Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes

S. Maenhout\textsuperscript{1} \quad B. De Baets\textsuperscript{2} \quad G. Haesaert\textsuperscript{1}

\textsuperscript{1}Department of Plant Production
University College Ghent, Belgium

\textsuperscript{2}Department of Applied Mathematics, Biometrics and Process Control
Ghent University, Belgium

XIV meeting of the Biometrics in Plant Breeding Section
1 Introduction

2 WAIS estimator
   - Properties
   - Simulation study
   - Maize breeding data

3 CoCoa
1 Introduction

2 WAIS estimator
   - Properties
   - Simulation study
   - Maize breeding data

3 CoCoa
The coefficient of coancestry (CoC) between two individuals $i$ and $j$ is defined as the probability that at an allele drawn from both $i$ and $j$ at the same locus is identical by descent (ibd) from a recent common ancestor.

- is often used to model the covariance between polygenic backgrounds in breeding value estimation or association studies
- can be estimated from pedigree information (e.g. tabular method)
  - requires pedigree knowledge tracing back to the natural populations
  - assumes equal contribution of both parents (absence of selection, genetic drift, . . . )
The CoC can be estimated from similarities in molecular marker fingerprints but:

- an allele identity does not necessarily reflect an identity by descent
- assumptions made by estimators from population genetics are often violated in breeding pools
  - knowledge of allele frequencies
  - linkage equilibrium between markers
  - Hardy-Weinberg equilibrium
- resulting estimators do not always lie within the unit interval \([0, 1]\)
- resulting coancestry matrices \(A\) are not necessarily positive semi-definite
A matrix is psd if all its eigenvalues are greater or equal to zero or equivalently if

$$\mathbf{v}' \mathbf{A} \mathbf{v} \geq 0, \quad \forall \mathbf{v} \neq \mathbf{0}.$$  

- every variance-covariance matrix is psd by definition!
- if $\text{Var}(\mathbf{g}) = 2\sigma^2 \mathbf{g}$ then CoC matrix $\mathbf{A}$ must be psd as well
- linear mixed model packages rely on the psd property of user supplied coancestry matrices
Outline

1. Introduction

2. WAIS estimator
   - Properties
   - Simulation study
   - Maize breeding data

3. CoCoea
**Calculation of WAIS**

\[ A^{WAIS} = \frac{1}{4I} (XX' + Q), \]

- **W** is a diagonal matrix containing a weight for each allele \( z \):
  \[ \omega_z = P(\alpha_i = \alpha_j \mid \alpha_i = z, \alpha_j = z) \]
- **Q** diagonal matrix containing a correction factor for each genotype \( i \):
  \[ q_i = \sum_{z=1}^{p} x_{(i,z)}^2 y_{i,z} \]
The WAIS estimator is ideally suited for use in hybrid breeding programs:

- corrects the observed allele identities by means of allele-dependent weights
- weights are estimated from observed allele identities between unrelated genotypes (cf. Bernardo, 1993)
- allows for a mix of inbred and non-inbred genotypes
- all coancestry estimates are guaranteed to lie within the unit interval
- always produces psd coancestry matrices
- achieves a similar or better linear mixed model fit compared to other marker-based estimators
Simulated breeding data

Studying the behaviour of WAIS in breeding pools requires simulation of the entire breeding history including:

- artificial phenotypical selection
- linkage disequilibrium between marker loci
- varying genotype relatedness

We simulated 8 breeding cycles (cf. Stich et al. 2007, Hohenheim University maize breeding program):

- 250 QTL, allele number $\sim \text{Pois}(\lambda+2)$, $\lambda = 0 \ldots 12$
- allele frequencies in subpopulations $\sim \text{Dir}(a)$ where $a = \{p(\frac{1-Fst}{Fst})\}$ (Balding, 2003)
- 101 SSR markers (RAGT R2n linkage map)
After each breeding cycle the following CoC estimators were calculated:

- the actual ibd CoC
- pedigree-based CoC
- uncorrected alikeness in state (AIS)
- WAIS
- BNO (Bernardo, 1993)
- MLE, maximum likelihood estimator (Thompson, 1975)
- LOI (Loiselle et al., 1995)
Bias and variance

