universitätbonn

Institut für **N**utzpflanzenwissenschaften und
Ressourcenschutz
Professur für Pflanzenzüchtung

CTGACTCGTATATATACAGACG
Chair of Plant Breeding

Bayesian Adaptive Markov Chain Monte Carlo (MCMC) estimation of genetic parameters

Chair of Plant Breeding, University of Bonn, Germany

Boby Mathew

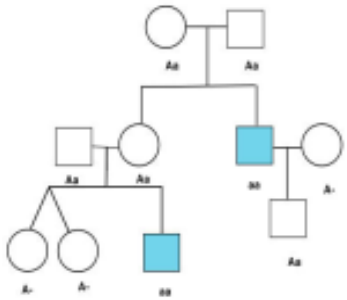
Workshop on “Animal models with applications in ecology”, University of Oulu, Finland

06.11.2014

Structure of the presentation

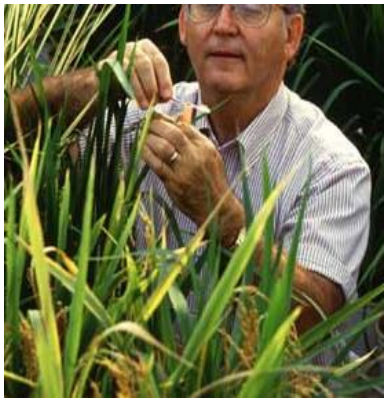
Chair of Plant Breeding

- Introduction
- Theory and Methods
- Results
- Conclusion



Introduction

- The goal of plant/animal breeding is to develop new varieties with desirable characteristics.



- The selection of plants/animals by looking at the performance record is always difficult.

- One solution is to look at the total **Phenotypic variation (Vp)**

Phenotypic variation

The total **Phenotypic variance (Vp)** can be expressed as:

$$V_P = \underbrace{V_A + V_D}_{V_G} + V_E$$

Additive genetic variance: measures the genetic variation associated with the average effects of substituting one allele for another at a given locus.

Dominance variance: is due to the interaction between alleles at the same locus .

Questions asked by a breeder…

How much of the observed phenotypic variation is due to genetic vs. environmental factors?

How much of the genetic variation is due to breeding value (additive genetic variance)?

Are there any genotype by environment interactions?

Molecular biology techniques with

statistical methods can help to answer some of these questions

Breeding Value (BV)

Once the total amount of genetic variation responsible for a trait is obtained, then the heritability **(h^2)** of trait can be calculated as:

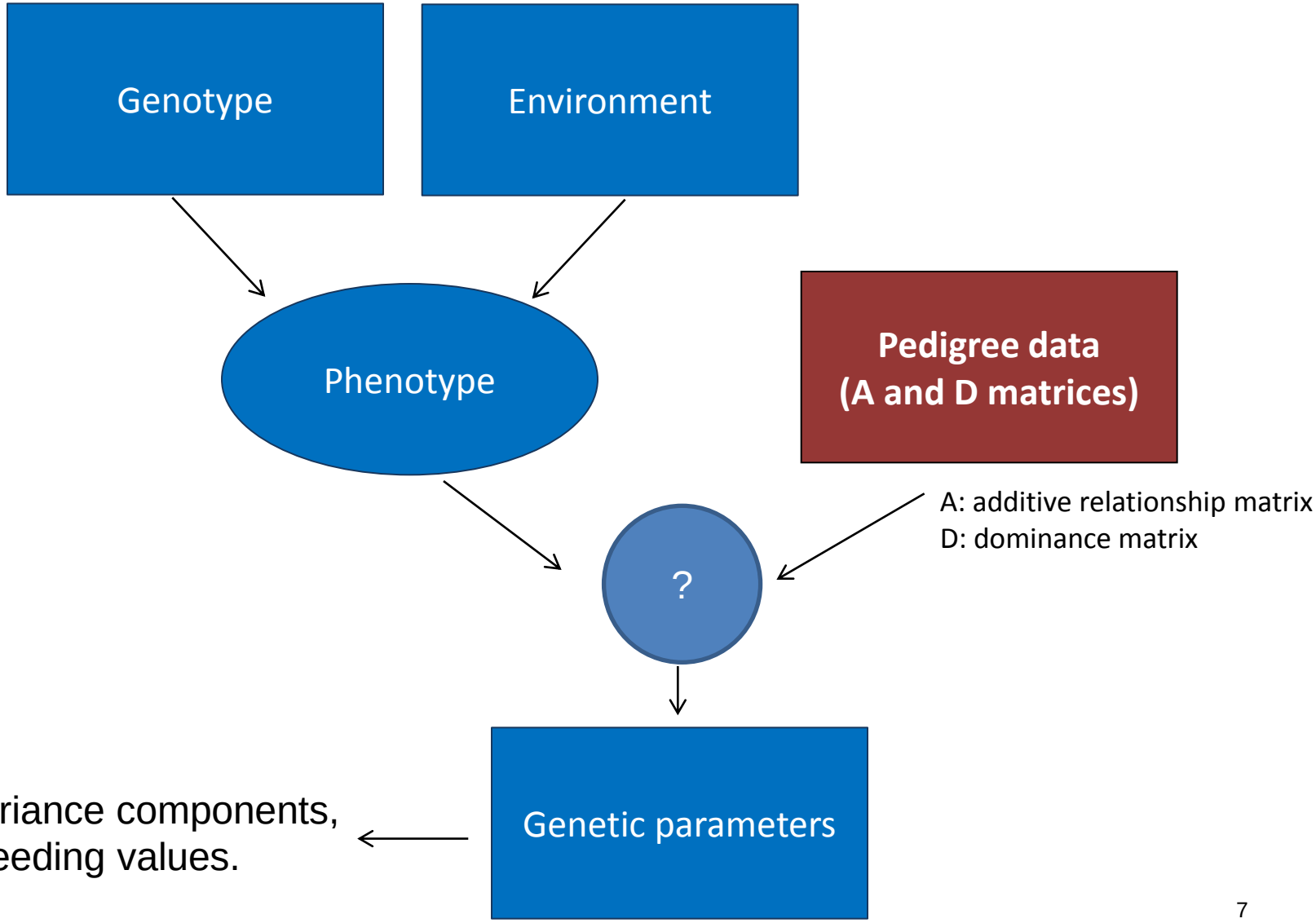
$$h^2 = V_G/V_P$$

Breeding value of a line = $h^2 * (y_i - \mu_{population})$

V_G=genetic variance, V_P=phenotypic variance, y_i=phenotypic value of the i^{th} line

Estimated breeding values can help to pick up the best individuals in a population.

Estimation of genetic parameters



Methods for variance component estimation

1. Maximum Likelihood (ML) and Restricted Maximum Likelihood(REML)

- *e.g.* ASReml package

2. Bayesian methods

- MCMC based method: *e.g.* MCMCglmm package
- Integrated Nested Laplace Approximation: *e.g.* R-inla package

Bayesian method are able to incorporate the prior information

Hybrid Gibbs sampler

- **Blocked Gibbs sampler (Garcia-Cortés and Sorensen, 1996):**

  Slow but Fast mixing of the chain. Also the blocked has faster convergence and better mixing when the parameters are correlated (eg, family relations) .

- **Single-site Gibbs sampler (Sorensen and Gianola, 2002):**

  Fast but slow mixing of the chain.

- **Hybrid Gibbs sampler (Waldmann *et al.* 2008) :**

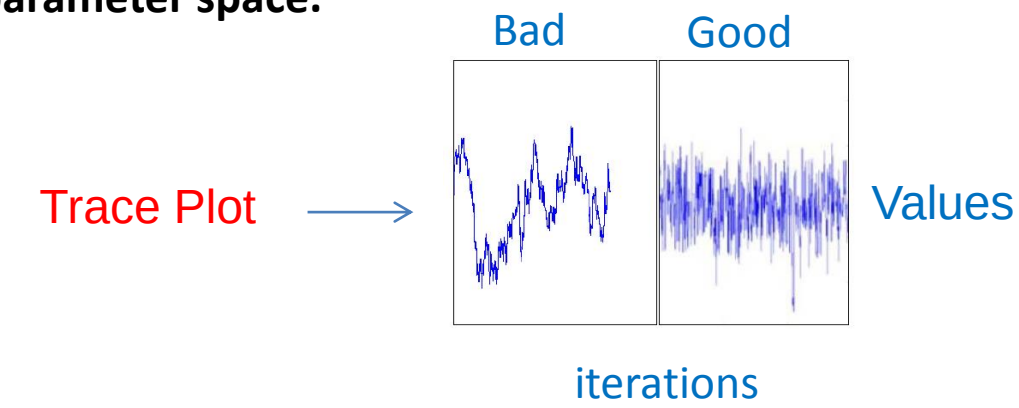
  Is a combination of both single-site and block Gibbs sampler with block update every 50th iteration.

