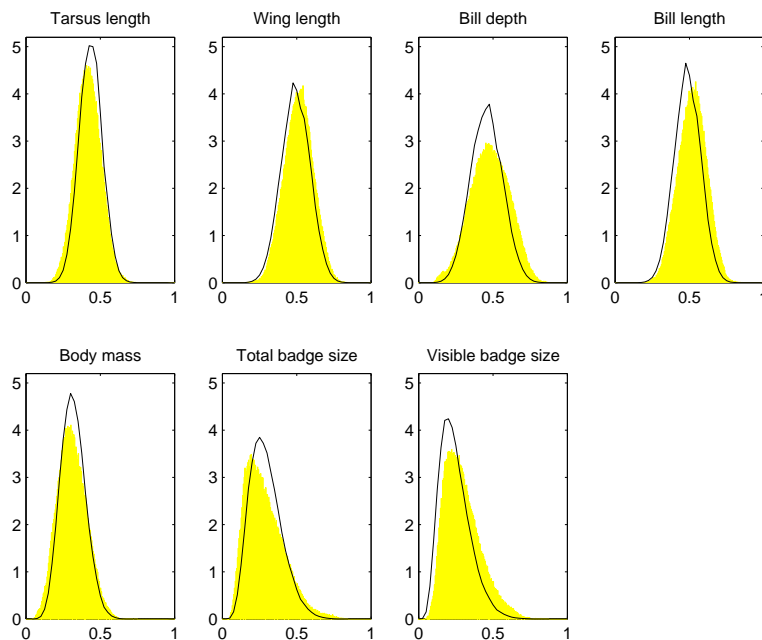
Figure 3: *Posterior heritability for seven morphological traits in house sparrows from the multiple-trait model (histograms) and single-trait models (solid lines).*
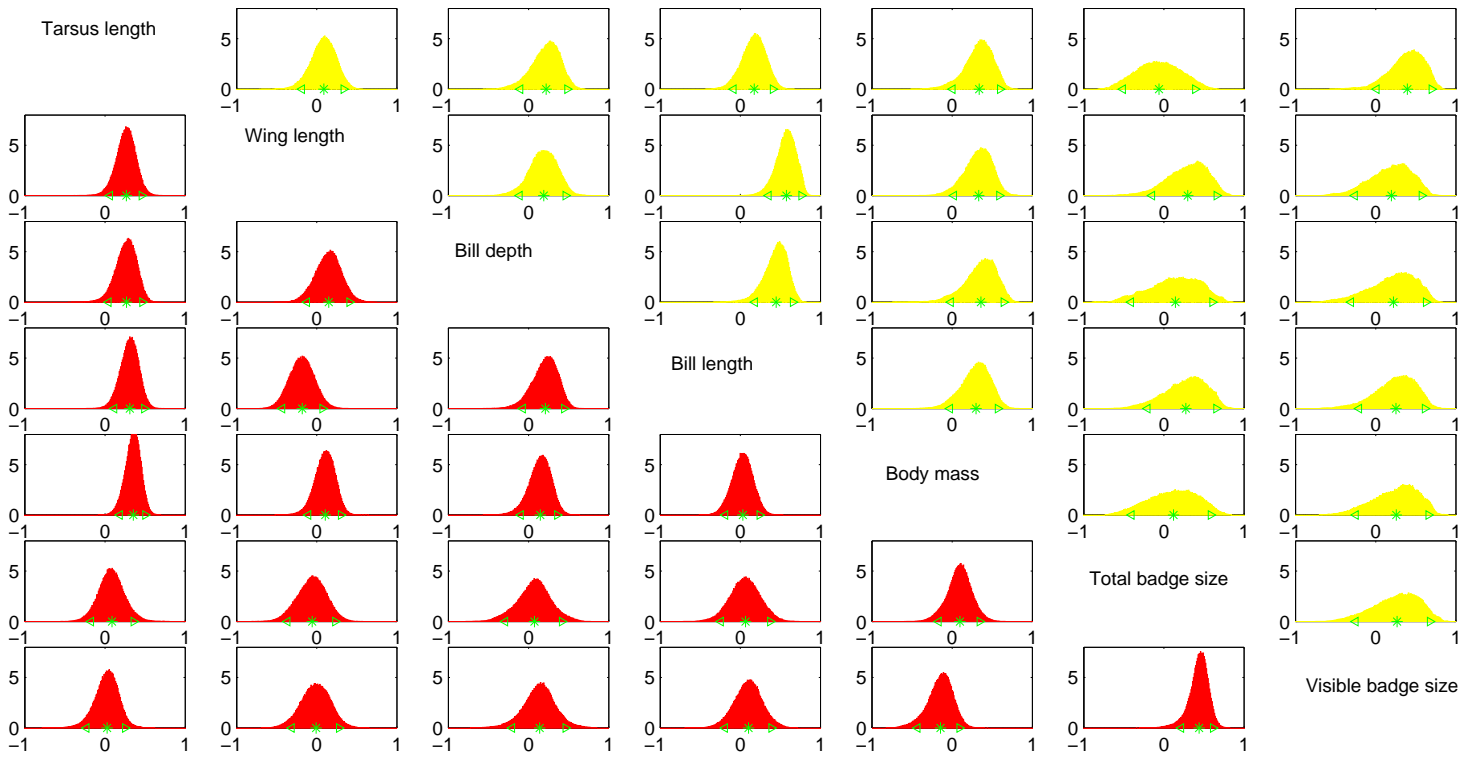
Figure 4: *Posterior genetic (above the diagonal) and environmental (below the diagonal) correlations between morphological traits in house sparrows. Mean ($*$), 5% quantile ($<$) and 95% quantile ($>$) marked.*
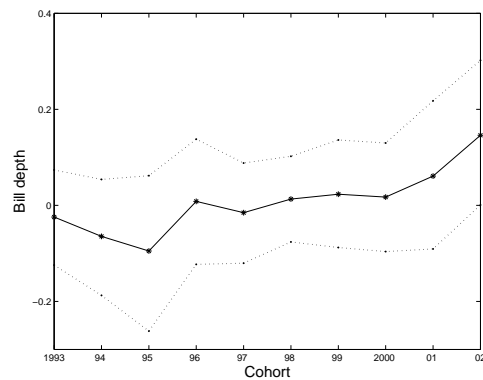

Figure 5: *Posterior mean ($*$) and 90% credibility interval for average bill depth breeding values for house sparrow cohorts that hatched in years 1993-2002.*
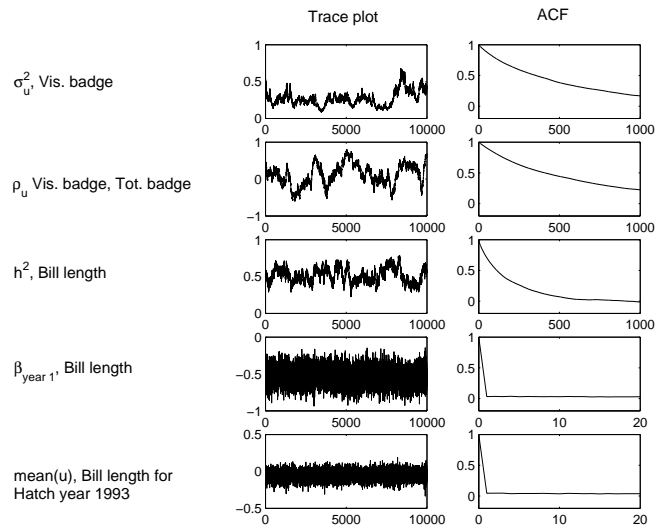
