PCA in R

OR: SOPHISTICATED SIMPLICITY

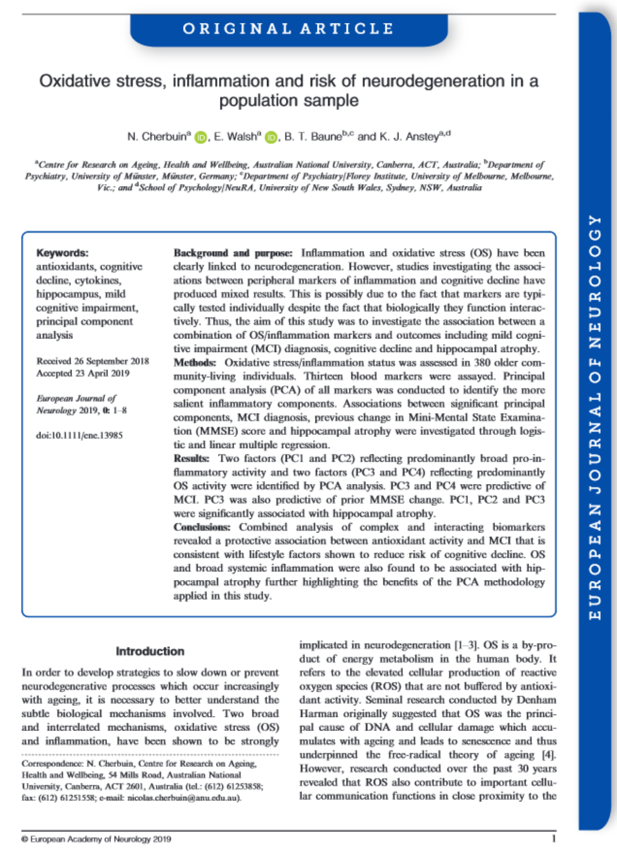
DR ERIN I WALSH | 21/08/2019

The plan for today

- Why this is useful
- How to think in *n* dimensions
- From raw score to principal component
- PCA in general
- PCA in R

Image source: dead____artist on unsplash.com

Why this is useful



ORIGINAL ARTICLE

EUROPEAN JOURNAL OF NEUROLOGY

Oxidative stress, inflammation and risk of neurodegeneration in a population sample

N. Cherbuin[a], E. Walsh[a], B. T. Baune[b,c] and K. J. Anstey[a,d]

[a]*Centre for Research on Ageing, Health and Wellbeing, Australian National University, Canberra, ACT, Australia;* [b]*Department of Psychiatry, University of Münster, Münster, Germany;* [c]*Department of Psychiatry/Florey Institute, University of Melbourne, Melbourne, Vic.; and* [d]*School of Psychology/NeuRA, University of New South Wales, Sydney, NSW, Australia*

Keywords:
antioxidants, cognitive decline, cytokines, hippocampus, mild cognitive impairment, principal component analysis

Received 26 September 2018
Accepted 23 April 2019

European Journal of Neurology 2019, **0:** 1–8

doi:10.1111/ene.13985

Background and purpose: Inflammation and oxidative stress (OS) have been clearly linked to neurodegeneration. However, studies investigating the associations between peripheral markers of inflammation and cognitive decline have produced mixed results. This is possibly due to the fact that markers are typically tested individually despite the fact that biologically they function interactively. Thus, the aim of this study was to investigate the association between a combination of OS/inflammation markers and outcomes including mild cognitive impairment (MCI) diagnosis, cognitive decline and hippocampal atrophy.
Methods: Oxidative stress/inflammation status was assessed in 380 older community-living individuals. Thirteen blood markers were assayed. Principal component analysis (PCA) of all markers was conducted to identify the more salient inflammatory components. Associations between significant principal components, MCI diagnosis, previous change in Mini-Mental State Examination (MMSE) score and hippocampal atrophy were investigated through logistic and linear multiple regression.
Results: Two factors (PC1 and PC2) reflecting predominantly broad pro-inflammatory activity and two factors (PC3 and PC4) reflecting predominantly OS activity were identified by PCA analysis. PC3 and PC4 were predictive of MCI. PC3 was also predictive of prior MMSE change. PC1, PC2 and PC3 were significantly associated with hippocampal atrophy.
Conclusions: Combined analysis of complex and interacting biomarkers revealed a protective association between antioxidant activity and MCI that is consistent with lifestyle factors shown to reduce risk of cognitive decline. OS and broad systemic inflammation were also found to be associated with hippocampal atrophy further highlighting the benefits of the PCA methodology applied in this study.

Introduction

In order to develop strategies to slow down or prevent neurodegenerative processes which occur increasingly with ageing, it is necessary to better understand the subtle biological mechanisms involved. Two broad and interrelated mechanisms, oxidative stress (OS) and inflammation, have been shown to be strongly implicated in neurodegeneration [1–3]. OS is a by-product of energy metabolism in the human body. It refers to the elevated cellular production of reactive oxygen species (ROS) that are not buffered by antioxidant activity. Seminal research conducted by Denham Harman originally suggested that OS was the principal cause of DNA and cellular damage which accumulates with ageing and leads to senescence and thus underpinned the free-radical theory of ageing [4]. However, research conducted over the past 30 years revealed that ROS also contribute to important cellular communication functions in close proximity to the

Correspondence: N. Cherbuin, Centre for Research on Ageing, Health and Wellbeing, 54 Mills Road, Australian National University, Canberra, ACT 2601, Australia (tel.: (612) 61253858; fax: (612) 61251558; e-mail: nicolas.cherbuin@anu.edu.au).

Cherbuin, N., Walsh, E., Baune, B. T., & Anstey, K. J. (2019). Oxidative stress, inflammation and risk of neurodegeneration in a population sample. *European journal of neurology*.

Abstract

Background and purpose

Inflammation and oxidative stress (OS) have been clearly linked to neurodegeneration. **However, studies investigating the associations between peripheral markers of inflammation and cognitive decline have produced mixed results. This is possibly due to the fact that markers are typically tested individually despite the fact that biologically they function interactively.** Thus, the aim of this study was to investigate the association between a combination of OS/inflammation markers and outcomes including mild cognitive impairment (MCI) diagnosis, cognitive decline and hippocampal atrophy.

Methods

Oxidative stress/inflammation status was assessed in 380 older community-living individuals. Thirteen blood markers were assayed. **Principal component analysis (PCA) of all markers was conducted to identify the more salient inflammatory components.** Associations between significant principal components, MCI diagnosis, previous change in Mini-Mental State Examination (MMSE) score and hippocampal atrophy were investigated through logistic and linear multiple regression.

