Quantitative Genetics, House Sparrows and a Multivariate Gaussian Markov Random Field Model

Ingelin Steinsland & Henrik Jensen

ingelins@math.ntnu.no

Norwegian University of Science and Technology

Quantitative Genetics

- *Quantitative genetics is the study of quantitative characters. It is based on the assumption that such characters are determined by genes (...)*
- Trait = Genetic + environmental effects.
- Much used in animal and plant breeding.
- In this study:
  - Wild life population
  - Several traits simultaneously.

House Sparrows

Multivariate Gaussian Markov Random Field Model (MGMRF)

Multivariate Gaussian Markov Random Field Model

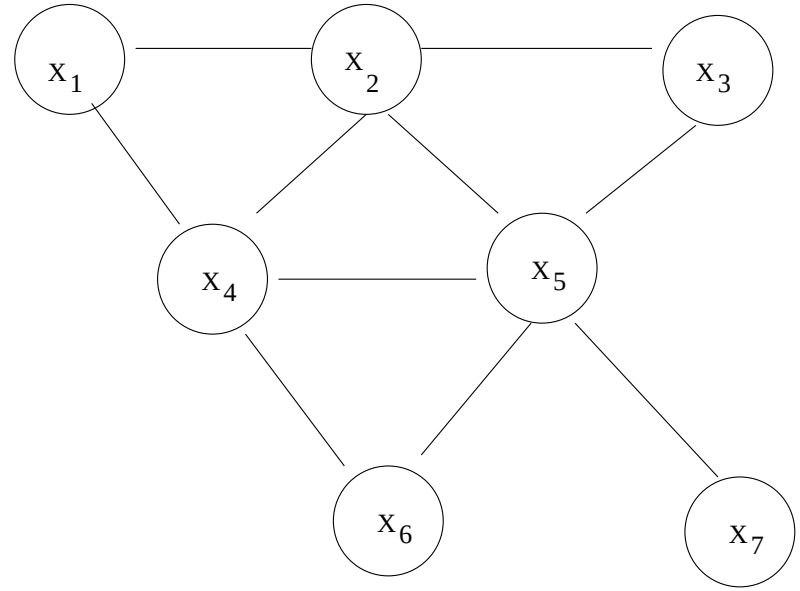
Multivariate Gaussian Markov Random Field Model (MGMRF)

- Multivariate Gaussian Model, $x \sim N(\mu, \Sigma)$

Multivariate Gaussian Markov Random Field Model

Multivariate Gaussian Markov Random Field Model (MGMRF)

- Conditional independence structure

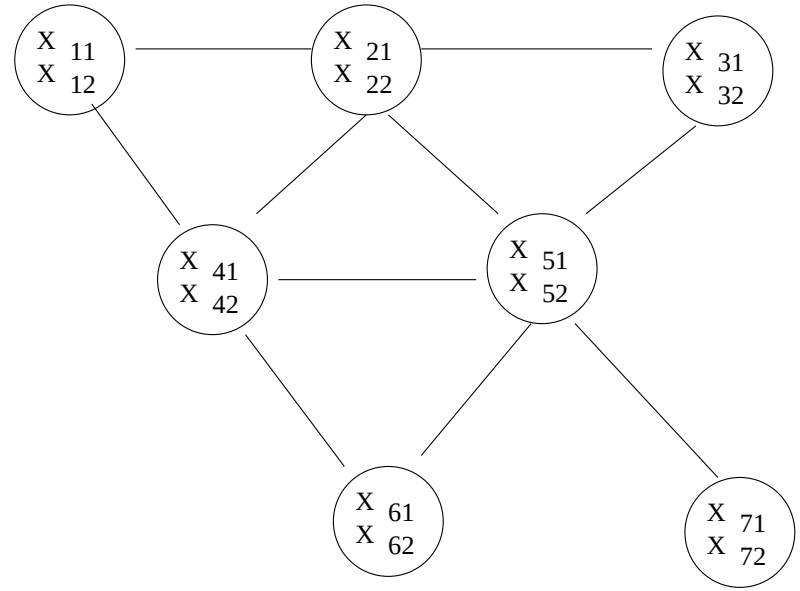


- Gives non-zero structure of $Q = \Sigma^{-1}$.

Multivariate Gaussian Markov Random Field Model

Multivariate Gaussian Markov Random Field Model (MGMRF)

- Each node is multivariate Gaussian (without Markov property)

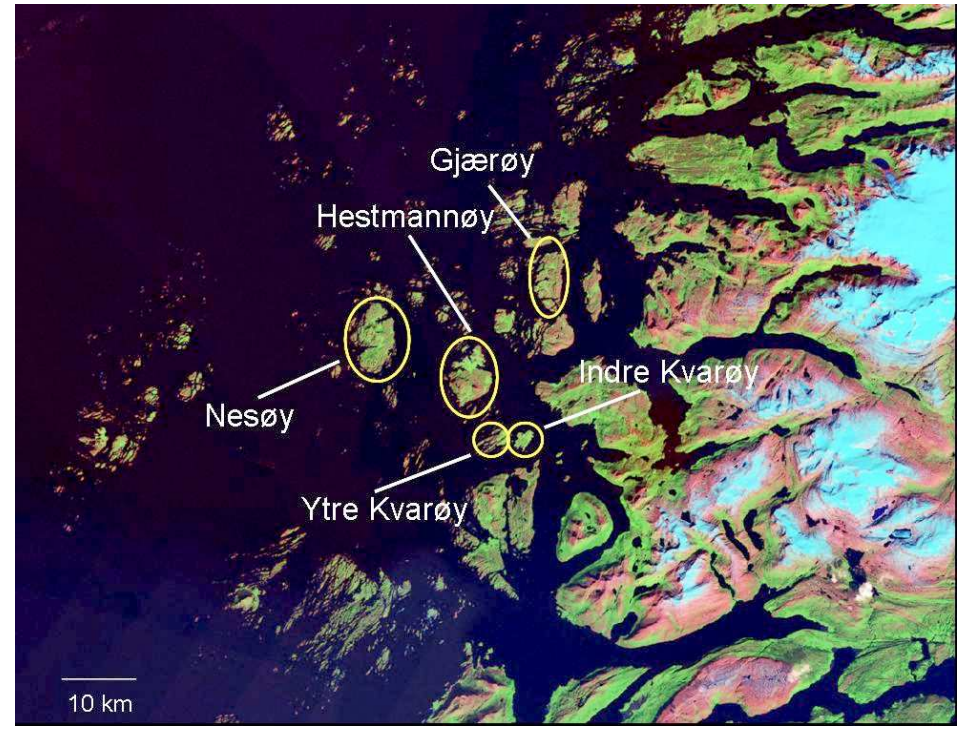


Outline

- Data
- Biological motivation
- Model
- Gibbs sampler
- Fast sampling of GMRF
- Results
- Summary and further work

Data, islands

- Most house sparrows of five islands on Helgeland, off the coast of Northern Norway, are registered since 1993.



Data, islands

- Most house sparrows of five islands on Helgeland, off the coast of Northern Norway, are registered since 1993.

Hestmannøy

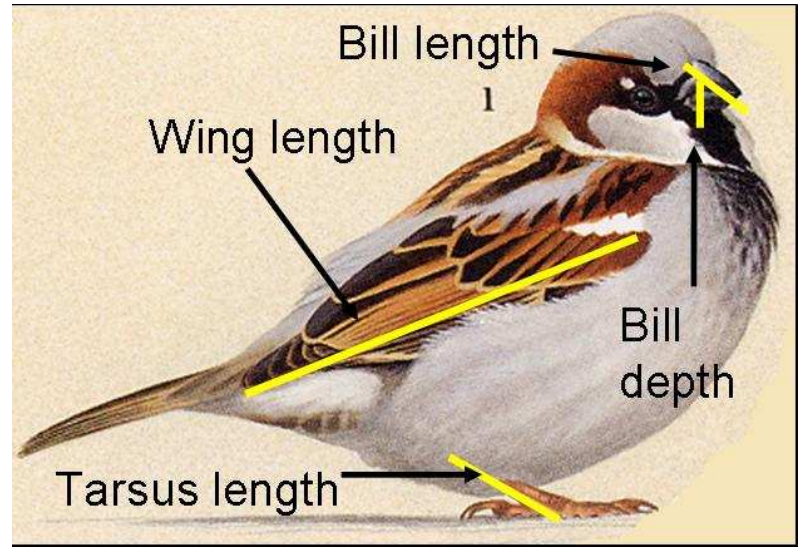
Data Collection

- Almost all fledglings are marked and blood sample collected in the nest.
- Most of the population is caught with nets during the summer and morphological traits are collected.

