Pre-selection of markers for genomic selection

Torben Schulz-Streeck,
Joseph Ogutu & Hans-Peter Piepho

Bioinformatics Unit
University of Hohenheim

Germany

Poznań, 18.05.2010

Table of contents

1. Mixed models for GS

Ridge regression BLUP model for genomic selection:

$$y_i = \mu + g_i + e_i \quad (i = 1,...,G)$$

y_i = the phenotype record of the i-th individual
μ = overall mean
g_i = genotypic value of the i-th individual
e_i = the residual error (e_i~N(0, σ_e^2))

$$g_i = \sum_{k=1}^{M} u_k z_{ik}$$

u_k = regression coefficients for the k-th marker
z_{ik} = regressor variable for the i-th genotype and k-th marker

z_{ik} = 1 for A1A1
z_{ik} = -1 for A2A2
z_{ik} = 0 for A1A2

$\mathrm{var}(\boldsymbol{e}_i) = \sigma_e^2$ and $\mathrm{var}(\boldsymbol{u}_k) = \sigma_u^2$ $\Rightarrow$ Estimate by REML! (Ruppert et al., 2003)

2. Pre-selection of markers

Method 1:

Each SNP was tested by a linear regression (Macciotta et al., 2009)

$$y_i = \mu + u_k z_{ik} + e_i$$

y_i is the phenotype record of the i-th individual
μ is the intercept
z_{ik} is the genotype of the i-th individual for the k-th marker
u_k is the slope for the linear regression on the k-th marker
e_i is the residual error (e_i~N(0,σ_e^2))

Pre-selection of marker for consistent effects between crosses

Method 2:

Each SNP was analysed for consistent effects across crosses.

$$y_{ic} = \mu + u_k z_{ik} + Cross_c + \beta_{ck} z_{ik} + e_{ic}$$

y_{ic} is the phenotype record of the i-th individual
μ is the intercept
z_{ik} is the genotype of the i-th individual for the k-th marker
u_k is the slope for the linear regression on the k-th marker
$Cross_c$ is the random effect of the c-th cross
β_{ck} is the slope of c-th cross for the random linear regression on the k-th marker.
e_{ic} is the residual error (e_{ic}~N(0, σ_e^2))
The variance-covariance matrix of the random regression is assumed to be unstructured,
i.e. $\begin{pmatrix} Cross_c \\ \beta_{ck} \end{pmatrix} \sim BVN(0, \Sigma)$, where Σ is an unstructured 2*2 variances-covariance matrix.

