

Reducing redundancy using machine learning in genomic predictions from high-density SNP genotypes

O. González-Recio¹, K. Weigel², D. Gianola^{2,3}, H. Naya⁴,
G.J.M. Rosa²

¹Dpto Mejora Genética Animal, INIA, Madrid; ²Dpt Dairy Science, UW-Madison;
³Dpt Animal Science, UW-Madison; ⁴Inst. Pasteur de Montevideo, Montevideo

Session 28. EAAP 2009, Barcelona.

Outline

- 1 Motivation
- 2 Objective
- 3 Material and Methods
 - Methods
 - Simulated data
 - Real data
- 4 Remarks

MOTIVATION.

High dense SNP genotypes

- High density genotyping technologies:
 - Cattle (54K SNPs); Equine (54K SNPs); Sheep (50K SNPs); Swine (60K SNPs).
- Great interest for the international scientific community.
- Efforts in prediction of genome-enhanced breeding values.

Statistical methods

- Genomic BLUP, Bayes A, Bayes B, RKHS, Bayesian LASSO,...
- High marker density: Satisfactory results in terms of predictive ability with most of them (Meuwissen et al., 2001; Gianola et al., 2006; González-Recio et al., 2008; Hayes et al., 2009; Van Raden et al., 2009)
- Affordable individual cost, but expensive for large scale genotyping within a population.

MOTIVATION.

High dense SNP genotypes

- High density genotyping technologies:
 - Cattle (54K SNPs); Equine (54K SNPs); Sheep (50K SNPs); Swine (60K SNPs).
- Great interest for the international scientific community.
- Efforts in prediction of genome-enhanced breeding values.

Methods development

- Genomic BLUP, Bayes A, Bayes B, RKHS, Bayesian LASSO,...
 - High marker density: Satisfactory results in terms of predictive ability with most of them (Meuwissen et al., 2001; Gianola et al., 2006; González-Recio et al., 2008; Hayes et al., 2009; Van Raden et al., 2009)
- Affordable individual cost, but expensive for large scale genotyping within a population.

MOTIVATION.

High dense SNP genotypes

- High density genotyping technologies:
 - Cattle (54K SNPs); Equine (54K SNPs); Sheep (50K SNPs); Swine (60K SNPs).
- Great interest for the international scientific community.
- Efforts in prediction of genome-enhanced breeding values.

Methods development

- Genomic BLUP, Bayes A, Bayes B, RKHS, Bayesian LASSO,...
- High marker density: Satisfactory results in terms of predictive ability with most of them (Meuwissen et al., 2001; Gianola et al., 2006; González-Recio et al., 2008; Hayes et al., 2009; Van Raden et al., 2009)
- Affordable individual cost, but expensive for large scale genotyping within a population.

MOTIVATION

Reducing genotyping cost in genomic selection

- For preselecting parents of next generation. Genotyping whole population.
- Detect informative SNPs. Reducing redundancy of high dense assays.
- Increase predictive ability of methods used to predict G-BV for low marker density.
 - More differences are expected to exist between methods.

Selection of informative SNPs

Bayesian LASSO (de los Campos *et al.*, 2008; Weigel *et al.*, 2009)

Machine Learning (Long *et al.*, 2007; Gonzalez-Recio *et al.*, 2008)

- Boosting (Freund and Schapire, 1999; Friedman, 2001)
 - Useful for high dimensional regression problems doing some sort of variable selection (Bühlmann and Yu, 2003)

MOTIVATION

Reducing genotyping cost in genomic selection

- For preselecting parents of next generation. Genotyping whole population.
- Detect informative SNPs. Reducing redundancy of high dense assays.
- Increase predictive ability of methods used to predict G-BV for low marker density.
 - More differences are expected to exist between methods.

Selection of informative SNPs

Bayesian LASSO (de los Campos *et al.*, 2008; Weigel *et al.*, 2009)

Machine Learning (Long *et al.*, 2007; Gonzalez-Recio *et al.*, 2008)

- Boosting (Freund and Schapire, 1999; Friedman, 2001)
 - Useful for high dimensional regression problems doing some sort of variable selection (Bühlmann and Yu, 2003)

MOTIVATION

Reducing genotyping cost in genomic selection

- For preselecting parents of next generation. Genotyping whole population.
- Detect informative SNPs. Reducing redundancy of high dense assays.
- Increase predictive ability of methods used to predict G-BV for low marker density.
 - More differences are expected to exist between methods.

