Kevin Wang

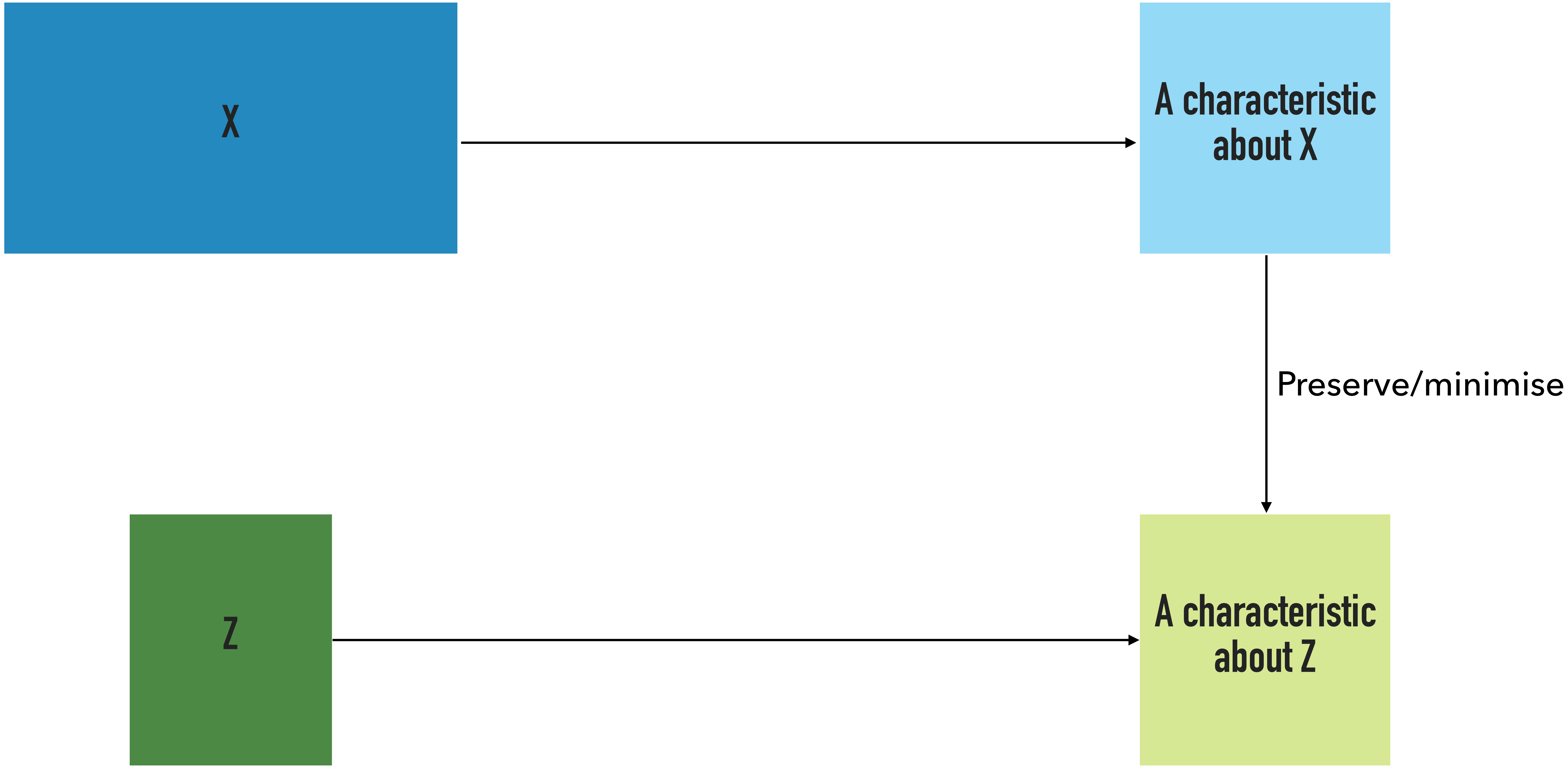
Dimensional Reduction

Acknowledgement

- A large proportion of this material was adapted from the Honours thesis of Nelson Ma, formerly at the School of Mathematics and Statistics, the University of Sydney.

