

Mixed model procedures for genome-wide selection

Hans-Peter Piepho & Torben Schulz-Streeck

Bioinformatics Unit
Universität Hohenheim
Germany

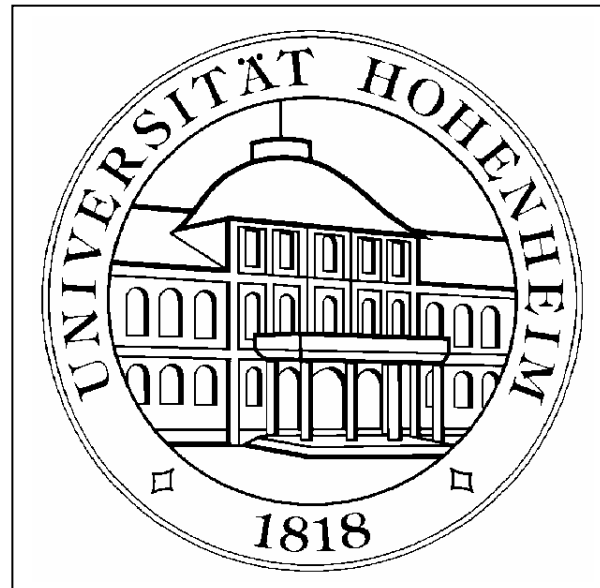


Table of contents

1. Background

2. Mixed models for GS

2.1 Ridge regression

2.2 Partitioning the genetic effect

2.3 Spatial models for GS

2.4 BayesA and BayesB

2.5 Extension of models to genotype-environment data

3. An example with maize

4. Summary

5. References

1. Background

Crop variety trials and plant breeding trials:

- Test performance for **target region**
- Trials in **large number of environments** (ideally random sample)

Standard trial designs for large number of treatments:

- Lattice designs, α -designs, row-column designs (Williams and John, 1995)
- Designs with spatial analysis in mind (Cullis et al., 2006; Williams et al., 2006)

“We should never forget who is the elephant in the room”

(Alan Schulman, plant geneticist, at a recent meeting within EU COST action TD0801 “Statistical challenges on the 1000€ genome sequences in plants”)

Estimation of genetic values

- Classical plant breeding based on phenotypic data alone (field trials)
- Hunting for single genes:
 - ⇒ Use of marker data for mapping of quantitative trait loci (QTL) in simple segregating populations, linkage mapping
 - ⇒ Association mapping in larger populations with diverse structure (multiple crosses, diverse breeding material, gene bank data)
 - ⇒ Marker-assisted selection (MAS) based on detected QTL
- Giving up the hunt:
 - ⇒ Just try to improve estimate of genotypic value (breeding value)
using all markers

Key idea of genomic selection (GS)

Predict genotypic value g_i of i -th genotype by regression on marker types

$$g_i = \sum_{k=1}^M u_k z_{ik} \quad (i = 1, 2, \dots, G)$$

where

z_{ik} = regressor variable for the i -th genotype and k -th marker ($k = 1, \dots, M$)

u_k = regression coefficients

Biallelic marker with alleles A_1 and A_2 , DH lines:

$$z_{ik} = 1 \quad \text{for } A_1A_1$$

$$z_{ik} = -1 \quad \text{for } A_2A_2$$

$$z_{ik} = 0 \quad \text{for } A_1A_2 \text{ or when the marker genotype is missing}$$

2. Mixed models for GS

2.1 Ridge regression

$$\mathbf{g} = \mathbf{Z}\mathbf{u} \quad ,$$

where $\mathbf{g}' = (g_1, g_2, \dots, g_G)$, $\mathbf{Z} = \{z_{ik}\}$, and $\mathbf{u}' = (u_1, u_2, \dots, u_M)$.

Simplest case:

- A single observation y_i per genotype (mean-centered!)
- Independent residual errors e_i having zero mean and variance σ_e^2

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \mathbf{e} \quad ,$$

where $\mathbf{y}' = (y_1, y_2, \dots, y_G)$ and $\mathbf{e}' = (e_1, e_2, \dots, e_G)$.

