Indirect predictions based on SNP effects from ssGBLUP in large genotyped populations

Daniela Lourenco
A. Legarra, S. Tsuruta, Y. Masuda, T. Lawlor, I. Misztal

ADSA 2018
Why Indirect predictions (IP)?

- Interim evaluations
  - Between official runs

- Not all genotyped animals are in the evaluations
  - Animals with incomplete pedigree increase bias and lower $R^2$

- Commercial products
  - e.g. GeneMax for non-registered animals
Indirect predictions in ssGBLUP

\[
\begin{bmatrix}
X'X & X'W \\
W'X & W'W + H^{-1}\lambda
\end{bmatrix}
\begin{bmatrix}
\hat{b} \\
\hat{u}
\end{bmatrix} =
\begin{bmatrix}
X'y \\
W'y
\end{bmatrix}
\]

\[
H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}
\]

\[
\hat{a} = \lambda D Z' G^{-1} \hat{u}
\]

GEBV_{young} = w_1PA + w_2GP - w_3PP

GEBV_{young} \approx GP

\text{COR(GEBV}_{young}, Z\hat{a}) > 0.99

GEBV_{young} \approx GP = Z\hat{a}

Lourenco et al., 2015
Problems with Indirect predictions

\[ \text{COR}(\hat{\text{GEBV}}_{\text{young}}, \hat{Z}) > 0.99 \]

\[ \text{Avg}(\hat{\text{GEBV}}) \approx 100 \quad \neq \quad \text{Avg}(\hat{Z}) \approx 0 \]
Objectives

1) Fine-tune indirect predictions to be compatible with GEBV

2) Investigate whether SNP effects are accurate when APY is used
   • Possibly use subset of core animals
Dataset

- US Holstein 2014 type data
  - 8.3M animals in pedigree
  - No UPG
  - 9.2M Udder Depth (UD)
    - \( h^2 = 0.33 \)
  - 9.2M Foot Angle (FA)
    - \( h^2 = 0.11 \)
  - 105k genotyped
    - Training: 100k
    - Validation: 5k
- Complete
  - Phenotypes up to 2010
  - Genotypes up to 2010 (105k)
- Reduced
  - Phenotypes up to 2010
  - NO Genotypes for validation (100k)
- SNP effects and IP from Reduced
  - Compare with GEBV from Complete
Accuracy of SNP effects

\[ \hat{a}_G = \lambda D Z' G^{-1} \hat{u} \]

\[ \hat{a}_{G_{APY}}^{-1} = \lambda D Z' G_{APY\_high\_reliability}^{-1} \hat{u}_{APY} \]

\[ \hat{a}_{G_{APY}R}^{-1} = \lambda D Z' G_{APY\_random}^{-1} \hat{u}_{APY} \]

\[ \hat{a}_{G_{cc}}^{-1} = \lambda D Z' G_{cc\_high\_reliability}^{-1} \hat{u}_{APY} \]

\[ \hat{a}_{G_{cc}R}^{-1} = \lambda D Z' G_{cc\_random}^{-1} \hat{u}_{APY} \]
Statistics for SNP effects - $G_{APY}^{-1}$

Udder Depth

Foot Angle
Statistics for SNP effects - $G_{CC}^{-1}$

Udder Depth

Foot Angle
Statistics for $\mathbf{Z\hat{a}} - \mathbf{G}^{-1}_{CC}$

<table>
<thead>
<tr>
<th></th>
<th>$G^{-1}$</th>
<th>0.99</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_{CC_High}$</td>
<td>$G^{-1}_{CC_High}$</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>$G_{CC_Rand}$</td>
<td>$G^{-1}_{CC_Rand}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Udder Depth

Foot Angle
Fine-tuning indirect predictions from ssGBLUP

Understanding genetic and genomic bases

• Base of BLUP: *founders of the pedigree*

• Base of GBLUP: *genotyped* animals

• Base of SSGBLUP: Vitezica et al. (2011) modeled as a mean for genotyped

  • $p(u_g) = N(1\mu, G)$

  • $\mu = \text{(Pedigree base)} - \text{(Genomic base)}$
Fine-tuning indirect predictions from ssGBLUP

1) Formula in Legarra (2017)
\[
\hat{u}_{ip} = \hat{\mu} + 0.95 \hat{Z}\hat{a} + 0.05 \hat{u}_{parents}
\]

2) Double fitting
   a) fit a regression using genotyped animals in the evaluation
   \[
   \text{GEBV}_{eval} = b_0 + b_1 \hat{Z}\hat{a}
   \]
   b) apply regression for indirectly predicted animals
   \[
   \hat{u}_{ip} = b_0 + b_1 \hat{Z}\hat{a}
   \]

3) Add average GEBV
   \[
   \hat{u}_{ip} = \frac{\text{GEBV}_{eval}}{\hat{Z}\hat{a}}
   \]
Bias of indirect predictions

Udder Depth

Foot Angle
Correlation between GEBV and indirect predictions

Correlation(GEBV, u_p)

- Za: 0.98
- Legarra_2017: 0.98
- Double_Fitting: 0.98
- Average_GEBV: 0.98

FA: blue
UD: green
Fine-tuning indirect predictions in ssGBLUP

\[ E(\hat{u}|\hat{a}) = \mu + Z \frac{1}{2 \sum p(1-p)} \left( I \frac{1}{2 \sum p(1-p)} \right)^{-1} (\hat{a} - 0) \]

\[ E(\hat{u}|\hat{a}) = \mu + Z\hat{a} \]

\[ \approx \]

\[ E(\hat{u}|\hat{a}) = GEBV + Z\hat{a} \]
Final Remarks

• SNP effects can be calculated based on APY $G^{-1}$ or core $G^{-1}$
  • Slightly less accurate with core

• Indirect predictions based on core animals are accurate
  • Reduction in computing time compared to $G^{-1}$
  • Similar computing time as APY $G^{-1}$

• Indirect predictions are unbiased after corrections
  • Average GEBV, Legarra (2017) or double fitting
  • Can be used as interim evaluation
Acknowledgements