Dimensional reduction

- High dimensional data are tricky:
  - Correlation between variables could contain redundant information
  - Humans eyes are not great beyond 3 dimensions
  - Humans brains are not great at handling non-linear relationships
- Reduce the dimension of our data, while **preserving** one key characteristic



X

A characteristic about X

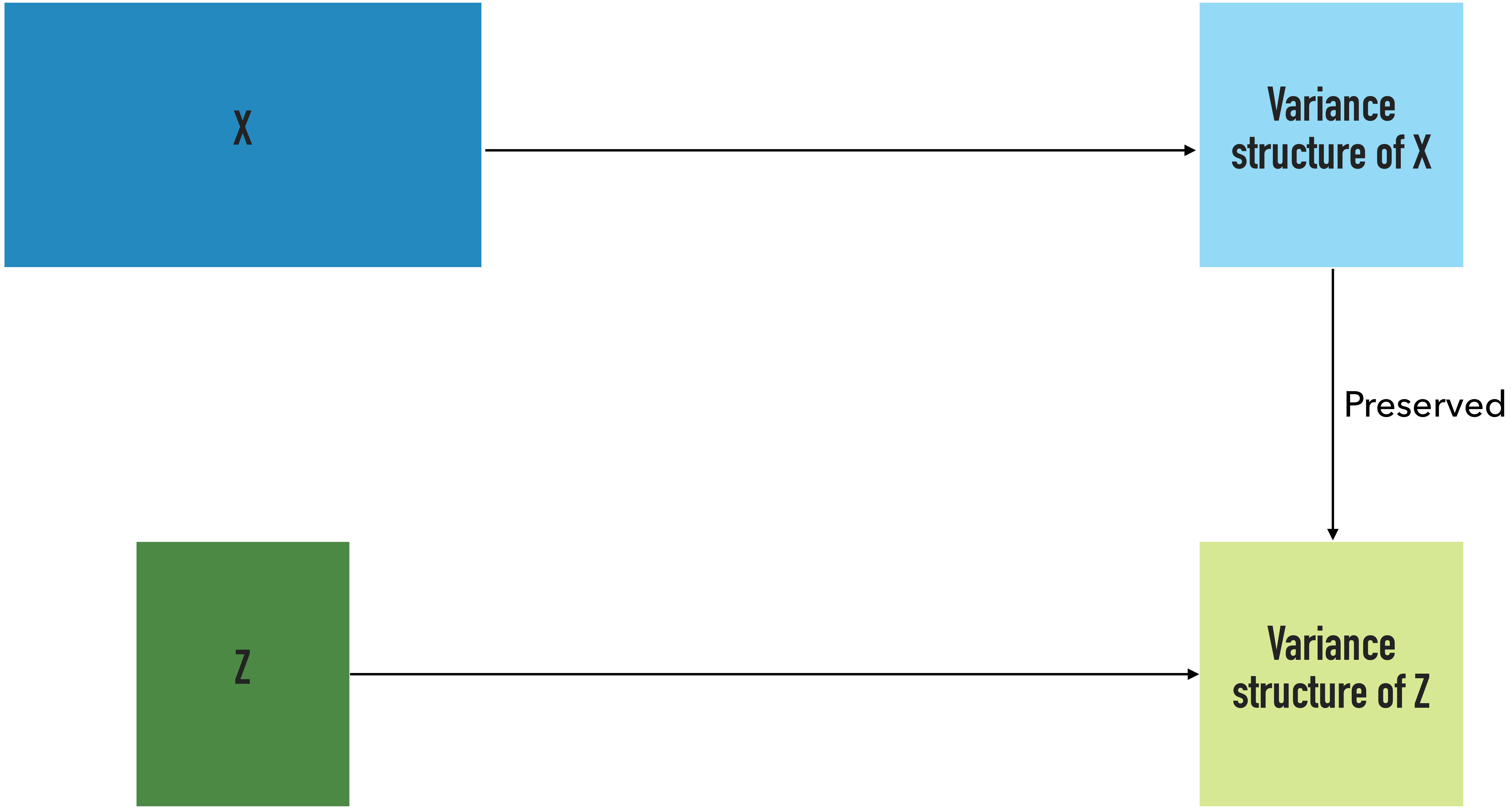
Preserve/minimise

Z

A characteristic about Z

PCA

- Decompose the correlation matrix $\Sigma = U \Lambda U^\top$
- Create a score matrix: $Z = XU$
- The score matrix has the **same amount of variance** as the original data matrix
- Columns of score matrix **successively** inherit the maximum possible variance from X
- This is why the first few columns of the score matrix can be used for visualisation: they already captured a large amount of variation in the original data.


