Estimating heritability and predictive accuracy of genomic prediction in plant breeding programs

Hans-Peter Piepho

Biostatistics Unit
Universität Hohenheim
Germany

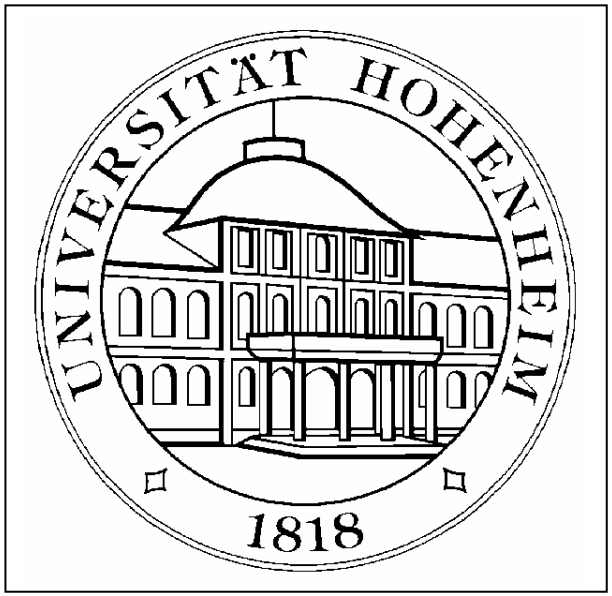


Table of contents

1. Introduction

Broad-sense heritability

$$H^2 = \sigma_g^2 / \sigma_p^2$$

σ_g^2 = genotypic variance

Narrow-sense heritability

$$h^2 = \sigma_a^2 / \sigma_p^2$$

σ_a^2 = additive genetic variance

σ_p^2 = phenotypic variance

Uses of heritability

- Descriptive measure of precision of trial
- Compute response to selection ($R = h^2 S$, where S = selection differential)
- Compute predictive accuracy in genomic prediction

1. Introduction

What is the 'phenotype' here?

- Single plot observation?
- Genotype mean in a trial?
- Genotype mean in multi-environment trial (MET)?
- BLUP of genotypic value?

$\Rightarrow$ How to estimate the phenotypic variance σ_p^2 ?

2. Heritability for balanced data

Basic model for MET data

$$y_{ijk} = \mu + g_i + e_j + (ge)_{ij} + r_{jk} + \varepsilon_{ijk}$$

y_{ijk} = yield of the i^{th} genotype in the j^{th} location and k^{th} replicate

μ = overall mean

g_i = main effect of the i^{th} genotype; $\sim N(0, \sigma_g^2)$

e_j = main effect of the j^{th} environment

$(ge)_{ij}$ = ij^{th} genotype × environment interaction effect; $\sim N(0, \sigma_{ge}^2)$

r_{jk} = k^{th} replicate effect in j^{th} environment

ε_{ijk} = residual comprising both genotype × location × year interaction as well as the error of a mean; $\sim N(0, \sigma^2)$

Model assumes a randomized complete block design (RCBD) per environment

2. Heritability for balanced data

The phenotype

Assume balanced data from RCBD or CRD & complete G x E classification $\Rightarrow$

$$\bar{y}_{i\bullet\bullet} = \mu + g_i + \bar{e}_{\bullet} + \overline{(ge)}_{i\bullet} + \bar{r}_{\bullet\bullet} + \bar{\varepsilon}_{i\bullet\bullet}$$

