

Strategies for MCMC computation in quantitative genetics

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

Aim

Discuss MCMC computational strategies for complex (non-normal) models in quantitative genetics.

Special focus on high-dimensional vectors of genetic random effects.

Joint work with Noelia Ibánñez and Daniel Sorensen.

Generic set-up and notation

y: vector of observations of a trait (litter size/animal weight/...)

a: vector of genetic random effects

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

where σ_a^2 additive genetic variance and A additive genetic relationship matrix.

$f(\mathbf{y}|\mathbf{a}; \mu)$: sampling density of **y** given **a**.

Normal likelihood $L(\mu, \sigma_a^2) = f(\mathbf{y}; \mu, \sigma_a^2)$ if $f(\mathbf{y}|\mathbf{a}; \mu)$ density for linear normal model.

Normal case computationally rather straightforward (whether frequentist or Bayes).

This talk: computational strategies for Bayesian inference in the non-normal case.

Non-normal or non-linear models

Suppose $f(\mathbf{y}|\mathbf{a}; \mu)$ not normal or non-linear model involving \mathbf{a} .

Then likelihood

$$f(\mathbf{y}; \mu, \sigma_a^2) = \int f(\mathbf{y}|\mathbf{a}, \mu) p(\mathbf{a}; \sigma_a^2) d\mathbf{a}$$

not available in closed form.

Example (generalized linear mixed model): observation y_i Poisson with mean $\exp(\mu + \mathbf{z}_i^T \mathbf{a})$

Example (genetic variance heterogeneity): y_i normal with variance depending on additive genetic values (more details later).

Bayesian inference using MCMC

Introduce prior $p(\mu, \sigma_a^2)$ for unknown parameters and explore posterior

$$p(\mathbf{a}, \mu, \sigma_a^2 | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{a}; \mu) p(\mathbf{a}; \sigma_a^2) p(\mu, \sigma_a^2)$$

using MCMC sample: $(\mathbf{a}^1, \mu^1, \sigma_a^{2,1}), (\mathbf{a}^2, \mu^2, \sigma_a^{2,2}), \dots$

Metropolis-Hastings update of \mathbf{a} :

Current value: \mathbf{a}^k and proposal $\mathbf{a}^{\text{prop}} \sim q(\mathbf{a}^{\text{prop}} | \mathbf{a}^k)$

where q proposal density. With probability

$$\min \left\{ 1, \frac{p(\mathbf{a}^{\text{prop}}, \mu, \sigma^2 | \mathbf{y}) q(\mathbf{a}^k | \mathbf{a}^{\text{prop}})}{p(\mathbf{a}^k, \mu, \sigma^2 | \mathbf{y}) q(\mathbf{a}^{\text{prop}} | \mathbf{a}^k)} \right\}$$

new state $\mathbf{a}^{k+1} = \mathbf{a}^{\text{prop}}$; otherwise $\mathbf{a}^{k+1} = \mathbf{a}^k$.

Problem: efficient update of highdimensional \mathbf{a} .

Choice of proposal density q

Gibbs sampler: q conditional density of \mathbf{a} given $(\mu, \sigma_a^2, \mathbf{y})$ - only available for standard linear mixed model.

Random walk:

$$\mathbf{a}^{\text{prop}} \sim N(\mathbf{a}^k, hI)$$

- small acceptance rates when \mathbf{a} highdimensional.

Langevin-Hastings (use gradient information):

$$\mathbf{a}^{\text{prop}} \sim N(\mathbf{a}^k + h\nabla \log p(\mathbf{a}|\mathbf{y}, \mu, \sigma_a^2)/2, hI)$$

- better acceptance rates than random walk in high dimensions.

Reparametrization: apply Langevin-Hastings to transformed random effects

$$\boldsymbol{\gamma} = \sigma_a^{-1} A^{-1/2} \mathbf{a} \sim N(0, I) \text{ (a priori)}$$

Example: genetic variance heterogeneity

Genetic random effects \mathbf{a} and \mathbf{a}^* influencing mean and variance of y_i .