X

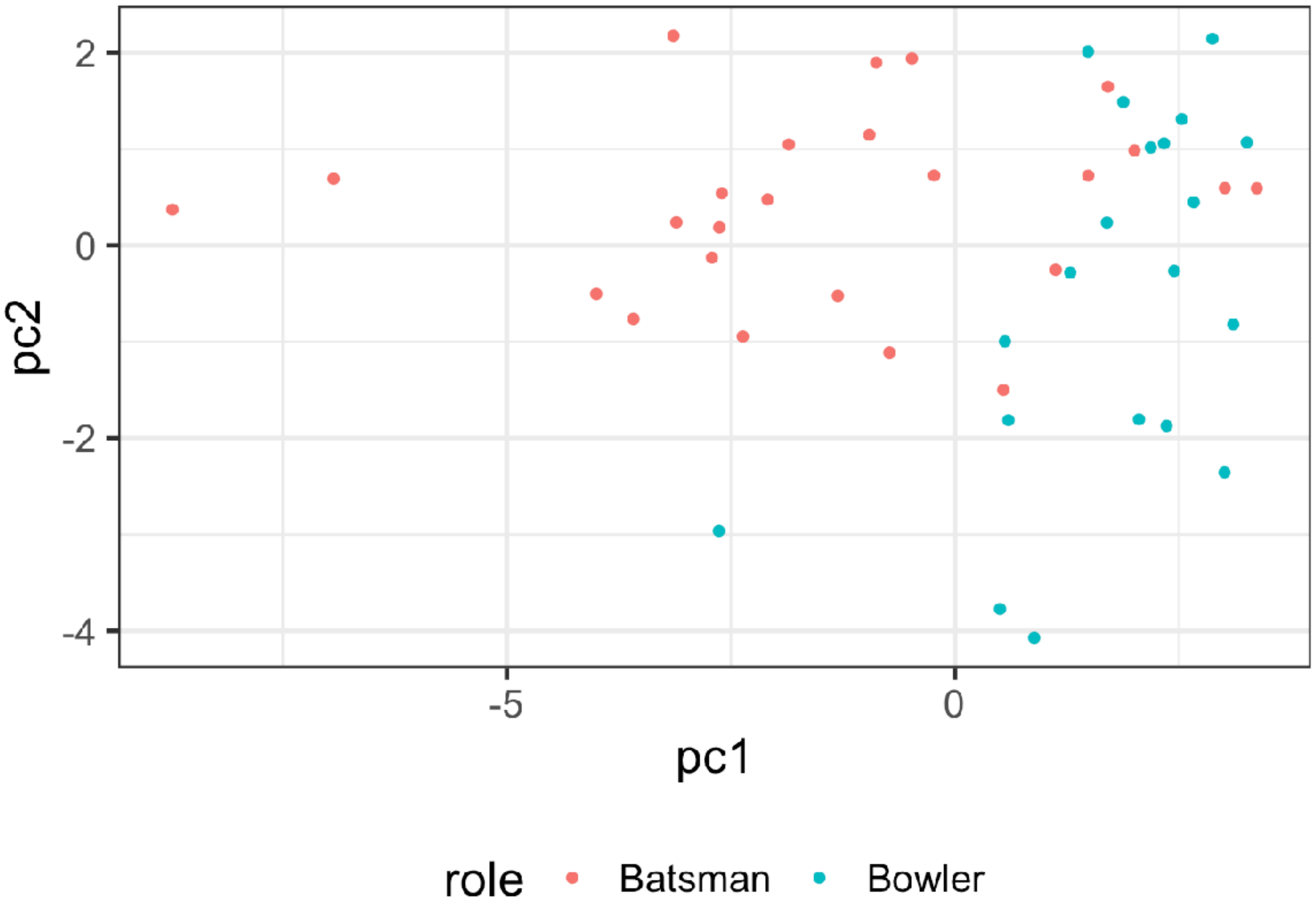
Variance structure of X

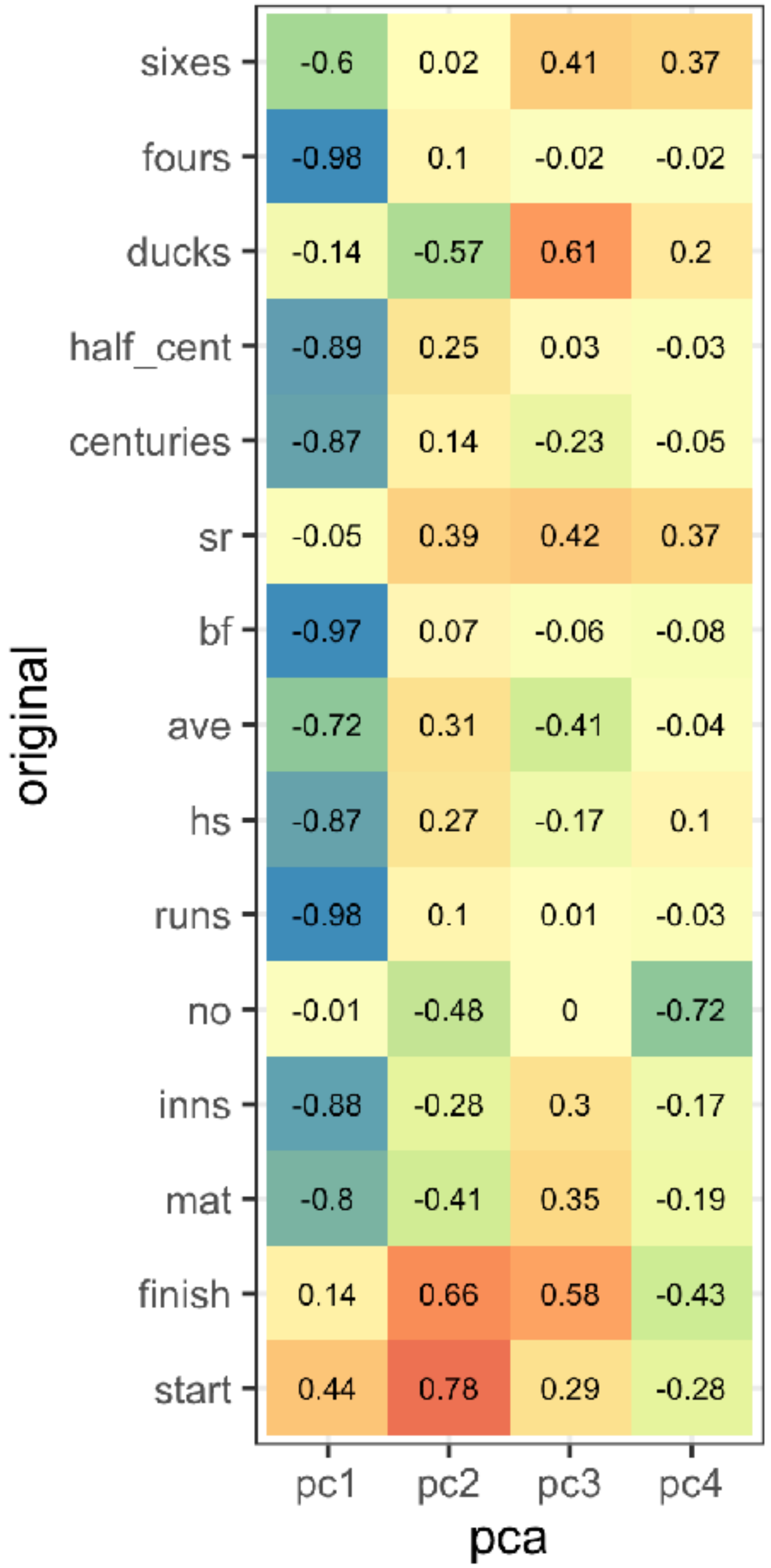
Preserved

Z

Variance structure of Z

PCA visualisation



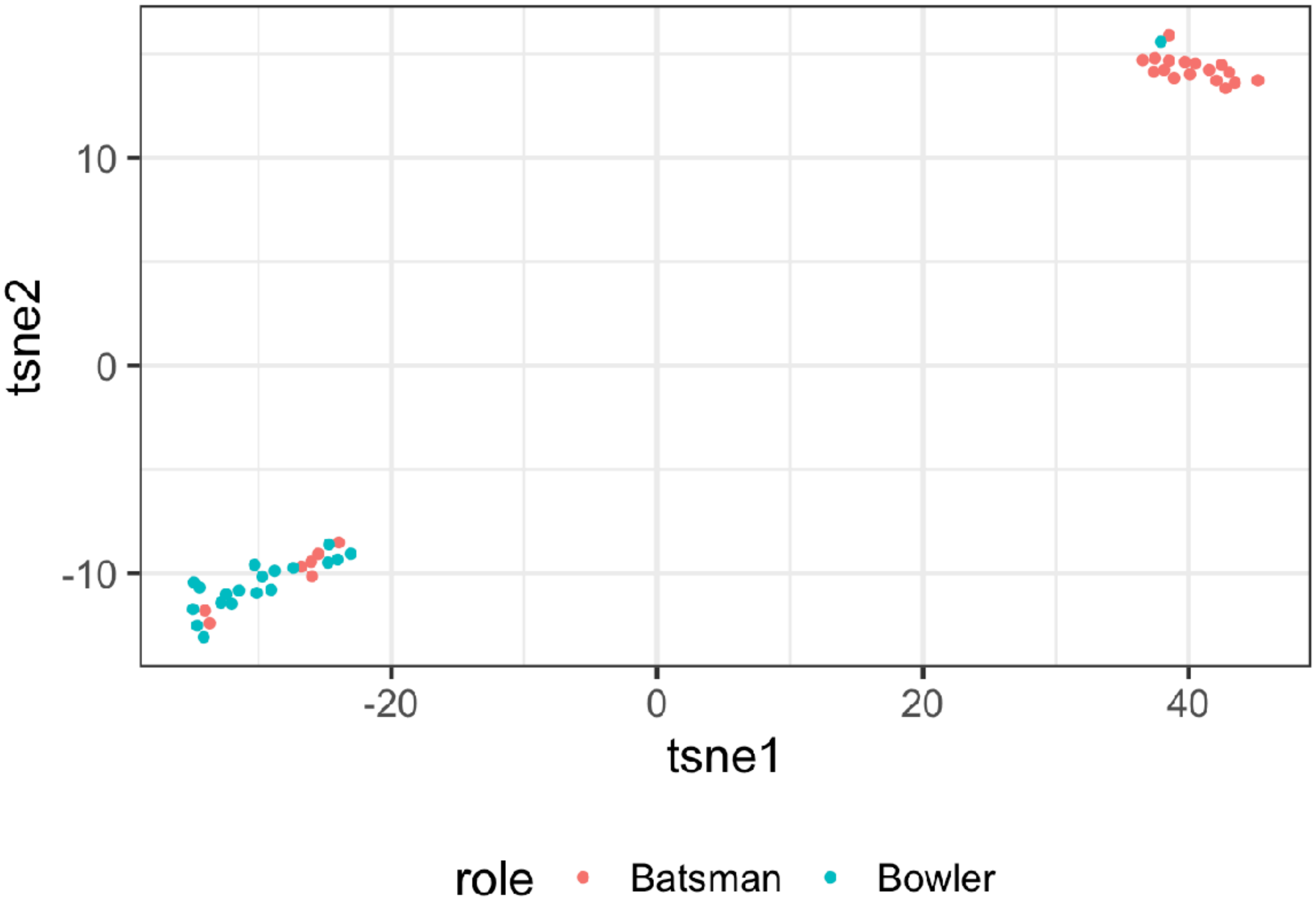


original	pc1	pc2	pc3	pc4
sixes	-0.6	0.02	0.41	0.37
fours	-0.98	0.1	-0.02	-0.02
ducks	-0.14	-0.57	0.61	0.2
half_cent	-0.89	0.25	0.03	-0.03