Morphological data

Traits measured for both sexes:

- Tarsus length (t)
- Wing length (w)
- Bill depth (bd)
- Bill length (bl)
- Body mass (bm)



Morphological data

For males only:

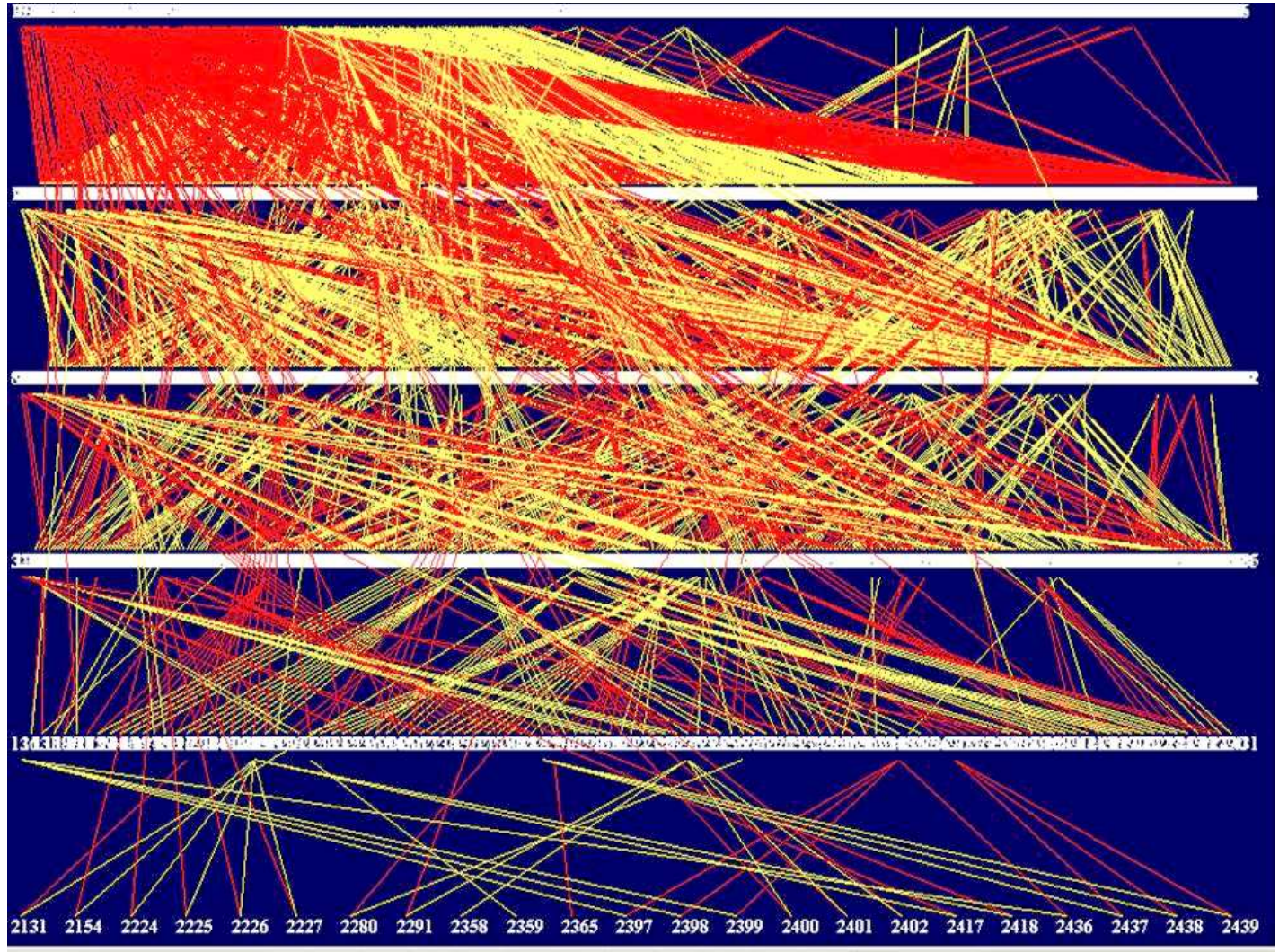
- Total badge size (tb)
- Visual badge size (vb)

Morphological data

- Tarsus length (t)
- Wing length (w)
- Bill depth (bd)
- Bill length (bl)
- Body mass (bm)
- Total badge size (tb), missing for all females
- Visual badge size (vb), missing for all females

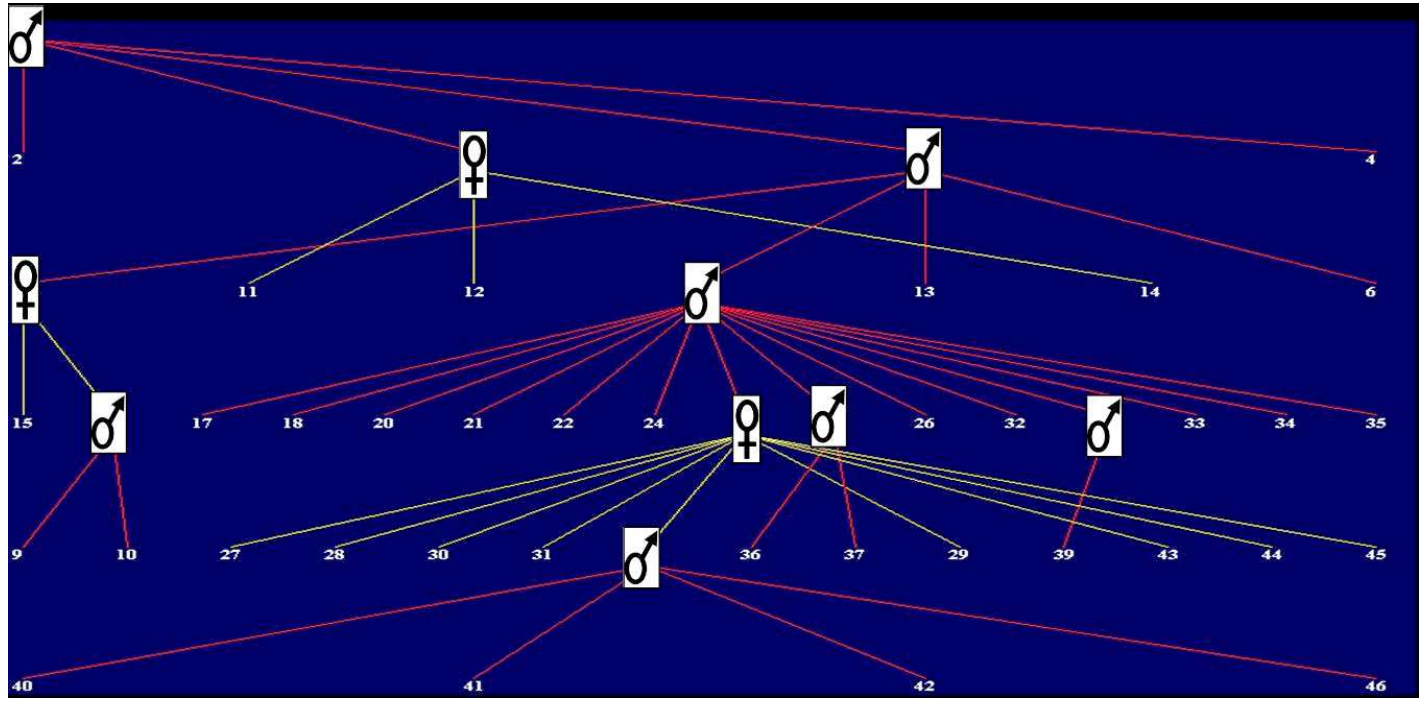
Pedigree data

- Blood samples taken in nest ⇒ DNA ⇒ Pedigree



Pedigree data

- Blood samples taken in nest ⇒ DNA ⇒ Pedigree



Summary Data

- 2563 birds in the pedigree (from 1993 - 1999)
- 385 with measured traits as one year olds.
  - 194 males
  - 191 females
- Have 7 traits for adult birds.
  - There are missing data, e.g. females do not have badge.
- Sex, hatch year and island is known for all the birds.

