Application of generalized BayesA and BayesB in the analysis of genomic data

Ismo Strandén, Raphael Mrode, Donagh Berry

[1] MTT Agrifood Research Finland, Biotechnology and Food Research, Jokioinen, Finland

[2] Scottish Agricultural College, Sir Stephen Watson Building, Bush, Penicuik, EH260PH, UK

[3] Moorepark Dairy Production Research Center, Fermoy, Co. Cork, Ireland

Introduction

- Model: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}, \quad \mathbf{e} \sim N\ \mathbf{0}, \mathbf{I}\sigma_e^2$

- BayesA & BayesB (Meuwissen et al. 2001)
  - Gaussian marker effects with individual variance

$$g_j \sim N\ 0, s_j^2\ , \quad s_j^2 \sim {\nu\sigma_t^2}\Big/{\chi_\nu^2}, \quad j = 1, \ldots, m$$

- BayesA/B is **Student-t** (Gianola et al. 2009):
  - Prior densities of the marker effects are Student-t with known degrees of freedom and dispersion:

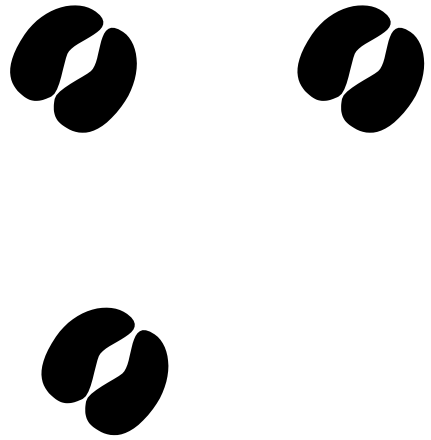
$$g_j \sim t_\nu\ 0, \sigma_t^2\ , \quad j = 1, \ldots, m$$

Objectives of this study

- Generalize BayesA/B with a parametrization similar to standard Gaussian model with unknown variances
  - Degrees of freedom and marker dispersion unknown

$$g_j \sim t_\nu \ \ 0, \sigma_t^2 \ \ , \quad j = 1, ..., m$$

- Analyze Irish Holstein dairy cattle data with BayesA/B and the generalized alternatives



Material and methods

Generalized BayesA (GtA)

Model: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}, \quad \mathbf{e} \sim N\ \mathbf{0}, \mathbf{R}\sigma_e^2$

Priors: $p\ \mu\ \propto 1$

$g_j \sim t_\nu\ 0, \sigma_t^2\ , \quad j = 1, ..., m$

$\sigma_t^2 \sim \tau T \Big/ \chi_\tau^2$ Scaled inverse chi-square

$\nu \sim \text{Gamma}\ \alpha, \beta$

Equivalent Gaussian mixture for the markers:

$g_j \sim N\ 0, \sigma_t^2 s_j^2\ , \quad s_j^2 \sim \nu \Big/ \chi_\nu^2$

GtA vs. BayesA

GtA: $g_j \sim N\ 0, \sigma_t^2 s_j^2\ ,\quad s_j^2 \sim \nu / \chi_\nu^2 \qquad \nu \sim \text{Gamma}\ \alpha, \beta$

$$\sigma_t^2 \sim \tau T / \chi_\tau^2$$

BayesA: $g_j \sim N\ 0, s_j^2\ ,\quad s_j^2 \sim \nu\sigma_t^2 / \chi_\nu^2 ,\quad j = 1, ..., m$

GtA vs. BayesA

GtA: $g_j \sim N\ 0, \sigma_t^2 s_j^2\ ,\quad s_j^2 \sim \nu / \chi_\nu^2 \qquad \nu \sim \text{Gamma}\ \alpha, \beta$

$$\sigma_t^2 \sim \tau T / \chi_\tau^2$$

BayesA: $g_j \sim N\ 0, s_j^2\ ,\quad s_j^2 \sim \nu\sigma_t^2 / \chi_\nu^2, \quad j = 1, ..., m$

$$\text{Var}\ g_j \mid \nu, \sigma_t^2\ = \frac{\nu}{\nu - 2}\sigma_t^2, \quad \nu > 2$$

The common variance Gaussian (Gc) or Bayesian ridge regression model:

$$g_j \sim N\ 0, \sigma_g^2 \qquad \sigma_g^2 \sim \tau T / \chi_\tau^2 \qquad \text{Var}\ g_j \mid \sigma_g^2\ = \sigma_g^2$$

Generalized BayesB: GtB

Model: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}, \quad \mathbf{e} \sim N\ \mathbf{0}, \mathbf{R}\sigma_e^2$