centuries	-0.87	0.14	-0.23	-0.05
sr	-0.05	0.39	0.42	0.37
bf	-0.97	0.07	-0.06	-0.08
ave	-0.72	0.31	-0.41	-0.04
hs	-0.87	0.27	-0.17	0.1
runs	-0.98	0.1	0.01	-0.03
no	-0.01	-0.48	0	-0.72
inns	-0.88	-0.28	0.3	-0.17
mat	-0.8	-0.41	0.35	-0.19
finish	0.14	0.66	0.58	-0.43
start	0.44	0.78	0.29	-0.28

tSNE: t-distributed stochastic neighbor embedding

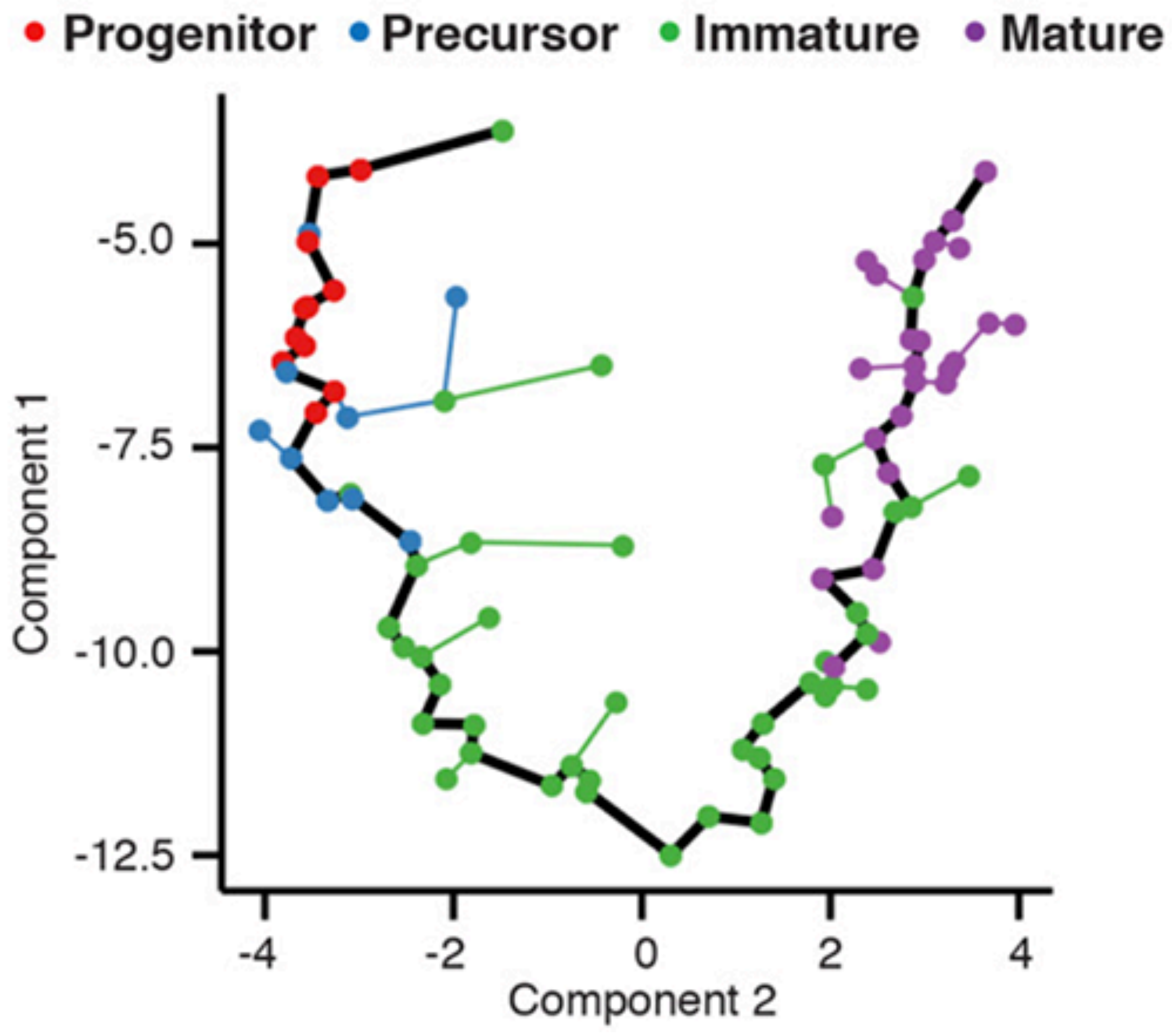
- tSNE was invented in 2008 as a non-linear alternative of PCA
- Unlike PCA, the output matrix of tSNE does not have an interpretation, but its major advantage is in the visualisation
- (Speaking from personal experience) For complex data in my research, tSNE tends to produce more separation of clusters

