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: $y = 1\mu + Zg + e, \quad e \sim N(0, I\sigma_e^2)$

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

$$g_j \sim N(0, s_j^2), \quad s_j^2 \sim \frac{\nu\sigma_t^2}{\chi^2_v}, \quad j = 1, \ldots, m$$

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:

$$g_j \sim t_v(0, \sigma_t^2), \quad j = 1, \ldots, m$$
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
    \[ g_j \sim t_v, \ 0, \sigma_t^2, \ j = 1, \ldots, m \]
- Analyze Irish Holstein dairy cattle data with BayesA/B and the generalized alternatives
Material and methods
Generalized BayesA (GtA)

Model: \( y = \mathbf{1} \mu + \mathbf{Z} \mathbf{g} + \mathbf{e}, \quad \mathbf{e} \sim N(0, \mathbf{R} \sigma_e^2) \)

Priors: \( p, \mu \propto 1 \)

\( g_j \sim t_v, 0, \sigma_t^2, \quad j = 1, ..., m \)

\( \sigma_t^2 \sim \frac{\tau T}{\chi_r^2} \) Scaled inverse chi-square

\( \nu \sim \text{Gamma} \quad \alpha, \beta \)

Equivalent Gaussian mixture for the markers:

\( g_j \sim N(0, \sigma_t^2 s_j^2), \quad s_j^2 \sim \frac{\nu}{\chi_v^2} \)
GtA vs. BayesA

GtA: \[ g_j \sim N(0, \sigma_t^2 s_j^2), \quad s_j^2 \sim \frac{\nu}{\chi^2_v}, \quad \nu \sim \text{Gamma } \alpha, \beta \]
\[ \sigma_t^2 \sim \frac{T}{\chi^2_T} \]

BayesA: \[ g_j \sim N(0, s_j^2), \quad s_j^2 \sim \frac{\nu \sigma_t^2}{\chi^2_v}, \quad j = 1, \ldots, m \]
GtA vs. BayesA

GtA: \[ g_j \sim N \ 0, \sigma_t^2 s_j^2, \quad s_j^2 \sim \frac{\nu}{\chi_v^2} \quad \nu \sim \text{Gamma} \ \alpha, \beta \]
\[ \sigma_t^2 \sim \frac{\nu}{\chi_\tau^2} \]

BayesA: \[ g_j \sim N \ 0, \ s_j^2, \quad s_j^2 \sim \frac{\nu \sigma_t^2}{\chi_v^2}, \quad j = 1, \ldots, m \]

\[ \text{Var} \ g_j | \nu, \sigma_t^2 = \frac{\nu}{\nu - 2} \sigma_t^2, \quad \nu > 2 \]

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

\[ g_j \sim N \ 0, \sigma_g^2 \quad \sigma_g^2 \sim \frac{\nu}{\chi_\tau^2} \quad \text{Var} \ g_j | \sigma_g^2 = \sigma_g^2 \]
Generalized BayesB: GtB

Model: \( y = \beta + Xg + e, \quad e \sim N(0, \Sigma) \)

GtB: \[
g_j \sim \begin{cases} 
0 & \text{with probability } \pi \\
t_v \ 0, \sigma_i^2 & \text{with probability } 1-\pi 
\end{cases}
\]

\[
\sigma_i^2 \sim \frac{\tau T}{\chi^2} \quad \text{Scaled inverse chi-square}
\]

\[
\nu \sim \text{Gamma } \alpha, \beta
\]

BayesB:

\[
g_j \sim \begin{cases} 
0 & \text{with probability } \pi \\
t_v \ 0, \sigma_i^2 & \text{with probability } 1-\pi 
\end{cases}
\]
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

<table>
<thead>
<tr>
<th></th>
<th>Markers</th>
<th>Dispersion</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gc</td>
<td>Gaussian</td>
<td>Estimated</td>
<td>N/A</td>
</tr>
<tr>
<td>BayesA</td>
<td>Student-t</td>
<td>Known</td>
<td>4.01</td>
</tr>
<tr>
<td>GtA(4.01)</td>
<td>Student-t</td>
<td>Estimated</td>
<td>4.01</td>
</tr>
<tr>
<td>GtA</td>
<td>Student-t</td>
<td>Estimated</td>
<td>Estimated</td>
</tr>
<tr>
<td>BayesB</td>
<td>Student-t</td>
<td>Known</td>
<td>4.01</td>
</tr>
<tr>
<td>GtB(4.01)</td>
<td>Student-t</td>
<td>Estimated</td>
<td>4.01</td>
</tr>
<tr>
<td>GtB</td>
<td>Student-t</td>
<td>Estimated</td>
<td>Estimated</td>
</tr>
</tbody>
</table>