Biological motivation

- From Kruuk et al (2002), Evolution *Antler size in red deer: Heritability and selection, but no evolution*
- One hypotheses: *Genetic correlation between a given trait and other traits under selection will constrain the direction and pace of any evolutionary change*
- Speed and direction of evolution.
- Important for ability to respond to environmental changes, e.g. climate changes.

Model, one bird

For bird i ($i = 1, \ldots, 385$):

observed traits	=fixed	+genetic	+environmental
y_i	$=\beta_i$	$+u_i$	$+\epsilon_i$

- y_i: observations (traits),
  $y_i = (y_l, y_w, y_{bd}, y_{bl}, y_{bm}, y_{tb}, y_{vb})_i^T$
- β_i: "*fixed effects*" (sex, hatch year and island),
  $\beta \sim N(0, \sigma_\beta^2 I)$
- u_i: genetic effects, $u_i \sim N(0, \Sigma_u)$
- ϵ_i: Environmental effects: $\epsilon_i \sim N(0, \Sigma_\epsilon)$
- Conjugate prior for Σ_u and Σ_ϵ (inverted Wishart).

Animal model, for the population

$$y = X\beta + Zu + \epsilon = W \begin{pmatrix} u \\ \beta \end{pmatrix} + \epsilon$$

- $y = (y_1, y_2, \ldots, y_{mdata})^T$
- X and Z: incidence matrices, $W = (Z, X)^T$.
- u: genetic effect, $u = (u_1, u_2, \ldots, u_{nind})$.
  $u \sim N(0, \Sigma_u \otimes A)$
- A: relationship matrix
- ϵ: environmental effect $\epsilon \sim N(0, \Sigma_\epsilon \otimes I)$

Relationship matrix A

- $A_{ij} = 2\theta_{ij}$, θ_{ij}: coefficient of coancestry.

Relationship matrix A

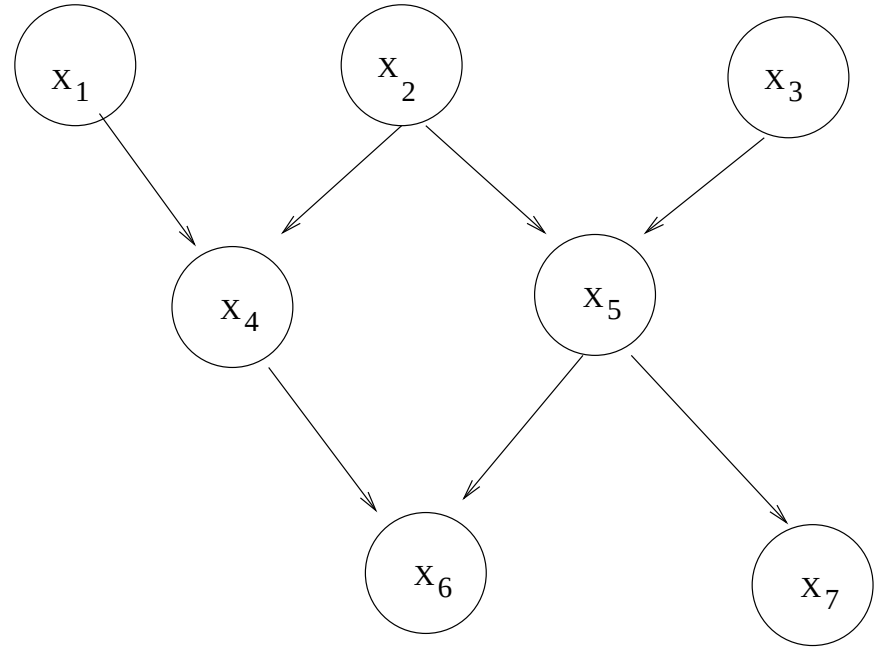
- $A_{ij} = 2\theta_{ij}$, θ_{ij}: coefficient of coancestry.
  - Identical by descent: Genes that are direct descendents of a specified gene carried in some ancestral individual.
  - Assume genes drawn randomly from two individuals (one gene from each).
  - θ_{ij}: the probability that these two genes are identical by descent.

Relationship matrix A

- $A_{ij} = 2\theta_{ij}$, θ_{ij}: coefficient of coancestry.
- A is nearly a full matrix.

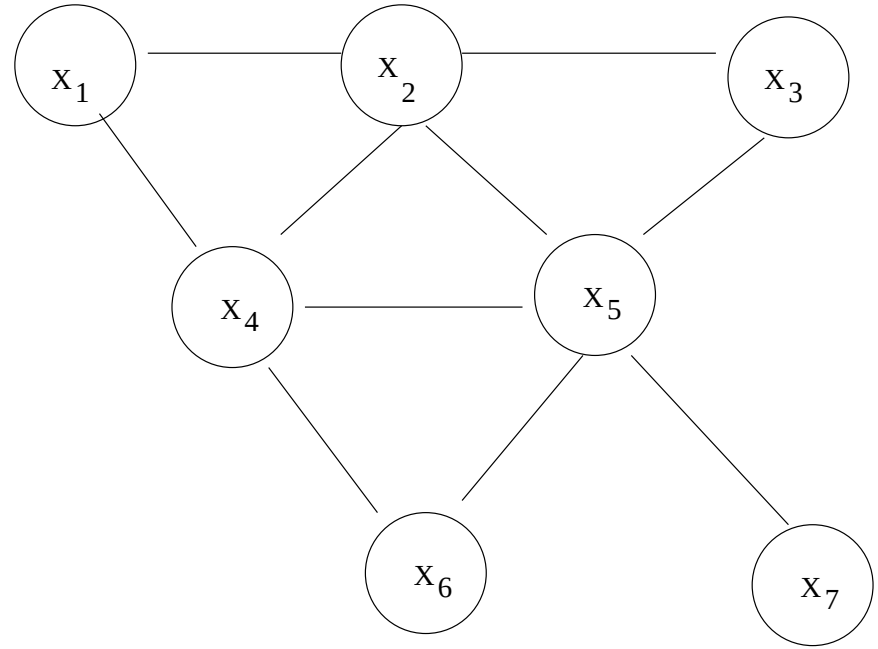
Relationship matrix A

- $A_{ij} = 2\theta_{ij}$, θ_{ij}: coefficient of coancestry.
- Pedigree = DAG (Directed Acyclic Graph)