Classical least squares:

$$\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y} \quad \text{minimizes} \quad \|\mathbf{y} - \mathbf{Z}\mathbf{u}\|^2$$

Ridge regression:

$$\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z} + \lambda^2 \mathbf{I}_G)^{-1} \mathbf{Z}'\mathbf{y} \quad \text{minimizes} \quad \|\mathbf{y} - \mathbf{Z}\mathbf{u}\|^2 + \lambda^2 \mathbf{u}'\mathbf{u}$$

where λ^2 is a penalty parameter

Determining the penalty parameter

- Cross validation
- Bayesian methods
- BLUP

Assume $\text{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$ and $\text{var}(\mathbf{u}) = \mathbf{I}\sigma_u^2$

$$\Rightarrow \text{BLUP}(\mathbf{u}) = \left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2} \mathbf{I}_G \right)^{-1} \mathbf{Z}'\mathbf{y}$$

$$\Rightarrow \lambda^2 = \frac{\sigma_e^2}{\sigma_u^2}$$

\Rightarrow Estimate penalty by REML! (Ruppert et al., 2003)

2.2 Partitioning the genetic effect

$$h_i = g_i + v_i$$

h_i = total genotypic effect

g_i = component explained by the markers

v_i = polygenic component not captured by the markers

- Our main objective is to estimate h_i
- It is assumed throughout that g_i and v_i are independent of one another
- It is important to account for v_i in order to avoid overfitting

Model for polygenic residual ('nugget')

For a **single unstructured population**, for example a population of DH lines generated from a single cross, we have

$$\text{var}(\mathbf{v}) = \sigma_v^2 \mathbf{I}_G ,$$

where $\mathbf{v}' = (v_1, v_2, \dots, v_G)$.

For **structured populations**, $\text{var}(\mathbf{v})$ may involve covariances among relatives

Conditional models for genetic effect g

All conditional models will be of the form

$$\text{var}(g / \mathbf{Z}) = \sigma_u^2 \mathbf{\Gamma}$$

for some matrix $\mathbf{\Gamma}$ that is a function of \mathbf{Z}

Ridge regression: $\mathbf{\Gamma} = \mathbf{Z}\mathbf{Z}'$

Pedigree-based BLUP: $\mathbf{\Gamma} = 2\mathbf{A}$ (\mathbf{A} is the numerator relationship matrix)

Can we still assume independent genotypes?

Conditional variance:

$$\text{var}(\mathbf{g} / \mathbf{Z}) = \sigma_u^2 \mathbf{Z}\mathbf{Z}'$$

Unconditional variance:

$$\text{var}(\mathbf{g}) = E_{\mathbf{Z}} \text{var}(\mathbf{g} / \mathbf{Z}) + \text{var}_{\mathbf{Z}} E(\mathbf{g} / \mathbf{Z}) = \sigma_u^2 E_{\mathbf{Z}} (\mathbf{Z}\mathbf{Z}')$$

with $E_{\mathbf{Z}}$ and $\text{var}_{\mathbf{Z}}$ representing the expectation and variance over \mathbf{Z}

Doubled Haploid (DH) population derived from a single cross

$$E_Z(\mathbf{Z}\mathbf{Z}') = M[p\mathbf{I}_G + (1-p)\mathbf{J}_G]$$

p = probability that a marker is segregating in the underlying cross

\mathbf{J}_G = a $G \times G$ matrix of ones

$\Rightarrow \hat{\sigma}_u^2 = M^{-1}\hat{\sigma}_g^2$ is a reasonable estimate of σ_u^2 , when $p = 1$ and $\sigma_v^2 = 0$

(Bernardo and Yu, 2007)

But: It may be better to estimate σ_u^2 directly by REML based on the ridge regression model, because this allows accounting for $\sigma_v^2 > 0$ and/or $p < 1$.

2.3 Simple spatial mixed models

$$\mathbf{\Gamma} = \{f(d_{ii'})\} ,$$

$d_{ii'}$ = Euclidean distance of genotypes i and i' = $\|z_i - z_{i'}\|$

z'_i = i -th row of \mathbf{Z}

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

Table 1: Genotypic covariance models of the form $\Gamma = \{f(d_{ii'})\}$.

