To have your steak and eat it: Genetic principal component analysis for beef cattle data

Karin Meyer

Animal Genetics and Breeding Unit, University of New England, Armidale
kmeyer@didgeridoo.une.edu.au

Motivation

- Multiple, correlated random effects
  - several traits, random regression coefficients
- Covariance matrix generally assumed ‘unstructured’
  - \(k\) variables \(\rightarrow k(k+1)/2\) covariances
- Recent interest in imposing ‘structure’ \(\rightarrow\) parsimony
  - Constrain selected components or their functions
  - Variance function + parametric correlation structure
    - auto-regressive, structured ante-dependence, etc. (Gilmour & Thompson, 2006)
  - Alternative: parameterisation based on
    - eigen-decomposition \(\rightarrow\) principal components (PCs)
    - factor analytic structure (e.g. Jennrich & Schluchter, 1986)
    - reduced rank models

Objectives

- So far: Two-step procedure
  - Estimate unstructured covariance matrix \(\rightarrow\) decompose
  - Transform data to PCs (phenotypic SS/CP) \(\rightarrow\) estimate parameters of new 'traits'
- Better: Directly estimate leading PCs only
  - feasible within standard linear mixed model framework
  - requires simple re-parameterisation only

This paper

- Review direct estimation of leading principal components
- Show application to beef cattle carcass traits
### Introduction

- **Dimension reduction**
- **Factor analysis**

### Basics of Principal Components

- **Dimension reduction**
- **Factor analysis**

- **Set of** $k$ correlated variables $v$ with covariance matrix $\Sigma$
  - traits
  - random regression coefficients
- **Principal components** are the set of $k$ variables which are
  - linear functions of original effects $v$
  - uncorrelated with each other
  - successively explain maximum variation
- **Eigen-decomposition**: $\Sigma = E \Lambda E' = \sum_{i=1}^{k} \lambda_i e_i' e_i$
  - $E E' = I$
  - assume $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$
  - eigenvector $e_i$ gives direction $\rightarrow P_i = e_i' v$
  - eigenvalue $\lambda_i$ gives variance explained

### Toy example

$\Sigma = \begin{pmatrix} 2 & -1.05 \\ -1.05 & 1 \end{pmatrix}$

$\Lambda = \begin{pmatrix} 2.66 & 0 \\ 0 & 0.34 \end{pmatrix}$

### Dimension reduction

- **Principal components**
  - summarise information
  - widely used to reduce dimensions $\rightarrow$ no. variables
- $P_i$ explains maximum variation given $P_1, \ldots, P_{i-1}$
- $\text{Var}(P_{m+1}) = \lambda_{m+1}$ close to zero
  - $P_{m+1}, \ldots, P_k$ provide negligible information
  - $P_{m+1}, \ldots, P_k$ can be ignored
  - Dimension reduced from $k$ to $m$
- Consider first $m$ PCs only $\rightarrow \Sigma^* = \sum_{i=1}^{m} \lambda_i e_i' e_i = E_m \Lambda_m E'_m$
  - $\Sigma^*$ has reduced rank $m$
  - $\Sigma^*$ has $m(2k - m + 1)/2$ parameters
  - not $m + mk$ as $e_i' e_i = 1$ and $e_i' e_j = 0$
Factor analysis

Different concept
- PCA → identify variables explaining maximum variance
- FA → find common factors which explain covariances

Fit latent model: \( v = Fz + \epsilon \)
- \( F = E_m A_m^{1/2} \)
- \( \text{Var}(z) = I_m \)
- \( \text{Var}(\epsilon) = \Psi = \text{Diag}\{\sigma_i^2\} \)
  - \( \sigma_i^2 \): specific variances \( i = 1, \ldots, k \)
- \( \text{Var}(v) = \Sigma^+ = E_m A_m E_m^T + \Psi = \Sigma^* + \Psi \)
  - \( \Sigma^* \) generally has full rank \( k \)
  - \( \Sigma^* \) involves \( m(2k - m + 1)/2 + k \) parameters
  - \( \leq k(k + 1)/2 \) → limit on \( m \)

Factor analysis

- Different concept
- PCA → identify variables explaining maximum variance
- FA → find common factors which explain covariances

Fit latent model: \( v = Fz + \epsilon \)
- \( F = E_m A_m^{1/2} \)
- \( \text{Var}(z) = I_m \)
- \( \text{Var}(\epsilon) = \Psi = \text{Diag}\{\sigma_i^2\} \)
  - \( \sigma_i^2 \): specific variances \( i = 1, \ldots, k \)
- \( \text{Var}(v) = \Sigma^+ = E_m A_m E_m^T + \Psi = \Sigma^* + \Psi \)
  - \( \Sigma^* \) generally has full rank \( k \)
  - \( \Sigma^* \) involves \( m(2k - m + 1)/2 + k \) parameters
  - \( \leq k(k + 1)/2 \) → limit on \( m \)