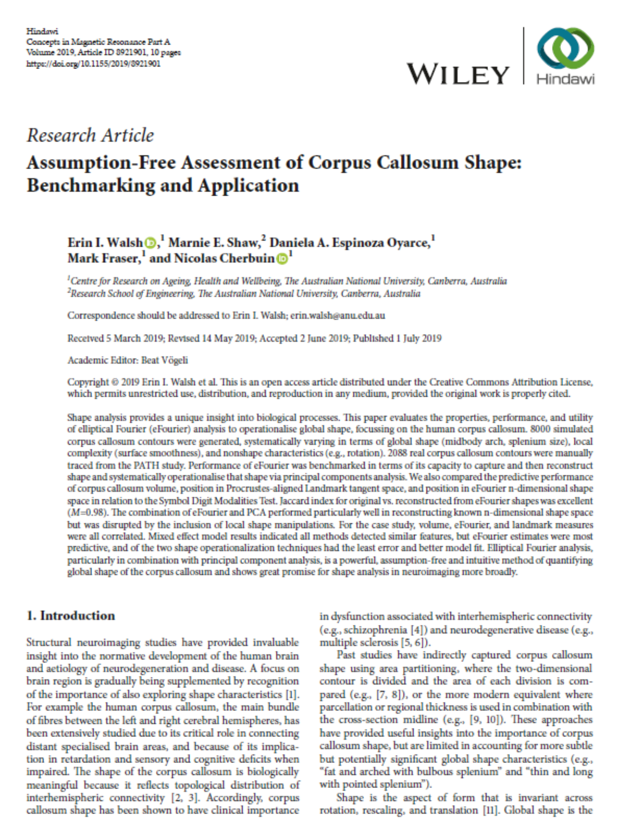
Results

Two factors (PC1 and PC2) reflecting predominantly broad pro-inflammatory activity and two factors (PC3 and PC4) reflecting predominantly OS activity were identified by PCA analysis. PC3 and PC4 were predictive of MCI. PC3 was also predictive of prior MMSE change. PC1, PC2 and PC3 were significantly associated with hippocampal atrophy.

Conclusions

Combined analysis of complex and interacting biomarkers revealed a protective association between antioxidant activity and MCI that is consistent with lifestyle factors shown to reduce risk of cognitive decline. **OS and broad systemic inflammation were also found to be associated with hippocampal atrophy further highlighting the benefits of the PCA methodology applied in this study.**

Why this is useful



Hindawi
Concepts in Magnetic Resonance Part A
Volume 2019, Article ID 8921901, 10 pages
https://doi.org/10.1155/2019/8921901

WILEY | Hindawi

Research Article

Assumption-Free Assessment of Corpus Callosum Shape: Benchmarking and Application

Erin I. Walsh,[1] Marnie E. Shaw,[2] Daniela A. Espinoza Oyarce,[1] Mark Fraser,[1] and Nicolas Cherbuin[1]

[1] *Centre for Research on Ageing, Health and Wellbeing, The Australian National University, Canberra, Australia*
[2] *Research School of Engineering, The Australian National University, Canberra, Australia*

Correspondence should be addressed to Erin I. Walsh; erin.walsh@anu.edu.au

Received 5 March 2019; Revised 14 May 2019; Accepted 2 June 2019; Published 1 July 2019

Academic Editor: Beat Vögeli

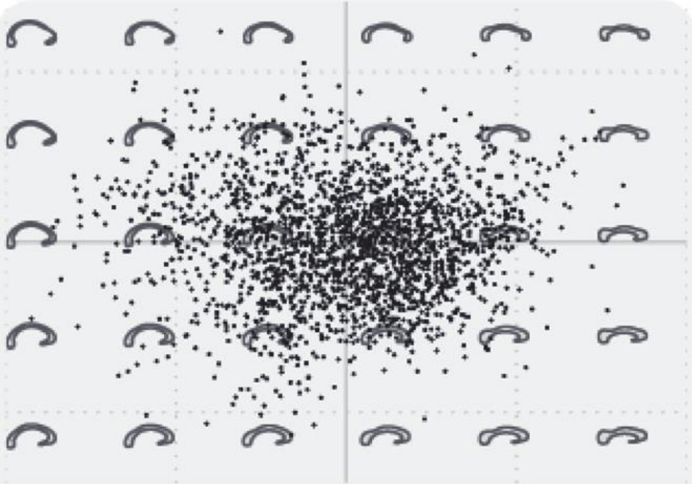
Shape analysis provides a unique insight into biological processes. This paper evaluates the properties, performance, and utility of elliptical Fourier (eFourier) analysis to operationalise global shape, focussing on the human corpus callosum. 8000 simulated corpus callosum contours were generated, systematically varying in terms of global shape (midbody arch, splenium size), local complexity (surface smoothness), and nonshape characteristics (e.g., rotation). 2088 real corpus callosum contours were manually traced from the PATH study. Performance of eFourier was benchmarked in terms of its capacity to capture and then reconstruct shape and systematically operationalise that shape via principal components analysis. We also compared the predictive performance of corpus callosum volume, position in Procrustes-aligned Landmark tangent space, and position in eFourier n-dimensional shape space in relation to the Symbol Digit Modalities Test. Jaccard index for original vs. reconstructed from eFourier shapes was excellent (M=0.98). The combination of eFourier and PCA performed particularly well in reconstructing known n-dimensional shape space but was disrupted by the inclusion of local shape manipulations. For the case study, volume, eFourier, and landmark measures were all correlated. Mixed effect model results indicated all methods detected similar features, but eFourier estimates were most predictive, and of the two shape operationalization techniques had the least error and better model fit. Elliptical Fourier analysis, particularly in combination with principal component analysis, is a powerful, assumption-free and intuitive method of quantifying global shape of the corpus callosum and shows great promise for shape analysis in neuroimaging more broadly.

