

# Discussion of Dimension Reduction...

## by Dennis Cook

Peter McCullagh

Department of Statistics  
University of Chicago

Anthony Atkinson 70th birthday celebration  
London, December 14, 2007





# The renaissance of PCA?

Copyright@Saakshi O. Juneja



# State of affairs regarding PCA

In a regression model used for prediction, only the vector space  $\text{span}(X)$  matters, not the individual vectors.

Non-invariance: PCA of  $X$  versus PCA of  $XL$   
(either diagonal  $L$  or general  $L$ )

No logical reason why the smallest PC should not be best for predicting  $Y$ .

Logical conclusion: PCA cannot be useful.

Yet it continues to be used by research workers. And we are their methodological leaders so we must follow them!

N.B. logical argument = coordinate-free linear algebra argument



# State of affairs regarding PCA

In a regression model used for prediction, only the vector space  $\text{span}(X)$  matters, not the individual vectors.

Non-invariance: PCA of  $X$  versus PCA of  $XL$   
(either diagonal  $L$  or general  $L$ )

No logical reason why the smallest PC should not be best for predicting  $Y$ .

Logical conclusion: PCA cannot be useful.

Yet it continues to be used by research workers. And we are their methodological leaders so we must follow them!

N.B. logical argument = coordinate-free linear algebra argument



# State of affairs regarding PCA

In a regression model used for prediction, only the vector space  $\text{span}(X)$  matters, not the individual vectors.

Non-invariance: PCA of  $X$  versus PCA of  $XL$   
(either diagonal  $L$  or general  $L$ )

No logical reason why the smallest PC should not be best for predicting  $Y$ .

Logical conclusion: PCA cannot be useful.

Yet it continues to be used by research workers. And we are their methodological leaders so we must follow them!

N.B. logical argument = coordinate-free linear algebra argument



# State of affairs regarding PCA

In a regression model used for prediction, only the vector space  $\text{span}(X)$  matters, not the individual vectors.

Non-invariance: PCA of  $X$  versus PCA of  $XL$   
(either diagonal  $L$  or general  $L$ )

No logical reason why the smallest PC should not be best for predicting  $Y$ .

Logical conclusion: PCA cannot be useful.

Yet it continues to be used by research workers. And we are their methodological leaders so we must follow them!

N.B. logical argument = coordinate-free linear algebra argument



# State of affairs regarding PCA

In a regression model used for prediction, only the vector space  $\text{span}(X)$  matters, not the individual vectors.

Non-invariance: PCA of  $X$  versus PCA of  $XL$   
(either diagonal  $L$  or general  $L$ )

No logical reason why the smallest PC should not be best for predicting  $Y$ .

Logical conclusion: PCA cannot be useful.

Yet it continues to be used by research workers. And we are their methodological leaders so we must follow them!

N.B. logical argument = coordinate-free linear algebra argument



# State of affairs regarding PCA

In a regression model used for prediction, only the vector space  $\text{span}(X)$  matters, not the individual vectors.

Non-invariance: PCA of  $X$  versus PCA of  $XL$   
(either diagonal  $L$  or general  $L$ )

No logical reason why the smallest PC should not be best for predicting  $Y$ .

Logical conclusion: PCA cannot be useful.

Yet it continues to be used by research workers. And we are their methodological leaders so we must follow them!

N.B. logical argument = coordinate-free linear algebra argument



# When/where can PCA be successful?

Example: Sub-population structure in genetics

$Y_i$  ethnic origin of individual  $i$

$x_{ir}$  allele at SNP locus  $r$  on individual  $i$

$n = x000$ ,  $p = 10 * n$ , both large

Every allele is a potential discriminator

No single allele (or pair) is a good discriminator

First few PCs  $X_\xi$  of  $X'X$  sufficient to identify sub-groups

sub-groups related to certain measurable characteristics

Sub-group structure in  $X$  may also be related to  $Y$ .



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# What is different about genetics example?

Natural to regard  $X$  as random

Each column (locus) of  $X$  measured on same scale

Natural to regard  $X$  as a process indexed by loci (columns) and individuals (rows)

Exchangeable rows for individuals, possibly exchangeable columns for loci.

What models do we have for exchangeable arrays?

What interesting models do we have for patterned covariance matrices?

A class intermediate between  $I_p$  and general  $\Sigma$ ?

Relation to PCA?



# Rooted trees as covariance matrices

Defn: A symmetric matrix of order  $n$  such that

$$\Sigma_{rs} \geq \min\{\Sigma_{rt}, \Sigma_{st}\} \geq 0$$

spherical if  $\Sigma_{ij} = \Sigma_{jj}$ .