Common dataset
14th QTL-MAS workshop

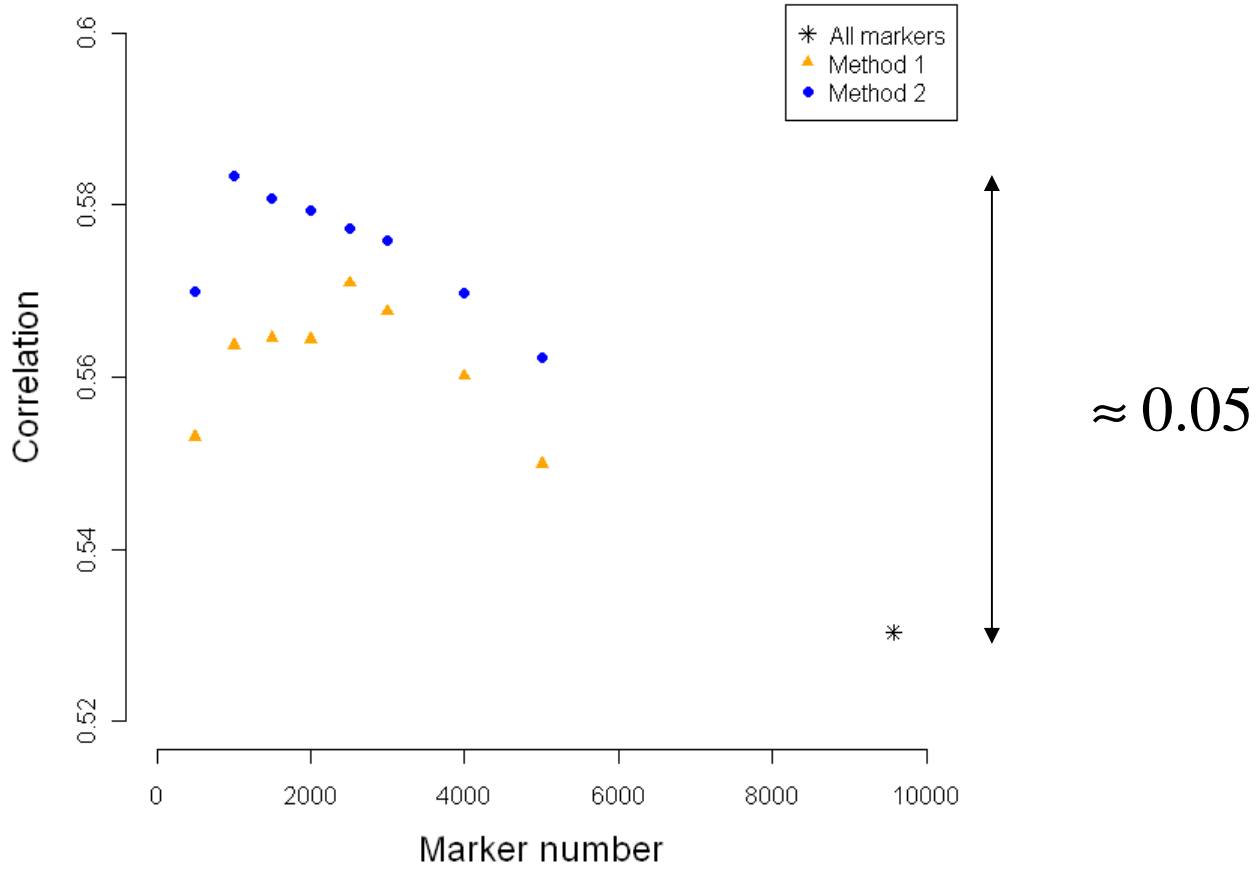
Dataset:

- 2306 individuals
- Three generations
- 75 crosses
- 10031 SNP
- The trait is quantitative

5-fold cross-validation (CV):

- Dataset was randomly split in five subsamples
  - Each subsample contains 15 crosses
- Five CV-rounds
  - In each round the phenotype records of one of the subsamples were discarded, which is used as a validation set
  - Each subsample was only discarded once

Accuracy of the prediction for different marker numbers



3. Spatial mixed models

All conditional models will be of the form:

$$\mathrm{var}(\boldsymbol{g} \mid \boldsymbol{Z}) = \sigma_s^2 \boldsymbol{\Gamma}$$