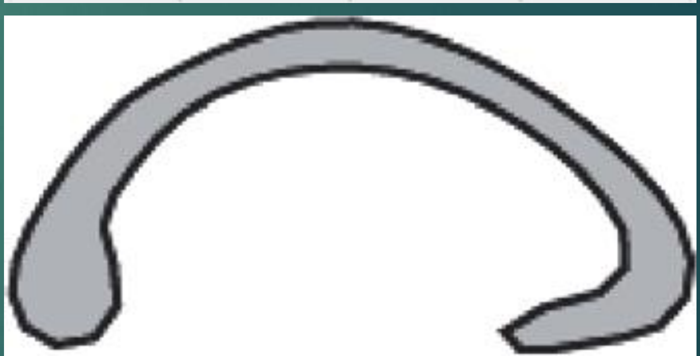
1. Introduction

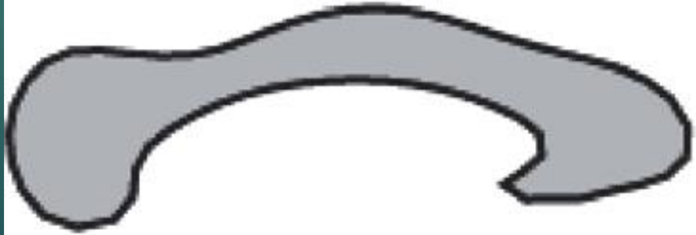
Structural neuroimaging studies have provided invaluable insight into the normative development of the human brain and aetiology of neurodegeneration and disease. A focus on brain region is gradually being supplemented by recognition of the importance of also exploring shape characteristics [1]. For example the human corpus callosum, the main bundle of fibres between the left and right cerebral hemispheres, has been extensively studied due to its critical role in connecting distant specialised brain areas, and because of its implication in retardation and sensory and cognitive deficits when impaired. The shape of the corpus callosum is biologically meaningful because it reflects topological distribution of interhemispheric connectivity [2, 3]. Accordingly, corpus callosum shape has been shown to have clinical importance in dysfunction associated with interhemispheric connectivity (e.g., schizophrenia [4]) and neurodegenerative disease (e.g., multiple sclerosis [5, 6]).

Past studies have indirectly captured corpus callosum shape using area partitioning, where the two-dimensional contour is divided and the area of each division is compared (e.g., [7, 8]), or the more modern equivalent where parcellation or regional thickness is used in combination with the cross-section midline (e.g., [9, 10]). These approaches have provided useful insights into the importance of corpus callosum shape, but are limited in accounting for more subtle but potentially significant global shape characteristics (e.g., "fat and arched with bulbous splenium" and "thin and long with pointed splenium").

Shape is the aspect of form that is invariant across rotation, rescaling, and translation [11]. Global shape is the







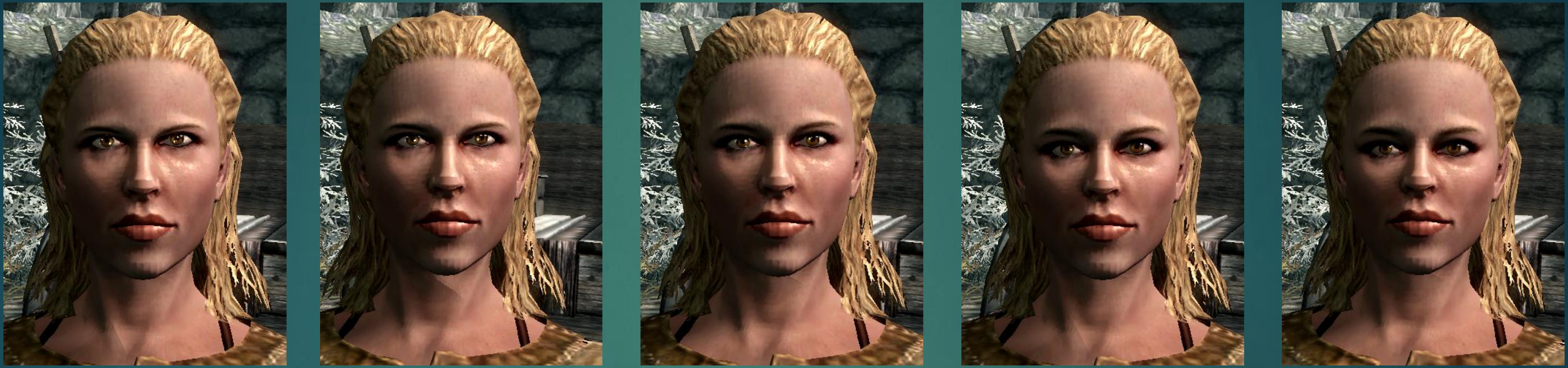
Walsh, E. I., Shaw, M. E., Oyarce, D. A. E., Fraser, M., & Cherbuin, N. (2019). Assumption-Free Assessment of Corpus Callosum Shape: Benchmarking and Application. *Concepts in Magnetic Resonance Part A, 2019.*

How to think in *n* dimensions

- Faces have many possibly salient features that vary
  - e.g. eyes

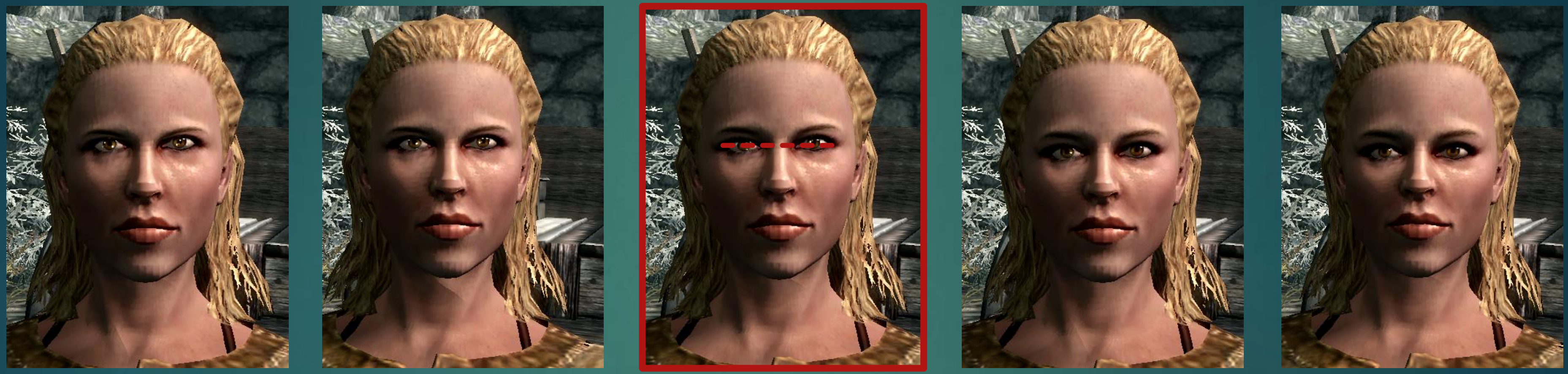
How to think in *n* dimensions

- Each of these features can vary in some way
  - e.g. eyes can vary in position. This is a *dimension*: eye position