Figure 6: *Trace-plots (left) and estimated autocorrelation functions (ACFs, right) for (from the top and downwards) the genetic variance of visible badge size, genetic correlation between bill length and visible badge size, the heritability of bill length, group level effect for hatch year 1993 of bill length and mean breeding values for bill length in year 1993.*
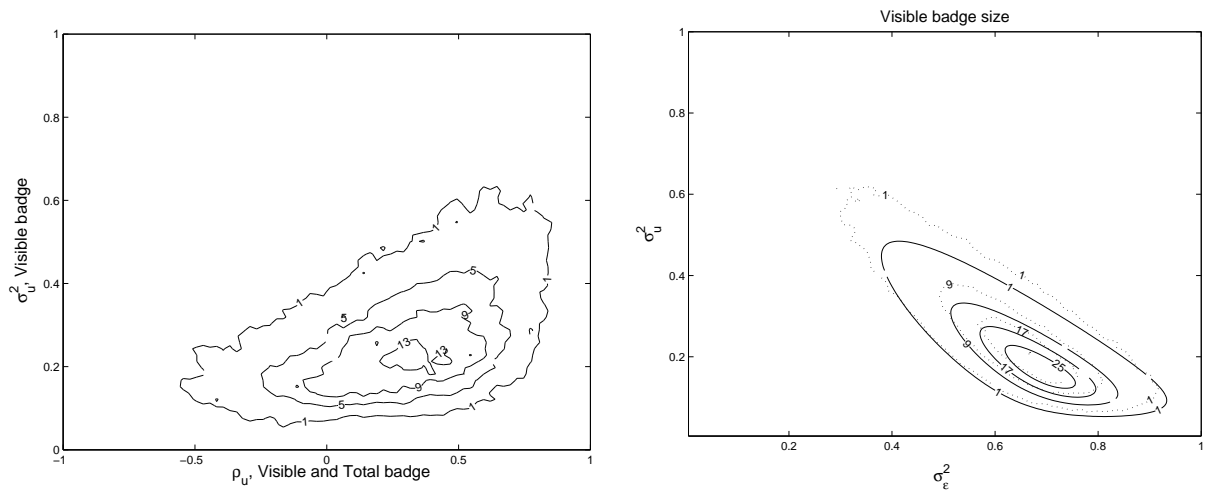

Figure 7: *Left: The joint posterior distribution of the genetic variance of the visible badge and the genetic correlation between the badges. Right: The marginal posterior for the genetic and environmental variance for visible badge with single trait model (solid lines) and multiple-trait model (dotted lines).*