Ridge regression: $\boldsymbol{\Gamma} = ZZ'$

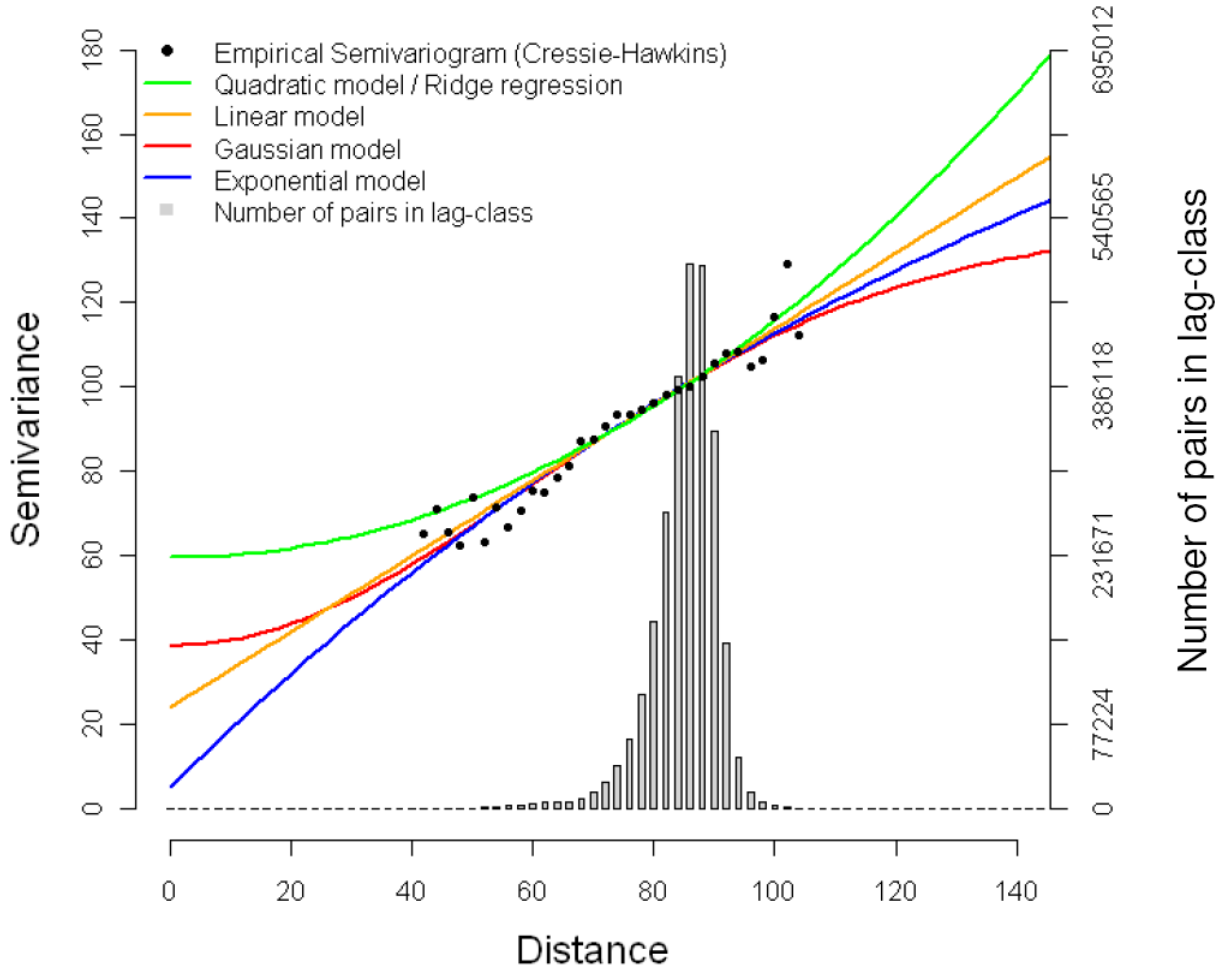
Spatial models: $\boldsymbol{\Gamma} = \{f(d_{ii'})\}$

$d_{ii'}$ = Euclidean distance of two genotypes depending on the marker profiles

$f(d)$ = some monotonically decreasing function of d

Name	Equation
Linear	$f(d) = 1 - \theta d$
Quadratic	$f(d) = 1 - \theta d^2$ ($\Leftrightarrow$ ridge regression)
Exponential	$f(d) = \exp(-d / \theta)$
Gaussian	$f(d) = \exp(-d^2 / \theta^2)$

Semivariogram



Results

	Correlation 5-fold CV*			
	Pre-selection of markers by *Method 2*			
Model for g_i	500 markers	1000 markers	2000 markers	9570 markers
Ridge Regression	0.570	0.583	0.579	0.530
Gaussian	0.569	0.583	0.580	0.530
Exponential	0.572	0.583	0.582	0.530
Linear	0.572	0.584	0.582	

* Mean of 5 CV-rounds (GEBV-Validation set vs. observed values)

4. Heterogeneous marker variance

The genotypic value (g) was predicted by the regression on the maker types:

$$g_i = \sum_{k=1}^{M} u_k z_{ik}$$

z_{ik} = regressor variable for the i-th genotype and k-th marker
u_k = regression coefficients for the k-th marker

Heterogeneous marker variance:

$$u_{km} \sim N\left(0, \sigma^2_{u_m}\right), \quad (m = 1,2)$$

m=1: n (n= 50, 100, 250) most significant markers

m=2: remaining markers

Results

Model for g_i	**Correlation***
Pre-selection of SNPs (*Method 2*)	
RR (1000 markers) (*n*=0)	0.583
RR (1000 markers) (*n*=50)	**0.587**
RR (1000 markers) (*n*=100)	0.586
RR (1000 markers) (*n*=250)	0.584

* Mean of 5 CV-rounds (GEBV-Validation set vs. observed values)

Summary

- Pre-selecting of markers for GS was of benefit
- Spatial models provide alternative methods for GS
- Semivariograms can be used as a tool to identify suitable models for GS
- Modeling heterogeneous variances between the most significant markers and the remaining markers increased the accuracy of prediction slightly
- Model selection criteria such as AIC and GCV may be used instead of cross-validation to reduce computing time