Prerequisites of a good MCMC sampler

Some of the desired characteristics of a good Markov chain Monte Carlo (MCMC) sampling algorithm are:

- **Fast convergence: It quickly converges to its stationary distribution (posterior distribution).**
- **Good mixing properties:**

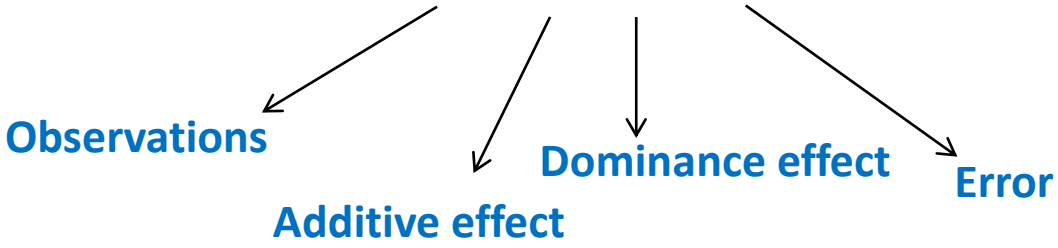
  A well mixing chain will be able to explore the entire parameter space.



- **And the total Computation time.**

Basic model and simulation

Model: $Y = a + d + e$



$$a \sim MVN\left(0, A\sigma_a^2\right) \quad d \sim MVN\left(0, D\sigma_d^2\right) \quad e \sim MVN\left(0, I\sigma_e^2\right)$$

A is the additive relationship matrix,

D is the dominance relationship matrix,

I is the identity matrix.

Cholesky decomposition of the matrices A and D were used to draw samples from the above distributions.

Initial analysis

We analyzed the simulated dataset with 3175 lines with the hybrid Gibbs sampler which uses the single-site update for the variance components.

Looked at the trace plots and fond mixing of the MCMC chain was bad, also signs there exists multiple modes.

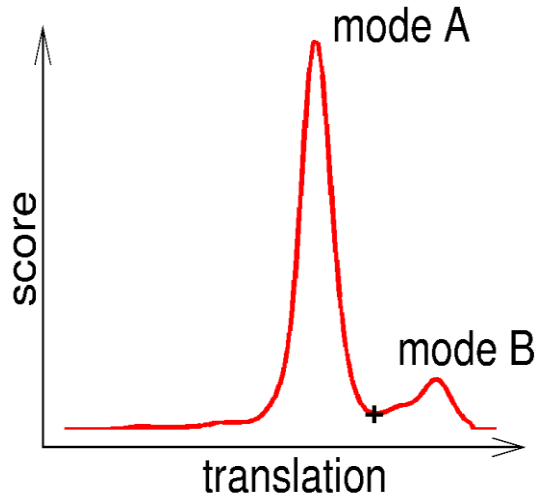
We looked at the correlation and found that there was posterior correlation between dominance and error variance components.

We decided to used block-update for the variance components.

Multimodal distributions

Multimodal distributions are probability distribution with different modes.

The existing estimation methods may have difficulties in "jumping" between different modes in the posterior distribution.



In Bayesian analysis multimodal distributions can lead to slow convergence and bad mixing

Datasets

1.Simulated datasets with no inbreeding.

- ***Bimodal data:*** With dominance variance

  (a total of 3175 lines with σ_a^2 = 800, σ_d^2 = 600, σ_e^2 = 3025)

- ***Unimodal data:*** With dominance variance

  (a total of 3640 lines with σ_a^2 = 800, σ_d^2 = 600, σ_e^2 = 3025)

- ***QTLMAS XII workshop data:*** With zero dominance (a total of 4665 lines)

2. Field dataset

82 spring barley (*H. vulgare* L.) lines from Dikopshof (NRW, Germany) for three different years: 2001, 2002, and 2003. Each year was considered a different environment.

Also inbreeding was accounted in the analysis.

Drawbacks of the existing methods

REML (Restricted Maximum Likelihood)

- Fast and computationally less complex **but** fails to detect different modes.

Bayesian analysis via Gibbs sampler

- Gibbs sampling can give the whole posterior distribution of the parameter of interest whereas REML provide the point estimates.
- **BUT** may have difficulties in detecting multimodality.
- Single-site updates for the variance components.
- Slow convergence (stationary distribution) of the algorithm.