GtB:

$$g_j \sim \begin{cases} 0 & \text{with probability } \pi \\ t_\nu\ 0, \sigma_t^2 & \text{with probability } 1\text{-}\pi \end{cases}$$

$\sigma_t^2 \sim \tau T / \chi_\tau^2$ Scaled inverse chi-square

$\nu \sim \text{Gamma}\ \alpha, \beta$

BayesB:

$$g_j \sim \begin{cases} 0 & \text{with probability } \pi \\ t_\nu\ 0, \sigma_t^2 & \text{with probability } 1\text{-}\pi \end{cases}$$

Data

- Same data as presented in WCGALP 2010
- 1009 genotyped bulls with 41,739 SNP's
  - bulls genotyped using the Illumina Bovine SNP50 BeadChip (Illumina, San Diego, CA)
- Phenotypic data deregressed PTAs (DRP) of milk, fat, and protein
  - Training data: 755 bulls, Validation: 254 bulls

Marker effect prior densities

	Markers	Dispersion	Degrees of freedom
Gc	Gaussian	Estimated	N/A
BayesA	Student-t	Known	4.01
GtA(4.01)	Student-t	Estimated	4.01
GtA	Student-t	Estimated	Estimated
BayesB	Student-t	Known	4.01
GtB(4.01)	Student-t	Estimated	4.01
GtB	Student-t	Estimated	Estimated

BayesB & GtB: $\pi = \frac{2}{3}$

All dispersion priors: $p(\sigma^2) \propto 1$

Prior density for d.f.: $\nu \sim \text{Gamma}(\alpha = 2, \beta = 0.02)$, $\nu > 4$

Prior values for BayesA/BayesB

$$\hat{\sigma}_g^2 = \frac{\hat{\sigma}_s^2}{(1-\pi)\sum_{j=1}^{m} 2p_j\ 1-p_j}$$

BayesA: $\pi = 0$; BayesB: $\pi = 2/3$

$$g_j \sim t_\nu\ 0, \sigma_t^2 \quad \Rightarrow\ Var\ g_j\ = \frac{\nu}{\nu-2}\sigma_t^2 \Rightarrow \hat{\sigma}_t^2 = \frac{\nu-2}{\nu}\hat{\sigma}_g^2$$

BayesA parameters	**Milk**	**Fat**	**Protein**
Sire variance, σ_s^2	**56385**	**75**	**44**
Marker variance, σ_m^2	**3.81**	**0.00507**	**0.00297**
Dispersion, σ_t^2	**1.91**	**0.00254**	**0.00149**

Analyses

- DRP: weighted analysis with w= $r^2/(1- r^2)$
- MCMC for Bayesian analysis
  - 400,000 or 800,000 iterations (5% burn-in)
  - Effective sample sizes (ESS) and trace plots
- Model comparison statistics using genomic breeding values (DGV) in the validation set:
  - Validation reliability: $R^2 = corr^2(DRP, DGV)$
  - Regression coefficient b_1 in

$$DRP = b_0 + b_1 DGV$$

Results

	Protein		Milk		Fat	
	R^2	b_1	R^2	b_1	R^2	b_1
Gc	**0.50**	0.85	0.49	0.82	0.47	**0.86**
BayesA	0.49	0.81	0.48	0.79	0.49	0.77
GtA(4.01)	**0.50**	0.85	**0.54**	0.86	0.51	0.81
GtA	**0.50**	0.85	0.49	0.82	0.51	0.81
BayesB	**0.50**	0.84	0.52	0.85	0.51	0.81
GtB(4.01)	**0.50**	**0.88**	**0.54**	**0.90**	**0.52**	**0.85**
GtB	**0.50**	**0.88**	**0.54**	**0.90**	**0.52**	0.84