How to think in *n* dimensions

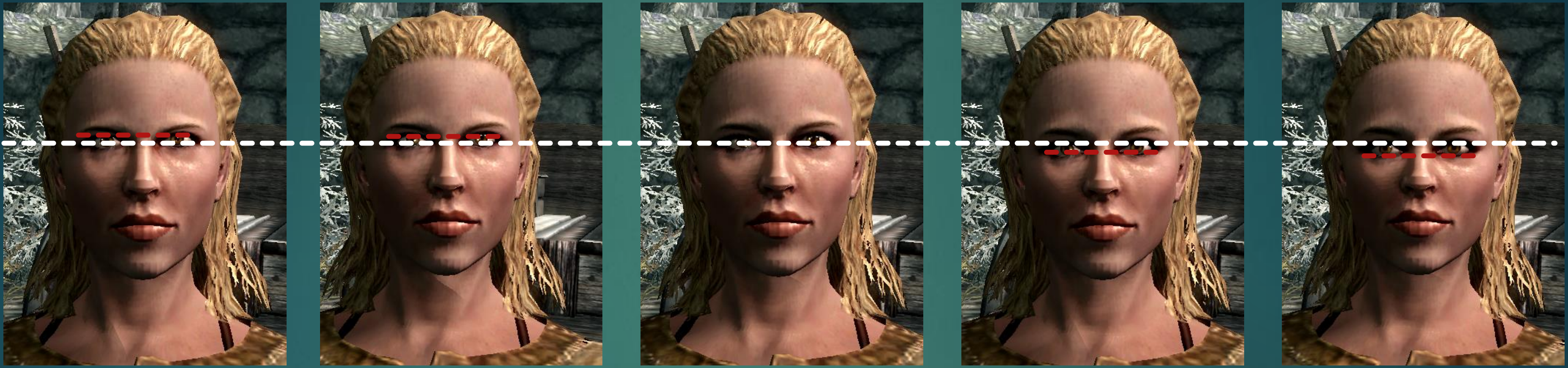
- Each of these features can vary in some way
  - e.g. eyes can vary in position. This is a *dimension*: eye position



- In any sample, there is an average on every dimension
  - e.g. mean eye position is at the **centre** of the dimension
    - Not necessarily a real individual, could be somewhere in between

How to think in *n* dimensions

- Each of these features can vary in some way
  - e.g. eyes can vary in position. This is a *dimension*: eye position

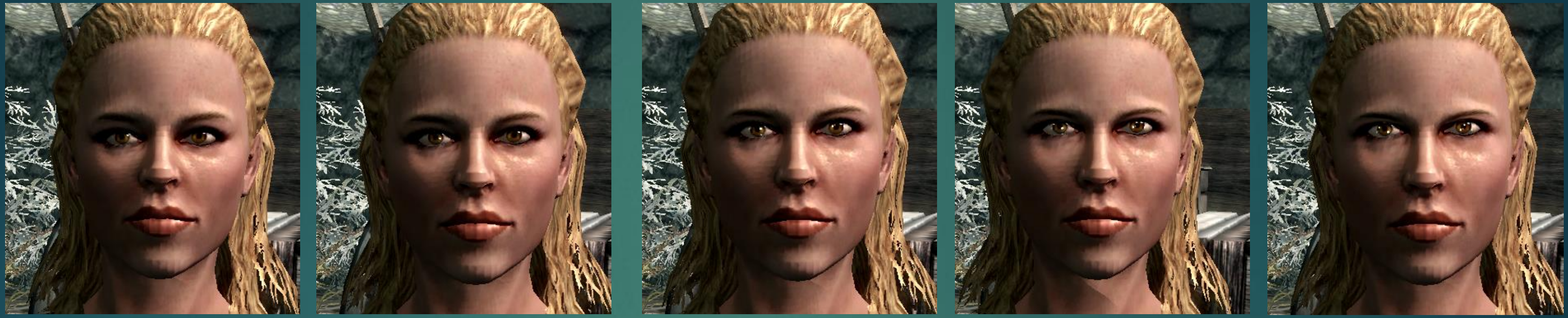


- In any sample, there is an average on every dimension
  - e.g. mean eye position is at the **centre** of the dimension
    - Not necessarily a real individual, could be somewhere in between
- Every feature of an individual in the population can be characterised by their distance from this average
  - This is the position on the dimension
  - Has distribution and variability like any other variable