Objective of the new proposed algorithm

- Improve the mixing and convergence of the Markov chain.
- Identify different modes if they are present in the posterior distribution.
- Improve the estimation accuracy.
- Improve the total computation time for the Bayesian analysis.
- Use block-update for the variance components.

Outline of the new algorithm

Learning Phase

Step 1

Use hybrid Gibbs sampler to learn the covariance structure of the variance components

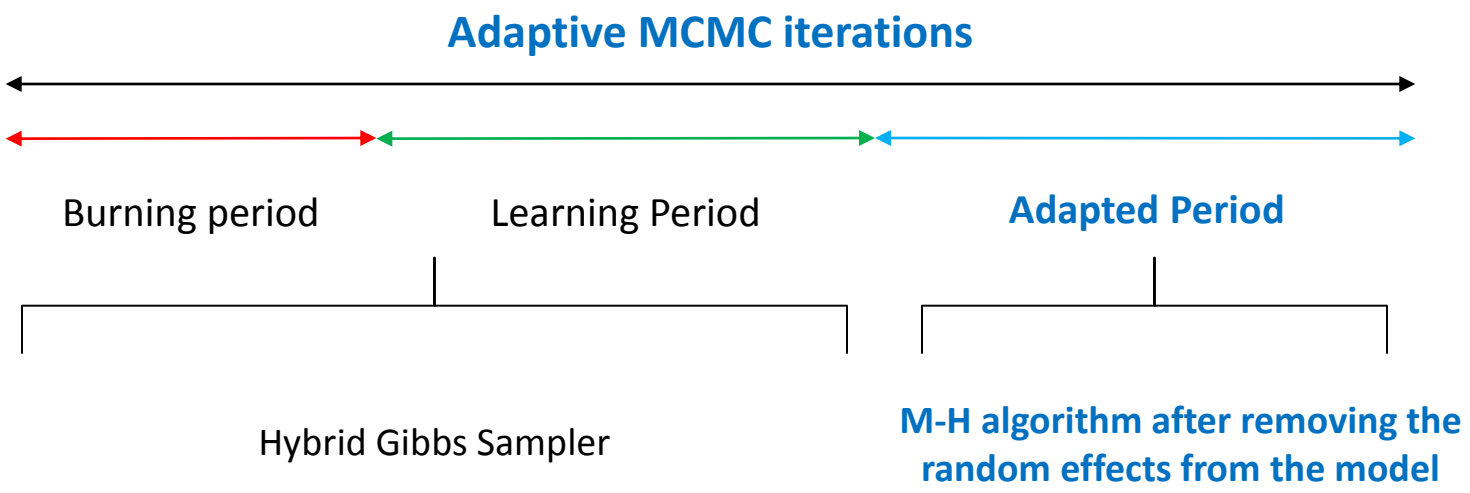
Adaptive Phase

Step 2 and 3

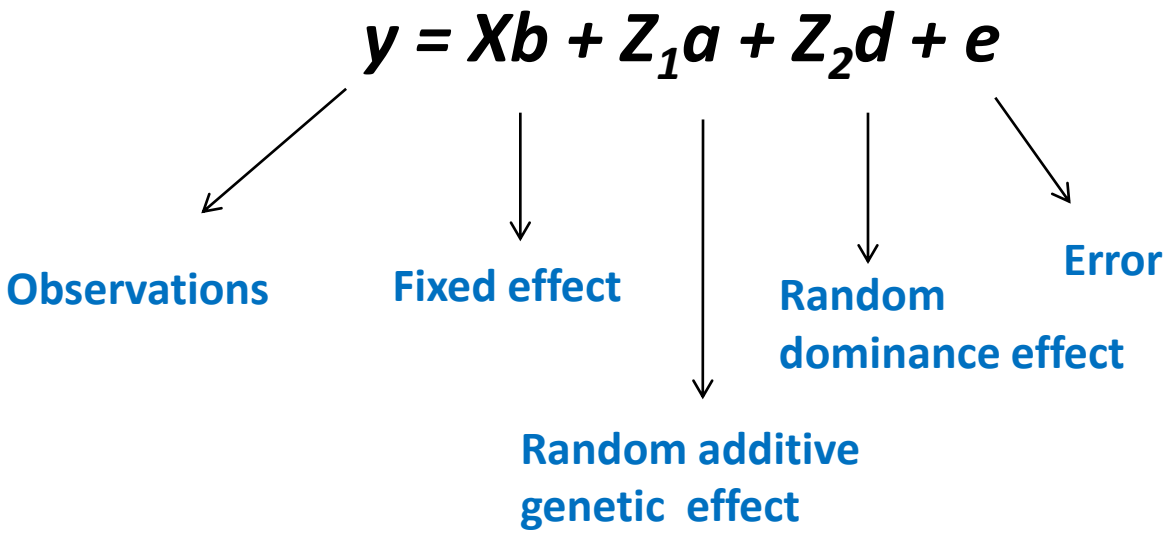
- Propose new variance components based on the learned covariance structure from a multivariate normal distribution.
- Integrate out the random effects from the likelihood and accept/ reject the new proposed values based on the Metropolis Hastings (M-H) ratio.