$$\sigma_p^2 = \sigma_g^2 + \underbrace{\sigma_{ge}^2 / m + \sigma^2 / (rm)}$$

variance of a mean = ½ variance of a difference

m = number of environments

r = number of replicates per environment

3. Heritability for unbalanced data

The phenotype : mostly from unbalanced data / designs

***Crop variety trials and plant breeding trials*:**

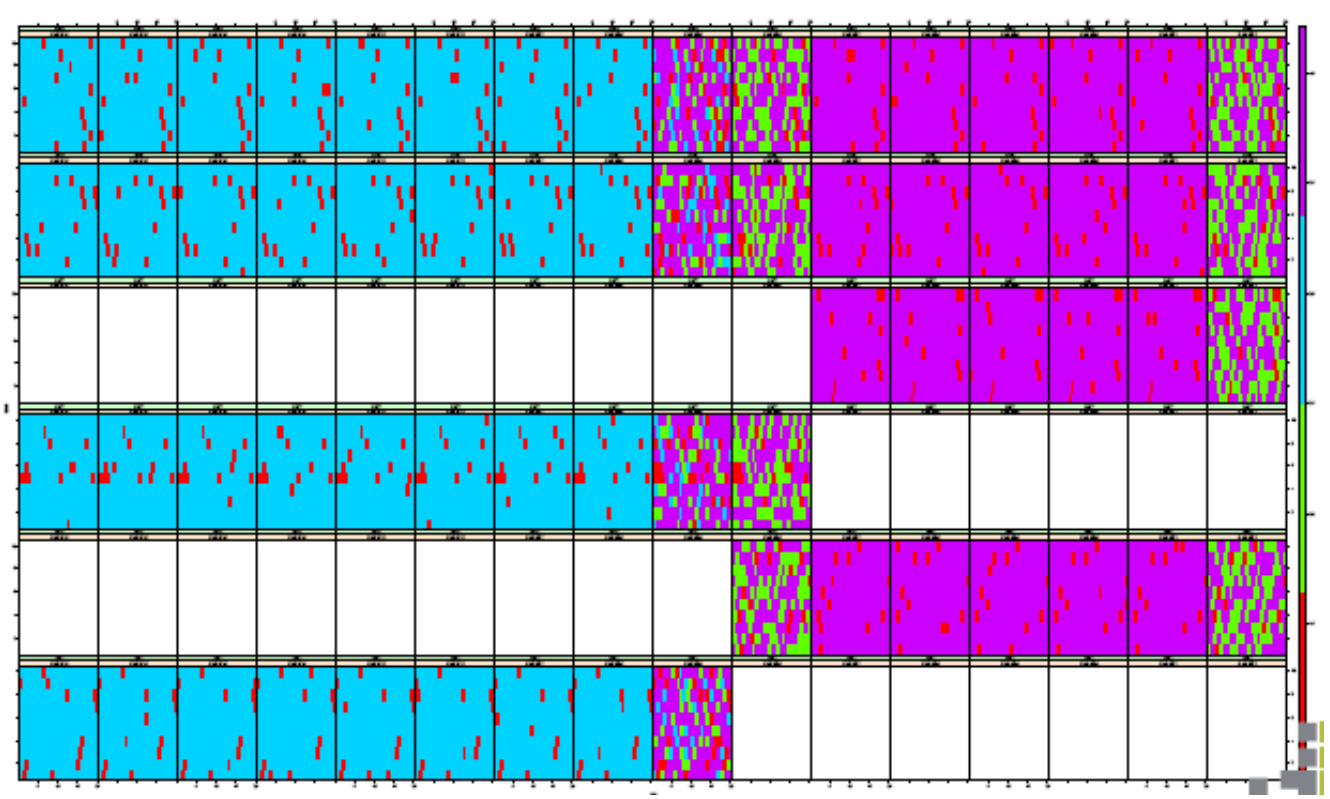
- Test performance for target region
- Trials in large number of environments (ideally random sample from target)

***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)
- Unreplicated designs with checks, p-rep designs, augmented p-rep designs (Cullis et al., 2006; Williams et al., 2011, 2013)

3. Heritability for unbalanced data

A multi-location trial in a maize breeding programme (KWS)



6 locations

16 series (laid out as 10 x 10 lattices)

(T. Albrecht, TUM)

3. Heritability for unbalanced data

Holland et al. (2003, p.64)

Balanced case:

$$\sigma_p^2 = \sigma_g^2 + \sigma_{ge}^2 / m + \sigma^2 / (rm)$$

Divisor of σ_{ge}^2 = no. of environments m

Divisor of σ^2 = no. of plots $p = rm$

Expected genetic gain (EGG):

$$EGG = i \sigma_g H$$

3. Heritability for unbalanced data

Unbalanced (used with all kinds of incomplete block design):

$$\sigma_p^2 = \sigma_g^2 + \sigma_{ge}^2 / m_h + \sigma^2 / p_h$$

$$m_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{m_i}} \quad , \quad p_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{p_i}}$$

m_i = no. of environments for i^{th} genotype

p_i = no. of plots for i^{th} genotype

n = no. of genotypes

(Holland et al. 2003)

3. Heritability for unbalanced data

Problems:

- $\bar{y}_{i\bullet\bullet}$ is not the best 'phenotype' (BLUE of $\mu + g_i$) with unbalanced data
- $\sigma^2_{ge} / m_h + \sigma^2 / p_h$ is not ½ variance of a difference,

  no matter if $\bar{y}_{i\bullet\bullet}$ or the BLUE of $\mu + g_i$ is used

  $\Leftrightarrow$ some or many genotype-environment combinations may be missing!

