Overall Accuracy of Prediction in Cross-Validation

Jian Zeng

April 3, 2014
Cross-Validation

- Cross-Validation
  - Rotation estimation

<table>
<thead>
<tr>
<th>Fold 1</th>
<th>T</th>
<th>T</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 2</td>
<td>T</td>
<td>V</td>
<td>T</td>
</tr>
<tr>
<td>Fold 3</td>
<td>V</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

- Applications
  - Breeding value prediction
  - **Accuracy of prediction**
  - Grid search for hyperparameters
Accuracy of Prediction

• Pearson correlation coefficient between GEBV and phenotypes (for fold $i$ of CV)

$$r_i = \text{corr}(\hat{g}_i, y_i) = \frac{\text{cov}(\hat{g}_i, y_i)}{\sqrt{\text{var}(\hat{g}_i)} \sqrt{\text{var}(y_i)}}$$

• Basic assumptions – Independent pairs of data

$$\begin{bmatrix} y \\ \hat{g} \end{bmatrix} \sim iid \begin{bmatrix} \mu_y \\ \hat{\mu}_g \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \sigma_{y\hat{g}} \\ \sigma_{\hat{g}y} & \sigma_{\hat{g}}^2 \end{bmatrix}$$

• Likely to be violated if
  – Systematic effects exist
  – Animals are related
Accuracy of Prediction

• For the data set as a whole
  – Problem may become more severe when data is partitioned based on some clustering criteria
  – Means, variances and covariances may be different between contemporary groups
Question

- How to estimate the overall accuracy of prediction?
  - Simple correlation as in a single fold may not be good
  - Genetic correlation from pooled data
  - Saatchi et al. (2011, 2012)
Pooled Genetic Correlation

• Overall accuracy

\[ r = \frac{\text{cov}(\hat{g}, g)}{\text{var}(\hat{g}) \text{ var}(g)} \]

• Variance and covariance components, assuming to be the same for all groups, are estimated from a bivariate animal model
Bivariate Animal Model

\[
\begin{bmatrix}
\hat{y}
\end{bmatrix} = \begin{bmatrix}
X_1 & 0
\end{bmatrix} \begin{bmatrix}
b_1
\end{bmatrix} + \begin{bmatrix}
X_2
\end{bmatrix} \begin{bmatrix}
b_2
\end{bmatrix} + \begin{bmatrix}
Z_1 & 0
\end{bmatrix} \begin{bmatrix}
a_1
\end{bmatrix} + \begin{bmatrix}
e_1
\end{bmatrix} + \begin{bmatrix}
Z_2
\end{bmatrix} \begin{bmatrix}
a_2
\end{bmatrix} + \begin{bmatrix}
e_2
\end{bmatrix}
\]

**Fixed effects**: Account for differences between groups in mean

**Random effects – breeding values**: Account for differences between groups in variance and covariance

\[
\text{var} \begin{bmatrix}
a_1
\end{bmatrix} = A \otimes G = \begin{bmatrix}
A\sigma_{a_1}^2 & A\sigma_{a_1a_2}
\end{bmatrix}
\]

\[
G = \begin{bmatrix}
\sigma_{a_1}^2 & \sigma_{a_1a_2}
\sigma_{a_1a_2} & \sigma_{a_2}^2
\end{bmatrix}
\]

\[
\text{var} \begin{bmatrix}
e_1
\end{bmatrix} = \begin{bmatrix}
W\sigma_e^2 & 0
0 & I\sigma_e^2
\end{bmatrix}
\]

\[
r = \frac{\sigma_{a_1a_2}}{\sigma_{a_1}\sigma_{a_2}}
\]
What is A?

A is the block diagonal pedigree relationship matrix, e.g. in a 3-fold CV

\[
A = \begin{bmatrix}
A_{11} & 0 & 0 \\
0 & A_{22} & 0 \\
0 & 0 & A_{33}
\end{bmatrix}
\]

- The off-block diagonals are zeroed out such that, in REML, variances and covariance are estimated by pooling the information from each group (the overall likelihood is the product of that for each group)

- Or an intuitive explanation is in cross-validation the estimation of random effects in one group do not depend on that in other groups
Example

<table>
<thead>
<tr>
<th>Method</th>
<th>Simple Correlation</th>
<th>Pooled Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.473</td>
<td>0.445</td>
</tr>
<tr>
<td>2</td>
<td>0.417</td>
<td>0.473</td>
</tr>
<tr>
<td>3</td>
<td>0.410</td>
<td>0.390</td>
</tr>
<tr>
<td>4</td>
<td>0.413</td>
<td>0.389</td>
</tr>
</tbody>
</table>

**Correlation** 0.411

- The simple correlation for cross-validation may be bias.
- It is suggested to use the pooled correlation.