Adaptive MCMC approach

We proposed **adaptive MCMC** algorithm which combines Gibbs sampler and Metropolis–Hastings (M-H) algorithm.

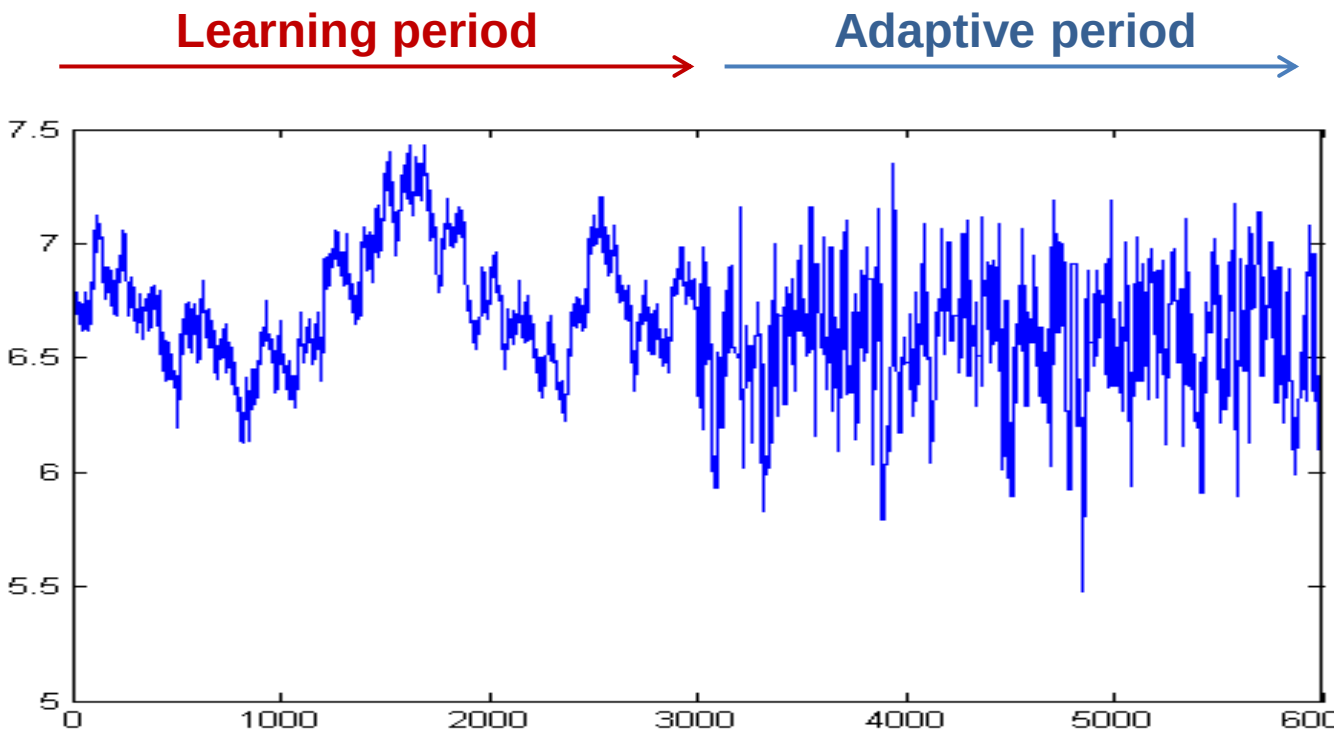


Linear mixed model for simulated data



X, Z_1 and Z_2 are design matrices, which relate records to fixed and random effects, respectively.

