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

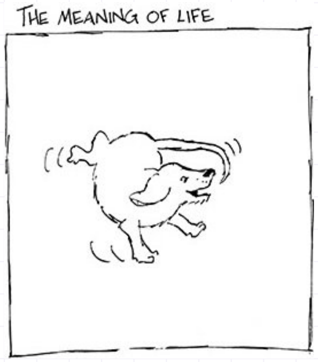
kmeyer@didgeridoo.une.edu.au

AGBU ANIMAL GENETICS AND BREEDING UNIT A joint unit of NSW DPI and UNE

mla MEAT & LIVESTOCK AUSTRALIA

Outline

1. **Introduction**
2. **Basics of Principal Components**
   - Dimension reduction
   - Factor analysis
3. **PCs in Mixed Models**
   - Parameterisation
   - Reduced rank estimation
4. **Application**
   - Data & Analyses
   - Results
5. **Discussion**



Motivation

- Multiple, correlated random effects
  - several traits, random regression coefficients
- Covariance matrix generally assumed ‘unstructured’
  - k variables → $k(k+1)/2$ covariances
- Recent interest in imposing ‘structure’ → 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 → principal components (PCs)
    - factor analytic structure (e.g. Jennrich & Schluchter, 1986)
    - reduced rank models

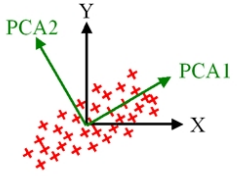
Objectives

- So far : Two-step procedure
  - Estimate unstructured covariance matrix → decompose
  - Transform data to PCs (phenotypic SS/CP) → 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

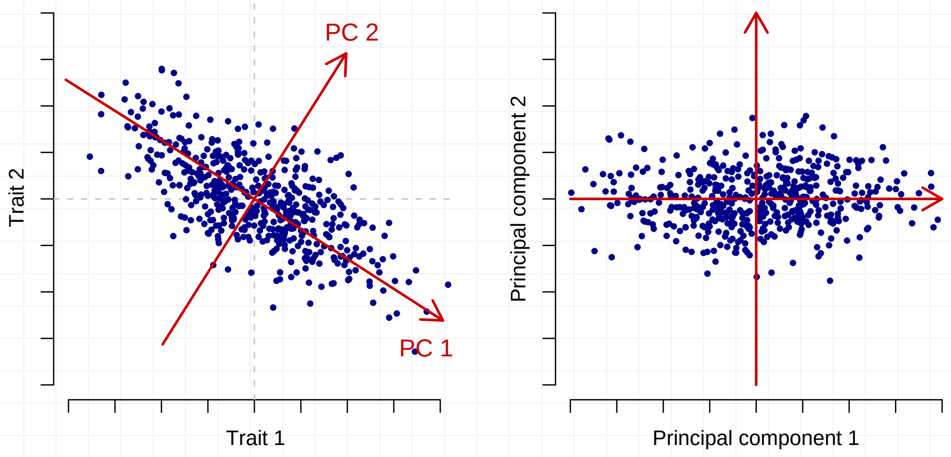
- Review direct estimation of leading principal components
- Show application to beef cattle carcass traits

1. Introduction
2. **Basics of Principal Components**
   - Dimension reduction
   - Factor analysis
3. PCs in Mixed Models
4. Application
5. Discussion



Toy example

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & -1.05 \\ -1.05 & 1 \end{pmatrix} \qquad \mathbf{\Lambda} = \begin{pmatrix} 2.66 & 0 \\ 0 & 0.34 \end{pmatrix}$$



What are PCs ?

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

Dimension reduction

- Principal components
  - summarise information
  - widely used to reduce dimensions → no. variables
- $\mathcal{P}_i$ explains maximum variation given $\mathcal{P}_1, \ldots, \mathcal{P}_{i-1}$
- Var $(\mathcal{P}_{m+1}) = \lambda_{m+1}$ close to zero
  - $\mathcal{P}_{m+1}, \ldots, \mathcal{P}_k$ provide negligible information
  - $\mathcal{P}_{m+1}, \ldots, \mathcal{P}_k$ can be ignored
  - Dimension reduced from k to m
- Consider first m PCs only → $\mathbf{\Sigma}^\star = \sum_{i=1}^{m} \lambda_i \mathbf{e}_i'\mathbf{e}_i = \mathbf{E}_m\mathbf{\Lambda}_m\mathbf{E}_m'$
  - $\mathbf{\Sigma}^\star$ has reduced rank m
  - $\mathbf{\Sigma}^\star$ has $m(2k-m+1)/2$ parameters
    - not $m+mk$ as $\mathbf{e}_i'\mathbf{e}_i = 1$ and $\mathbf{e}_i'\mathbf{e}_j = 0$