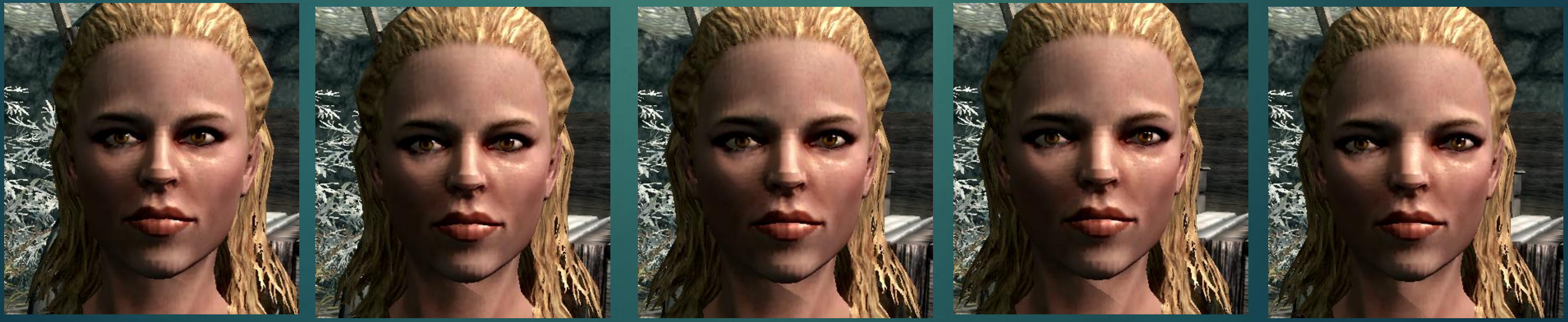
How to think in *n* dimensions

- You can physically arrange them along one salient dimension at a time

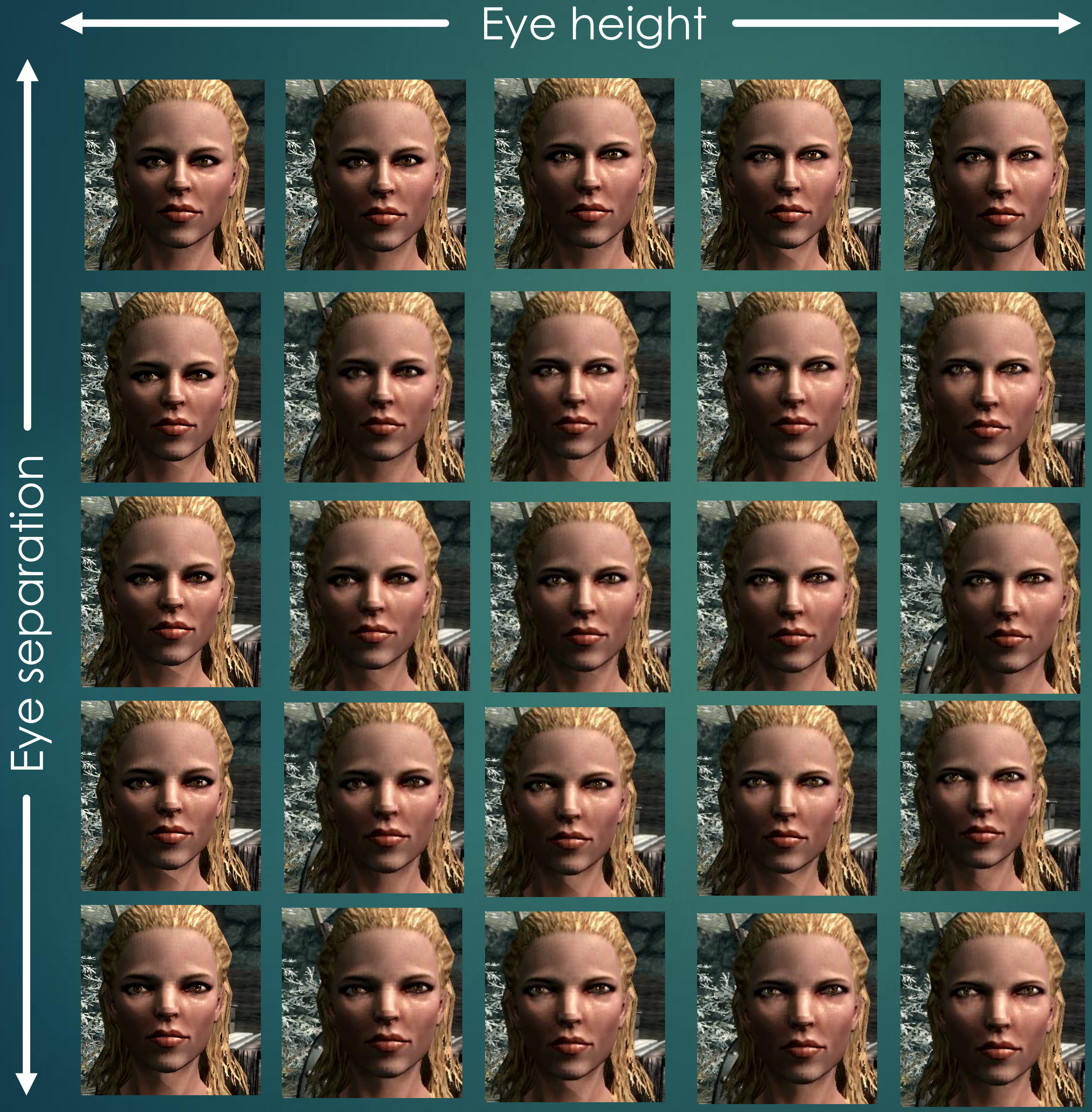
Eye height



Eye separation



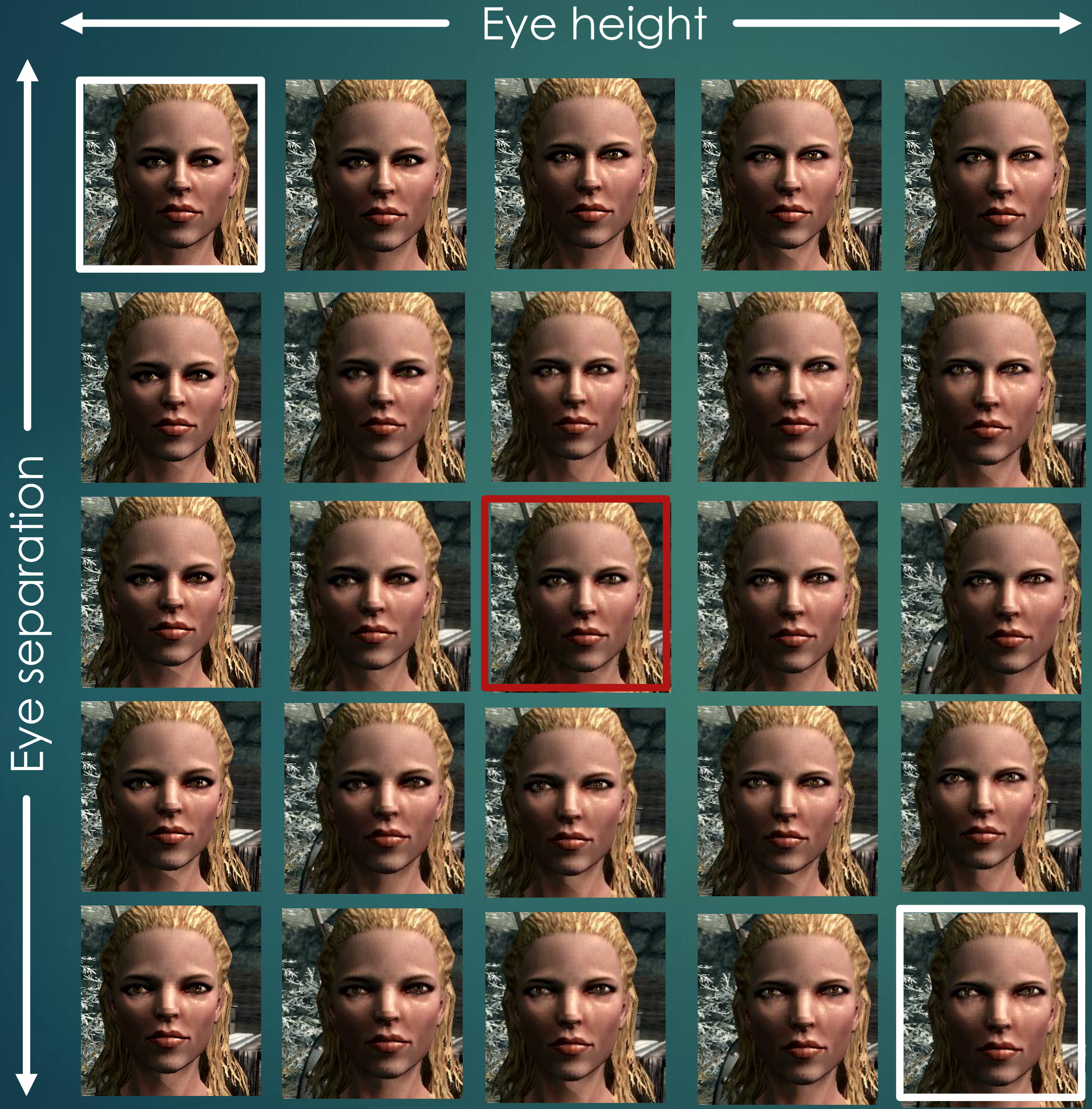
- Or arrange across two dimensions at once