Trace Plots from learning and adaptive phases



Mathew et al. 2012 Heredity

Trace plot shows that MCMC chain **mixes more rapidly** in the adaptive period

Effective Sample Size (ESS)

Effective Sample Size (ESS) is a popular measure of how good the mixing of the MCMC chain.

	ESS for different variance components in Learning Phase			ESS for different variance components in Adapted Phase		
	σ^2_a	σ^2_d	σ^2_e	σ^2_a	σ^2_d	σ^2_e
Unimodal data	103. 6	8.76	29.56	177	318	192
Bimodal data	12.8	12.01	11.27	241	160	176
Workshop data	41.2	3.58	11.62	243	93.4	205

ESS is calculated for 3000 iterations from each phase and shows higher values in the adaptive period.

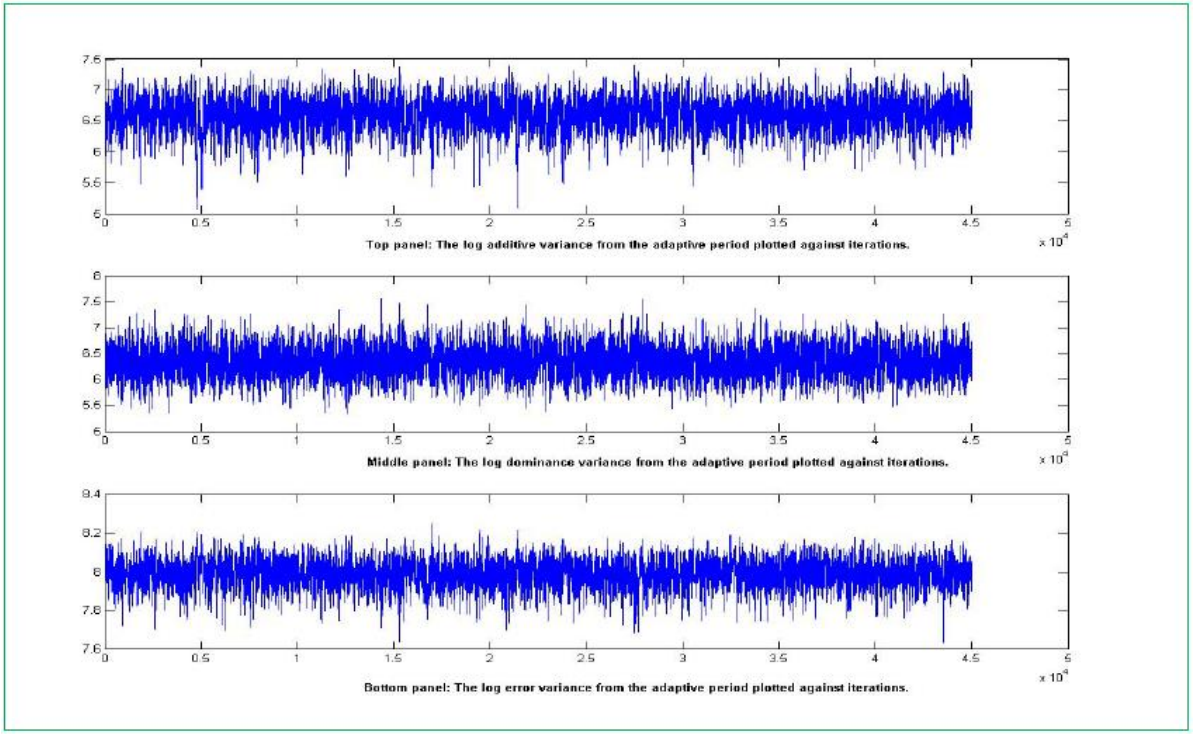
Trace plots for the unimodal data

Chair of Plant Breeding

Additive variance

Dominance variance

Error variance



Mathew et al. 2012 Heredity

Iterations

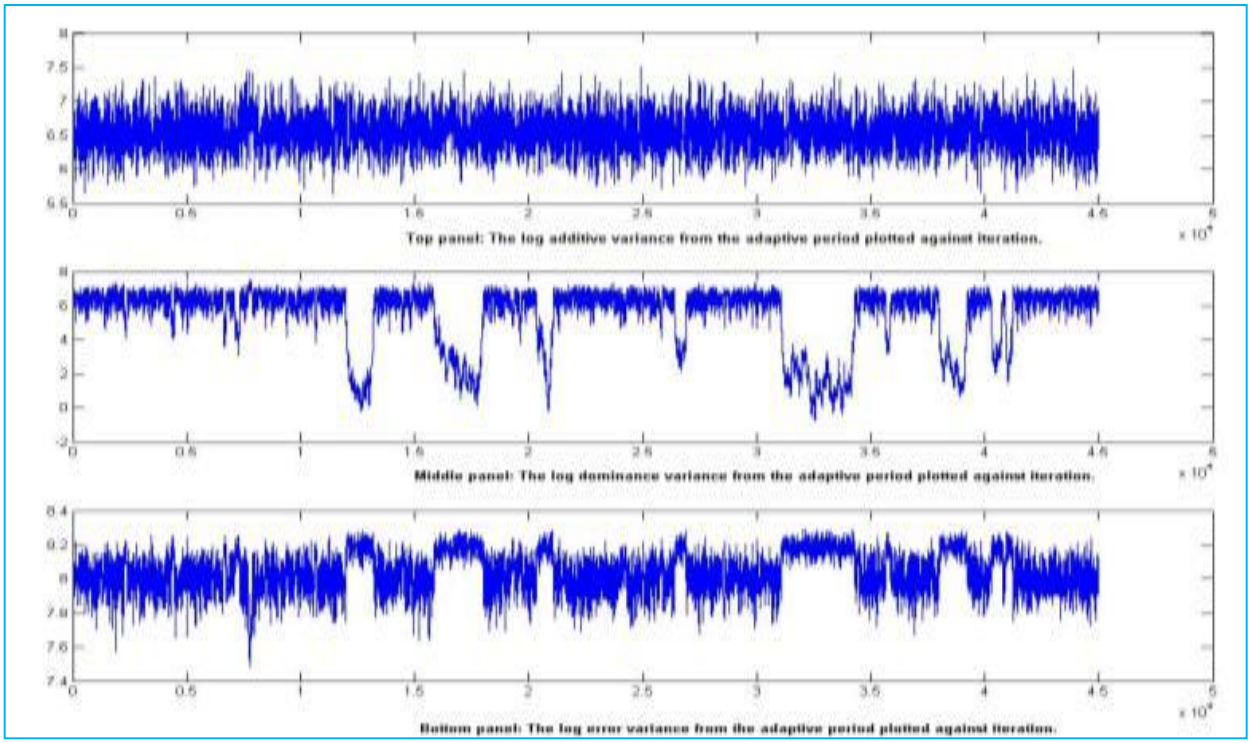
The trace plots shows that only one mode is present in the posterior distribution.

Different modes in the bimodal data

Additive variance

Dominance variance

Error variance

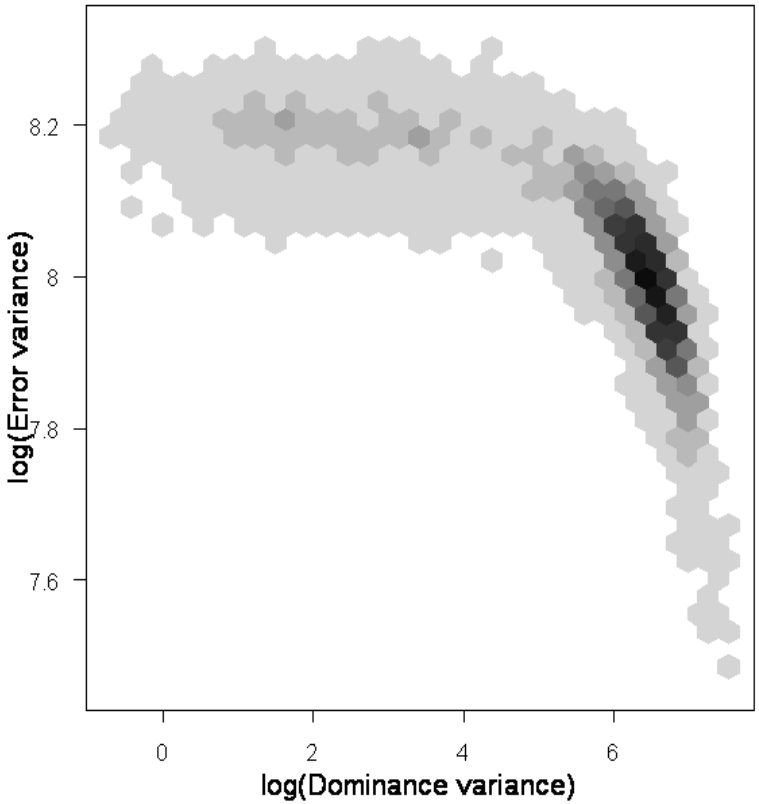


Mathew et al. 2012 Heredity

Iterations

The trace plots shows that there are two modes present in the posterior distribution.