Factor analysis

- Different concept
  - PCA → identify variables explaining maximum variance
  - FA → find common factors which explain covariances
- Fit latent model : $\mathbf{v} = \mathbf{F}\mathbf{z} + \boldsymbol{\epsilon}$
  - $\mathbf{F} = \mathbf{E}_m \boldsymbol{\Lambda}_m^{1/2}$
  - $\text{Var}(\mathbf{z}) = \mathbf{I}_m$
  - $\text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \text{Diag}\left\{\sigma_i^2\right\}$
    - σ_i^2 : specific variances $(i = 1, \ldots, k)$
- $\text{Var}(\mathbf{v}) = \boldsymbol{\Sigma}^+ = \mathbf{E}_m \boldsymbol{\Lambda}_m \mathbf{E}'_m + \boldsymbol{\Psi} = \boldsymbol{\Sigma}^\star + \boldsymbol{\Psi}$
  - $\boldsymbol{\Sigma}^+$ generally has full rank k
  - $\boldsymbol{\Sigma}^+$ involves $m(2k - m + 1)/2 + k$ parameters
    - $\leq k(k+1)/2$ → limit on m

1 Introduction

2 Basics of Principal Components

3 PCs in Mixed Models
- Parameterisation
- Reduced rank estimation

4 Application

5 Discussion

Reparameterising the linear mixed model

- ‘Standard’, full rank model

  $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ with $\text{Var}(\mathbf{u}) = \boldsymbol{\Sigma} \otimes \mathbf{A}$

- Reparameterise

  $$\begin{aligned}\mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{Z}\left(\mathbf{Q} \otimes \mathbf{I}_N\right)\left(\mathbf{Q}^{-1} \otimes \mathbf{I}_N\right)\mathbf{u} + \boldsymbol{\epsilon} \\ &= \mathbf{X}\mathbf{b} + \mathbf{Z}^\star\mathbf{u}^\star + \boldsymbol{\epsilon}\end{aligned}$$

- For $\mathbf{Q} = \mathbf{E}$ → equivalent models
  - $\mathbf{u}^\star$ → vector of (genetic) PCs
  - $\text{Var}(\mathbf{u}^\star) = \boldsymbol{\Lambda} \otimes \mathbf{A}$
- For $\mathbf{Q} = \mathbf{E}_m$ → fit leading m PCs only
  - $\mathbf{u}^\star$ has m elements per animal
  - backtransform : $\hat{\mathbf{u}} = \left(\mathbf{E}_m \otimes \mathbf{I}\right)\hat{\mathbf{u}}^\star$

Reduced rank estimation

Alternative forms for variance component estimation

- $\mathbf{Q} = \mathbf{E}_m \boldsymbol{\Lambda}_m^{1/2}$
  - $\text{Var}(\mathbf{u}^\star) = \mathbf{I}_m \otimes \mathbf{A}$
  - FA model with zero specific variances
  - Linear equations determine elements given by orthogonality constraints on $\mathbf{E}$
  - Estimate $\hat{\lambda}_i = \hat{\mathbf{q}}'_i \hat{\mathbf{q}}_i$
- $\mathbf{Q} = \mathbf{L}_m$
  - $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ → Cholesky factor
  - Singular value decomp. $\mathbf{L} = \mathbf{E}\boldsymbol{\Lambda}^{1/2}\mathbf{T}$ (e.g. Harville, 1997)

  ⇒ Estimate $\mathcal{P}_1$ to $\mathcal{P}_m$ of $\boldsymbol{\Sigma} \equiv$ estimate columns 1 to m of $\mathbf{L}$
    - $\mathbf{T}\mathbf{T}' = \mathbf{I}$ → orthogonal rotation of parameter space
    - non-zero elements → correct no. of parameters
    - Cholesky form → good convergence rates

REML estimation for PC model

$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}^\star\mathbf{u}^\star + \epsilon$