Table 1a. Viewer preference correlations for 10 programmes (Ehrenberg, 1981)

WoS	1.000	0.581	0.622	0.505	0.296	0.140	0.187	0.145	0.093	0.078
MoD	0.581	1.000	0.593	0.473	0.326	0.121	0.131	0.082	0.039	0.049
GrS	0.622	0.593	1.000	0.474	0.341	0.142	0.181	0.132	0.070	0.085
PrB	0.505	0.473	0.474	1.000	0.309	0.124	0.168	0.106	0.065	0.092
RgS	0.296	0.327	0.341	0.309	1.000	0.121	0.147	0.064	0.051	0.097
24H	0.140	0.122	0.142	0.124	0.121	1.000	0.524	0.395	0.243	0.266
Pan	0.187	0.131	0.181	0.168	0.147	0.524	1.000	0.352	0.200	0.197
ThW	0.145	0.082	0.132	0.106	0.064	0.395	0.352	1.000	0.270	0.188
ToD	0.093	0.039	0.070	0.065	0.051	0.243	0.200	0.270	1.000	0.155
LnU	0.078	0.049	0.085	0.092	0.097	0.266	0.197	0.188	0.155	1.000

Table 1b. Fitted tree for viewer preference correlations

WoS	0.99	0.59	0.61	0.48	0.32	0.10	0.10	0.10	0.10	0.10
MoD	0.59	1.01	0.59	0.48	0.32	0.10	0.10	0.10	0.10	0.10
GrS	0.61	0.59	0.99	0.48	0.32	0.10	0.10	0.10	0.10	0.10
PrB	0.48	0.48	0.48	1.00	0.32	0.10	0.10	0.10	0.10	0.10
RgS	0.32	0.32	0.32	0.32	1.00	0.10	0.10	0.10	0.10	0.10
24H	0.10	0.10	0.10	0.10	0.10	0.96	0.51	0.36	0.25	0.20
Pan	0.10	0.10	0.10	0.10	0.10	0.51	1.01	0.36	0.25	0.20
ThW	0.10	0.10	0.10	0.10	0.10	0.36	0.36	0.99	0.25	0.20
ToD	0.10	0.10	0.10	0.10	0.10	0.25	0.25	0.25	1.03	0.20
LnU	0.10	0.10	0.10	0.10	0.10	0.20	0.20	0.20	0.20	1.01



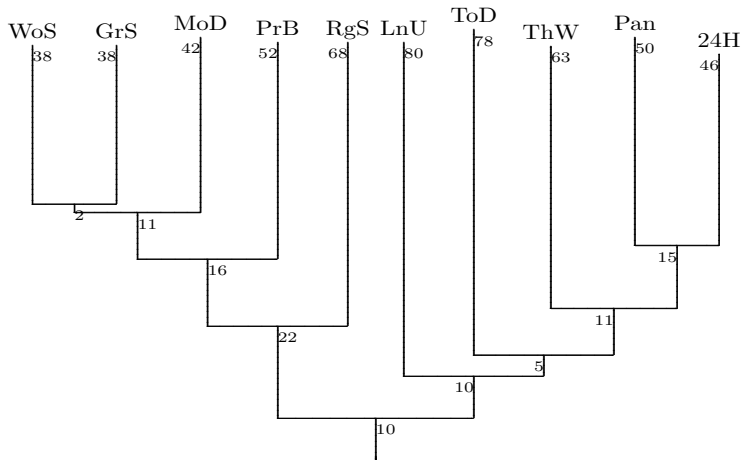


Figure 1: Rooted tree illustrating an approximate correlation matrix. For each pair of variables, the correlation in percent is the distance from the root to the junction.



# Matrix random-effects models

$B$  a partition of  $[n]$  means  $B_{ij} = 1$  if  $i, j$  in same block  
( $B$  may be random);  $\#B$  = number of blocks or clusters  
 $X$  a random matrix of order  $n \times p$

- (i)  $X \sim N(0, I_n \otimes I_p + B \otimes \Sigma)$ ,  $\Sigma$  arbitrary  
equivalent to  $X_{ij} = \epsilon_{ij} + \eta_{b(i),j}$  with  $\eta \sim N_p(0, \Sigma)$   
Implies  $X'X/n \simeq I_p + \Sigma$  with eigenvalues  $1 + \lambda$   
 $Z = X\xi \sim N_n(0, I_n + \lambda B)$  equivalent to  $Z_i = \epsilon_i + \lambda^{1/2}\eta_{b(i)}$   
with  $\lambda_1$  maximizing rms distance between blocks
- (ii) More elaborate version with  $B$  a random partition (Ewens distn)
- (iii) Replace stump  $B$  with a fully branched tree
- (iv) ...or an exchangeable random tree
- (v) More principled version:  $X$  a process on loci (stationary  $\Sigma$ )
- (vi) Relation between partition/tree  $B$  and response  $Y$ ?  
plausibly related but no necessary connection!