Eye height

Eye separation

- Or arrange across two dimensions at once

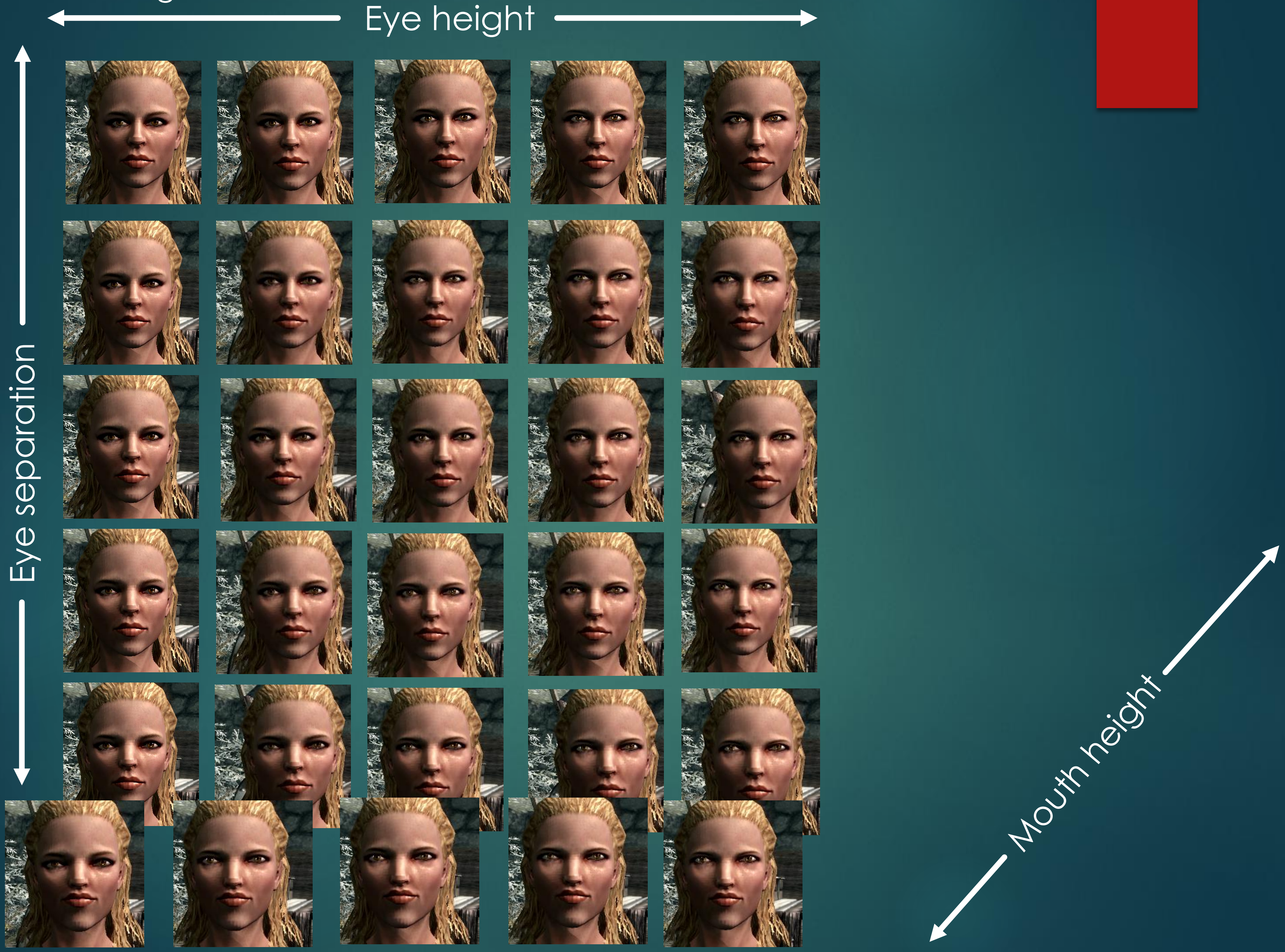


Low and close

Average

High and wide

- Or arrange across three…

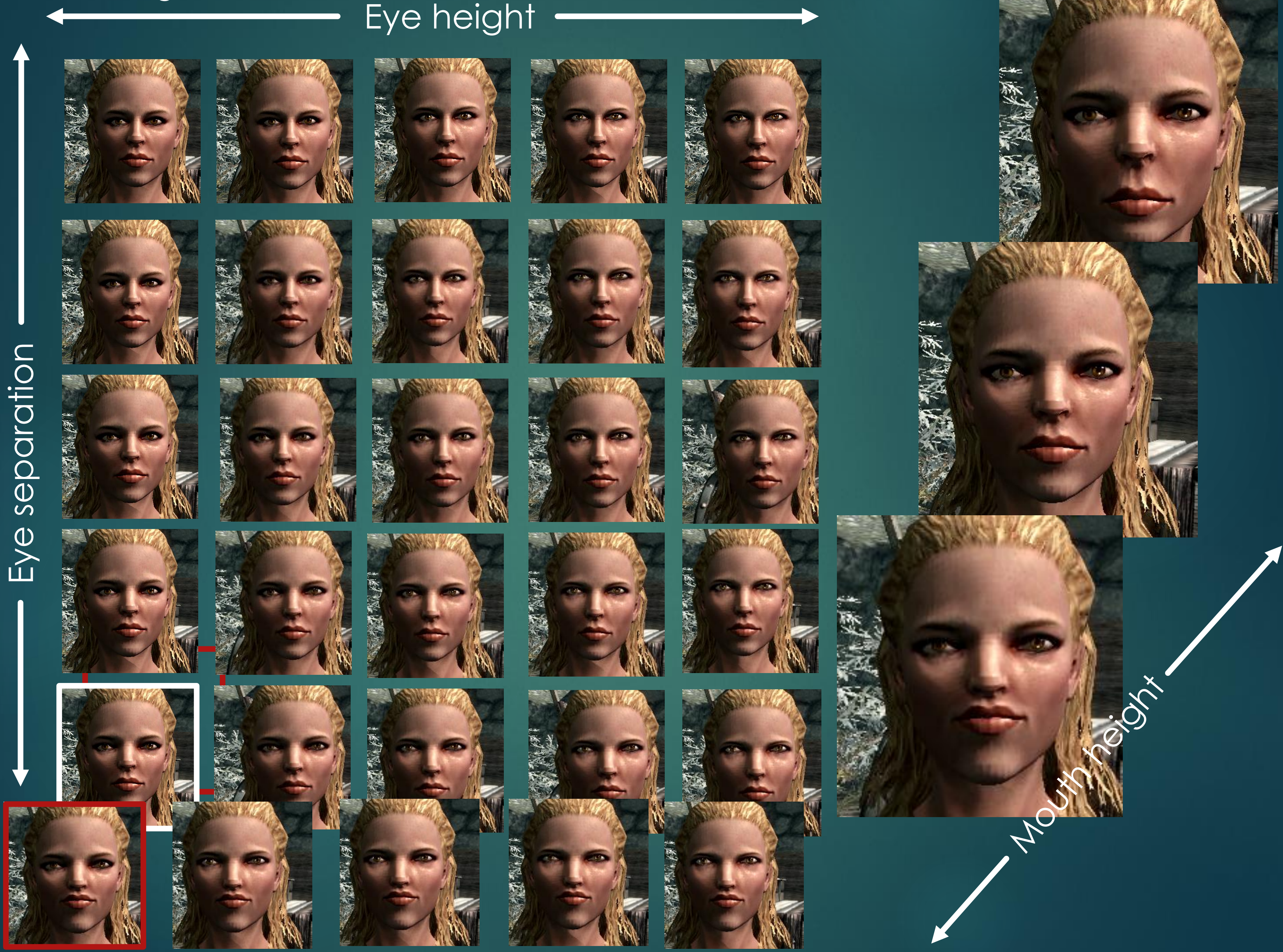


Eye height

Eye separation