- Standard REML algorithms readily adapted
  - Parameters to be estimated part of design matrix
    $$\partial \mathbf{Z}^\star/\partial q_{ij} = \mathbf{Z}\,(\partial \mathbf{Q}/\partial q_{ij} \otimes \mathbf{I}_N)$$
- '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

1 Introduction

2 Basics of Principal Components

3 PCs in Mixed Models

4 **Application**
- Data & Analyses
- Results

5 Discussion

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 at slaughter

1	Carcass weight	C.WT
2	Retail beef yield	C.RBY
3	Eye muscle area	C.EMA
4	Intra-muscular fat	C.IMF
5	Rump fat depth	C.P8
6	Rib fat depth	C.RIB

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

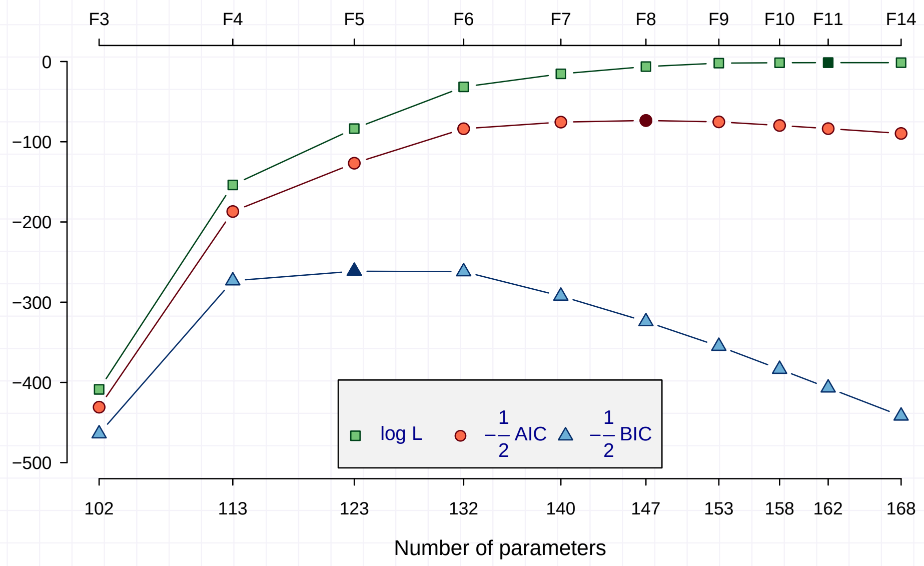
- Records for Angus cattle
- Carcass traits
  - Data from meat quality research project
  - Progeny test records (C.WT, C.P8 & C.RIB)
- Live ultra-sound scan traits
  - Field data → accredited operators
  - 300 to 700 days of age
  - Select animals in herds of origin of carcass traits
- 121 924 records 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

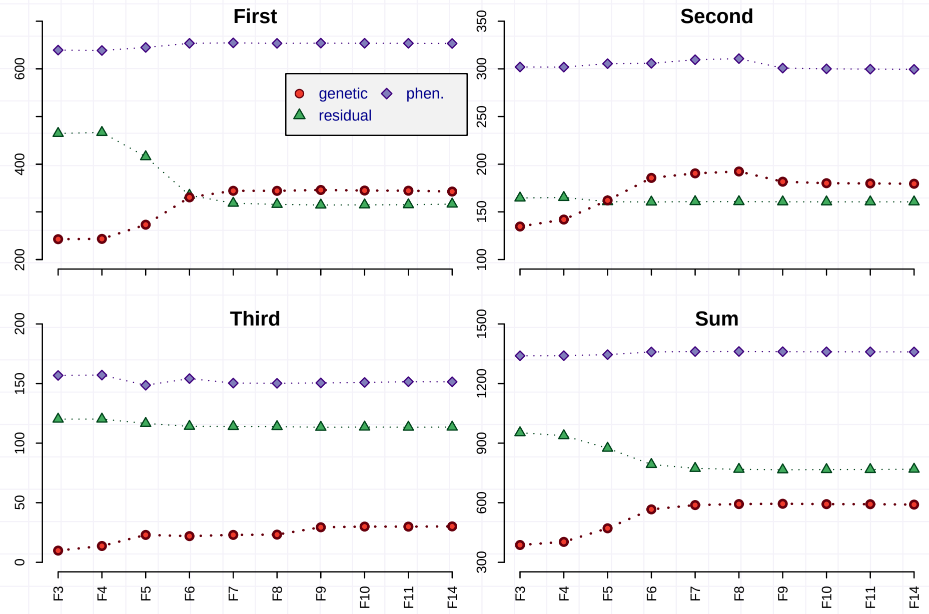
- Estimate covariance components using REML (WOMBAT)
- 14-trait multi-variate analyses
- Standard fixed effects
- Simple animal model; 45 928 animals in pedigree
- Genetic covariance matrix
  - Full rank → F14 with 168 parameters
  - Reduced rank fitting *m* PCs → F3 to F11
- Residual covariance matrix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