3. Heritability for unbalanced data

Piepho & Möhring (2007)

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + 0.5\overline{vd}}$$

Rationale:

- $\overline{vd}$ is the average variance of a difference between adjusted means (BLUE of $\mu + g_i$) based on an analysis with fixed genotype effects
  $\Rightarrow$ proportional to *effective error mean square*

- For balanced data:

$$0.5\overline{vd} = \sigma_{ge}^2 / m + \sigma^2 / (rm)$$

3. Heritability for unbalanced data

When 'phenotype' is a BLUP

- So far, phenotype was a "mean" (ideally the BLUE)
- When genotypes are random, BLUP of g_i is usually the better estimator

$$H^2 = 1 - \frac{\bar{v}_{BLUP}}{2\sigma_g^2}$$

$\bar{v}_{BLUP}$ = mean variance of a difference of the BLUP of g_i

Rationale:

- Mean of $[corr(g_i, \hat{g}_i)]^2$
- Concept of effective error variance
- For balanced data $EGG = i\sigma_g H$ (Cullis et al., 2006)

3. Heritability for unbalanced data

When genotypic effects are correlated

- So far, genotypic effects were assumed to be i.i.d.
- Often, use pedigree or markers to model genetic covariance

$\Rightarrow$ Generalized heritability (Oakey et al. 2006)
$\Rightarrow$ Simulation (Piepho & Möhring, 2007)

3. Heritability for unbalanced data

Generalized heritability (Oakey et al., 2006)

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_g\boldsymbol{g} + \ldots$$

$\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \boldsymbol{V})$ (observed data vector)

$\boldsymbol{g} = (g_1, g_2, \ldots, g_n)^T \sim MVN(\boldsymbol{0}, \boldsymbol{G})$

Consider contrast $\boldsymbol{c}^T\boldsymbol{g}$ where $\boldsymbol{c}$ is a contrast vector

$$H^2 = \frac{\left[\mathrm{cov}\left(\boldsymbol{c}^T\boldsymbol{g}, \boldsymbol{c}^T\hat{\boldsymbol{g}}\right)\right]^2}{\mathrm{var}\left(\boldsymbol{c}^T\boldsymbol{g}\right)\mathrm{var}\left(\boldsymbol{c}^T\hat{\boldsymbol{g}}\right)} \qquad \text{where} \quad \hat{\boldsymbol{g}} = BLUP(\boldsymbol{g})$$

$\Rightarrow$ Find $\boldsymbol{c}$ such that H^2 is maximized

3. Heritability for unbalanced data

$$H^2 = \frac{\left[\mathrm{cov}\left(\boldsymbol{c}^T\boldsymbol{g}, \boldsymbol{c}^T\hat{\boldsymbol{g}}\right)\right]^2}{\mathrm{var}\left(\boldsymbol{c}^T\boldsymbol{g}\right)\mathrm{var}\left(\boldsymbol{c}^T\hat{\boldsymbol{g}}\right)} = \frac{\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{Z}_g^T\boldsymbol{P}_v\boldsymbol{Z}_g\boldsymbol{G}\boldsymbol{c}}{\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{c}}$$

$$\boldsymbol{P}_v = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$$

Constraint: $\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{c} = 1$

Method of Lagrange multipliers:

Maximize $\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{Z}_g^T\boldsymbol{P}_v\boldsymbol{Z}_g\boldsymbol{G}\boldsymbol{c} - \lambda\left(\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{c} - 1\right)$

λ = first eigenvalue of $\boldsymbol{Z}_g^T\boldsymbol{P}_v\boldsymbol{Z}_g\boldsymbol{G}$

$\boldsymbol{c}$ = corresponding eigenvector

3. Heritability for unbalanced data

$$\max_{c} H_c^2 = \lambda$$

is a component of the full heritability

Full set of non-zero eigenvalues: $\lambda_1, \lambda_2, \ldots, \lambda_s$

$\Rightarrow$ full heritability

$\Rightarrow$ eigenvectors $c_1, c_2, \ldots, c_s$ = full set of orthogonal genotype contrasts

Generalized heritability

$$H^2 = \frac{\sum_{h=1}^{s} \lambda_h}{s}$$ (Oakey et al., 2006)

3. Heritability for unbalanced data

Questions

- Would a breeder select for $\boldsymbol{c}^T\boldsymbol{g}$?
- Why would we allow contrast vector $\boldsymbol{c}$ to be determined by the data?
- What information is contained in the s orthogonal contrasts corresponding to $\lambda_1, \lambda_2, ..., \lambda_s$?

$\Rightarrow$ average of $\dfrac{\left[\mathrm{cov}\left(\boldsymbol{c}^T\boldsymbol{g}, \boldsymbol{c}^T\hat{\boldsymbol{g}}\right)\right]^2}{\mathrm{var}\left(\boldsymbol{c}^T\boldsymbol{g}\right)\mathrm{var}\left(\boldsymbol{c}^T\hat{\boldsymbol{g}}\right)}$ across all orthogonal contrasts?

