GENOME-WIDE PREDICTION OF COMPLEX TRAITS UNDER POPULATION STRATIFICATION AND HIDDEN RELATEDNESS

O. González-Recio, A. Pun, S. Forni, D. Gianola, K. Weigel, G.J.M. Rosa, X-L. Wu

Depto. Mejora Genética Animal, INIA; Depto Producción Animal, Fac.Veterinaria (UCM); Genus-PIC; University of Wisconsin-Madison

EAAP, Norway. STAVANGER, AUGUST 29TH 2011.

Outline

Background

- Genomic information ► improves predictive ability
- Challenge when subpopulations are included in the learning sample (reference population)(Goddard and Hayes, 2009; Hayes et al., 2009; Ibañez-Escriche et al., 2009; Toosi et al., 2010)
  - multi-breed/line
  - multi-environment
  - across-country

Background

- Genomic information ► improves predictive ability
- Challenge when subpopulations are included in the learning sample (reference population)(Goddard and Hayes, 2009; Hayes et al., 2009; Ibañez-Escriche et al., 2009; Toosi et al., 2010)
  - multi-breed/line
  - multi-environment
  - across-country

Background

- Difficulties:
  - population stratification
  - heterogeneous additive, epistatic or genotype x environment effects
  - different LD levels or LD phases across populations
  - different genotypic frequencies in learning and validation samples
  - some genotypic configurations not covered by the training model (important with non-additive variation)
- Possible consequences
  - false positives
  - mathematical artifacts
  - over-estimation of SNP effects
  - degraded predictive ability

Background

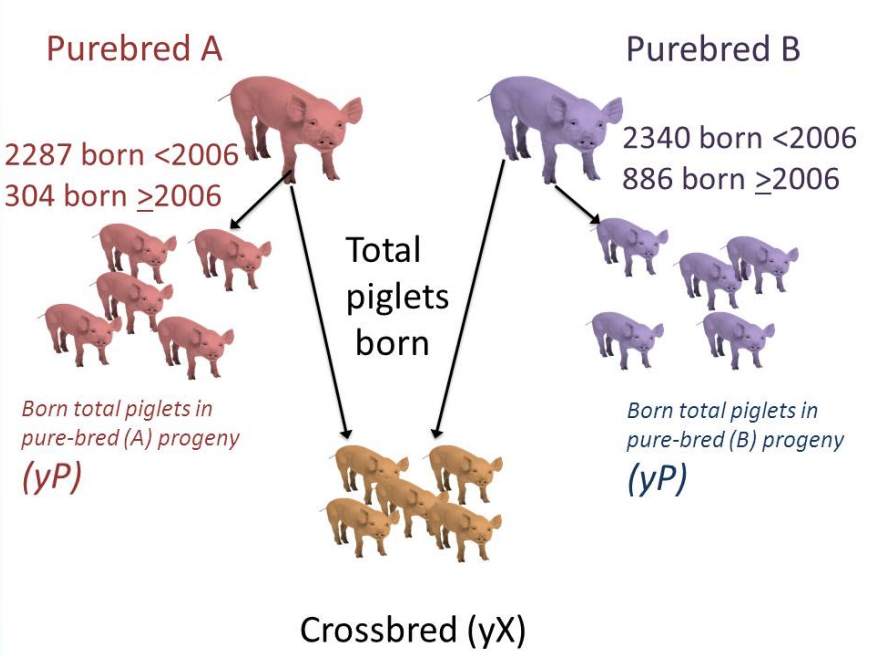
- Difficulties:
  - population stratification
  - heterogeneous additive, epistatic or genotype x environment effects
  - different LD levels or LD phases across populations
  - different genotypic frequencies in learning and validation samples
  - some genotypic configurations not covered by the training model (important with non-additive variation)
- Possible consequences
  - false positives
  - mathematical artifacts
  - over-estimation of SNP effects
  - degraded predictive ability

Proposal/Goal

- Propose a non-parametric model considering performance in each sub-population as a different trait
- Comparison with a single trait model trained in only one sub-population (Bayesian LASSO)

Outline

Materials & Methods



DATA *(provided by Genus-PIC)*

- Progeny adjusted average: Total piglets born in purebred (yP) and crosbred (yX) matings ($y^* = y - \mathbf{X}\beta$)
  - environmental effects: Farm-line-parity, Farm-year-Number of services, Farm-month, age at first farrowing