Selection of informative SNPs

Bayesian LASSO (de los Campos *et al.*, 2008; Weigel *et al.*, 2009)

Machine Learning (Long *et al.*, 2007; Gonzalez-Recio *et al.*, 2008)

- Boosting (Freund and Schapire, 1999; Friedman, 2001)
 - Useful for high dimensional regression problems doing some sort of variable selection (Bühlmann and Yu, 2003)

MOTIVATION

Reducing genotyping cost in genomic selection

- For preselecting parents of next generation. Genotyping whole population.
- Detect informative SNPs. Reducing redundancy of high dense assays.
- Increase predictive ability of methods used to predict G-BV for low marker density.
 - More differences are expected to exist between methods.

Selection of informative SNPs

Bayesian LASSO (de los Campos *et al.*, 2008; Weigel *et al.*, 2009)

Machine Learning (Long *et al.*, 2007; Gonzalez-Recio *et al.*, 2008)

- Boosting (Freund and Schapire, 1999; Friedman, 2001)
 - Useful for high dimensional regression problems doing some sort of variable selection (Bühlmann and Yu, 2003)

OBJECTIVE

TO TEST THE PERFORMANCE OF MACHINE LEARNING
ALGORITHMS (L_2 BOOSTING) TO INCREASE
PREDICTIVE ABILITY OF PRESELECTION OF SNPs
FOR LOW MARKER DENSITY

Outline

- 1 Motivation
- 2 Objective
- 3 Material and Methods
 - Methods
 - Simulated data
 - Real data
- 4 Remarks

Methods

Ensemble methods: L_2 BOOSTING (Freund and Schapire, 1999)

- Forms a “committee” of M “weak” learners or predictors, each is trained based on the performance of the previous one (AdaBoost).
- Extended to regression by Friedman (2001).
- Used for high dimension problems by Bühlmann and Yu (2003), using an L_2 Loss function, and doing some sort of covariate selection.
- May be interpreted as functional gradient descent technique.
- May be viewed as a sequence of Hilbert spaces.

Methods

Boosting

Algorithm

- 1 *Initialization.* $m = 0$. Given data, set $\mathbf{r}_m = \mathbf{y}$
- 2 Increase m by 1. Fit the “weak” learner to $\mathbf{r}_{m-1}$ using all covariates separately

$$\mathbf{r}_{m-1} = g_p(\mathbf{x}_p) + e$$

- 3 Do one-dimensional numerical search for the best predictor $f(\mathbf{x}_p)$, where

$$p = \operatorname{argmin}_p \sum_{i=1}^n (r_{(m-1)i} - g(x_{i,p}))^2$$

- 4 Set $\mathbf{r}_m = \mathbf{r}_{m-1} - f(\mathbf{x}_p)$, and repeat steps 2-4 until a stop criterion is reached (Bühlmann, 2006).

Methods

Boosting

- Yields an additive model whose terms are fitted in a stagewise fashion.

$$\mathbf{r}_{m-1} = g_p(\mathbf{x}_p) + e$$

- $g_p(\mathbf{x}_p)$ = non-parametric kernel regression (Nadaraya-Watson, 1964; Gianola et al., 2006).

$$g(\mathbf{x}) = \frac{\int yp(\mathbf{x}, y) dy}{p(\mathbf{x})}$$

with $\int yp(\mathbf{x}, y) dy = \frac{1}{nh} \sum_{i=1}^n y_i K_h(X - x_i)$, and $p(\mathbf{x}) = \frac{1}{nh} \sum_{i=1}^n K_h(X - x_i)$

Methods

Bayessian LASSO (Park and Casella, 2008)

- Conditional Laplace prior distribution on covariate estimates (λ = shrinkage parameter).

$$p(\beta|\sigma_e^2) = \prod_{j=1}^q \frac{\lambda}{2\sqrt{\sigma_e^2}} e^{(-\lambda|\beta_j|/\sigma_e)}$$

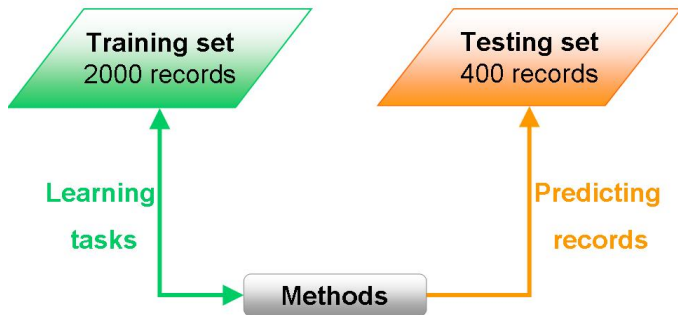
- SNPs were selected by larger absolute value estimate.

Each method offers a different bias-variance trade off.