3. Heritability for unbalanced data

Monte-Carlo simulation

- Can estimate variance and covariance of $\boldsymbol{g}$ and $\hat{\boldsymbol{g}} = BLUP(\boldsymbol{g})$
- From this can simulate many realizations of $(\boldsymbol{g}, \hat{\boldsymbol{g}})$
- Simulate H^2 = squared correlation of $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}$
- Simulate response to selection!
- Simulate anything else you would want to use H^2 for in the balanced case!

(Piepho & Möhring, 2007)

Advantages

- Completely flexible
- Can handle any covariance structure
- Can directly simulate any statistic of interest

3. Heritability for unbalanced data

Example

- Rapeseed variety trials in Germany
- 120 cultivars (G) tested in 4 years (Y) and at 4 locations (L)
- At some locations, several trials (T) were performed
- The series was rather unbalanced
- Trials were laid out in randomized complete blocks
- Trial means were analyzed by the variance components model

L.Y.T : G + G.L + G.Y + G.L.Y + G.L.Y.T

3. Heritability for unbalanced data

TABLE 3

Variance component estimates (REML) for example 2 based on model (21)

Term	Variance
G	3.8299
$G \cdot L$	0.2094
$G \cdot Y$	1.9245
$G \cdot L \cdot Y$	5.8699
$G \cdot L \cdot Y \cdot T$	1.8617

3. Heritability for unbalanced data

TABLE 5

Response to selection for example 2 computed by simulation based on (16), based on an *ad hoc* approach using $\overline{H}^2$, and based on $\overline{H}^2_C$ (CULLIS *et al.* 2006) for different selection fractions $p = n/I$ with $I = 120$

	Response to selection		
n	Based on simulation	*Ad hoc* $\overline{H}^2$	$\overline{H}^2_C$
1	3.364	2.632	2.979
2	3.056	2.455	2.777
3	2.850	2.327	2.635
4	2.693	2.227	2.522
5	2.565	2.145	2.428
10	2.153	1.858	2.103
15	1.899	1.669	1.887
20	1.711	1.520	1.720
30	1.433	1.291	1.461
40	1.221	1.108	1.254
50	1.045	0.952	1.077
60	0.889	0.811	0.918

3. Heritability for unbalanced data

Example

- Sugar beet
- 26 breeding trials (6 x 6 simple lattices)
- Connected by checks
- 825 entries
- 33 crosses
- Pedigree data available (ad hoc measures do not apply)

T.R + C + Ts : T.R.B + X.G

T = trial, R = replicate, B = block
C = factor separating individual checks from entries
Ts = tester
G = genotype, X = dummy variable (1 for entries, 0 for checks)

3. Heritability for unbalanced data

TABLE 6

Model fit for two genotypic variance–covariance structures (example 3)

Term	Form of var(g)	
	$\mathbf{I}\sigma_g^2$	$\mathbf{A}\sigma_a^2$
$T \cdot R \cdot B$	901.90	858.88
$X \cdot G$ (independent)	2,745.84	—
$X \cdot G$ (pedigree)	—	1,010.46
Residual	3,134.02	3,626.05
$r^2(g_i, \hat{g}_i)$	0.5970[a]	0.5328
−2 restricted log-likelihood	20,969.4	20,997.2

[a] Up to the fourth decimal place, the same estimate was obtained for $\overline{H}_C^2$ (Cullis *et al.* 2006).

