Reproducing kernel Hilbert spaces regression on SNPs for genomic selection

Oscar Gonzalez-Recio, D. Gianola, N. Long, K.A. Weigel, G.J.M. Rosa and S. Avendaño

Session 04, abstract 2831

e-mail: oscar.grecio@upm.es, ogonzalez2@wisc.edu

THE UNIVERSITY of WISCONSIN MADISON

Aviagen™

Genomic information

- Large amount of genomic information is now available (SNPs)
- Opens possibilities for predicting 'EGV' in complex traits using dense molecular information
- Large *p*, small *n* problem

Genomic studies

Big question ?

- Ignore SNP information

- Filter it

- Use it all

Genomic information

- Using SNP information
  - Whole genome
    - Meuwissen et al. (2001)
    - Gianola, Perez-Enciso and Toro (2003)
    - Xu (2003)
  - Pre-selection of SNP
    - Information gain and Bayesian Learning (Long et al., 2007)
    - Lasso (Tibshirani, 1996)
    - t-test

Genomic information

- Using the SNP information
  - Whole genome
    - Xu (2003)
  - Pre-selection of SNP
    - Information gain and Bayesian Learning (Long et al., 2007)

    +
    - Semi-parametric approach (RKHS)

- Objective

  - Asses predictive ability and goodness of fit of three models with/without genomic information

Data

- Data from Aviagen Ltd.
  - Average progeny late mortality (lm) in low hygiene environment for 200 sires from a broiler line (12,167 birds).
  - Pre-corrected for hatch, age of dam and dam, and standardized.
  - 5523 SNPs from *phase 1 project* of Aviagen Ltd.

Methods

- BLUP (Henderson, 1975)
  - No genomic information
- Semiparametric regression (Gianola and Van Kaam, 2008)
  - 24 SNPs from a filter and wrapper strategy (Long et al., 2007)
- Bayesian regression (Xu, 2003)
  - 1000 SNPs randomly distributed along the genome

Methods

Reproducing Kernel Hilbert Spaces

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\beta + \mathbf{K}\alpha + \mathbf{e}$$

K =matrix of kernels that measures distances in a non-Euclidean space. (*distance=sequence alignment score using DP algorithm*) $K_h(\mathbf{x} - \mathbf{x}_i) = \exp\left[-\frac{dis\tan ce(\mathbf{x} - \mathbf{x}_i)^2}{h}\right]$

α =non-parametric coefficients.

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{K}_h \\ \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{X} & \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{K}_h + \frac{1}{\lambda^{-1}}\mathbf{K}_h \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\ddot{e},h} \\ \hat{\mathbf{a}}_{\ddot{e},h} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Methods

- Goodness of fit to observed data:
  - Compute deviance measure based on mean squared errors:
    - A) Regression of adjusted average progeny lm on sire's PTA or EGV
    - B) Regression of raw average progeny lm on sire's PTA or EGV

    Lowess regression (*R-development core team, 2007*)

    (Non-parametric locally weighted regression)

RESULTS

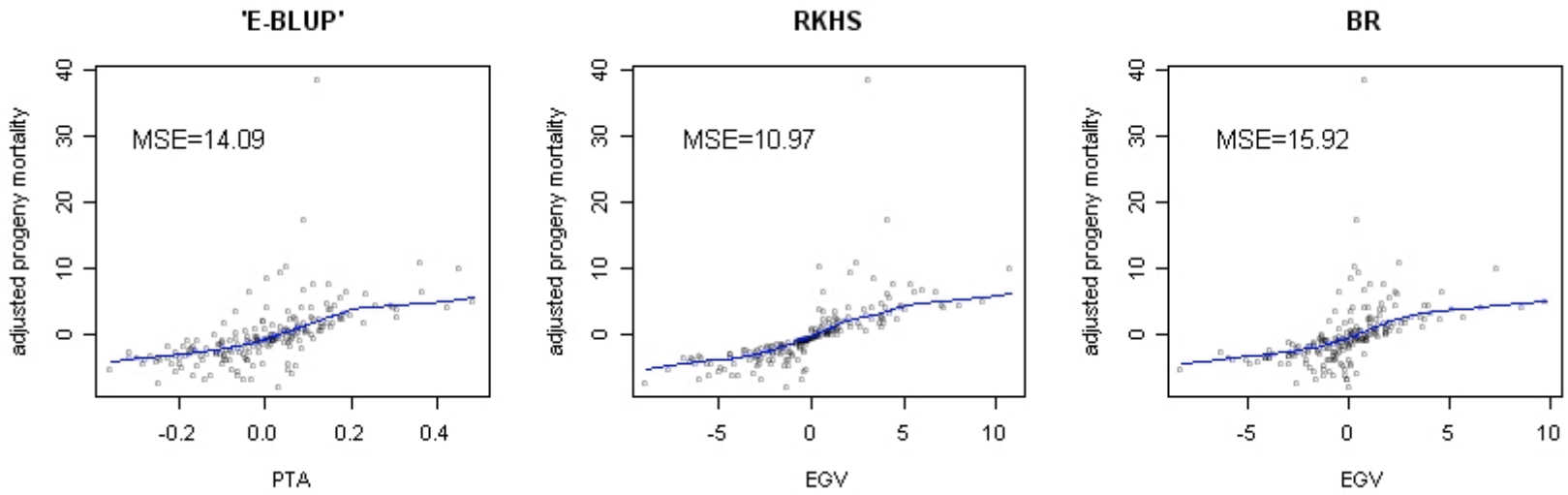
Variance components. Parameter estimates