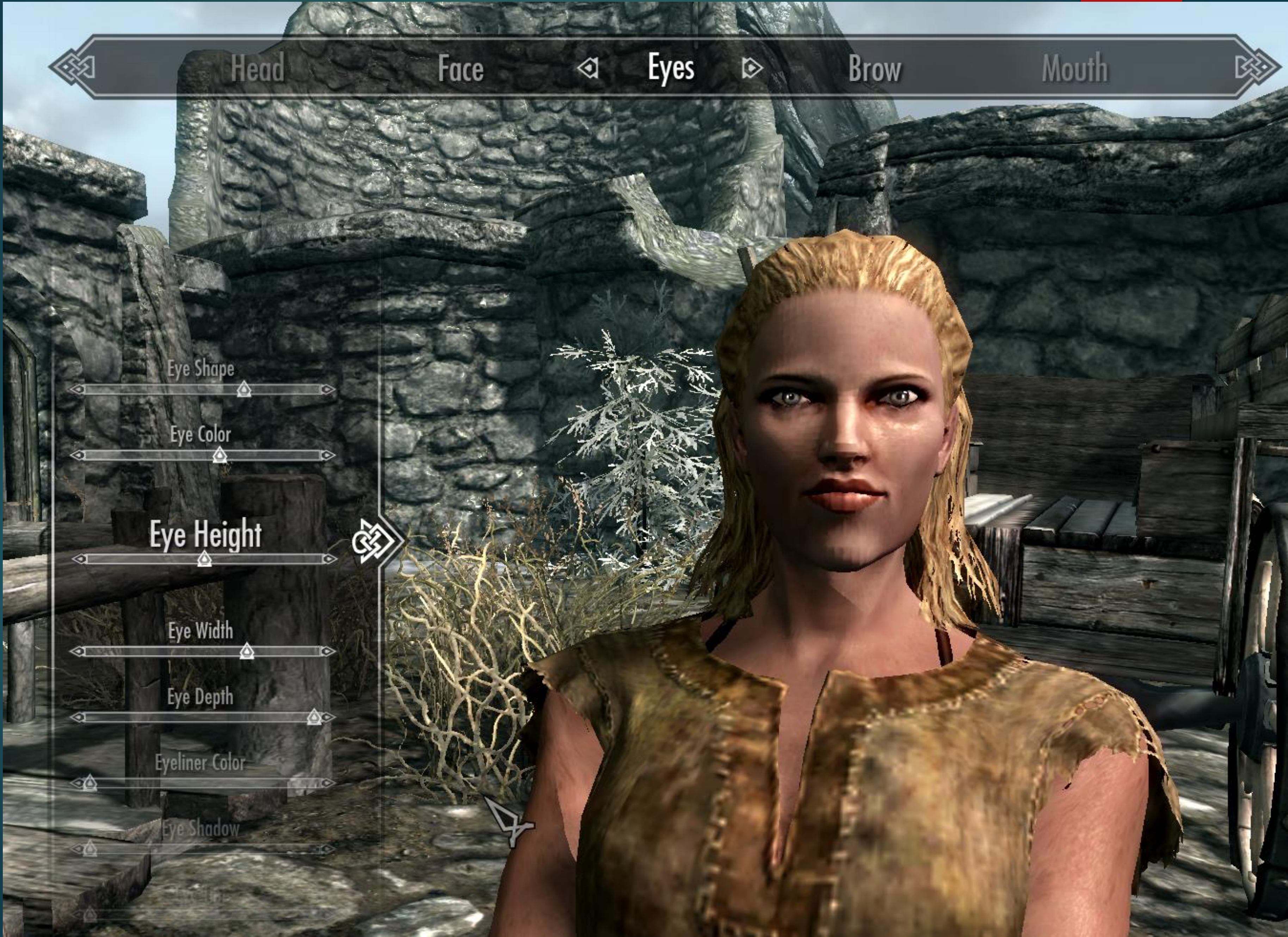
Mouth height

- Or arrange across three…



- But physical arrangements fall apart across more dimensions

- This was a constrained example (thanks, Skyrim character generator!)



- But face space is real (and very complex)!