3. Heritability for unbalanced data

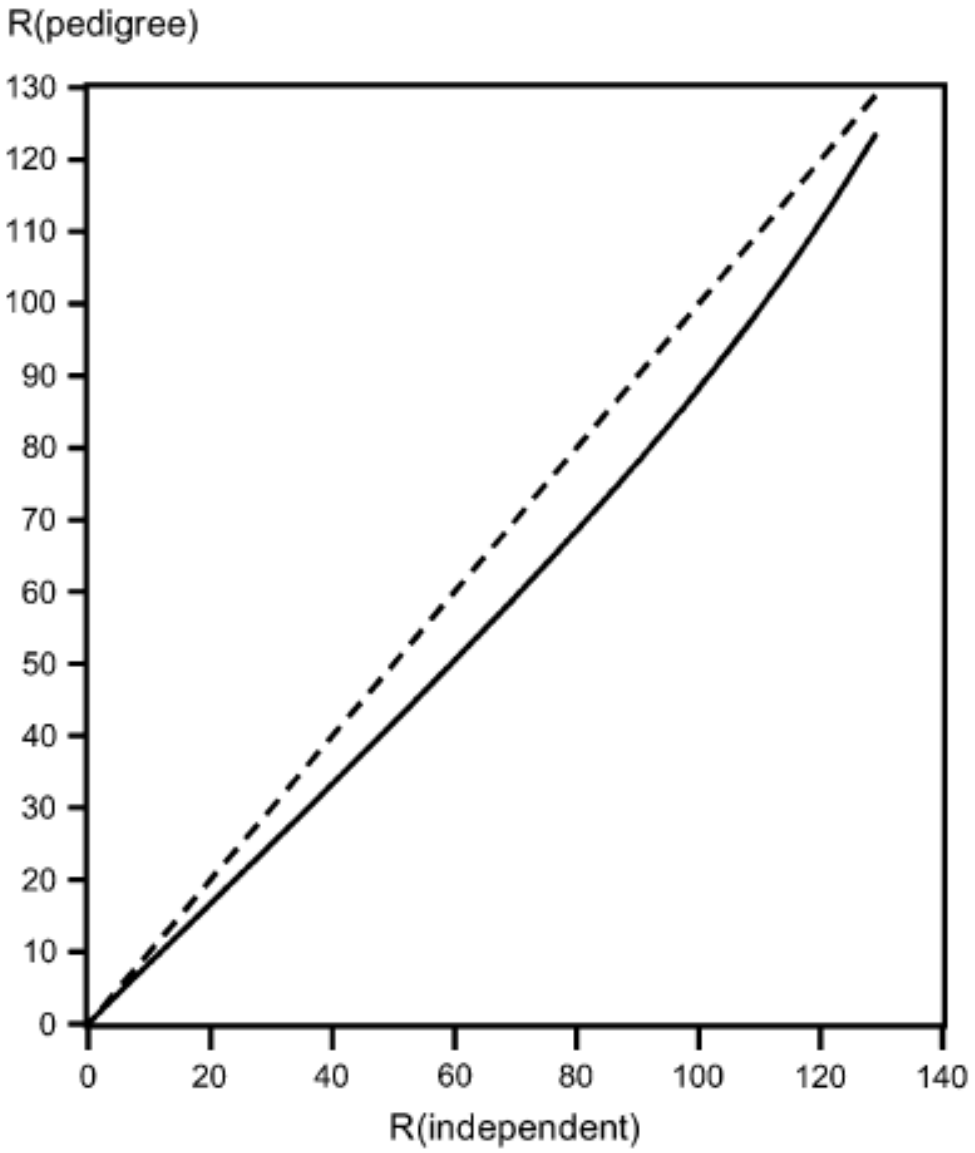


FIGURE 1.—Plot of simulated response to selection under pedigree-based [*R*(pedigree)] and independent [*R*(independent)] models for genotypic variance and for all selection fractions. The dashed diagonal indicates equal response to selection (example 3).

3. Heritability for unbalanced data

Sillanpää (2011)

$$H^2 = \frac{\sigma_p^2 - \sigma_e^2}{\sigma_p^2}$$

σ_p^2 = phenotypic variance

σ_e^2 = residual variance from whole genome random marker regression model

$$\hat{\sigma}_p^2 = \boldsymbol{p}_c^T \boldsymbol{Q} \boldsymbol{p}_c \big/ (n-1)$$

$\boldsymbol{p}_c$ = vector of mean-centered observed phenotypes (genotype means) = $\boldsymbol{P}_u\boldsymbol{p}$

n = number of genotypes

$\boldsymbol{Q}$ = identity matrix $\boldsymbol{I}$ *or* realized relationship estimated from markers

3. Heritability for unbalanced data

Count data

$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}$ (GLMM)

$E(\boldsymbol{y} \mid \boldsymbol{\eta}) = g^{-1}(\boldsymbol{\eta})$, where $g(.)$ denotes a link function

Example: Poisson distribution

$\lambda = E(y \mid \eta) = \exp(\eta)$

Challenge:

- Genetic variance on link-scale
- Error variance on observed scale (at least partly)

3. Heritability for unbalanced data

Foulley et al (1987)

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \overline{\lambda}^{-1}}$$

$\overline{\lambda}$ = average of the Poisson parameter λ across observations
(assumed no random effects here for simplicity only)

This is answer to question:

What is the value of a variance for an imaginative error term on the link scale that leads to the same variance on the back-transformed scale as that of a Poisson random variable with expectation parameter?