Results - **milk**

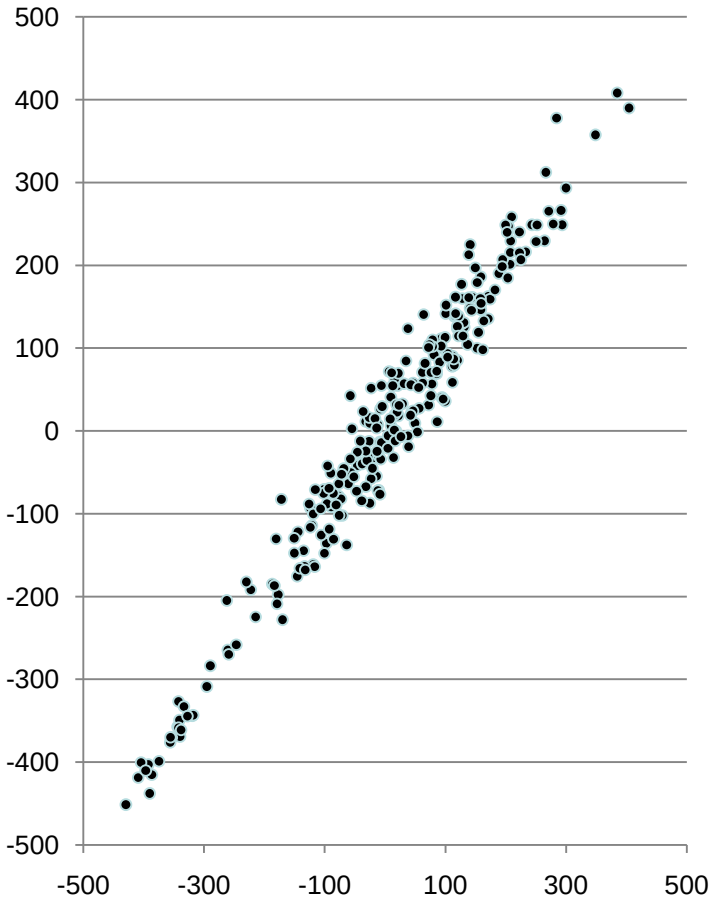
	Milk	
	R^2	b_1
Gc	0.49	0.82
BayesA	0.48	0.79
GtA(4.01)	**0.54**	0.86
GtA	0.49	0.82
BayesB	0.52	0.85
GtB(4.01)	**0.54**	**0.90**
GtB	**0.54**	**0.90**