Materials & Methods

Number of records per line and trait

	A	A	B	B
	yP	yX	yP	yX
Training	2287	282	2340	317
Testing*	63	20	354	78

**Only animals with progeny size >40 (line A) and >100 (line B)*

- Animals genotyped with PorcineSNP60 chip
  - 50284 SNPs after editing

Outline

Methods

- Genome-wide prediction for total piglets born (in purebreds and crossbreds)
  - Bayesian LASSO (BL)
    - univariate for yP
    - yP predicts yP and yX
  - Multitrait reproducing Kernel Hilbert spaces (`RKHSmulti`) regression
    - extends RKHS to multitrait (yP and yX)
    - yP predicts yP and yX
    - yX predicts yX

Methods

RKHSmulti

- $\mathbf{y} = \mathbf{X}\beta + \mathbf{K}\alpha + \mathbf{e}$;
- $\mathbf{y} = \left\{ \begin{array}{cc} y_P & y_X \end{array} \right\}$ vector of observations of total piglets born in each subpopulation
  - 88% missing value for yX.
    - $\mathbf{y}_{BTX} = \{\mathbf{y}_{Xo}, \mathbf{y}_{Xm}\}$
    - Data augmentation for missing yX: $\mathbf{y}_{X_m} \sim N(\mathbf{X}\beta + \mathbf{K}\alpha, \mathbf{R})$

Methods

RKHSmulti

- $\mathbf{y} = \mathbf{X}\beta + \mathbf{K}\alpha + \mathbf{e}$;
  - population means as systematic effects $\beta = (\mu_P \; \mu_X)$

Methods

RKHSmulti

- $\mathbf{y} = \mathbf{X}'\beta + \mathbf{K}\alpha + \mathbf{e}$;
- $\mathbf{K}\alpha = g(\mathbf{x})$ non-parametric function
  - $\mathbf{K}$ kernel matrix with elements $\{k_{ij}\}$; Gaussian kernel with global allelic similarity
  - $\alpha = \left\{ \begin{array}{cc} \alpha_P & \alpha_X \end{array} \right\}$ vector of non-parametric coefficients for each subpopulation (purebreds and crossbreds)

Methods

RKHSmulti

- Prior distributions
  - $\beta \sim U(-9999, 9999)$
  - $\alpha \sim N(0, \mathbf{K}^{-1} \otimes \mathbf{G})$, where $\mathbf{G} = \begin{bmatrix} \sigma^2_P & \sigma_{P,X} \\ \sigma_{X,P} & \sigma^2_X \end{bmatrix}$
  - $\mathbf{G} \sim IW$
  - $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{R})$, where $\mathbf{R} = \begin{bmatrix} \sigma^2_{e_P} & 0 \\ 0 & \sigma^2_{e_X} \end{bmatrix}$
  - $\sigma^2_{e_\cdot} \sim$ *Scaled Inverse Chi*^{2}

Results

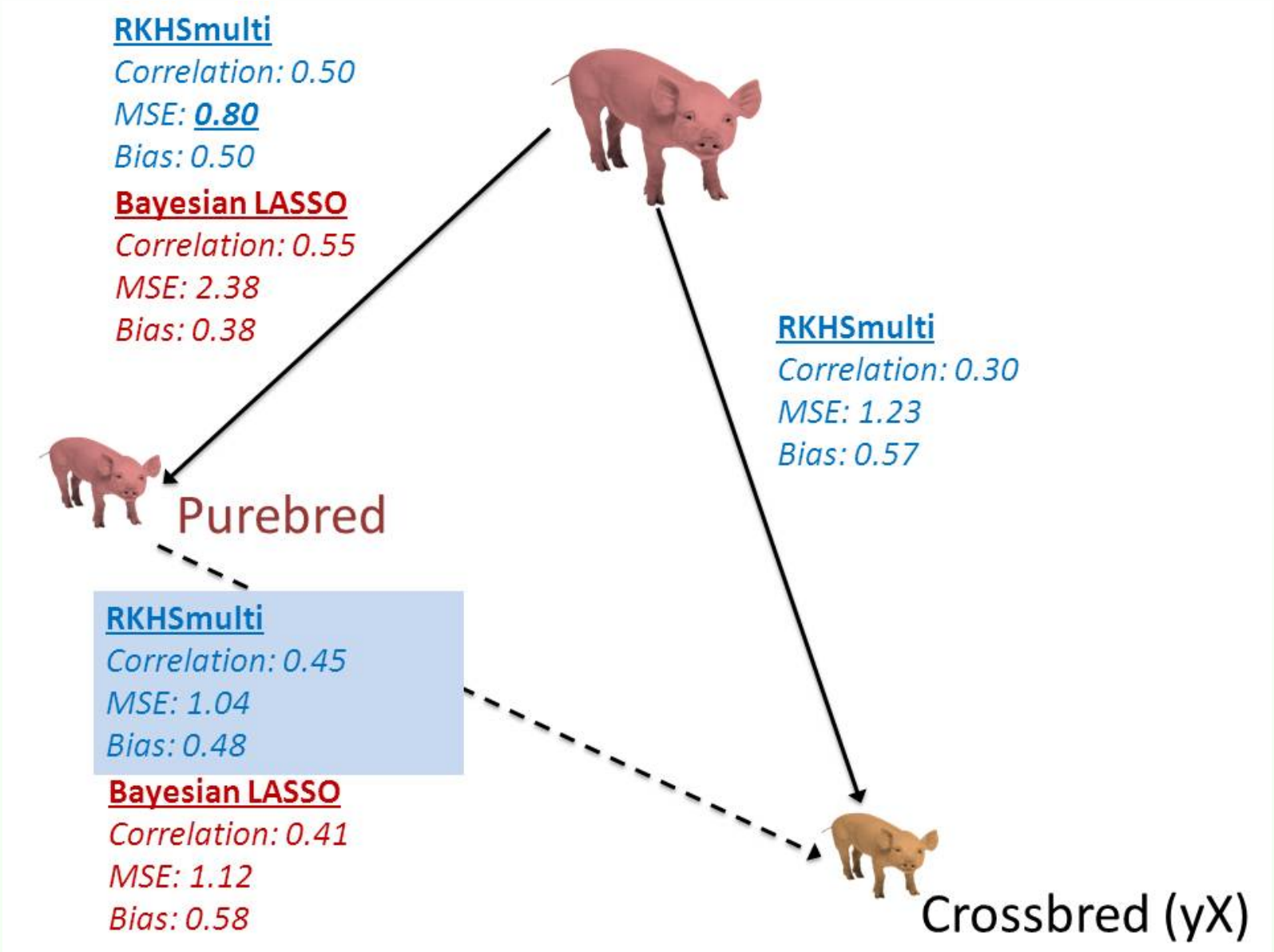
Correlations between purebreds and crossbreds from RKHSmulti