Hexagonal bins to visualize the modes



Bivariate histogram of the log-transformed dominance and error variance components using hexagonal bins.

Comparison of estimation accuracy

Chair of Plant Breeding

Unimodal data

Heritability 0.31	σ^2_a	σ^2_d	σ^2_e	Heritability
True Values	800	600	3025	0.31(0)
REML	781 (-18)	571 (-28)	2928 (-96)	0.31(0)
Hybrid Gibbs sampler				
mode	879 (+79)	658 (+58)	2894 (-130)	0.34(+0.03)
Adaptive MCMC				
mode	779 (-20)	579 (-20)	2960 (-64)	0.31(0)

Compare to the hybrid Gibbs sampler adaptive MCMC method was able to provide values more close to the true values

Estimated variance components the workshop data with no dominance

Chair of Plant Breeding

The main motivation to use QTLMAS workshop data was to check how well our new algorithm will perform in case of no dominance.

Heritability 0.31	σ^2_a	σ^2_d	σ^2_e	**Heritability**
True Values	1.36	**0.00**	3.20	0.30
REML	1.35(-0.01)	**0.00(0.00)**	3.12(-0.08)	0.30(0)
Hybrid Gibbs sampler				
mean	1.33(-0.03)	**0.09(-0.09)**	3.06(-0.06)	0.46(+0.16)
median	1.32(-0.04)	**0.10(+0.10)**	3.06(-0.06)	0.46(+0.16)
mode	1.10(-0.26)	**0.10(+0.10)**	2.84(-0.36)	0.46(+0.16)
Adaptive MCMC				
mean	1.34(-0.02)	**0.01(+0.01)**	3.13(-0.07)	0.29(-0.01)
median	1.33(-0.03)	**0.00(0.00)**	3.13(-0.07)	0.30(0)
mode	1.31(-0.04)	**0.00(0.00)**	3.15(-0.05)	0.29(-0.01)

Mathew et al. 2012 Heredity

Adaptive MCMC method was able to provide zero dominance variance.

Prior sensitivity analysis

- We a also checked the effect of prior distribution with the **zero dominance** (QTLMAS) dataset.

- ***Gamma*** prior for the precision parameters (k_i=1 and λ_i =0.001) was able to provide good mixing with realistic estimates of dominance variance in case of zero dominance.

Correlation of breeding values

Correlation coefficient (r) calculated between the estimated breeding value and the true genetic value using REML and adaptive MCMC method.

	REML	Adaptive MCMC
QTLMAS workshop data True genetic values	0.71	0.72

Both methods gave the same correlations.

Linear mixed model for the field data

$$y = Xb + Z_1a + Z_2(gxe) + e$$

- y: Observations
- Xb: Fixed effect
- Z_1a: Random additive genetic effect
- $Z_2(gxe)$: Random genotype by environment interaction effect
- e: Error

X, Z_1 and Z_2 are design matrices, which relate records to fixed and random effects, respectively.

Estimated variance components for the field data with GXE

Field data

	σ^2_a	σ^2_{gxe}	σ^2_e	Heritability (h^2)
True Values	N/A	N/A	N/A	N/A
REML	9.21	3.18	17.08	0.75
Adaptive MCMC				
mean	9.27	2.45	17.67	0.76
median	9.13	2.49	17.58	0.76
mode	9.08	2.60	17.50	0.76

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + (\sigma_{g\times e}^2/j) + (\sigma_e^2/j * k)}$$

Conclusions

Adaptive MCMC method was able to:

- Detects multiple modes in the posterior distribution.
- Improves convergence rate and mixing of Markov chain.
- Provide estimates near to the true values.
- Accounts different covariance structure in the analysis.
- And two times faster than normal hybrid Gibbs sampler.

Acknowledgements

Chair of Plant Breeding

Prof. Dr. Jens Léon,

Dr. Andrea Bauer,

and colleagues at the Chair of Plant Breeding, University of Bonn.

Prof. Dr. Mikko J. Sillanpää, Genetics and Biometry, University of Oulu.

Late Dr. Petri Koistinen, Department of Mathematics and Statistics, University of Helsinki, Finland.

Thanks for your attention!