Fit latent model: \( v = Fz + \epsilon \)
- \( F = E_m A_m^{1/2} \)
- \( \text{Var}(z) = I_m \)
- \( \text{Var}(\epsilon) = \Psi = \text{Diag}\{\sigma_i^2\} \)
  - \( \sigma_i^2 \): specific variances \( i = 1, \ldots, k \)
- \( \text{Var}(v) = \Sigma^+ = E_m A_m E_m^T + \Psi = \Sigma^* + \Psi \)
  - \( \Sigma^* \) generally has full rank \( k \)
  - \( \Sigma^* \) involves \( m(2k - m + 1)/2 + k \) parameters
  - \( \leq k(k + 1)/2 \) → limit on \( m \)

Reparameterising the linear mixed model

- 'Standard', full rank model
  \[ y = Xb + Zu + \epsilon \quad \text{with} \quad \text{Var}(u) = \Sigma \otimes A \]

- Reparameterise
  \[ y = Xb + Z (Q \otimes I_N) (Q^{-1} \otimes I_N) u + \epsilon \]
  \[ = Xb + Z^* u^* + \epsilon \]
- For \( Q = E \) → equivalent models
  - \( u^* \) → vector of (genetic) PCs
  - \( \text{Var}(u^*) = \Lambda \otimes A \)
- For \( Q = E_m \) → fit leading \( m \) PCs only
  - \( u^* \) has \( m \) elements per animal
  - backtransform: \( \hat{u} = (E_m \otimes I) \hat{u}^* \)

Reduced rank estimation

Alternative forms for variance component estimation

- \( Q = E_m A_m^{1/2} \)
  - \( \text{Var}(u^*) = I_m \otimes A \)
  - FA model with zero specific variances
  - Linear equations determine elements given by orthogonality constraints on \( E \)
  - Estimate \( \hat{\lambda}_i = \hat{q}_i q_i^T \)

- \( Q = L_m \)
  - \( \Sigma = LL' \) → Cholesky factor
  - Singular value decomp. \( L = E \Lambda^{1/2} T \) \( \text{(e.g. Harville, 1997)} \)
  - Estimate \( P_1 \) to \( P_m \) of \( \Sigma \) estimate columns 1 to \( m \) of \( L \)
    - \( TT' = I \) → orthogonal rotation of parameter space
    - non-zero elements → correct no. of parameters
    - Cholesky form → good convergence rates
### REML estimation for PC model

\[ \mathbf{y} = \mathbf{Xb} + \mathbf{Z}^* \mathbf{u}^* + \mathbf{\epsilon} \]

- Standard REML algorithms readily adapted
  - Parameters to be estimated part of design matrix
    \[ \frac{\partial \mathbf{Z}^*}{\partial q_{ij}} = \mathbf{Z} \left( \frac{\partial \mathbf{Q}}{\partial q_{ij}} \otimes \mathbf{I}_N \right) \]
- ‘Average information’ REML
  - Thompson *et al.* (2003) → invert coefficient matrix MME
  - Meyer & Kirkpatrick (2005) → automatic differentiation
- Expectation-Maximisation
  - ‘Parameter Expanded’ (PX-EM) → same form of reparameterisation of standard model
  - Reversed rôles of auxiliary & 'main' parameters
  - PX-EM algorithm (Foulley & van Dyk, 2000) almost directly gives estimators for PC model

### Traits

14 ‘carcass’ traits in genetic evaluation of beef cattle
- 6 carcass traits *per se* → report breeding values
- 8 live ultra-sound scan traits

**Measured on live animals**
- *Heifers or steers*
  - 7 Eye muscle area \( H.EMA \)
  - 8 Intra-muscular fat \( H.IMF \)
  - 9 Rump fat depth \( H.P8 \)
  - 10 Rib fat depth \( H.RIB \)
- *Bulls*
  - 11 Eye muscle area \( B.EMA \)
  - 12 Intra-muscular fat \( B.IMF \)
  - 13 Rump fat depth \( B.P8 \)
  - 14 Rib fat depth \( B.RIB \)

### Data

- Records for Angus cattle
- Carcass traits
  - Data from meat quality research project
  - Progeny test records \( C.WT, C.P8 \& C.RIB \)
- Live ultra-sound scan traits
  - Field data → accredited operators
  - 300 to 700 days of age
  - Select animals in herds of origin of carcass traits
- 121 924 records on 30 427 animals
  - 883 \( C.RBY \) to 3 780 \( C.WT \) records for carcass
  - 7 686 \( B.IMF \) to 18 362 \( H.P8 \) records for scan
Analyses

- Estimate covariance components using REML (WOMBAT)
- 14-trait multi-variate analyses
- Standard fixed effects
- Simple animal model; 45,928 animals in pedigree
- Genetic covariance matrix
  - Full rank → F14 with 168 parameters
  - Reduced rank fitting $m$ PCs → F3 to F11
- Residual covariance matrix
  - Full rank throughout → 63 non-zero components

Likelihood & information criteria

Which model fits best?