Sampling distribution of y_i given $(\mathbf{a}, \mathbf{a}^*)$:

$$y_i \sim N(\mu + \mathbf{z}_i^T \mathbf{a}, \exp(\mu^* + \mathbf{z}_i^T \mathbf{a}^*))$$

where

$$(\mathbf{a}, \mathbf{a}^*) \sim N(\mathbf{0}, G \otimes A)$$

with

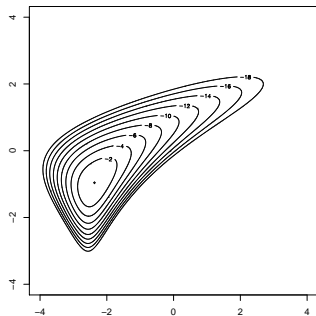
$$G = \begin{bmatrix} \sigma_a^2 & \rho\sigma_a\sigma_{a^*} \\ \rho\sigma_a\sigma_{a^*} & \sigma_{a^*}^2 \end{bmatrix}$$

Very challenging from a computational point of view.

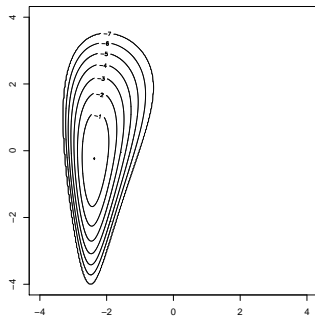
Illustration of MCMC strategies for toy example

\mathbf{a} and \mathbf{a}^* each one-dimensional (only one animal in pedigree),
simulated data $\mathbf{y} = (-2.62, -2.42)$.

Posterior of $(\mathbf{a}, \mathbf{a}^*)$

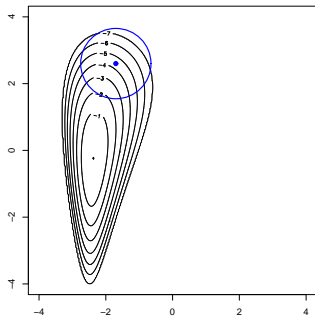


Posterior of (γ, γ^*)

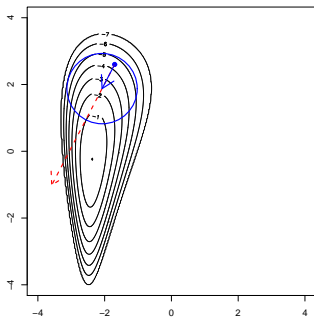


Random walk and Langevin-Hastings updates for (γ, γ^*) (blue dot is current value)

Random walk



Langevin-Hastings



Normal approximation

Idea: approximate posterior of \mathbf{a} (or \mathbf{a}, \mathbf{a}^*) using second order

Taylor expansion:

$$\log p(\mathbf{a}|\mathbf{y}) \approx \log p(\hat{\mathbf{a}}|\mathbf{y}) + (\mathbf{a} - \hat{\mathbf{a}}) \nabla \log p(\hat{\mathbf{a}}|\mathbf{y})^T - \frac{1}{2} (\mathbf{a} - \hat{\mathbf{a}}) H(\hat{\mathbf{a}}) (\mathbf{a} - \hat{\mathbf{a}})^T$$

Hence

$$\mathbf{a}^{\text{prop}} \sim N(\hat{\mathbf{a}} + \nabla \log p(\hat{\mathbf{a}}|\mathbf{y}) H(\hat{\mathbf{a}})^{-1}, H(\hat{\mathbf{a}})^{-1})$$

Normal approximation

Idea: approximate posterior of \mathbf{a} (or \mathbf{a}, \mathbf{a}^*) using second order Taylor expansion:

$$\log p(\mathbf{a}|\mathbf{y}) \approx \log p(\hat{\mathbf{a}}|\mathbf{y}) + (\mathbf{a} - \hat{\mathbf{a}}) \nabla \log p(\hat{\mathbf{a}}|\mathbf{y})^\top - \frac{1}{2} (\mathbf{a} - \hat{\mathbf{a}}) H(\hat{\mathbf{a}}) (\mathbf{a} - \hat{\mathbf{a}})^\top$$