Parameter	Posterior features	E-BLUP	RKHS	BR (Xu's)
σ_e^2	μ (s.d)	24.38 (3.88)	17.07 (3.02)	20.75 (2.91)
	HPD (95%)	16.88-32.04	11.78-23.64	15.62-27.09
σ_u^2	μ (s.d)	0.10 (0.06)	…	1.03 (0.71)
	HPD (95%)	0.03-0.24	…	0.67-1.95
σ_α^2	μ (s.d)	…	0.40 (0.07)	
	HPD (95%)	…	0.28-0.55	
h^2	μ (s.d)	0.02 (0.01)	…	…
	HPD (95%)	0.004-0.050	…	…

RESULTS

Fitness to the data

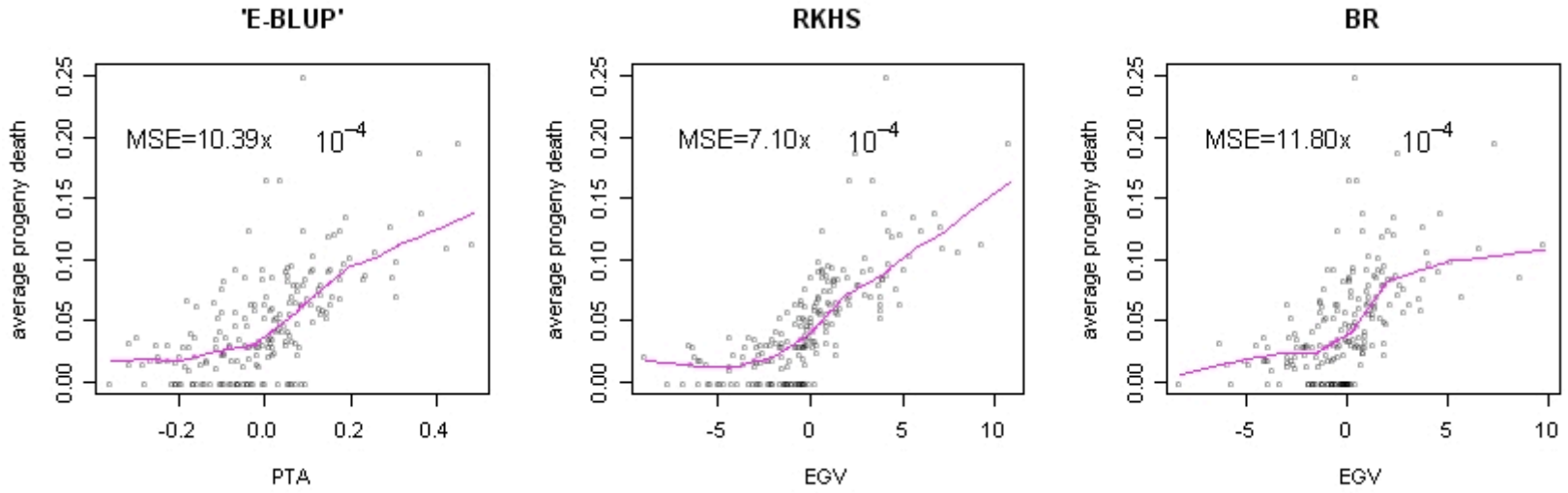
- A) Regression of adjusted average progeny lm on sire's PTA or EGV



RESULTS

Fitness to the data

- B) Regression of adjusted raw progeny lm on sire's PTA or EGV



RESULTS

- Less dispersion in non-parametric model

- Still….what model predicts the data best ?

Predictive ability

- Cross validation

1. 5 subsets, letting 20% sires phenotypes missing each time at random
2. Estimate their PTA or EGV using missing values.
3. Calculate Pearson correlations between actual and estimated average progeny LM, for each method within subset.

RESULTS

Predictive ability

Subset	E-BLUP	RKHS	BR
1st	0.03	**0.27**	0.13
2nd	0.18	**0.37**	0.12
3rd	**0.18**	-0.01	0.17
4th	-0.04	**0.28**	0.15
5th	0.17	0.15	**0.25**
GLOBAL	0.10	**0.20**	0.16

RESULTS

Predictive ability

- Although small amount of markers (24 SNPs) RKHS showed better predictive ability
  - 100% higher reliability than BLUP
  - 25% higher reliability than LR

RESULTS

Predictive ability

- Current work:
  - RKHS with whole genome info (5,000 SNPs) vs. Bayes A.
    - Similar predictive ability
  - RKHS with 400 filtered SNPs vs. Bayes A.
    - Better predictive ability
    - Encouraging results for filtering + RKHS

REMARKS

- SNP chips provide huge amount of genomic information
  - RKHS accommodates large amount of information and different SNP combinations (non-additive effects).
  - Deal with noisy, crude, incomplete and redundant information
- Customized kernels
- Smoothing parameter assessment must be done cautiously

THANK YOU