3. Heritability for unbalanced data

Sketch of derivation (for over-dispersed Poisson: variance = $\phi\lambda$)

$$\eta^* = \eta + e$$

What is a suitable variance of e (σ_e^2) on the link scale, such that the conditional variance of $g^{-1}(\eta^*)$ given η, $\mathrm{var}\left[g^{-1}(\eta^*)\,|\,\eta\right]$, approximately equals that of the observed data y?

$$\mathrm{var}\left[g^{-1}(\eta^*)\,|\,\eta\right] = \mathrm{var}\left[\exp(\eta^*)\,|\,\eta\right] \approx \left(\frac{\partial\lambda^*}{\partial\eta^*}\right)^2_{\eta^*=\eta} \times \sigma_e^2$$

$$= \left[\exp(\eta^*)\right]^2_{\eta^*=\eta} \times \sigma_e^2 = \lambda^2\sigma_e^2 = \phi\lambda$$

$$\Rightarrow\ \sigma_e^2 = \phi\lambda^{-1}$$

(Delta method)

3. Heritability for unbalanced data

Over-dispersed binomial distribution

$$\text{var}\left[g^{-1}\left(\eta^*\right)|\,\eta\right] = \text{var}\left[\Phi\left(\eta^*\right)|\,\eta\right] \approx \left[\frac{\partial\Phi(\eta)}{\partial\eta}\right]^2_{\eta^*=\eta} \times \sigma_e^2 = \left[\varphi\left(\eta^*\right)\right]^2_{\eta^*=\eta} \times \sigma_e^2$$

$$= \left[\varphi(\eta)\right]^2 \sigma_e^2 = \frac{\phi\pi(1-\pi)}{n}$$

$$\Rightarrow \sigma_e^2 = \frac{\phi\pi(1-\pi)}{\left[\varphi(\eta)\right]^2 n}$$

(Bennewitz et al., 2013)

4. Predictive accuracy

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 (or most of the) markers

(Meuwissen et al., 2001)

4. Predictive accuracy

Key idea of genomic prediction (genomic selection)

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

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

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

u_k = regression coefficients

Example: Biallelic marker (SNP) with alleles A_1 and A_2, DH lines:

$z_{ik} = 1$ for A_1A_1

$z_{ik} = -1$ for A_2A_2

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

4. Predictive accuracy

Genomic prediction

$$\boldsymbol{g} = \boldsymbol{Zu}$$

$\boldsymbol{Z}$ = $\{z_{ik}\}$ = marker (SNP) covariate design matrix
$\boldsymbol{u}$ = vector of random SNP effects u_k

- Estimate $\boldsymbol{u}$ from training dataset with phenotyped genotypes
- Predict $\boldsymbol{g}$ for unphenotyped genotypes
- There are many alternative models / methods to predict $\boldsymbol{g}$
  - RR-BLUP / G-BLUP, Bayesian methods (ABC…)
  - Machine learning methods, Artificial neural networks (ANNs)
  - Reproducing Kernel Hilbert Spaces (RKHS) etc.
- Very successful in animal breeding
- Increasingly popular in plant breeding

4. Predictive accuracy

The basic ridge-regression (RR) BLUP model:

$$\boldsymbol{p} = \mathbf{1}_n \mu + \boldsymbol{Zu} + \boldsymbol{e}$$

where

$\boldsymbol{p}$ = adjusted genotype means (2-stage approach) = observed phenotype
$\mathbf{1}_n$ = n-vector of ones
μ = common intercept
$\boldsymbol{Z}$ = n x p matrix containing the SNP marker information
$\boldsymbol{u}$ = random SNP marker effects with $\boldsymbol{u} \sim N\left(\mathbf{0}, \boldsymbol{I}_p \sigma_u^2\right)$
$\boldsymbol{e}$ = residual error associated with $\boldsymbol{p}$, with $\boldsymbol{e} \sim N\left(\mathbf{0}, \boldsymbol{I}_n \sigma_e^2\right)$

$$\Rightarrow \quad \mathrm{var}\left(\boldsymbol{g} = \boldsymbol{Zu}\right) = \boldsymbol{G} = \boldsymbol{ZZ}^{\boldsymbol{T}} \sigma_u^2$$

4. Predictive accuracy

k-fold cross-validation

- Split data into k parts (folds)
- Use $k-1$ parts for estimation of model
- Use k-th part for validation

Predictive ability

- Correlation between $\hat{\boldsymbol{g}}$ and $\boldsymbol{p}$: $r_{p,\hat{g}}$

Predictive accuracy

- Correlation between $\hat{\boldsymbol{g}}$ and $\boldsymbol{g}$: $r_{g,\hat{g}}$

4. Predictive accuracy

Estimation of predictive accuracy

$$r_{g,\hat{g}} = \frac{r_{p,\hat{g}}}{\hat{H}}$$

(e.g. Dekkers, 2007)

4. Predictive accuracy

Rationale:

Assume $s_{g,\hat{g}} = s_{p,\hat{g}}$ and $H^2 = \frac{s_g^2}{s_p^2} \Rightarrow s_g^2 = H^2 s_p^2 \Rightarrow$

$$r_{g,\hat{g}} = \frac{s_{g,\hat{g}}}{\sqrt{s_g^2 s_{\hat{g}}^2}} = \frac{s_{p,\hat{g}}}{\sqrt{s_g^2 s_{\hat{g}}^2}} = \frac{s_{p,\hat{g}}}{\sqrt{H^2 s_p^2 s_{\hat{g}}^2}} = \frac{s_{p,\hat{g}}}{\sqrt{s_p^2 s_{\hat{g}}^2}} \times \frac{1}{H} = \frac{r_{\hat{g},p}}{H}$$

where

$$r_{\hat{g},p} = \frac{s_{\hat{g},p}}{\sqrt{s_{\hat{g}}^2 s_p^2}}$$

4. Predictive accuracy

Problems

- Many ways to estimate H!
- Models for estimating H not always commensurate with RR-BLUP
- Equations for r assume i.i.d. sampling (independent errors)

4. Predictive accuracy

Method 1 $$H^2_{\mathrm{m1}} = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2 / r}$$

Method 2 $$H^2_{m2} = \frac{\sigma_g^2}{\sigma_g^2 + \bar{v}/2}$$

Method 3 $$H^2_{m3} = 1 - \frac{\bar{v}_{\mathrm{BLUP}}}{2\sigma_g^2}$$

4. Predictive accuracy

Method 4

$$H^2_{m4} = \frac{E\left(s_g^2\right)}{E\left(s_p^2\right)}$$

$$s_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(g_i - \bar{g}\right)^2 = \boldsymbol{g}^{\boldsymbol{T}}\boldsymbol{P}_{\boldsymbol{u}}\boldsymbol{g} \quad \text{with} \quad \boldsymbol{P}_{\boldsymbol{u}} = \frac{1}{n-1}\left(\boldsymbol{I}_{\boldsymbol{n}} - \frac{1}{n}\boldsymbol{J}_{\boldsymbol{n}}\right)$$

$$s_p^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(p_i - \bar{p}\right)^2 = \boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{P}_{\boldsymbol{u}}\boldsymbol{p}$$

For example: $E\left(s_g^2\right) = trace\left(\boldsymbol{P}_{\boldsymbol{u}}\boldsymbol{G}\right)$, where $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}^{\boldsymbol{T}}\sigma_u^2$

$\Rightarrow$ Models for genomic selection and H are commensurate

4. Predictive accuracy

Method 5

$$E\left(r_{g,\hat{g}}\right) \approx \frac{E\left(s_{g,\hat{g}}\right)}{\sqrt{E\left(s_g^2\right)E\left(s_{\hat{g}}^2\right)}}$$

Plug in estimates of the three expected values

4. Predictive accuracy

Method 7

$$\rho_i^2 = \frac{[\mathrm{cov}(g_i, \hat{g}_i)]^2}{\mathrm{var}(g_i)\mathrm{var}(\hat{g}_i)}$$

$\Rightarrow$ estimate from mixed model equations (MME)

$$\hat{\rho}_{m7}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}_i^2$$

(Mrode and Thompson, 2005; Piepho and Möhring, 2007)

4. Predictive accuracy

Simulation of datasets

- Consider a single trial laid out as an α-design
- Simulate block and plot effects using the marker and error variances, obtained from real datasets (AgReliant, KWS)
- The true breeding values simulated as $\boldsymbol{g} = \boldsymbol{Zu}$ using $\boldsymbol{Z}$ and σ_u^2 from real data
- The phenotypic data was calculated as:
  genetypic value + rep + rep.block + plot error
- The correlation between the true and the predicted breeding value ($r_{g,\hat{g}}$) used as a benchmark
- Use Methods 1-7 to estimate $r_{g,\hat{g}}$