Likelihood & information criteria

Which model fits best ?

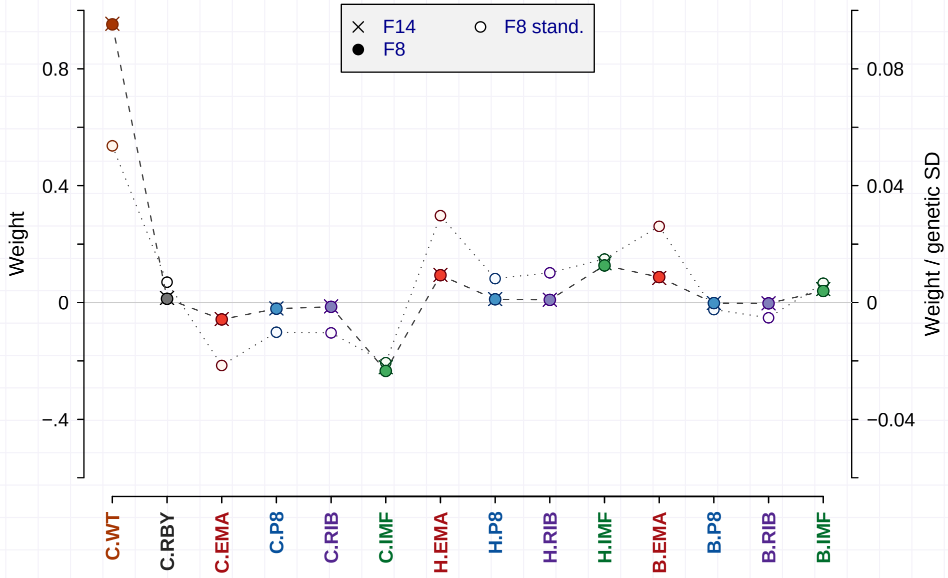


Estimates of eigenvalues



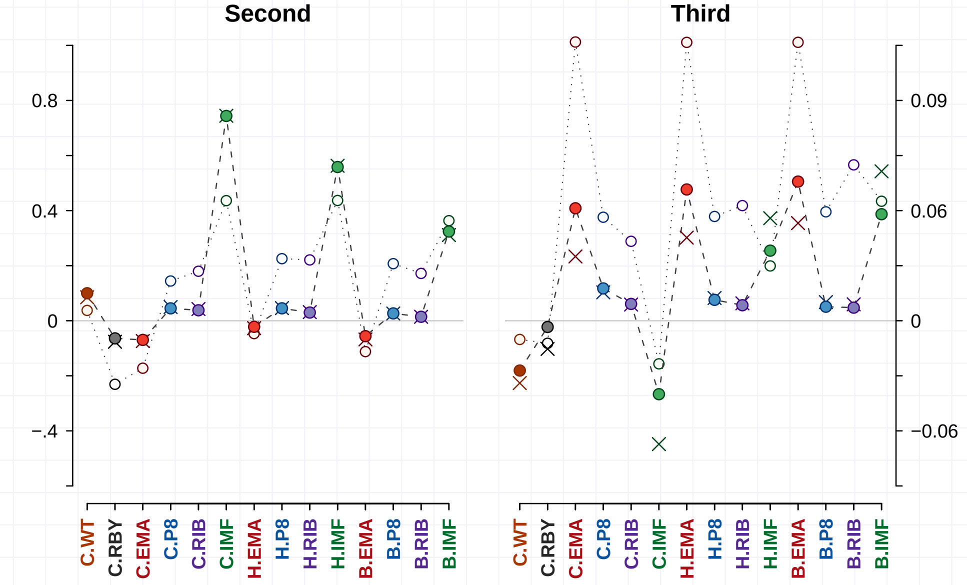
First genetic PC

Explains 58% of genetic variation

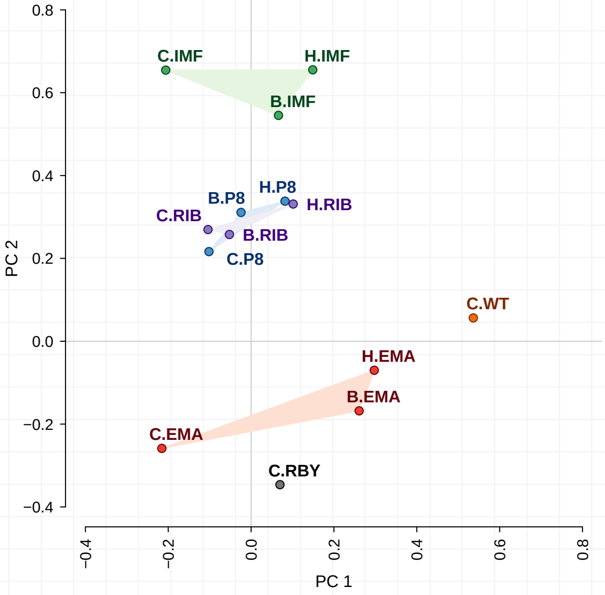


Second & third PC

Explain 32 % & 4% of variation



'Biplot'



- Summarise PCA
  - plot 'weights' in $\mathcal{P}_1$ vs. $\mathcal{P}_2$
  - cluster similar traits

Estimates of genetic parameters fitting 8 PCs

h^2 on, r_G below, r_E above diagonal (×100)

	Carcass						Heifers/steers				Bulls			
C.WT	**51**	86	-8	-19	-23	-22	28	5	10	-5	–	–	–	–
C.RBY	10	**75**	-21	-32	-14	-33	39	-7	3	–	–	–	–	–
C.EMA	-46	23	**22**	22	23	15	52	21	20	16	–	–	–	–
C.P8	-18	-52	-3	**38**	36	16	9	30	22	20	–	–	–	–
C.RIB	-18	-82	-21	83	**26**	18	2	16	23	8	–	–	–	–
C.IMF	-30	-43	-21	26	31	**58**	-7	4	15	–	–	–	–	–
H.EMA	51	1	**47**	-4	-11	-36	**31**	30	29	20	–	–	–	–
H.P8	17	-53	-18	**77**	73	28	19	**41**	71	35	–	–	–	–