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

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

Prior density for d.f.: $\nu \sim \text{Gamma} \quad \alpha = 2, \beta = 0.02 \quad , \quad \nu > 4$
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: \( \pi = 2/3 \)

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

<table>
<thead>
<tr>
<th>BayesA parameters</th>
<th>Milk</th>
<th>Fat</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sire variance, ( \sigma_s^2 )</td>
<td>56385</td>
<td>75</td>
<td>44</td>
</tr>
<tr>
<td>Marker variance, ( \sigma_m^2 )</td>
<td>3.81</td>
<td>0.00507</td>
<td>0.00297</td>
</tr>
<tr>
<td>Dispersion, ( \sigma_t^2 )</td>
<td>1.91</td>
<td>0.00254</td>
<td>0.00149</td>
</tr>
</tbody>
</table>
Analyses

- **DRP:** weighted analysis with \( w = \frac{r^2}{1 - r^2} \)

- **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: \( R^2 = \text{corr}^2(\text{DRP}, \text{DGV}) \)
  - Regression coefficient \( b_1 \) in
    \[
    \text{DRP} = b_0 + b_1 \text{DGV}
    \]
## Results

<table>
<thead>
<tr>
<th></th>
<th>Protein</th>
<th></th>
<th>Milk</th>
<th></th>
<th>Fat</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R^2</td>
<td>b_1</td>
<td>R^2</td>
<td>b_1</td>
<td>R^2</td>
<td>b_1</td>
</tr>
<tr>
<td>Gc</td>
<td>0.50</td>
<td>0.85</td>
<td>0.49</td>
<td>0.82</td>
<td>0.47</td>
<td>0.86</td>
</tr>
<tr>
<td>BayesA</td>
<td>0.49</td>
<td>0.81</td>
<td>0.48</td>
<td>0.79</td>
<td>0.49</td>
<td>0.77</td>
</tr>
<tr>
<td>GtA(4.01)</td>
<td>0.50</td>
<td>0.85</td>
<td>0.54</td>
<td>0.86</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>GtA</td>
<td>0.50</td>
<td>0.85</td>
<td>0.49</td>
<td>0.82</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>BayesB</td>
<td>0.50</td>
<td>0.84</td>
<td>0.52</td>
<td>0.85</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>GtB(4.01)</td>
<td>0.50</td>
<td>0.88</td>
<td>0.54</td>
<td>0.90</td>
<td>0.52</td>
<td>0.85</td>
</tr>
<tr>
<td>GtB</td>
<td>0.50</td>
<td>0.88</td>
<td>0.54</td>
<td>0.90</td>
<td>0.52</td>
<td>0.84</td>
</tr>
</tbody>
</table>
## Results - milk

<table>
<thead>
<tr>
<th></th>
<th>Milk</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>$b_1$</td>
<td></td>
</tr>
<tr>
<td>Gc</td>
<td>0.49</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>BayesA</td>
<td>0.48</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>GtA(4.01)</td>
<td>0.54</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>GtA</td>
<td>0.49</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>BayesB</td>
<td>0.52</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>GtB(4.01)</td>
<td>0.54</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>GtB</td>
<td>0.54</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Fat</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>Gc</td>
<td>0.47</td>
<td>0.86</td>
</tr>
<tr>
<td>BayesA</td>
<td>0.49</td>
<td>0.77</td>
</tr>
<tr>
<td>GtA(4.01)</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>GtA</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>BayesB</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>GtB(4.01)</td>
<td>0.52</td>
<td>0.85</td>
</tr>
<tr>
<td>GtB</td>
<td>0.52</td>
<td>0.84</td>
</tr>
</tbody>
</table>

GtA(4.01) and GtA same

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