parameter	A	B
Phenotypic correlation	0.05	0.42
Genomic (NP) correlation	0.42 (0.19)	0.75 (0.11)
Residual variance (yP)	1.72 (0.07)	0.90 (0.20)
Residual variance (yX)	2.34 (0.29)	0.94 (0.04)

- Forni et al. (2011) obtained additive correlations of 0.69 between pure and cross-bred total piglets born using BLUP (pedigree)

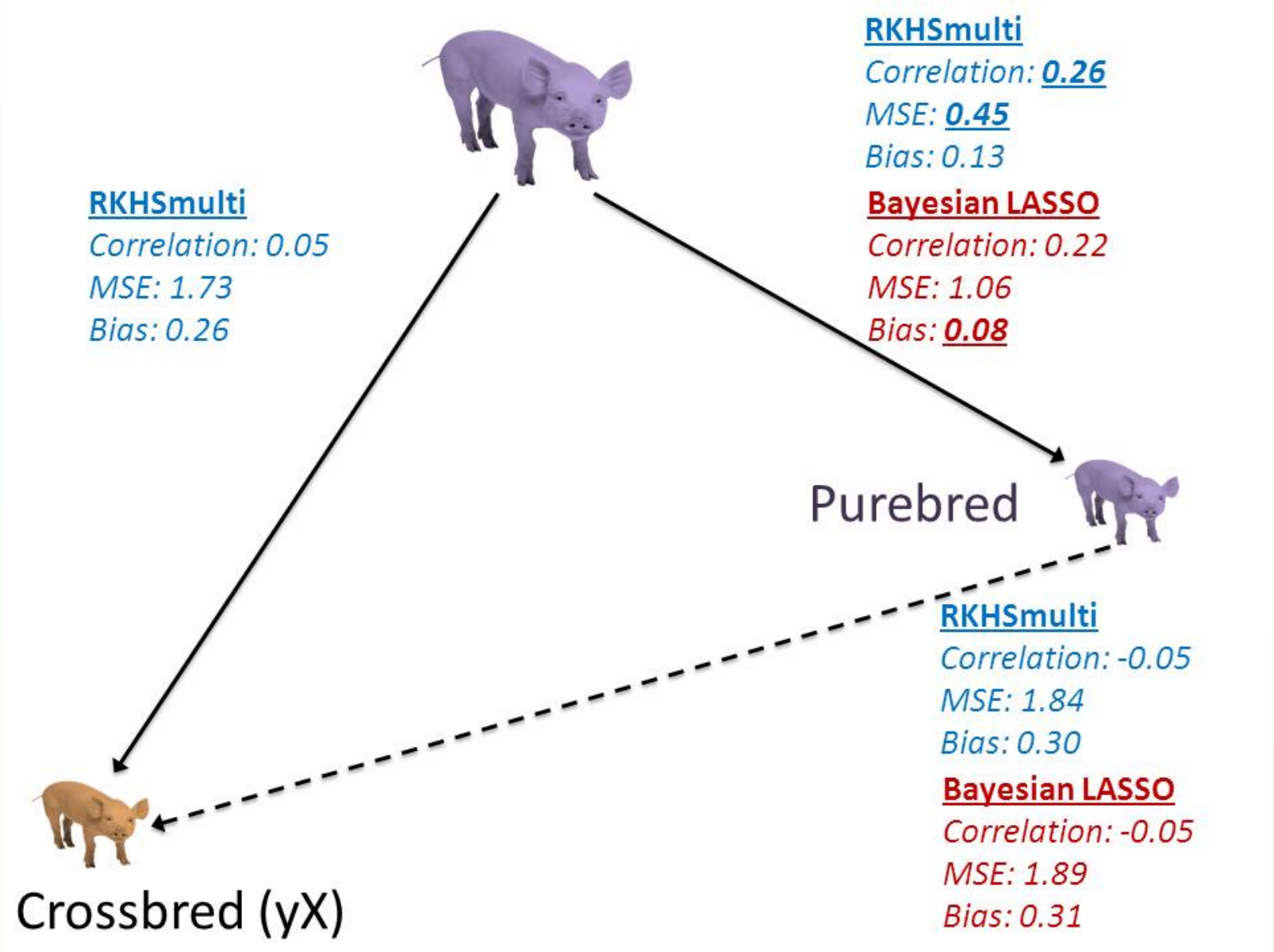
Results

Predictive ability in Line A



Results

Predictive ability in Line B



Results

Summary

1. Similar performance of RKHSmulti and BL in pure-bred progeny
2. RKHS showed slightly higher correlation in cross-bred progeny in both lines, but with different strategy (need uncertainty measurement of correlations).
3. Slightly larger bias with RKHSmulti
4. Promising behavior of RKHS in the cross-bred progeny, but inconclusive results
    1. Large proportion (88%) of missing data for yX
    2. No phenotypes of yX were used in BL (Multitrait model for SNP regression models should be tested; Calus and Veerkamp, 2011; Tsuruta et al., 2011)

Outline

Remarks

- RKHSmulti allows estimating genomic correlation non-parametrically
- Similar predictive ability of both methods for purebred animals (better in line A)
- Promising behavior of RKHS for multitrait analyses, deserves further research.
- Smaller proportion of missing record for yX may be more conclusive

Remarks

- RKHSmulti allows estimating genomic correlation non-parametrically
- Similar predictive ability of both methods for purebred animals (better in line A)
- Promising behavior of RKHS for multitrait analyses, deserves further research.
- Smaller proportion of missing record for yX may be more conclusive

Remarks

- RKHSmulti allows estimating genomic correlation non-parametrically