# Matrix random-effects models

$B$  a partition of  $[n]$  means  $B_{ij} = 1$  if  $i, j$  in same block  
( $B$  may be random);  $\#B$  = number of blocks or clusters  
 $X$  a random matrix of order  $n \times p$

- (i)  $X \sim N(0, I_n \otimes I_p + B \otimes \Sigma)$ ,  $\Sigma$  arbitrary  
equivalent to  $X_{ij} = \epsilon_{ij} + \eta_{b(i),j}$  with  $\eta \sim N_p(0, \Sigma)$   
Implies  $X'X/n \simeq I_p + \Sigma$  with eigenvalues  $1 + \lambda$   
 $Z = X\xi \sim N_n(0, I_n + \lambda B)$  equivalent to  $Z_i = \epsilon_i + \lambda^{1/2}\eta_{b(i)}$   
with  $\lambda_1$  maximizing rms distance between blocks
- (ii) More elaborate version with  $B$  a random partition (Ewens distn)
- (iii) Replace stump  $B$  with a fully branched tree
- (iv) ...or an exchangeable random tree
- (v) More principled version:  $X$  a process on loci (stationary  $\Sigma$ )
- (vi) Relation between partition/tree  $B$  and response  $Y$ ?  
plausibly related but no necessary connection!



# Matrix random-effects models

$B$  a partition of  $[n]$  means  $B_{ij} = 1$  if  $i, j$  in same block  
( $B$  may be random);  $\#B$  = number of blocks or clusters  
 $X$  a random matrix of order  $n \times p$

- (i)  $X \sim N(0, I_n \otimes I_p + B \otimes \Sigma)$ ,  $\Sigma$  arbitrary  
equivalent to  $X_{ij} = \epsilon_{ij} + \eta_{b(i),j}$  with  $\eta \sim N_p(0, \Sigma)$   
Implies  $X'X/n \simeq I_p + \Sigma$  with eigenvalues  $1 + \lambda$   
 $Z = X\xi \sim N_n(0, I_n + \lambda B)$  equivalent to  $Z_i = \epsilon_i + \lambda^{1/2}\eta_{b(i)}$   
with  $\lambda_1$  maximizing rms distance between blocks
- (ii) More elaborate version with  $B$  a random partition (Ewens distn)
- (iii) Replace stump  $B$  with a fully branched tree
- (iv) ...or an exchangeable random tree
- (v) More principled version:  $X$  a process on loci (stationary  $\Sigma$ )
- (vi) Relation between partition/tree  $B$  and response  $Y$ ?  
plausibly related but no necessary connection!



# Matrix random-effects models

$B$  a partition of  $[n]$  means  $B_{ij} = 1$  if  $i, j$  in same block  
( $B$  may be random);  $\#B$  = number of blocks or clusters  
 $X$  a random matrix of order  $n \times p$

- (i)  $X \sim N(0, I_n \otimes I_p + B \otimes \Sigma)$ ,  $\Sigma$  arbitrary  
equivalent to  $X_{ij} = \epsilon_{ij} + \eta_{b(i),j}$  with  $\eta \sim N_p(0, \Sigma)$   
Implies  $X'X/n \simeq I_p + \Sigma$  with eigenvalues  $1 + \lambda$   
 $Z = X\xi \sim N_n(0, I_n + \lambda B)$  equivalent to  $Z_i = \epsilon_i + \lambda^{1/2}\eta_{b(i)}$   
with  $\lambda_1$  maximizing rms distance between blocks
- (ii) More elaborate version with  $B$  a random partition (Ewens distn)
- (iii) Replace stump  $B$  with a fully branched tree
- (iv) ...or an exchangeable random tree
- (v) More principled version:  $X$  a process on loci (stationary  $\Sigma$ )
- (vi) Relation between partition/tree  $B$  and response  $Y$ ?  
plausibly related but no necessary connection!



# Matrix random-effects models

$B$  a partition of  $[n]$  means  $B_{ij} = 1$  if  $i, j$  in same block  
( $B$  may be random);  $\#B =$  number of blocks or clusters  
 $X$  a random matrix of order  $n \times p$

- (i)  $X \sim N(0, I_n \otimes I_p + B \otimes \Sigma)$ ,  $\Sigma$  arbitrary  
equivalent to  $X_{ij} = \epsilon_{ij} + \eta_{b(i),j}$  with  $\eta \sim N_p(0, \Sigma)$   
Implies  $X'X/n \simeq I_p + \Sigma$  with eigenvalues  $1 + \lambda$   
 $Z = X\xi \sim N_n(0, I_n + \lambda B)$  equivalent to  $Z_i = \epsilon_i + \lambda^{1/2}\eta_{b(i)}$   
with  $\lambda_1$  maximizing rms distance between blocks
- (ii) More elaborate version with  $B$  a random partition (Ewens distn)
- (iii) Replace stump  $B$  with a fully branched tree
- (iv) ...or an exchangeable random tree
- (v) More principled version:  $X$  a process on loci (stationary  $\Sigma$ )
- (vi) Relation between partition/tree  $B$  and response  $Y$ ?  
plausibly related but no necessary connection!