Gc and GtA same

All models overpredict variation in DRP but GtA/B less than BayesA/B

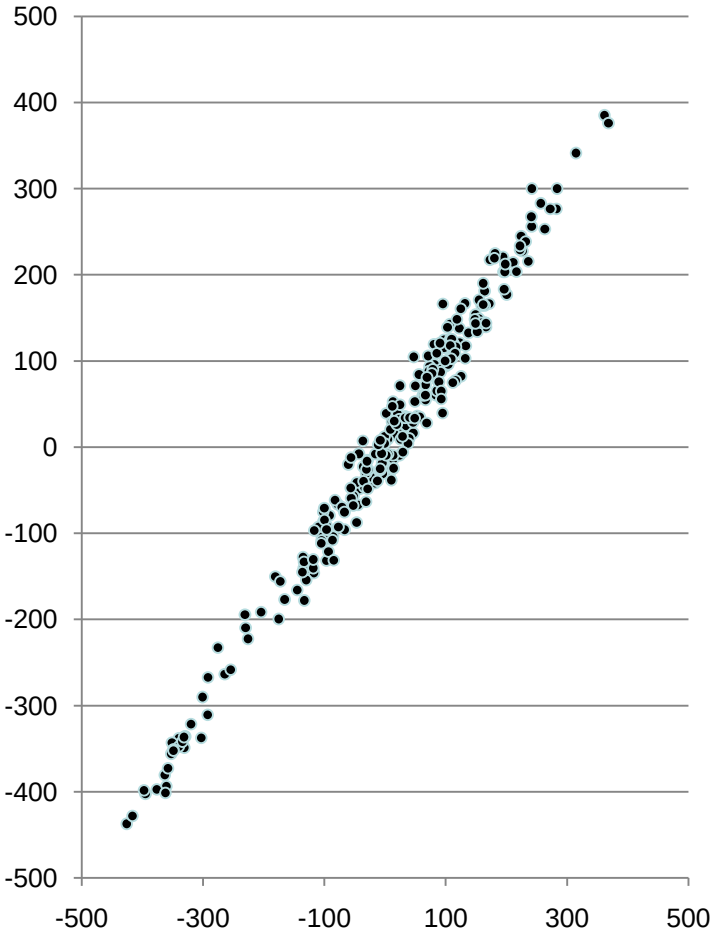
DGV solutions milk

BayesA



GtA, v=4.01

BayesB



GtB, v=4.01

Results - **fat**

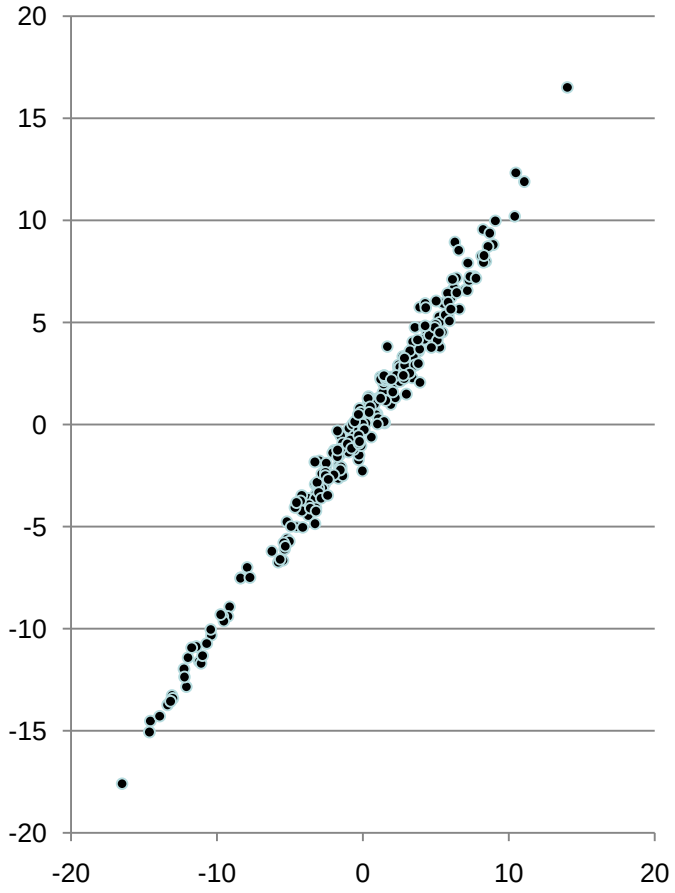
	Fat	
	R^2	b_1
Gc	0.47	**0.86**
BayesA	0.49	0.77
GtA(4.01)	0.51	0.81
GtA	0.51	0.81
BayesB	0.51	0.81
GtB(4.01)	**0.52**	**0.85**
GtB	**0.52**	0.84

GtA(4.01) and GtA same

Gc and GtB(4.01) showed less overprediction of DRP than others

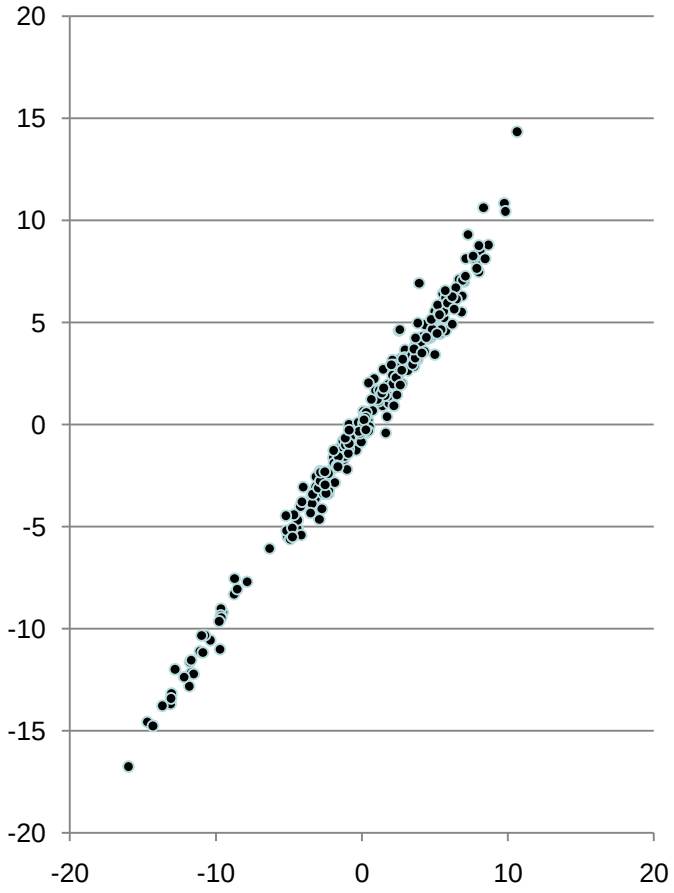
DGV solutions fat

BayesA



GtA, v=4.01

BayesB



GtB, v=4.01

Conclusions

- The generalized models (GtA, GtB)
  - outperform BayesA/B in flexibility
  - reduce to BayesA/B by highly informative priors
  - estimate more data parameters
- GtA/GtB performed well according to validation R^2
  - Higher values than by BayesA, BayesB or Gc