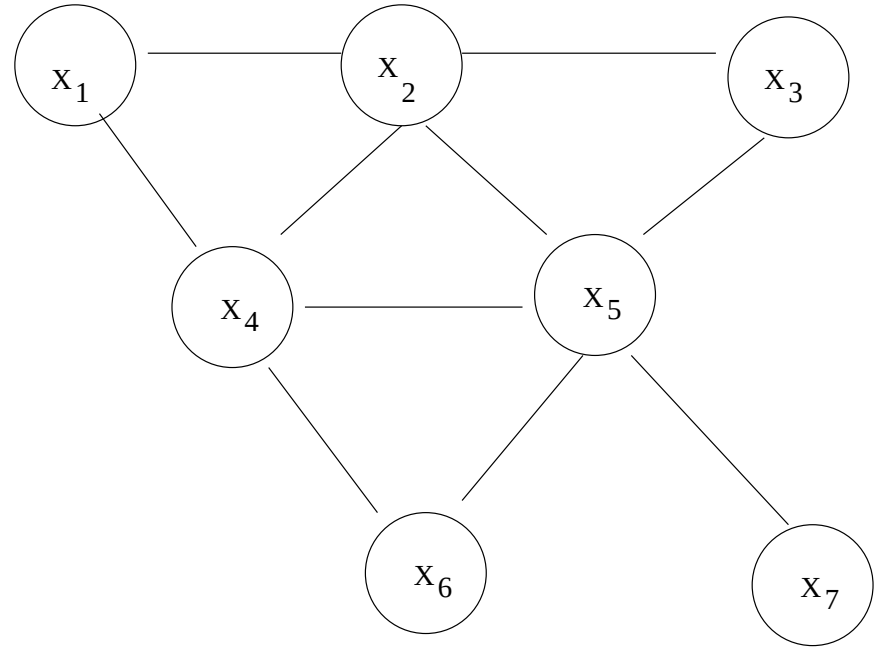
Relationship matrix A

- $A_{ij} = 2\theta_{ij}$, θ_{ij}: coefficient of coancestry.
- Pedigree = DAG (Directed Acyclic Graph)
- Structure of A^{-1} from moralising the pedigree



Relationship matrix A

- $A_{ij} = 2\theta_{ij}$, θ_{ij}: coefficient of coancestry.
- Pedigree = DAG (Directed Acyclic Graph)
- Structure of A^{-1} from moralising the pedigree



$\Rightarrow A^{-1}$ sparse.

Constraints etc.

- For all but one fixed effect (f =hatch year and island), for all traits:

$$\sum_{l=1}^{Nlevel} \beta_{flt}$$

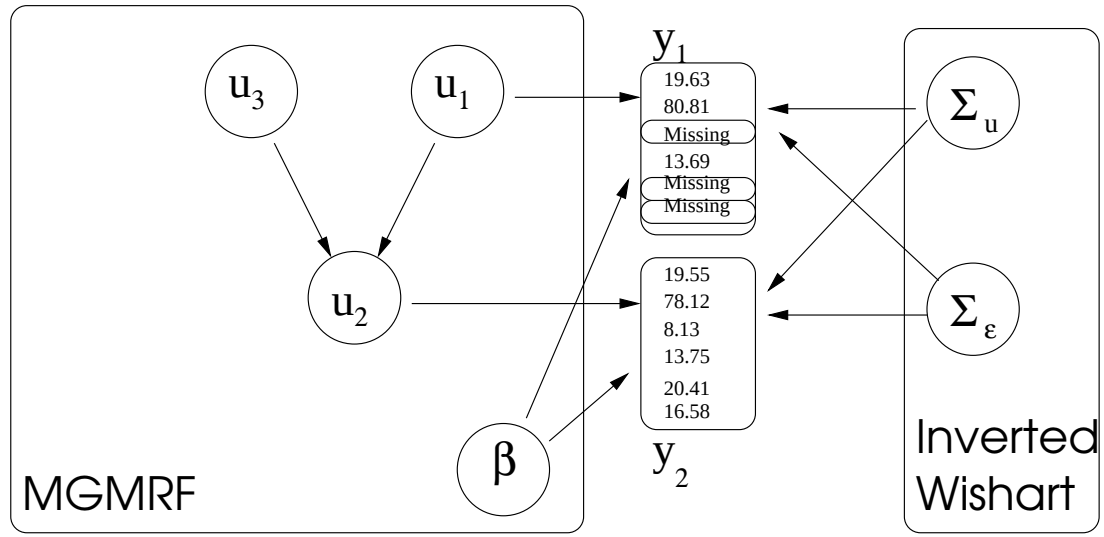
- For the breeding values, for all traits:

$$\sum_{i=1}^{N} u_{it} = 0$$

- For fixed effect sex: $\beta_{female,vb} = \beta_{female,tb} = 0$.

Parameters of interest

- u, β, Σ_u and Σ_ϵ
- Heritability; $h_j = \frac{\sigma^2_{uj}}{\sigma^2_{uj}+\sigma^2_{\epsilon j}}$,
  $j \in \{t, w, bd, bl, bm, vb, tb\}$



Gibbs sampler

Algorithm:

- For each iteration
  1. Sample from $\pi(\beta, u, y_{miss}|y_{obs}, \Sigma_u, \Sigma_\epsilon)$ of dimension ≈ 20000!
  2. Sample from $\pi(\Sigma_u, \Sigma_\epsilon|y, u, \beta)$

Blocking important for mixing

Gibbs sampler

Algorithm:

- For each iteration
  1. Sample from $\pi(\beta, u, y_{miss}|y_{obs}, \Sigma_u, \Sigma_\epsilon)$ of dimension ≈ 20000!
  2. Sample from $\pi(\Sigma_u, \Sigma_\epsilon|y, u, \beta)$

Blocking important for mixing

Conditional distributions

- $\beta, u, y_{miss}|y_{obs}, \Sigma_u, \Sigma_\epsilon \sim$ MGMRF, a multivariate Gaussian Markov Random Field.
- $\Sigma_u, \Sigma_\epsilon|y, \beta, u \sim$ Inverted Wishart

Gibbs sampler

Algorithm:

- For each iteration
  1. Sample from $\pi(\beta, u, y_{miss}|y_{obs}, \Sigma_u, \Sigma_\epsilon)$ of dimension ≈ 20000!
  2. Sample from $\pi(\Sigma_u, \Sigma_\epsilon|y, u, \beta)$

Blocking important for mixing

Conditional distributions

- $\beta, u, y_{miss}|y_{obs}, \Sigma_u, \Sigma_\epsilon \sim$ MGMRF, a multivariate Gaussian Markov Random Field.
- $\Sigma_u, \Sigma_\epsilon|y, \beta, u \sim$ Inverted Wishart
- Need to sample from a large MGMRF (structure given by A^{-1}).

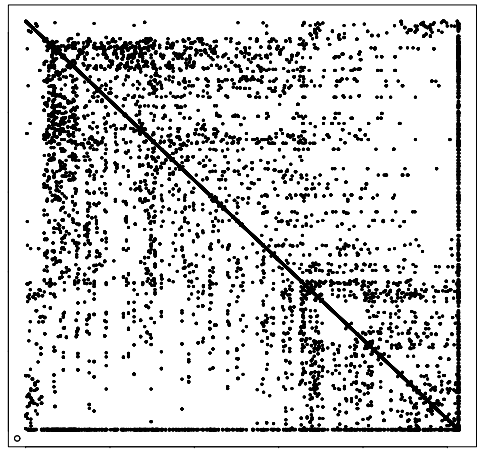
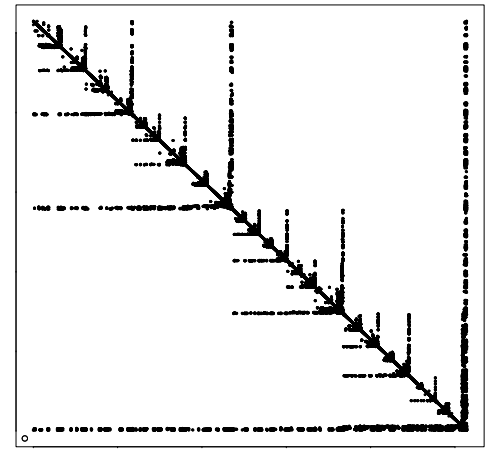
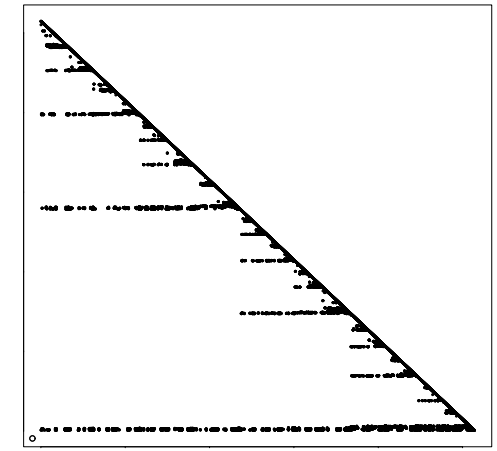
Fast sampling of GMRF

- Want to sample $x \sim N(0, Q^{-1})$.
- Find Choleskey factor; $Q = LL^T$, and sample $z \sim N(0, I)$. Solve $L^T x = z$.
- Complexity for general multivariate problem $\mathcal{O}(n^3)$.
- Sparse $Q \Rightarrow$ cheaper calculations.