Average root mean squared error over 100 iterations:

- Fst: 0.1
- Fst: 0.2
- Fst: 0.3
- Fst: 0.4

λ: 4

λ: 8

λ: 12
Proportion of non-psd matrices

MLE

LOI

Number of breeding cycles

Proportion

\( \lambda = 0 \) ○

\( \lambda = 4 \) △

\( \lambda = 8 \) +

\( \lambda = 12 \) ×

Steven Maenhout, Bernard De Baets, Geert Haesaert

WAIS: Weighted Alikeness In State
Application to maize breeding data

The pedigree and five marker-based CoC measures were used to model the covariance of GCA and SCA components in a linear mixed model analysis of a subset of the phenotypic evaluation data of the maize breeding program of RAGT R2n:

- 2,367 hybrids
- 92 Iodent and 105 Iowa Stiff Stalk Synthetic lines genotypes with 75 SSR markers
- 40,432 plots
- 1,280 multi-environment trials
- 110 locations spread over Europe
- 3 traits: grain yield, grain moisture content, days until flowering
Range of pairwise CoC estimates
### Restricted log-likelihoods

<table>
<thead>
<tr>
<th></th>
<th>yield</th>
<th>moisture %</th>
<th>flowering</th>
</tr>
</thead>
<tbody>
<tr>
<td>PED</td>
<td>$-222740.1 \ (4)$</td>
<td>$-194696.6 \ (1)$</td>
<td>$-55339.6 \ (1)$</td>
</tr>
<tr>
<td>AIS</td>
<td>$-222734.8 \ (2)$</td>
<td>$-194710.8 \ (2)$</td>
<td>$-55343.8 \ (2)$</td>
</tr>
<tr>
<td>BNO</td>
<td>$-222734.8 \ (1)$</td>
<td>$-194712.9 \ (3)$</td>
<td>$-55344.1 \ (3)$</td>
</tr>
<tr>
<td>WAIS</td>
<td>$-222739.2 \ (3)$</td>
<td>$-194715.3 \ (4)$</td>
<td>$-55347.7 \ (4)$</td>
</tr>
<tr>
<td>MLE</td>
<td>$-222743.2 \ (6)$</td>
<td>$-194716.2 \ (5)$</td>
<td>$-55357.0 \ (5)$</td>
</tr>
<tr>
<td>LOI</td>
<td>$-222741.0 \ (5)$</td>
<td>$-194725.6 \ (6)$</td>
<td>$-55361.9 \ (6)$</td>
</tr>
</tbody>
</table>

- BNO and LOI values were bounded within the unit interval
- MLE and the bounded LOI matrices were bended towards the closest psd matrix
Outline

1. Introduction
2. WAIS estimator
   - Properties
   - Simulation study
   - Maize breeding data
3. CoCoa
CoCoa: software for estimating the coefficient of coancestry from multilocus genotype data

- provides AIS, WAIS, BNO, MLE and LOI estimators
- uses the same input file format as Structure (Pritchard et al. 2000)
- allows to invert well-conditioned CoC matrices
- provides 2 bending routines for bending non-psd or (nearly) singular matrices
- allows to export (inverse) CoC matrices in file formats used by ASReml, SAS Proc Mixed, Wombat, Matlab, ...
- available for Linux and Windows OS
- freely available, open source (GNU General Public License)
CoCoa user interface
CoCoa homepage

http://webs.hogent.be/cocoa

provides:

- Windows installer
- 32 and 64 bit binaries for Linux
- manual
- C++ and Java source code
- ...
Summary

WAIS:
- is a new, marker-based CoC estimation procedure ideally suited for use in hybrid breeding programs
- estimators always lie within the unit interval
- resulting CoC matrices are guaranteed to be psd
- WAIS performs best in advanced cycle breeding pools

CoCoa:
- is a free software tool for estimating the CoC from molecular marker data
- provides 5 estimation procedures: AIS, WAIS, BNO, MLE and LOI
- provides several useful matrix manipulation tools: bending, bounding, inversion, ...