Hence

$$\mathbf{a}^{\text{prop}} \sim N(\hat{\mathbf{a}} + \nabla \log p(\hat{\mathbf{a}}|\mathbf{y}) H(\hat{\mathbf{a}})^{-1}, H(\hat{\mathbf{a}})^{-1})$$

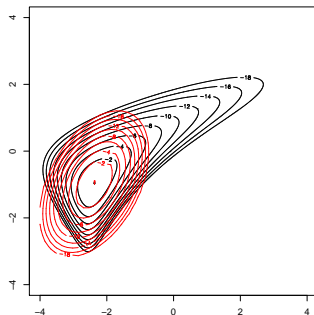
Possibilities for $\hat{\mathbf{a}}$:

- ▶ current value $\hat{\mathbf{a}} = \mathbf{a}^k$
- ▶ $\hat{\mathbf{a}}$: one-step Newton-Raphson from current value
- ▶ $\hat{\mathbf{a}}$ mode of $p(\mathbf{a}|\mathbf{y})$:

$$\mathbf{a}^{\text{prop}} \sim N(\hat{\mathbf{a}}, H(\hat{\mathbf{a}})^{-1})$$

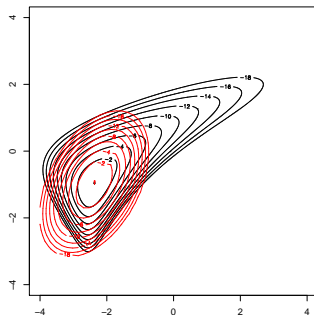
Normal approximation for toy example

Posterior and Normal
Approximation for $(\mathbf{a}, \mathbf{a}^*)$

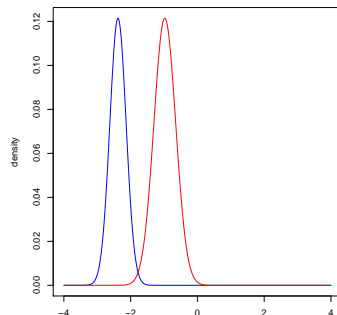


Normal approximation for toy example

Posterior and Normal Approximation for $(\mathbf{a}, \mathbf{a}^*)$



Conditional densities of $\mathbf{a}|\mathbf{a}^*, \mathbf{y}$ and $\mathbf{a}^*|\mathbf{a}, \mathbf{y}$ at mode



Use normal approximation for \mathbf{a} and \mathbf{a}^* separately (conditional distribution of \mathbf{a} given $(\mathbf{a}^*, \mathbf{y})$ exactly normal).

Rabbits case study

Data: ten generation divergent selection study for rabbit uterine capacity. Number of observations 2996 and 1161 animals in pedigree. Various fixed effects and permanent random effects.

Bayesian inference for model with genetically structured variance heterogeneity using either Langevin-Hastings (LH) or Normal approximation (NX) for $(\mathbf{a}, \mathbf{a}^*)$.

MCMC sample sizes needed to match precision of Monte Carlo estimate with sample of 100 independent draws from posterior

	LH	NX (current)	NX (one-step)	NX (mode)
σ_a^2	158800	8400	8700	7700
$\sigma_{a^*}^2$	342000	12900	7700	7200
ρ	1090000	8800	7800	6700

Up to 100 times larger samples needed with LH (more correlated samples than for NX). However, depending on implementation NX may be between 6-20 times slower pr. iteration than LH.

Simulated data - varying ρ

Estimated posterior mean for ρ : -0.74.