Name	Equation
Gaussian	$f(d) = \exp(-d^2/\theta)$
Power (exponential)	$f(d) = \theta^d$
Exponential	$f(d) = \exp(-d/\theta)$
Spherical	$f(d) = 1 - \frac{3d}{2\theta} + \frac{d^3}{2\theta^3} \quad (d < \theta)$
Linear	$f(d) = 1 - \theta d$
Quadratic	$f(d) = 1 - \theta d^2$ (\Leftrightarrow ridge regression)

Equivalence relationships among models

- ridge regression
- = quadratic model
- = use of Kinship matrix (Yu et al., 2006)

Under each of these models we can write

$$\text{var}(\mathbf{g} / \mathbf{Z}) = a\mathbf{S} + b\mathbf{J}_G, \quad \text{where}$$

$\mathbf{S} = \{s_{ii'}\}$, $s_{ii'}$ = simple matching coefficient of genotypes i and i'

\mathbf{J}_G = a $G \times G$ matrix of ones

BLUP based on spatial models – relations with other methods

Gaussian spatial model $f(d) = \exp(-d^2/\theta)$ is equivalent to

- **Least squares support vector machine** (LS-SVM) regression, when a Gaussian kernel is used (Suykens et al., 2002, p.106-107)
- **Reproducing kernel Hilbert spaces regression** (Gianola and van Kaam, 2008)
- Approximates **ridge regression** when d^2 / θ small: $\exp(-d^2/\theta) \approx 1 - d^2/\theta$

2.4 BayesA and BayesB (Meuwissen et al., 2001)

$$g_i = \sum_{k=1}^M \sqrt{\sigma_k^2} t_k z_{ik} ,$$

where $t_k \sim N(0,1)$ and σ_k^2 = variance for the k -th marker

A prior distribution is assumed for the variances σ_k^2 .

BayesA and BayesB cont'd

- The random regression coefficient $u_k = \sqrt{\sigma_k^2} t_k$ will have a symmetric non-normal marginal distribution whose specific form depends on the assumed prior for σ_k^2 .
- This marginal distribution for u_k has a constant variance, and so the only difference to ridge regression is that non-normality holds for u_k .

In light of these considerations, the good performance of BayesB relative to ridge regression in Meuwissen et al. (2001) is probably at least partly due to the strong impact of the assumed prior distribution, which was derived based on the model used to simulate the data.

A simple frequentist alternative to BayesB

- Regard σ_k^2 as fixed parameters
- Use REML

⇒ fit will typically yield many zero estimates

⇒ automatic selection of markers

2.5 Extension of models to genotype-environment data

Genotypes ($i = 1, 2, \dots, G$)

Environments ($j = 1, 2, \dots, E$)

$$h_{ij} = g_{ij} + v_{ij}$$

g_{ij} = marker-based effect

v_{ij} = polygenic effect

Partitioning into main effect and interaction

$$g_{ij} = g_i + f_{ij} \text{ and}$$

$$v_{ij} = v_i + w_{ij}$$

- Let $\mathbf{f}'_j = (f_{1j}, f_{2j}, \dots, f_{Gj})$ and $\mathbf{f}' = (\mathbf{f}'_1, \mathbf{f}'_2, \dots, \mathbf{f}'_E)$
- Let \mathbf{w} be similarly defined

Marker-based effects

Polygenic effects

$$\text{var}(\mathbf{g} / \mathbf{Z}) = \sigma_u^2 \mathbf{\Gamma}$$

$$\text{var}(\mathbf{v}) = \sigma_v^2 \mathbf{I}_G$$

$$\text{var}(\mathbf{f} / \mathbf{Z}) = \mathbf{\Sigma}_f \otimes \mathbf{\Gamma}$$

$$\text{var}(\mathbf{w}) = \mathbf{\Sigma}_w \otimes \mathbf{I}_G$$

Table 2: Models for variance-covariance among genotypes in different environments (Σ_q ; $q = f, w$).

Model	Short-hand	Equation
Independent	ID	$\sigma_1^2 \mathbf{I}_E$
Diagonal	DIAG	$\mathbf{D} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_E^2)$
Factor-analytic [§]	FA(P)	$\sum_{p=1}^P \lambda_p \lambda_p' + \mathbf{D}$

§ $\lambda_p' = (\lambda_{1p}, \lambda_{2p}, \dots, \lambda_{Ep})$

Two-stage analysis for genotype-environment data

Stage 1:

$$y_{ij} = \mu_j + h_i + w_{ij},$$

y_{ij} = adjusted mean of the i -th genotype in the j -th environment

μ_j = main effect of the j -th environment.