tSNE visualisation

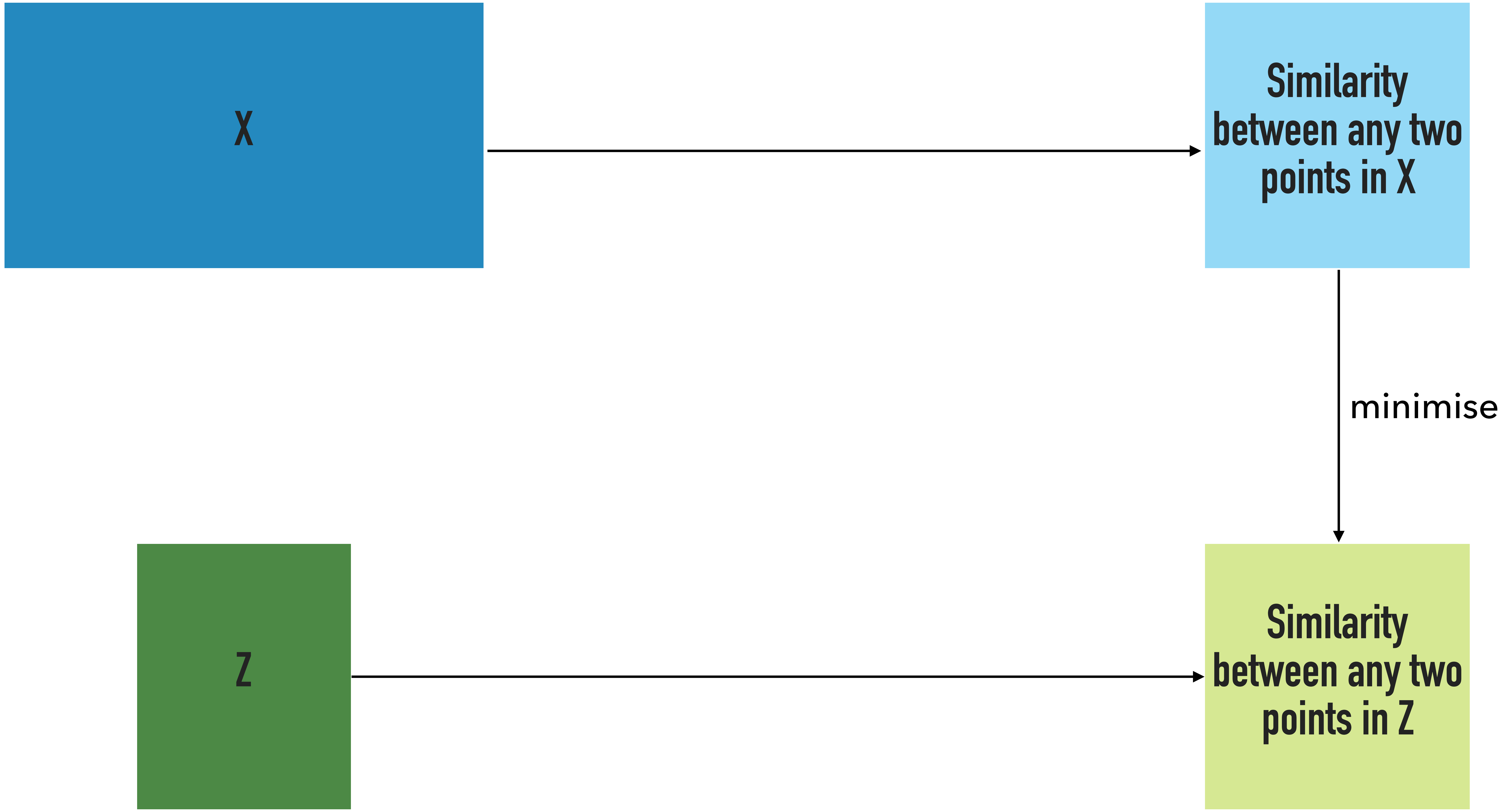


Points that are close to each other in the plot are also close in the original dimension

tSNE visualisation



In single-cell gene expression data, you can use tSNE to perform dimensional reduction before clustering and construct a trajectory of cell development.