4. Predictive accuracy

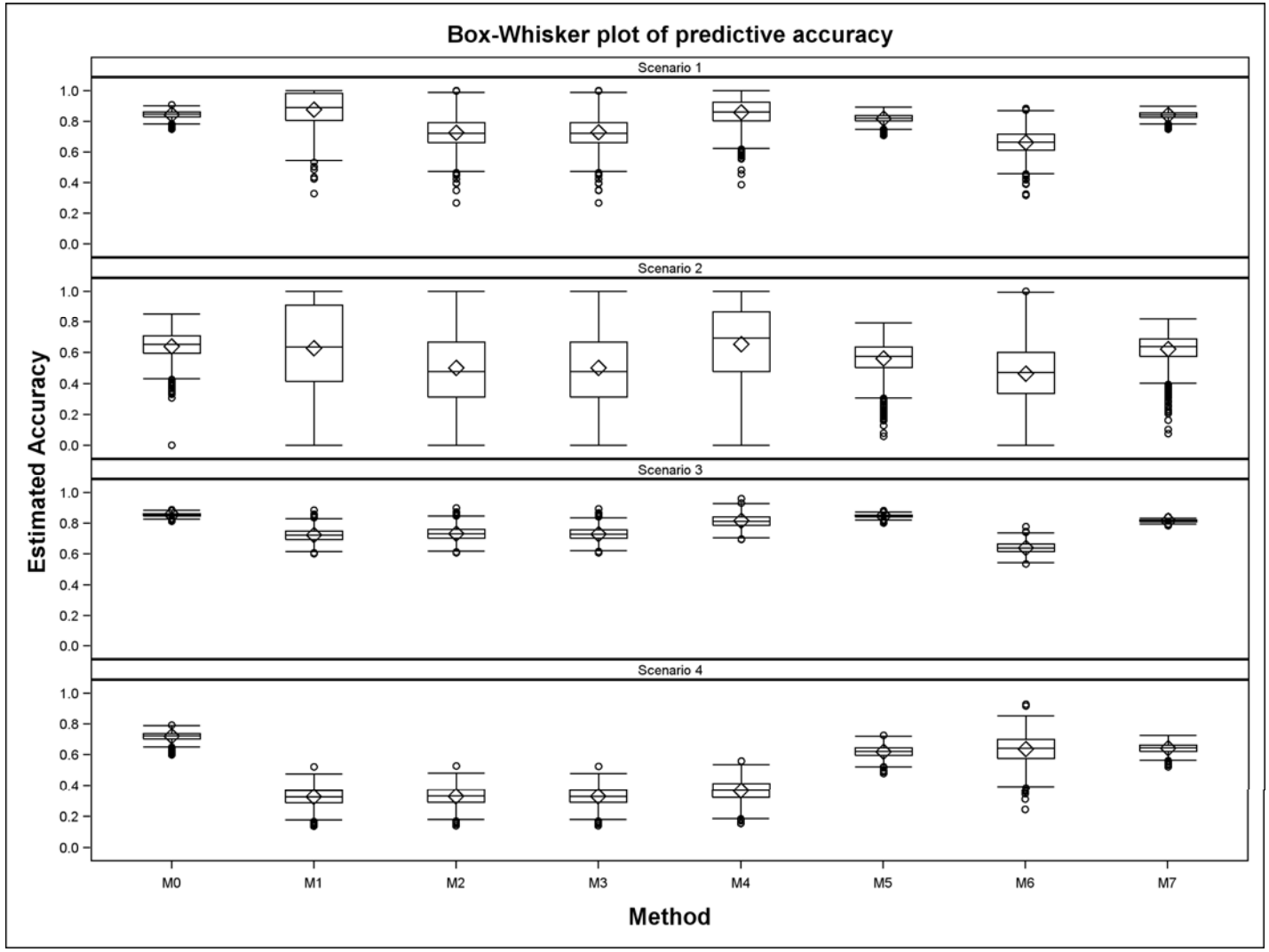


Figure: Predictive accuracy (estimates less than 0 were set to 0 whereas estimates greater than 1 were set to 1) for all the seven methods in each of the four scenarios.

4. Predictive accuracy

Table: The means of the estimated heritability for all simulated datasets. M0 is the square of the true correlation between the predicted and the true simulated breeding values.

	Methods					
Scenario	M0	M1	M2	M3	M4	M5
1	0.71	0.32	0.48	0.48	0.34	0.67
2	0.42	0.09	0.15	0.18	0.08	0.33
3	0.73	0.51	0.49	0.50	0.40	0.72
4	0.52	0.14	0.13	0.13	0.07	0.39

5. Summary

Heritability

- Heritability can be estimated by several *ad hoc* methods
- Response to selection can be approximated by plugging in *ad hoc* estimates
- But in more complex cases approximation poor or not available
- Can simulate any statistic of interest by simulation (parametric bootstrap)

Predictive accuracy (PA)

- Indirect methods use estimator of H in denominator
- None of the available estimators of H works very well for estimating PA
- Direct methods for estimating PA work best (New Method 5 and Method 7 from animal breeding preferred)

6. References

Bennewitz, J., Böglein, S., Stratz, P., Rodehutscord, M., Piepho, H.P., Kjaer, W., Bessei, W. (2014): Genetic parameters for feather pecking and aggressive behaviour in a large F2-cross of laying hens using generalized linear mixed models. *Poultry Science* **93**, 810-817.

Estaghvirou, B., Ogutu, J.O., Schulz-Streeck, T., Knaak, C., Ouzenova, M., Gordillo, A., Piepho, H.P. (2013): Evaluation of approaches for estimating prediction accuracy in genomic selection in plant and animal breeding. *BMC Genomics* **14**, 860.

Estaghvirou, B., Ogutu, J.O., Piepho, H.P. (2014): Influence of outliers on accuracy and robustness of genomic prediction. *G3* (online)

Piepho, H.P., Möhring, J. (2007): Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**, 1881-1888.

Thanks!