H.RIB	19	-56	-28	62	**78**	22	18	87	**36**	38	–	–	–	–
H.IMF	33	-42	-32	25	32	**69**	21	58	62	**31**	–	–	–	–
B.EMA	43	41	**56**	-23	-36	-44	**87**	0	-4	1	**26**	25	25	21
B.P8	-2	-62	-19	**63**	81	34	-4	**70**	64	32	-8	**41**	69	46
B.RIB	-9	-53	-12	62	**82**	25	-4	55	**68**	28	-6	90	**37**	42
B.IMF	17	-41	-24	41	51	**59**	5	40	46	**65**	4	70	75	**24**

1 Introduction

2 Basics of Principal Components

3 PCs in Mixed Models

4 Application

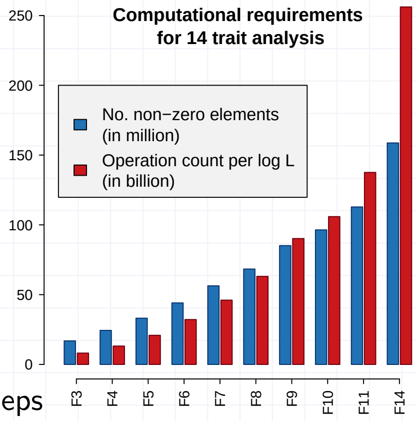
5 **Discussion**

PCs *versus* canonical transformation

- Canonical transformation
  - Diagonalise 2 matrices simultaneously
    $\mathbf{TVT}' = \boldsymbol{\Omega}$ and $\mathbf{TWT}' = \mathbf{I}$ with $\mathbf{W}^{-1}\mathbf{V} = \mathbf{T}\boldsymbol{\Omega}\mathbf{T}'$
  - 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

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 ??



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
- . . .

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