- Similar predictive ability of both methods for purebred animals (better in line A)
- Promising behavior of RKHS for multitrait analyses, deserves further research.
- Smaller proportion of missing record for yX may be more conclusive

Remarks

- RKHSmulti allows estimating genomic correlation non-parametrically
- Similar predictive ability of both methods for purebred animals (better in line A)
- Promising behavior of RKHS for multitrait analyses, deserves further research.
- Smaller proportion of missing record for yX may be more conclusive

Future work

- Kernel design study
- Multitrait-multi line analysis (yP_A, yP_B, yX_A, yX_B)
  - include pedigree matrix
- Selection of SNPs (check SNP's importance per line)
- Genetic distance between subpopulations (Predicting subpopulations with different genetic base may still be unfeasible)

Future work

- Kernel design study
- Multitrait-multi line analysis (yP_A, yP_B, yX_A, yX_B)
  - include pedigree matrix
- Selection of SNPs (check SNP's importance per line)
- Genetic distance between subpopulations (Predicting subpopulations with different genetic base may still be unfeasible)

Future work

- Kernel design study
- Multitrait-multi line analysis (yP_A, yP_B, yX_A, yX_B)
  - include pedigree matrix
- Selection of SNPs (check SNP's importance per line)
- Genetic distance between subpopulations (Predicting subpopulations with different genetic base may still be unfeasible)

Future work

- Kernel design study
- Multitrait-multi line analysis (yP_A, yP_B, yX_A, yX_B)
  - include pedigree matrix
- Selection of SNPs (check SNP's importance per line)
- Genetic distance between subpopulations (Predicting subpopulations with different genetic base may still be unfeasible)