Langevin-Hastings and Normal Approximation applied to simulated data with varying true value of ρ : -0.74, -0.3, 0, 0.3, 0.74

MCMC sample size required to match 100 independent draws:

		-0.74	-0.3	0	0.3	0.74
σ_a^2	NX	9300	5800	2900	4800	10100
	LH	182600	16900	2600	35100	95500
$\sigma_{a^*}^2$	NX	8600	3500	3800	4000	4800
	LH	6300	1300	820	1300	7000
ρ	NX	53200	5600	3000	5900	29200
	LH	131400	15000	6300	78300	134200

Up to 20 times longer sample size needed for Langevin-Hastings compared with normal approximation. However, normal approximation much slower so no clear winner.

Sampling from normal approximation

Normal approximation $N(\hat{\mathbf{a}}, H(\mathbf{a})^{-1})$ formally equivalent to conditional distribution of \mathbf{a} given $\tilde{\mathbf{y}} = Z\mathbf{a} + \tilde{\epsilon}$ for 'virtual' data $\tilde{\mathbf{y}}$.

Use García-Cortés & Sorensen algorithm based on

$$\mathbf{a} = (\mathbf{a} - E[\mathbf{a}|\tilde{\mathbf{y}}]) + E[\mathbf{a}|\tilde{\mathbf{y}}] = R + \hat{\mathbf{a}}$$

where 'prediction error' $R = (\mathbf{a} - E[\mathbf{a}|\tilde{\mathbf{y}}])$ and $\hat{\mathbf{a}} = E[\mathbf{a}|\tilde{\mathbf{y}}]$, $\tilde{\mathbf{y}}$ independent.

Hence if R_{sim} is a simulation of R then

$$\mathbf{a}_{\text{sim}} = R_{\text{sim}} + \hat{\mathbf{a}}$$

is a conditional simulation of \mathbf{a} given $\tilde{\mathbf{y}}$.

Generation of R_{sim} :

1. simulate $(\mathbf{a}_{\text{sim}}, \tilde{\mathbf{y}}_{\text{sim}})$ from joint distribution of $(\mathbf{a}, \tilde{\mathbf{y}})$ (use Henderson factorization $A = TDT^T$)
2. compute $\hat{\mathbf{a}}_{\text{sim}} = E[\mathbf{a}|\tilde{\mathbf{y}}_{\text{sim}}]$ (mixed model equations)
3. return $R_{\text{sim}} = \mathbf{a}_{\text{sim}} - \hat{\mathbf{a}}_{\text{sim}}$.

Sparse matrix methods

Use general sparse matrix Cholesky decomposition for hessian $H(\hat{\mathbf{a}})$ in normal approximation $N(\hat{\mathbf{a}}, H(\mathbf{a})^{-1})$.

GMRFLib (H. Rue): general software in c for MCMC computation in models with sparse precision matrix for random effects. E.g. routines for computing updates using normal approximation.

Lots of useful tricks and advice in book Rue & Knorr-Held (2005).






Summary

- ▶ updates based on normal approximation may reduce correlation in MCMC samples.
- ▶ advantage may partly be cancelled due to extra computational cost
- ▶ no definite recommendation - depends on application - experimenting required (GMRFLib helpful)

Further possibilities

- ▶ joint update of \mathbf{a} and σ_a^2 .
- ▶ joint update of permanent and genetic random effects.

References

-  Ibáñez, N., Sorensen, D., Waagepetersen, R. & Blasco, A. (2006). A study of canalization and response to selection for uterine capacity in rabbits. Submitted.
-  Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society B* **63**, 325–338.
-  Rue, H. & Knorr-Held, L. (2005). *Gaussian Markov random fields - theory and applications*. Chapman & Hall/CRC.
-  Sorensen, D. & Waagepetersen, R. (2003). Normal linear models with genetically structured variance heterogeneity: a case study. *Genetical Research* **82**, 207–222.
-  Steinsland, I. & Jensen, H. (2005). Making inference from Bayesian animal models utilising Gaussian Markov random field properties. *Statistics Preprint 10*, Norwegian University of Science and Technology.