First genetic PC

Explains 58% of genetic variation

- Weight / genetic SD
- log L
- $\frac{1}{2}$ AIC
- $\frac{1}{2}$ BIC

First PC

Second PC

Third PC

Sum


Residual covariance matrix

- Full rank throughout → 63 non-zero components

Estimates of eigenvalues

- Full rank throughout → 63 non-zero components

K. Meyer Genetic principal components

BWCGALP 18 / 28

K. Meyer Genetic principal components

BWCGALP 19 / 28

K. Meyer Genetic principal components

BWCGALP 20 / 28
Introduction  
PC basics  
PCs in MMs  
Application  
Discussion  
Data & Analyses  
Results

Second & third PC

Explain 32% & 4% of variation

\[ \text{Second} \]

\[ \text{Third} \]

Estimates of genetic parameters fitting 8 PCs

\[ h^2 \text{ on, } r_G \text{ below, } r_E \text{ above diagonal (x100)} \]

<table>
<thead>
<tr>
<th>Carcass</th>
<th>Heifers/steers</th>
<th>Bulls</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.WT</td>
<td>51 86 -19 -23 -22</td>
<td>28 5 10 -5</td>
</tr>
<tr>
<td>C.RBY</td>
<td>10 75 -32 -14 -33</td>
<td>39 7 3</td>
</tr>
<tr>
<td>C.EMA</td>
<td>-46 23 22 23 15</td>
<td>52 21 20 16</td>
</tr>
<tr>
<td>C.P8</td>
<td>-18 -52 -3 38 36</td>
<td>9 30 22 20</td>
</tr>
<tr>
<td>C.RIB</td>
<td>-18 -82 -21 83</td>
<td>2 16 23 8</td>
</tr>
<tr>
<td>C.IMF</td>
<td>-30 -43 -21 26 31</td>
<td>58 -7 4 15</td>
</tr>
<tr>
<td>H.EMA</td>
<td>51 47 -4 -11 -36</td>
<td>31 30 29 20</td>
</tr>
<tr>
<td>H.P8</td>
<td>17 37 77 73 31</td>
<td>41 69 46</td>
</tr>
<tr>
<td>H.RIB</td>
<td>19 67 62 28 20</td>
<td>78 19</td>
</tr>
<tr>
<td>H.IMF</td>
<td>33 69 31</td>
<td>69 21</td>
</tr>
<tr>
<td>B.EMA</td>
<td>43 41 56 -23 -36</td>
<td>87 0 -4 1 26</td>
</tr>
<tr>
<td>B.P8</td>
<td>-2 62 -19 63 81</td>
<td>34 70 64 19</td>
</tr>
<tr>
<td>B.RIB</td>
<td>-9 -53 -12 62</td>
<td>82 25</td>
</tr>
<tr>
<td>B.IMF</td>
<td>17 -41 -24 41</td>
<td>59 5 40 46 65</td>
</tr>
</tbody>
</table>

'Biplot'

- Summarise PCA
  - plot 'weights' in $P_1$ vs. $P_2$
  - cluster similar traits

1. Introduction
2. Basics of Principal Components
3. PCs in Mixed Models
4. Application
5. Discussion
PCs versus canonical transformation

- Canonical transformation
  - Diagonalise 2 matrices simultaneously
    \[ TVT' = \Omega \] and \[ TWT' = I \]
    with \[ W^{-1}V = T\Omega T' \]
  - Transform data
  - Reduce \( k \)–variate analysis to \( k \) univariate analyses
  - Restricted applicability
    - all traits recorded for all animals
    - equal design matrices

- PC parameterisation
  - Applied to one covariance matrix at a time
  - 'Transform' MME not data
  - Applicable to wide range of models
    - different rank for different random effects
    - decompose covariance matrix of correlated effects

Open questions

- How many PCs?
  - Bias versus sampling errors → MSE
  - Sampling properties
  - Repartitioning between sources of variation
  - Which criterion for model selection

- Shape of likelihood function?
  - Slow convergence for reduced rank REML
  - Last eigenvalue fitted tends to be underestimated
  - Alternative parameterisation
  - Better algorithm

Computational considerations

- PC model
  - Size of MME \( \propto m \) not \( k \)
  - No. of non-zero elements in coefficient matrix \( \propto m^2 \)
  - Operation count per log \( \mathcal{L} \) \( \propto m^x \) with \( x > 2 \)

- Small reduction in rank → big impact on computing required

- REML convergence
  - Less parameters but more AI steps
  - Gradual approach to max. log \( \mathcal{L} \)
  - Negate some comput. advantages
  - Reasons? Remedy??

Conclusions

- Direct estimation of PCs within mixed model analyses
  - is feasible
  - is highly appealing

- Advantages
  - Greater parsimony → more efficient use of data
    - genetic evaluation: fewer EBVs to be obtained
    - variance components: estimate fewer parameters
  - Decrease computational demands
    - facilitate analysis of larger data sets & more traits
    - Readily interpretable results
    - characterise patterns of covariances in multiple dimensions