Outline

- 1 Motivation
- 2 Objective
- 3 Material and Methods
 - Methods
 - **Simulated data**
 - Real data
- 4 Remarks

Data Design



Simulated data

- $y = \beta_1 x_1 + \beta_2 \sin(x_2) + \beta_3 \sin(x_2 \cdot x_3) + e$
- $\beta_1 = -\beta_2 = -\beta_3$
- Plus 17 noise covariates
- Two different broad sense heritability (**medium-low** and **high**) scenarios were simulated

Simulated data

Medium-low broad-sense heritability

Ranking of SNPs and MSE in the testing set

Bayesian LASSO

- SNP1 $\rightarrow$ rk1
- SNP2 $\rightarrow$ rk2
- SNP3 $\rightarrow$ rk20
- MSE in testing set: 1.28

Boosting

- SNP1 $\rightarrow$ rk2
- SNP2 $\rightarrow$ rk1
- SNP3 $\rightarrow$ rk7
- MSE in testing set: 1.15

Simulated data

Medium-low broad-sense heritability

Ranking of SNPs and MSE in the testing set

Bayesian LASSO

- SNP1 $\rightarrow$ rk1
- SNP2 $\rightarrow$ rk2
- SNP3 $\rightarrow$ rk20
- MSE in testing set: 1.28

Boosting

- SNP1 $\rightarrow$ rk2
- SNP2 $\rightarrow$ rk1
- SNP3 $\rightarrow$ rk7
- MSE in testing set: 1.15

Simulated data

High broad-sense heritability

Ranking of SNPs and MSE in the testing set

Bayesian LASSO

- SNP1 $\rightarrow$ rk1
- SNP2 $\rightarrow$ rk2
- SNP3 $\rightarrow$ rk4
- MSE in testing set: 2.31

Boosting

- SNP1 $\rightarrow$ rk1
- SNP2 $\rightarrow$ rk2
- SNP3 $\rightarrow$ rk3
- MSE in testing set: 1.19

Simulated data

High broad-sense heritability

Ranking of SNPs and MSE in the testing set

Bayesian LASSO

- SNP1 $\rightarrow$ rk1
- SNP2 $\rightarrow$ rk2
- SNP3 $\rightarrow$ rk4
- MSE in testing set: 2.31

Boosting

- SNP1 $\rightarrow$ rk1
- SNP2 $\rightarrow$ rk2
- SNP3 $\rightarrow$ rk3
- MSE in testing set: 1.19

Simulated data

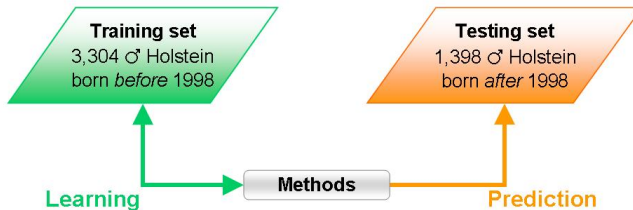
Remarks

- Boosting outperformed Bayesian LASSO in the simulations
 - Ranking of SNPs
 - Predictive ability
 - More relevant for large non-additive effects

Outline

- 1 Motivation
- 2 Objective
- 3 Material and Methods
 - Methods
 - Simulated data
 - Real data
- 4 Remarks

Real data Design



Real data

- Phenotypes ($\mathbf{y}$) = productive lifetime PTA
- Genotypes ($\mathbf{X}\beta$) = 32,611 SNPs

Provided by USDA-ARS Beltsville Agricultural Research Center

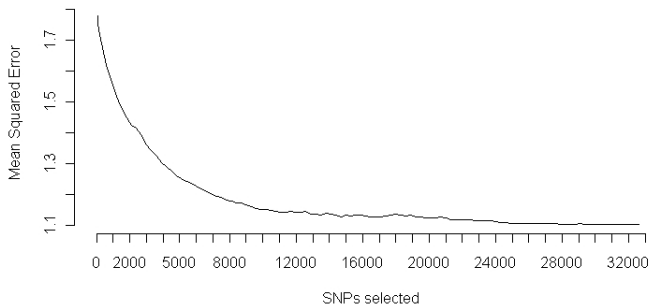


Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO.
 - MSE regarding number of SNPs selected to make predictions

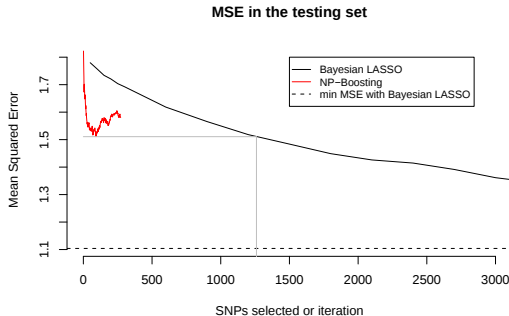
MSE in the testing set



Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO and Non-parametric Boosting.
 - MSE regarding number of SNPs selected to make predictions



optimal iteration for L_2 -Boosting = 92 (90 SNPs)

Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO showed larger predictive ability with larger number of markers.
 - Converge to the nadir with $>10,000$ SNPs.
- Boosting showed equal MSE with 90 SNPs than Bayesian LASSO with 1200 markers.
- Boosting presented a more rapid decrease on MSE with inclusion of subsequent markers.
- at equal amount of SNPs (90), Boosting reduced MSE by 14% regarding Bayesian LASSO with 90 SNPs.
- Bayesian LASSO with 32K SNPs reduced MSE by 27% regarding Boosting with 90 SNPs, but using 35,000% more markers.
- Bayesian LASSO showed some bias at small number of preselected SNPs.

Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO showed larger predictive ability with larger number of markers.
 - Converge to the nadir with $>10,000$ SNPs.
- **Boosting** showed equal MSE with 90 SNPs than Bayesian LASSO with 1200 markers.
- **Boosting** presented a more rapid decrease on MSE with inclusion of subsequent markers.
- at equal amount of SNPs (90), Boosting reduced MSE by 14% regarding Bayesian LASSO with 90 SNPs.
- Bayesian LASSO with 32K SNPs reduced MSE by 27% regarding Boosting with 90 SNPs, but using 35,000% more markers.
- Bayesian LASSO showed some bias at small number of preselected SNPs.

Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO showed larger predictive ability with larger number of markers.
 - Converge to the nadir with $>10,000$ SNPs.
- **Boosting** showed equal MSE with 90 SNPs than Bayesian LASSO with 1200 markers.
- **Boosting** presented a more rapid decrease on MSE with inclusion of subsequent markers.
- at equal amount of SNPs (90), Boosting reduced MSE by 14% regarding Bayesian LASSO with 90 SNPs.
- Bayesian LASSO with 32K SNPs reduced MSE by 27% regarding Boosting with 90 SNPs, but using 35,000% more markers.
- Bayesian LASSO showed some bias at small number of preselected SNPs.

Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO showed larger predictive ability with larger number of markers.
 - Converge to the nadir with $>10,000$ SNPs.
- **Boosting** showed equal MSE with 90 SNPs than Bayesian LASSO with 1200 markers.
- **Boosting** presented a more rapid decrease on MSE with inclusion of subsequent markers.
- at equal amount of SNPs (90), Boosting reduced MSE by 14% regarding Bayesian LASSO with 90 SNPs.
- Bayesian LASSO with 32K SNPs reduced MSE by 27% regarding Boosting with 90 SNPs, but using 35,000% more markers.
- Bayesian LASSO showed some bias at small number of preselected SNPs.

Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO showed larger predictive ability with larger number of markers.
 - Converge to the nadir with $>10,000$ SNPs.
- **Boosting** showed equal MSE with 90 SNPs than Bayesian LASSO with 1200 markers.
- **Boosting** presented a more rapid decrease on MSE with inclusion of subsequent markers.
- at equal amount of SNPs (90), Boosting reduced MSE by 14% regarding Bayesian LASSO with 90 SNPs.
- Bayesian LASSO with 32K SNPs reduced MSE by 27% regarding Boosting with 90 SNPs, but using 35,000% more markers.
- Bayesian LASSO showed some bias at small number of preselected SNPs.

Real data

Predictive ability. MSE in testing set.

- Bayesian LASSO showed larger predictive ability with larger number of markers.
 - Converge to the nadir with $>10,000$ SNPs.
- **Boosting** showed equal MSE with 90 SNPs than Bayesian LASSO with 1200 markers.
- **Boosting** presented a more rapid decrease on MSE with inclusion of subsequent markers.
- at equal amount of SNPs (90), Boosting reduced MSE by 14% regarding Bayesian LASSO with 90 SNPs.
- Bayesian LASSO with 32K SNPs reduced MSE by 27% regarding Boosting with 90 SNPs, but using 35,000% more markers.
- Bayesian LASSO showed some bias at small number of preselected SNPs.

Remarks

Applications in genomic selection

- Use Bayesian LASSO for genome-enhanced EBV with whole-genome genotypes.
- Boosting presents some advantages at low density markers.
 - May enhance predictive ability in small populations.
 - ...also in association studies.
 - Better bias-variance trade off.

Further considerations and future jobs

- Several stopping criteria exist for boosting, and respective behaviours could be tested.
- Boosting with non-parametric learner is, so far, highly computing time demanding, but more efficient computational strategies might be developed. Parallelization dream.
- Other weak learners may be used.
- L_2 -Boosting performance should be compared against other methods: RKHS, Bayes B and more traditional approaches.

Further considerations and future jobs

- Several stopping criteria exist for boosting, and respective behaviours could be tested.
- Boosting with non-parametric learner is, so far, highly computing time demanding, but more efficient computational strategies might be developed. Parallelization dream.
- Other weak learners may be used.
- L_2 -Boosting performance should be compared against other methods: RKHS, Bayes B and more traditional approaches.

Appendix

MSE in the testing set