X

Similarity between any two points in X

minimise

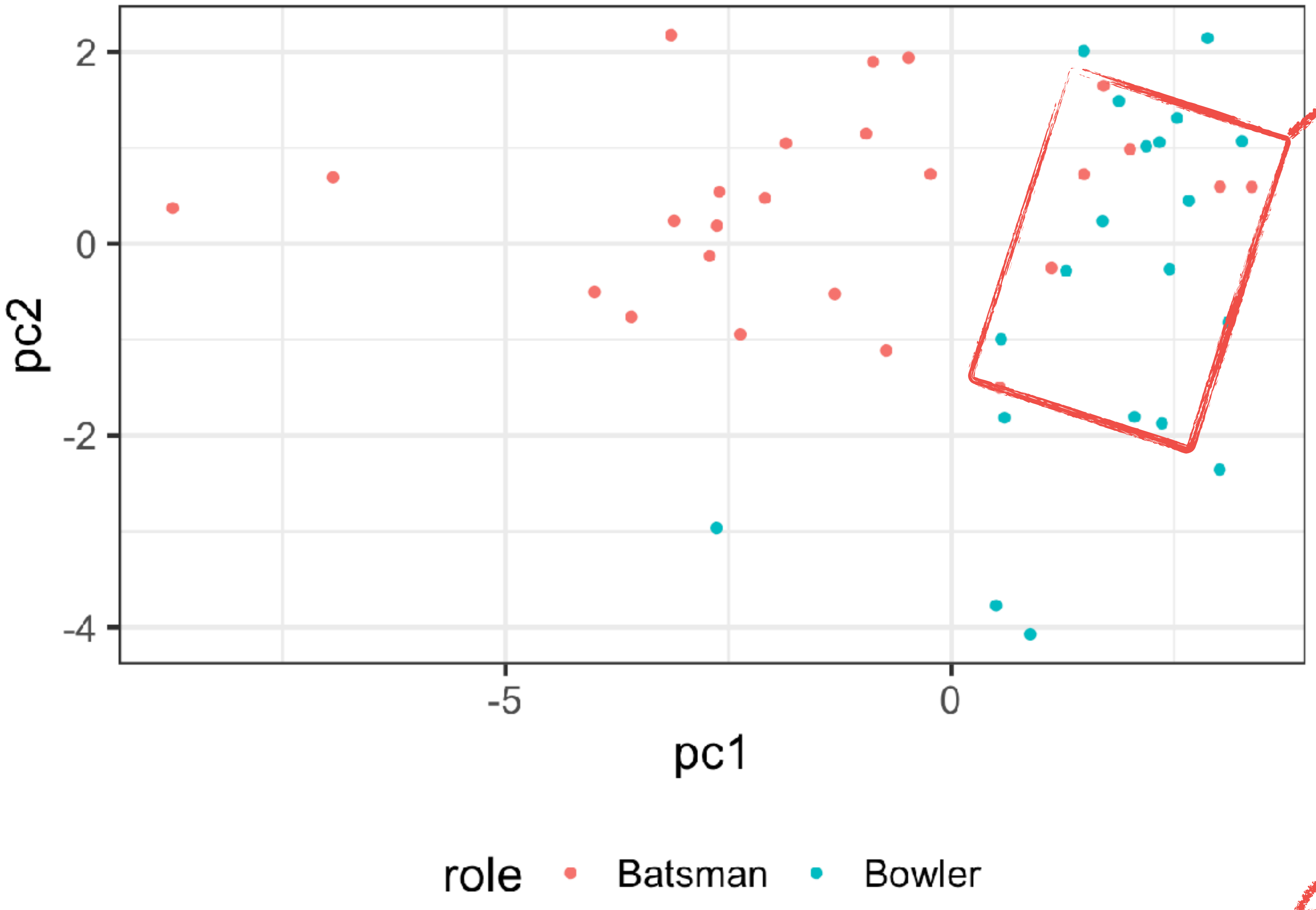
Z

Similarity between any two points in Z

Summary

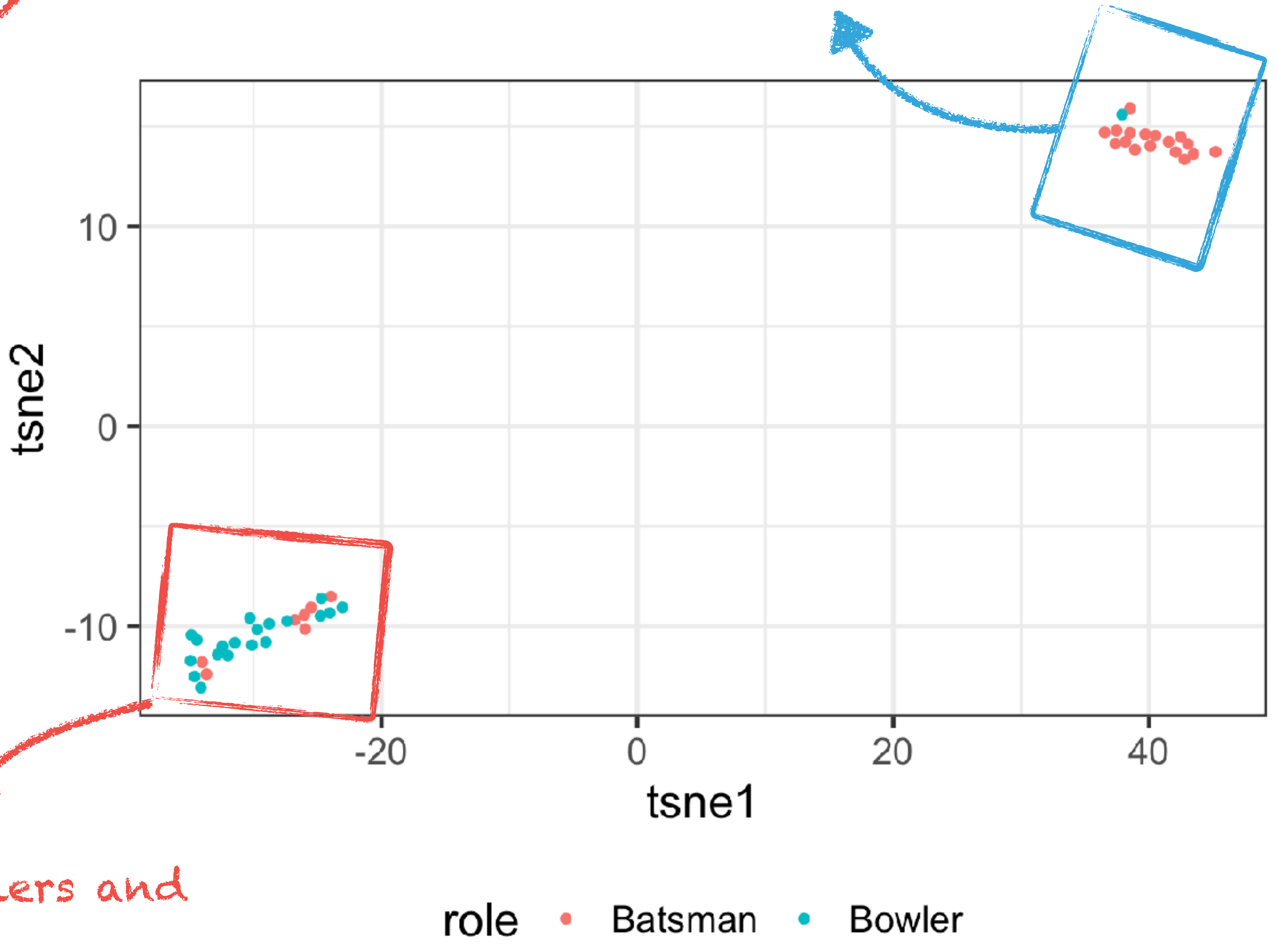
	PCA	tSNE
Relationship captured	linear	non-linear
What is preserved/ minimised between X and Z	variance	similarity between points
Interpretation of output numerical matrix	yes	no

PCA vs tSNE for Cricket data



Batsmen who did poorly or bowlers did well?

Typical batsmen and over-performed bowler (Shane Warne)



Typical bowlers and under-performed batsmen