⇒ estimate adjusted genotype means \bar{y}_i taking both μ_j and h_i as fixed

Stage 2:

$$\bar{y}_i = \mu + h_i + e_i$$

where

$$\text{var}(\mathbf{h} / \mathbf{Z}) = \text{var}(\mathbf{g} / \mathbf{Z}) + \text{var}(\mathbf{v}) \text{ with } \mathbf{h}' = (h_1, h_2, \dots, h_G)$$

$\text{var}(e_i)$ is fixed at the squared standard error of the adjusted mean \bar{y}_i

For comparison:

Merge e_i with v_i into an independent residual with constant variance.

3. An example with maize (kindly provided by KWS, 2008)

- 208 DH lines originating from a single cross of two inbred parental lines
- Tested in three series of trials over five locations
- In four locations, a lattice design with block size ten was employed
- In one location a complete block design was used
- In four locations, only a single replicate was planted, while in one location there were two replicates
- Seven check genotypes
- Adjusted means for entries with marker data subjected to mixed model analysis
- Trait: kernel dry weight per plot

Table 3: AIC for various variance-covariance structures Σ_w fitted to the phenotypic data (genotype-environment means). Models had fixed main effects for genotypes and environments.

Model (Σ_w)	Deviance	AIC
ID	2843.1	2845.1
DIAG	2753.7	2763.7
FA(1)	2744.4	2762.4
FA(2)	2743.5	2771.5

Table 4: Model fits of different genetic covariance models with the maize data. Error variance $\text{var}(e_i)$ not fixed (average error variance of a mean = 0.174).

Model for g_i	Deviance	AIC	Residual variance [§]
Independent [§]	372.8	374.8	0.3454
Ridge regression			
RR _{hom}	336.9	340.9	0.2272
RR _{het} (38 markers selected)	289.5	367.5	0.1635
RR _{hom2} [§] (38 markers)	303.3	307.3	0.1773
Spatial models			
Linear	335.6	339.6	0.1139
Quadratic	336.9	340.9	0.2272
Power	334.8	340.8	0.0020
Exponential	334.8	340.8	0.0018
Gaussian	333.9	339.9	0.0002
Spherical	334.3	340.3	<0.0001

§ Residual subsumes v_i and e_i , because $\text{var}(e_i)$ was not fixed.

\$ Heterogeneity of variance among selected markers was not significant according to a likelihood ratio test ($\alpha = 5\%$).

Table 5: Model fits of different genetic covariance models with the maize data. Error variance $\text{var}(e_i)$ fixed at value of squared standard error of a mean.

Model for g_i	Deviance	AIC	Polygenic variance (σ_v^2)
Independent	372.8	374.8	0.1712
Ridge regression			
RR _{hom}	336.9	340.9	0
RR _{het} (37 markers selected)	289.7	363.7	0
RR _{hom2} ^{\$} (37 markers)	301.9	305.9	0.0045
Spatial models			
Linear	337.1	339.1	0
Quadratic	336.9	340.9	0.0528
Power [#]	337.2	341.2	0
Exponential	337.1	341.1	0
Gaussian	335.2	339.2	0
Spherical	337.1	341.1	0

^{\$} Heterogeneity of variance among selected markers was not significant according to a likelihood ratio test ($\alpha = 5\%$); variance estimates were shrunken to the overall mean.

[#] autocorrelation converged to value close to unity.

4. Summary

- Ridge regression can be implemented as a mixed model analysis
- Can perform GS in a mixed model framework accounting for various sources of variation
- Spatial models provide alternative methods for GS
- Polygenic effect (nugget) may be required
- Independent estimate of error needed to avoid overfitting
- Can and probably should also model genotype-by-environment interaction in GS

5. References

- Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. *Crop Science* 47:1082-1090.
- Gianola, D. and J.B.C.H.M. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178: 2305-2313.
- Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819-1829.
- Piepho, H.P. 2009. Ridge regression and extensions for genome-wide selection in maize. *Crop Science* 49:1165-1176.
- Ruppert, D., M.P. Wand, and R.J. Carroll. 2003. *Semiparametric regression*. Cambridge University Press, Cambridge.
- Schulz-Streeck, T., Piepho, H.P. 2009. Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. *BMC Proceedings*