# Matrix random-effects models

$B$  a partition of  $[n]$  means  $B_{ij} = 1$  if  $i, j$  in same block  
( $B$  may be random);  $\#B$  = number of blocks or clusters  
 $X$  a random matrix of order  $n \times p$

- (i)  $X \sim N(0, I_n \otimes I_p + B \otimes \Sigma)$ ,  $\Sigma$  arbitrary  
equivalent to  $X_{ij} = \epsilon_{ij} + \eta_{b(i),j}$  with  $\eta \sim N_p(0, \Sigma)$   
Implies  $X'X/n \simeq I_p + \Sigma$  with eigenvalues  $1 + \lambda$   
 $Z = X\xi \sim N_n(0, I_n + \lambda B)$  equivalent to  $Z_i = \epsilon_i + \lambda^{1/2}\eta_{b(i)}$   
with  $\lambda_1$  maximizing rms distance between blocks
- (ii) More elaborate version with  $B$  a random partition (Ewens distn)
- (iii) Replace stump  $B$  with a fully branched tree
- (iv) ...or an exchangeable random tree
- (v) More principled version:  $X$  a process on loci (stationary  $\Sigma$ )
- (vi) Relation between partition/tree  $B$  and response  $Y$ ?  
plausibly related but no necessary connection!



Einstein quote 'nature tricky but not mean':  
useful quote but not relevant

Reasons that PCA is widely used in various fields:  
need to do something. why not PCA?  
Absence of alternative things to do

No shortage of testimonials to PCA:  
it sometimes gives interesting projections

$X$  needs to be a process in order to evade Cox's objection

Difficult to formulate a model in which the PC projection  
emerges as part of the likelihood solution.

Sufficient reductions:  
not convincing:  $R(X)$  not a statistic



# Conclusions

Einstein quote 'nature tricky but not mean':  
useful quote but not relevant

Reasons that PCA is widely used in various fields:  
need to do something. why not PCA?  
Absence of alternative things to do

No shortage of testimonials to PCA:  
it sometimes gives interesting projections

$X$  needs to be a process in order to evade Cox's objection

Difficult to formulate a model in which the PC projection emerges as part of the likelihood solution.

Sufficient reductions:  
not convincing:  $R(X)$  not a statistic



# Conclusions

Einstein quote 'nature tricky but not mean':  
useful quote but not relevant

Reasons that PCA is widely used in various fields:  
need to do something. why not PCA?  
Absence of alternative things to do

No shortage of testimonials to PCA:  
it sometimes gives interesting projections

*X* needs to be a process in order to evade Cox's objection  
Difficult to formulate a model in which the PC projection  
emerges as part of the likelihood solution.

Sufficient reductions:  
not convincing:  $R(X)$  not a statistic



# Conclusions

Einstein quote 'nature tricky but not mean':  
useful quote but not relevant

Reasons that PCA is widely used in various fields:  
need to do something. why not PCA?  
Absence of alternative things to do

No shortage of testimonials to PCA:  
it sometimes gives interesting projections

$X$  needs to be a process in order to evade Cox's objection

Difficult to formulate a model in which the PC projection  
emerges as part of the likelihood solution.

Sufficient reductions:  
not convincing:  $R(X)$  not a statistic



# Conclusions

Einstein quote 'nature tricky but not mean':  
useful quote but not relevant

Reasons that PCA is widely used in various fields:  
need to do something. why not PCA?  
Absence of alternative things to do

No shortage of testimonials to PCA:  
it sometimes gives interesting projections

$X$  needs to be a process in order to evade Cox's objection

Difficult to formulate a model in which the PC projection  
emerges as part of the likelihood solution.

Sufficient reductions:  
not convincing:  $R(X)$  not a statistic



Einstein quote 'nature tricky but not mean':  
useful quote but not relevant

Reasons that PCA is widely used in various fields:  
need to do something. why not PCA?  
Absence of alternative things to do

No shortage of testimonials to PCA:  
it sometimes gives interesting projections

$X$  needs to be a process in order to evade Cox's objection

Difficult to formulate a model in which the PC projection  
emerges as part of the likelihood solution.

Sufficient reductions:  
not convincing:  $R(X)$  not a statistic