Reordering

Reorder elements in Q to get sparse L.

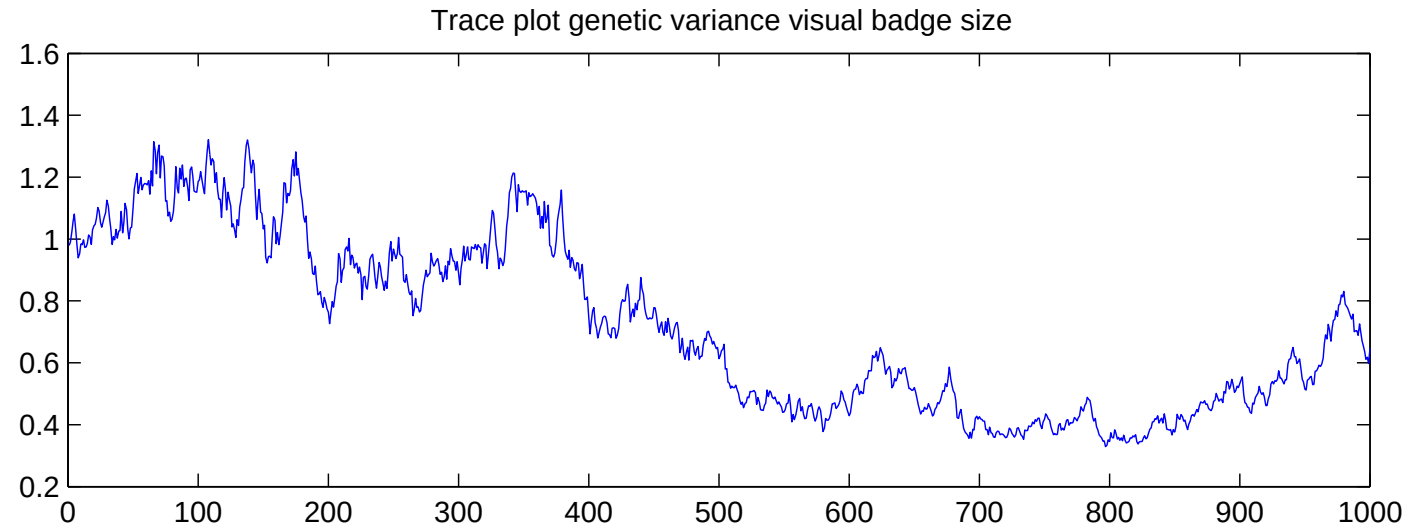
Original Q | Reordered Q | Choleskey factor

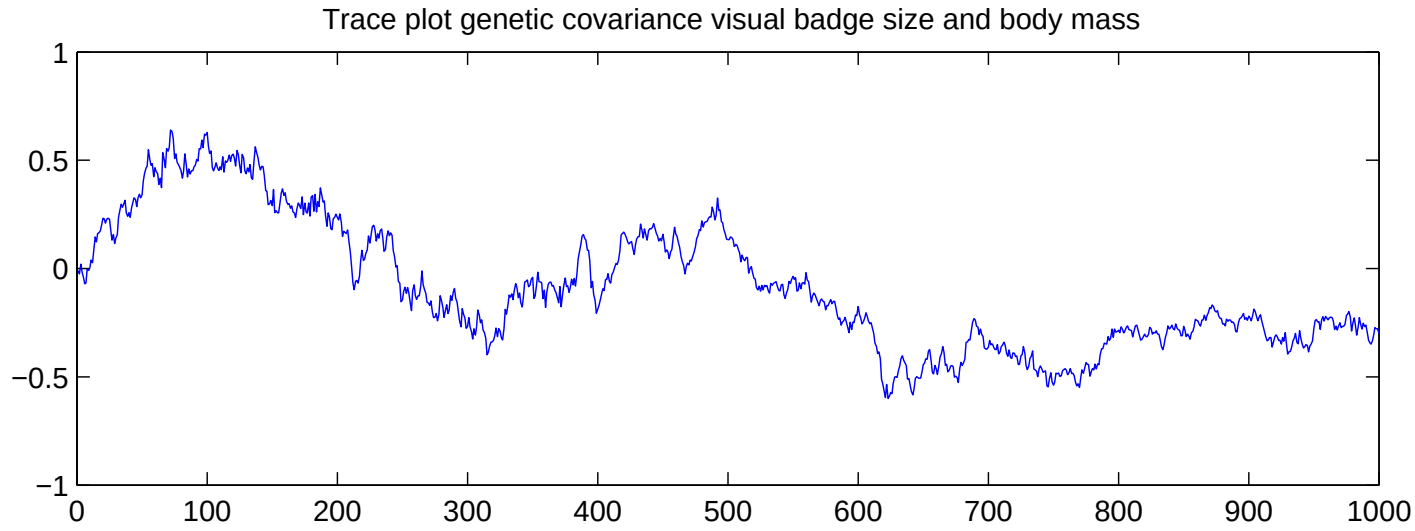




- Complexity $< \mathcal{O}(n^{1.5})$
- For constraints $\mathcal{O}(k^3)$, k: no. of constraints.
- Have used GMRFLib.

Results, MCMC

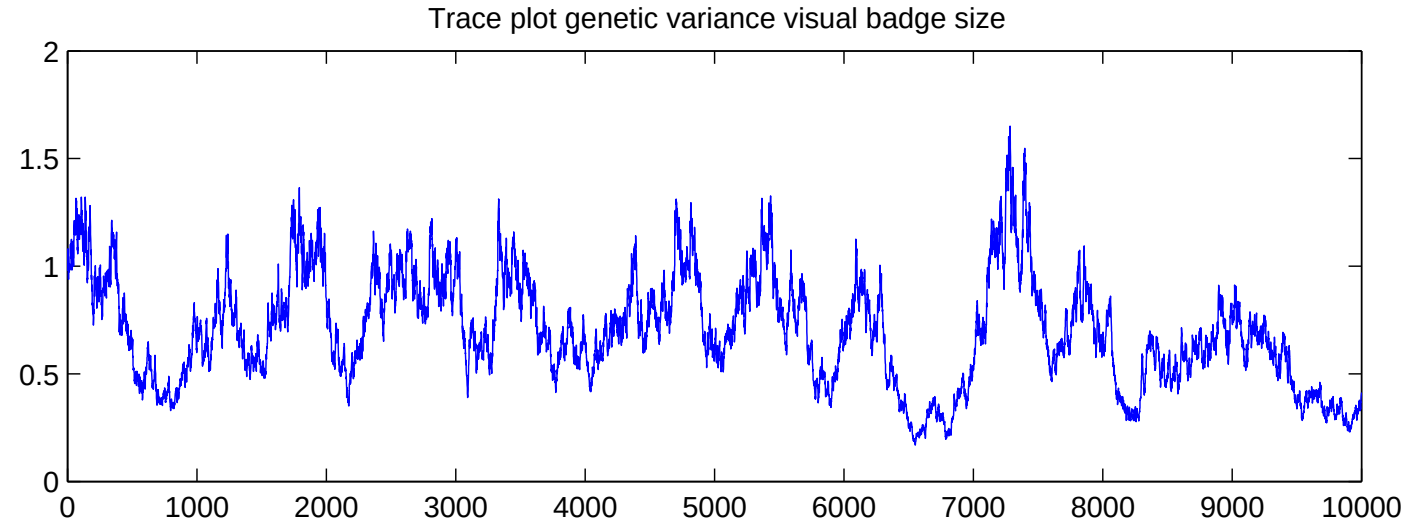
Trace-plots 1000 iterations:

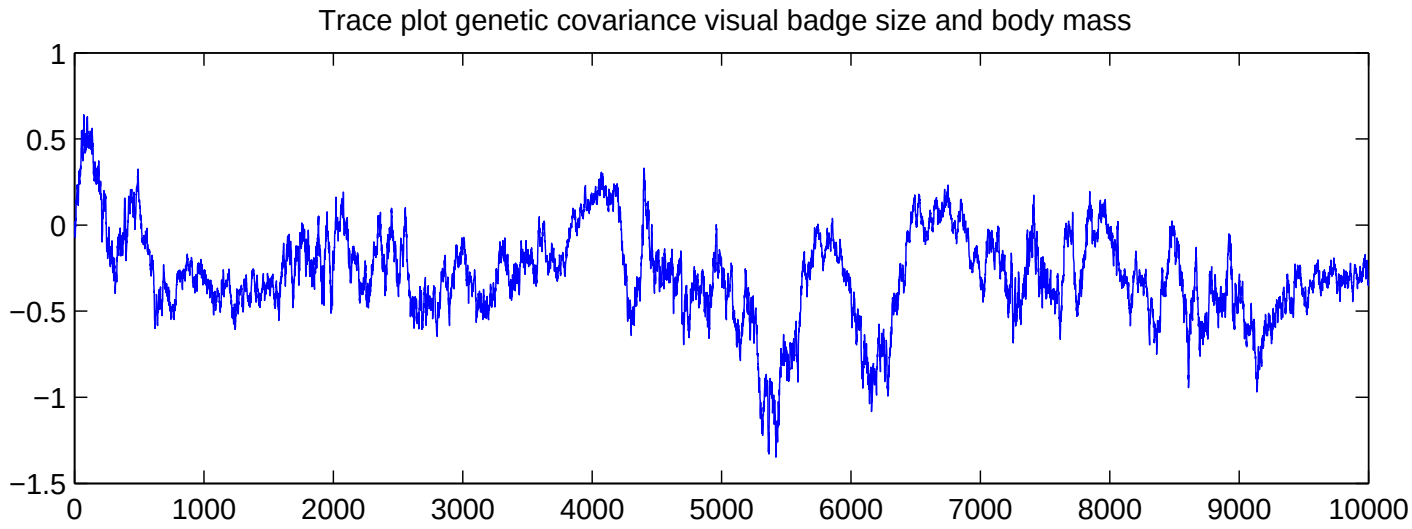




Results, MCMC

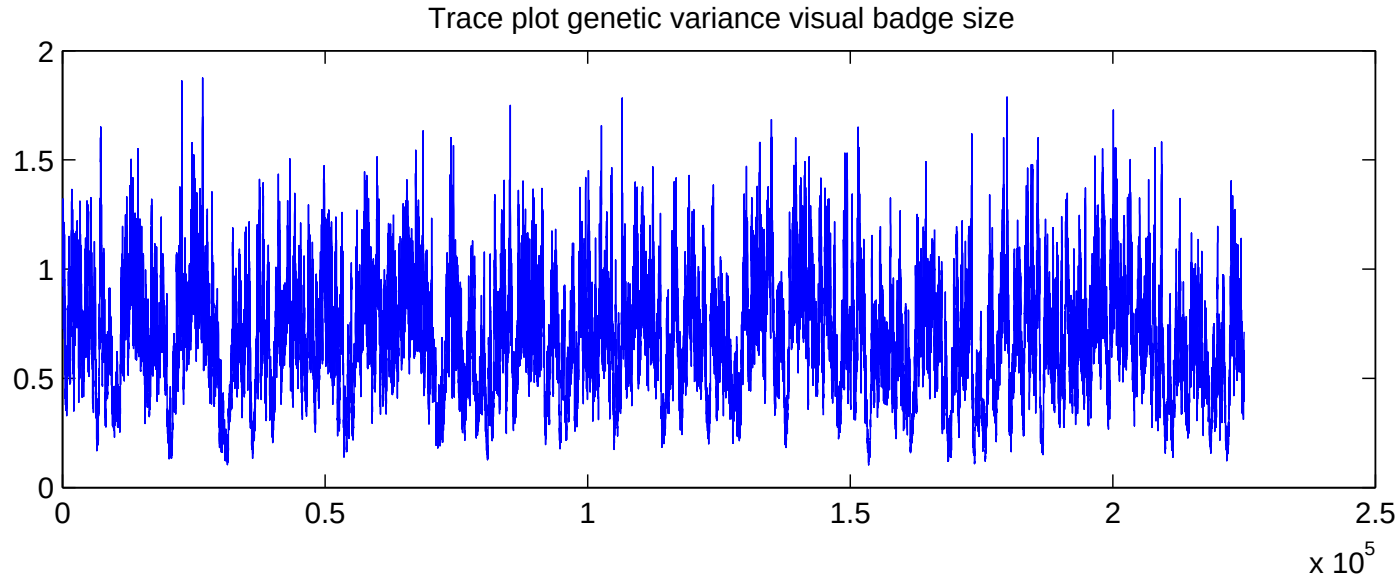
Trace-plots 10000 iterations:

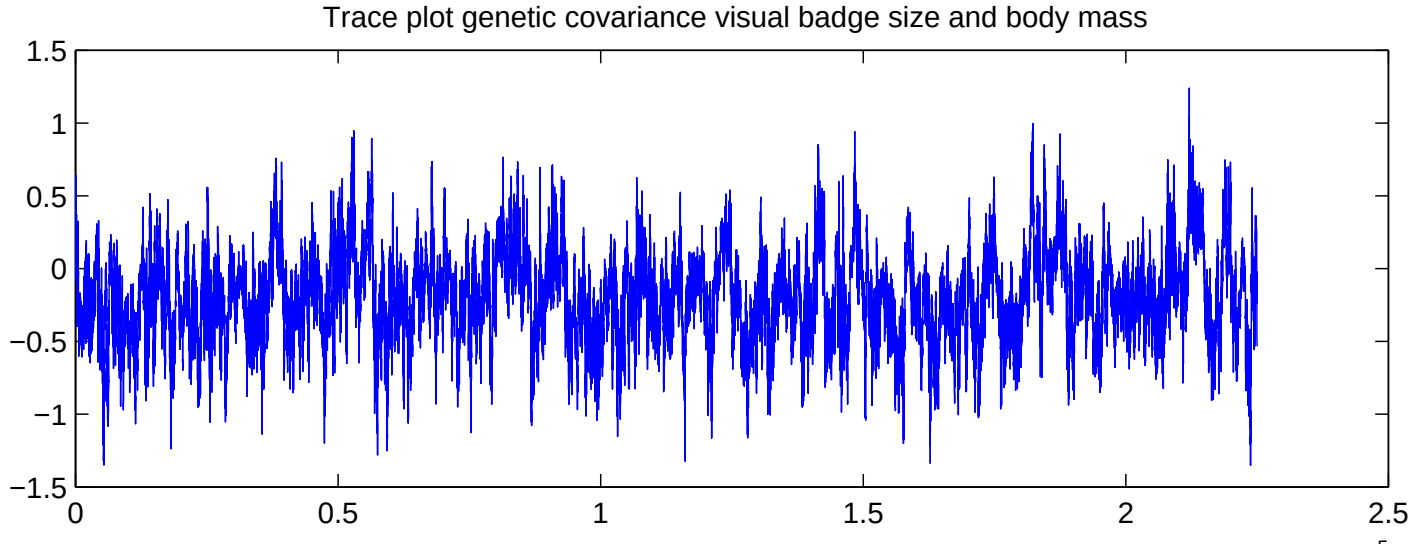




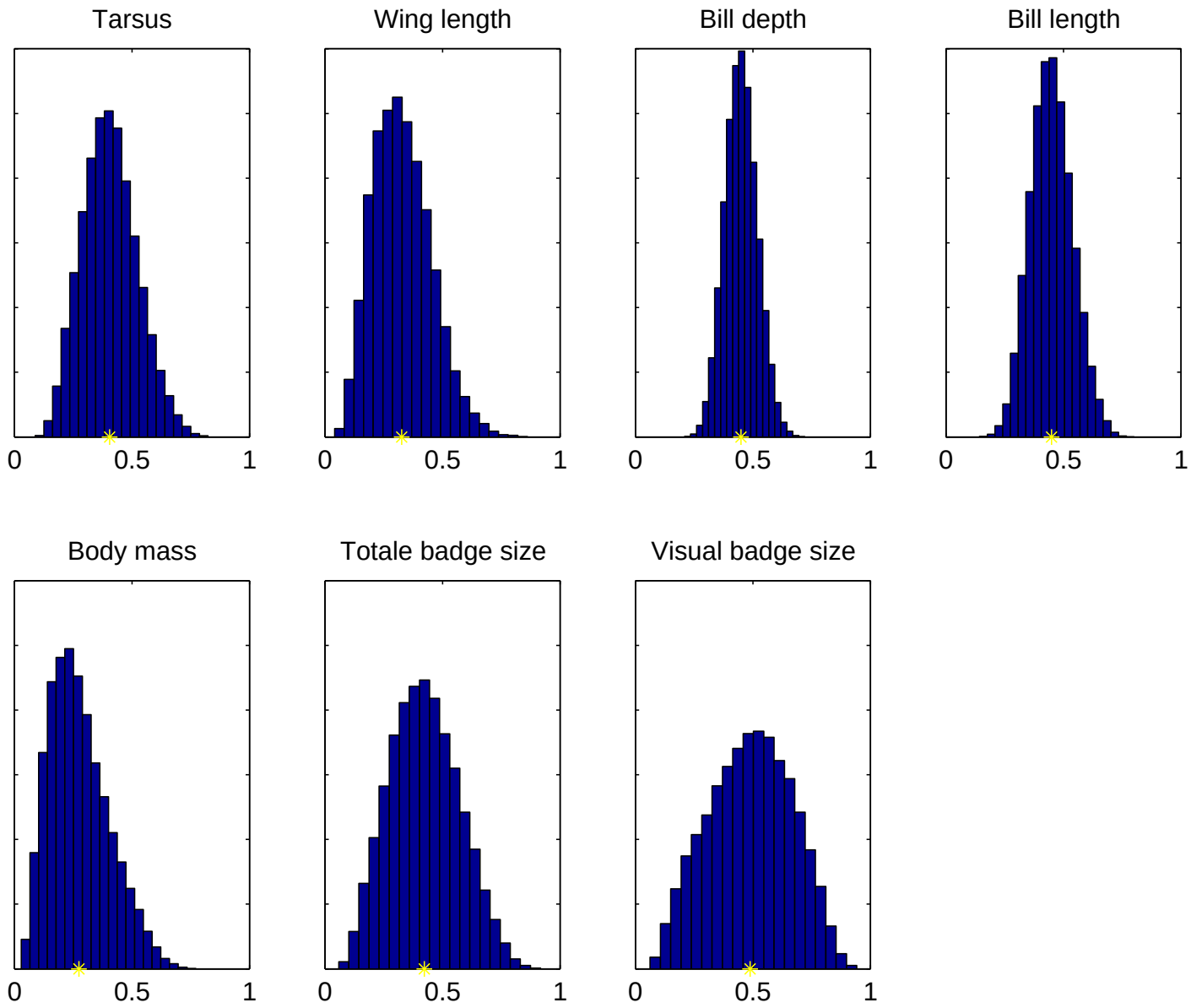
Results, MCMC

Trace-plots 225000 iterations:





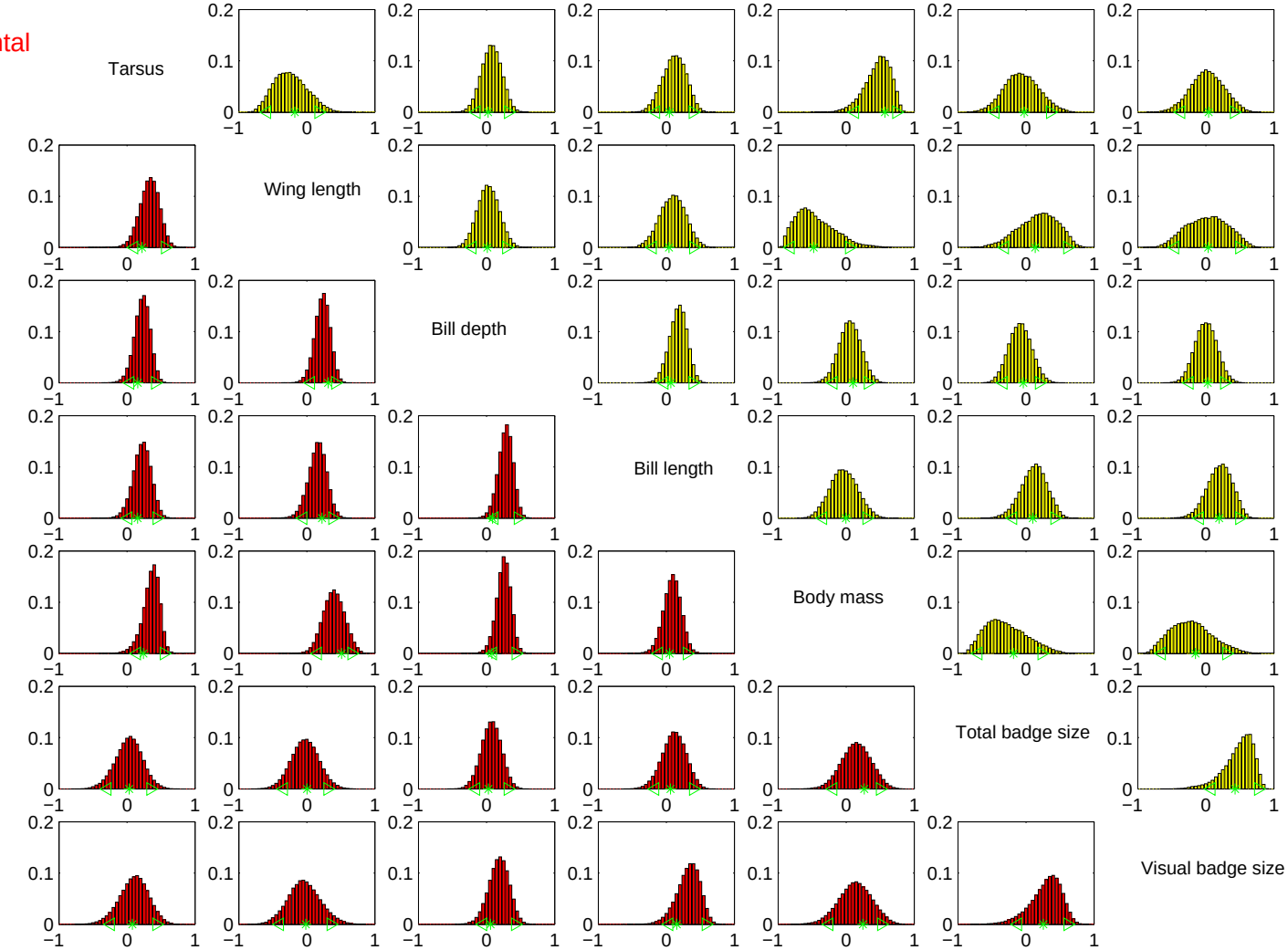
Heritability estimates



Correlation estimates

Genetic correlation

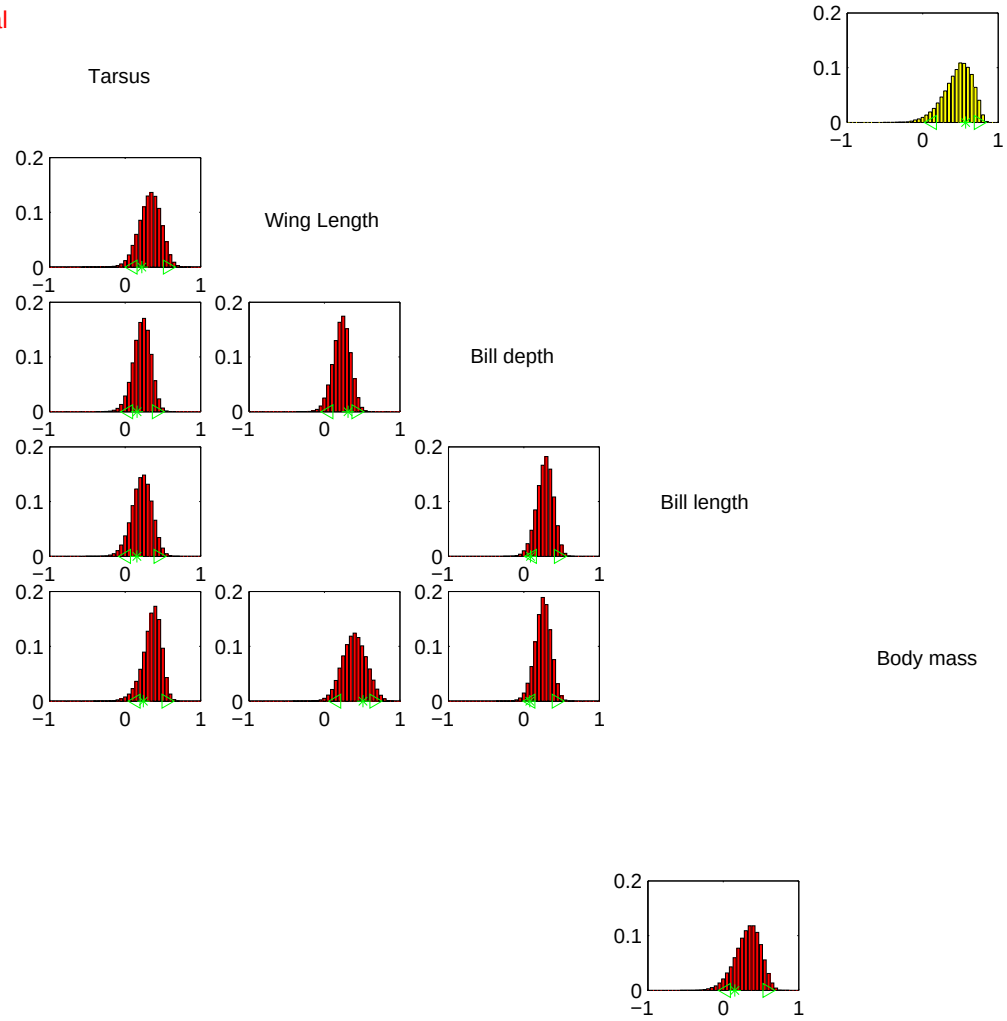
Enviromental correlation



Correlation estimates

Genetic correlation

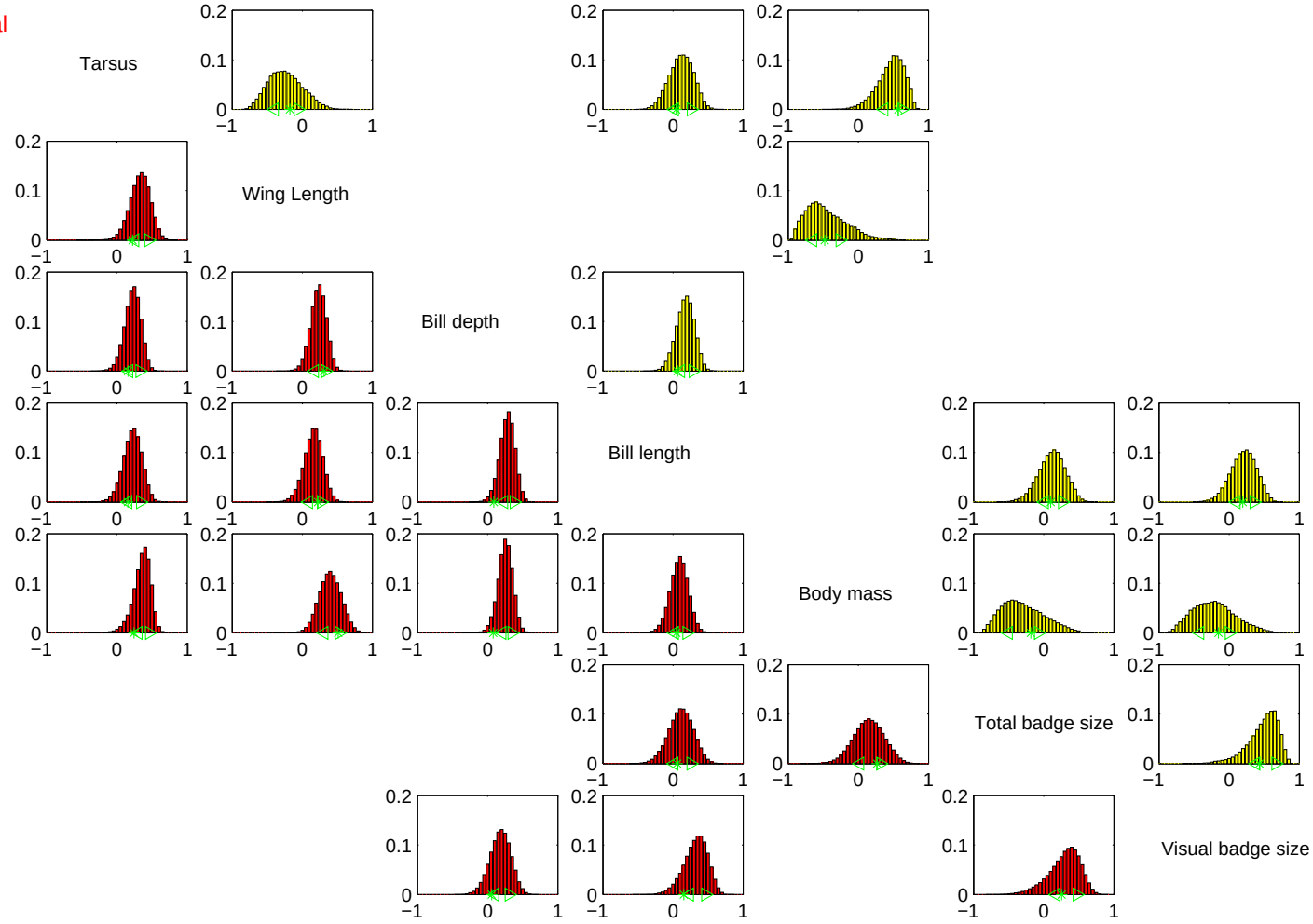
Enviromental correlation



Correlation estimates

Genetic correlation

Enviromental correlation



Summary

- Animal model of multiple traits = MGMRF model.
- Use a Gibbs sampler with two blocks
- Because of the sparse structure the pedigree impose, sampling from a MGMRF of dimension 20000 is fast.

What's new

- Statistics:
  - Sampling of MGMRF (with constraints) in one block.
- Biology:
  - Bayesian approach for animal model for wild life population.
  - Bayesian approach for animal model with multiple traits.

Further work

- Near future
  - One-block Metropolis-Hasting (Σ_u and Σ_ϵ together with u, β and y_{miss}).
  - Use data until 2004
  - Publish
- More serious extensions
  - Include life-history traits in the model.
  - Selection on breeding values?
  - Include environmental variables (e.g. weather) $\Rightarrow$ model selection.
  - Selection studies.