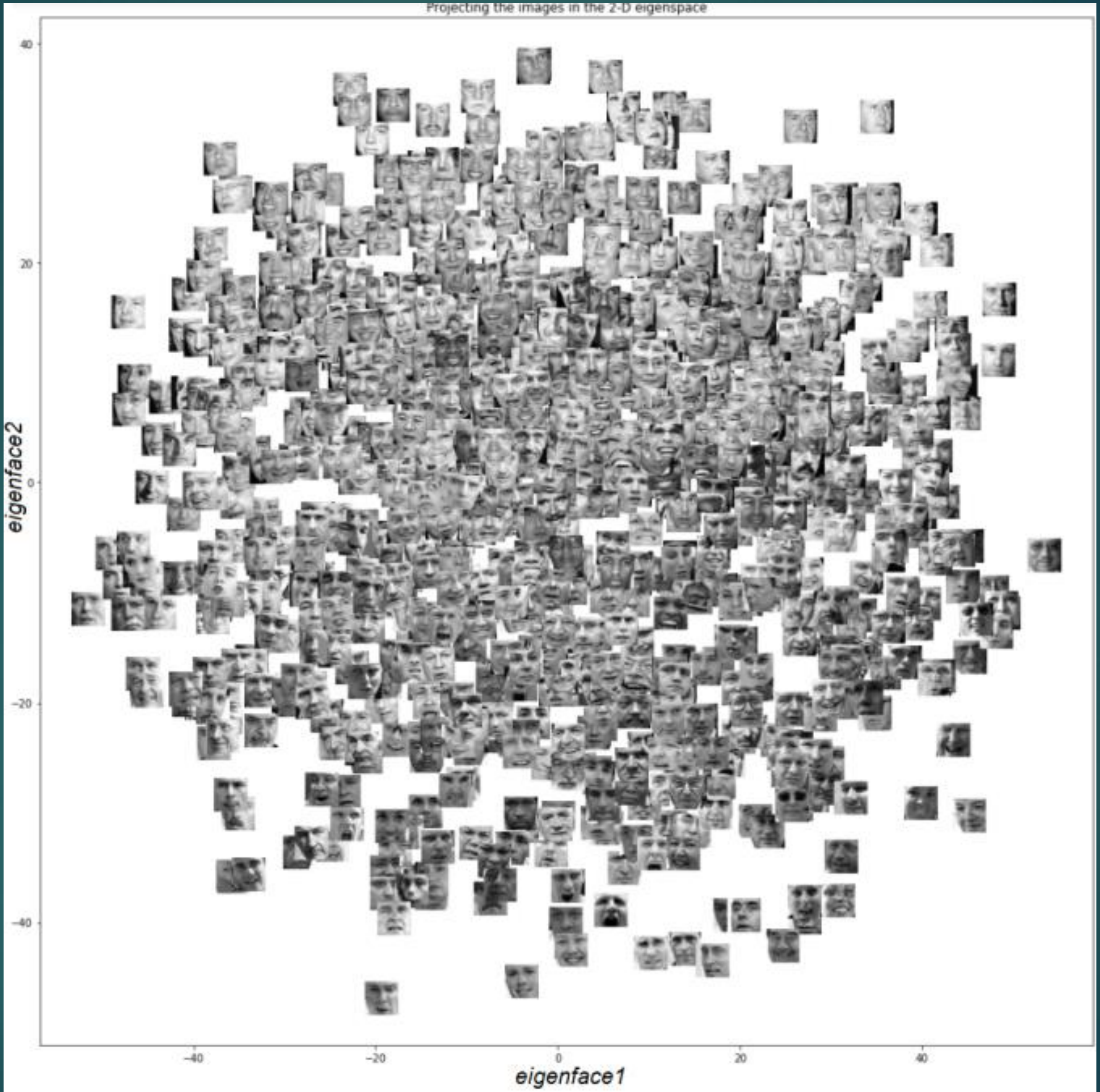


Image source: https://sandipanweb.files.wordpress.com

How to think in *n* dimensions

Faces are a good start, but there are many *n* dimensional constructs: everyone come up with a construct with more than 4 dimensions and share!

Image source: https://independent.co.uk/

Drawing the line: Raw score to PC

Conceptually, a **component** is a variable summarising the degree to which individuals in a sample vary on a target dimension of meaning.

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.

Drawing the line: Raw score to PC

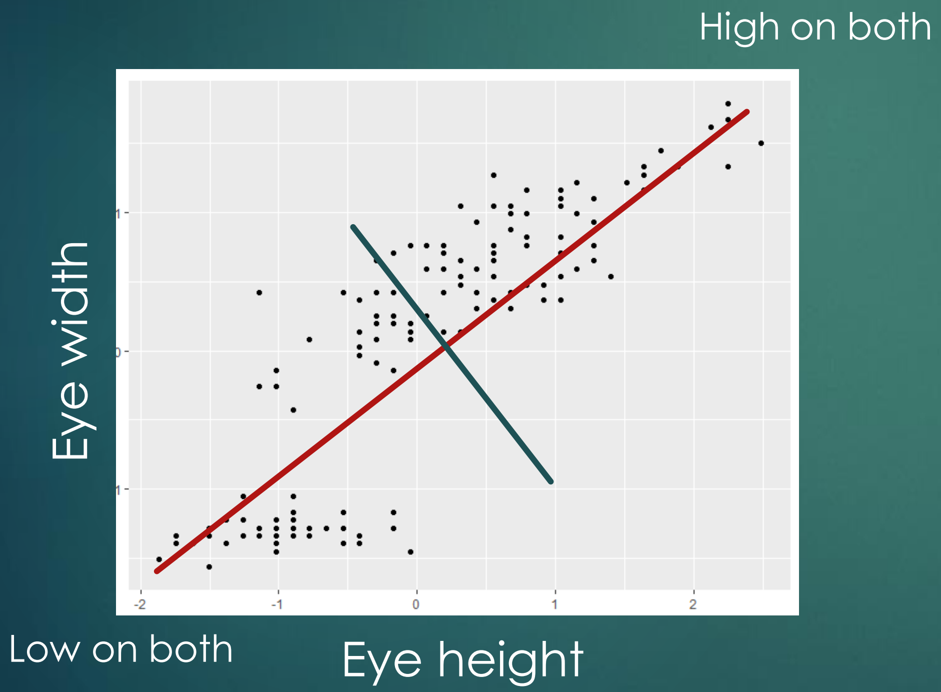
Raw scales could be considered a "component", but they…

- May not relate directly to the dimension of interest

Drawing the line: Raw score to PC

Raw scales could be considered a "component", but they…

- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
  - A principal component maximises the amount of variation captured.

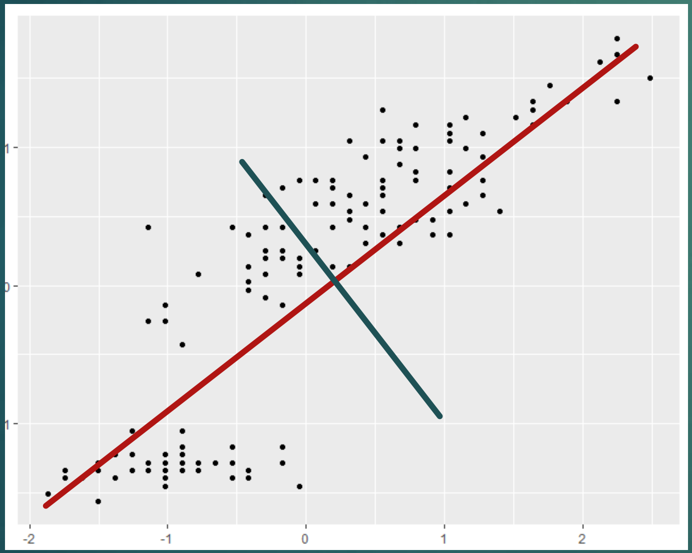


Drawing the line: Raw score to PC

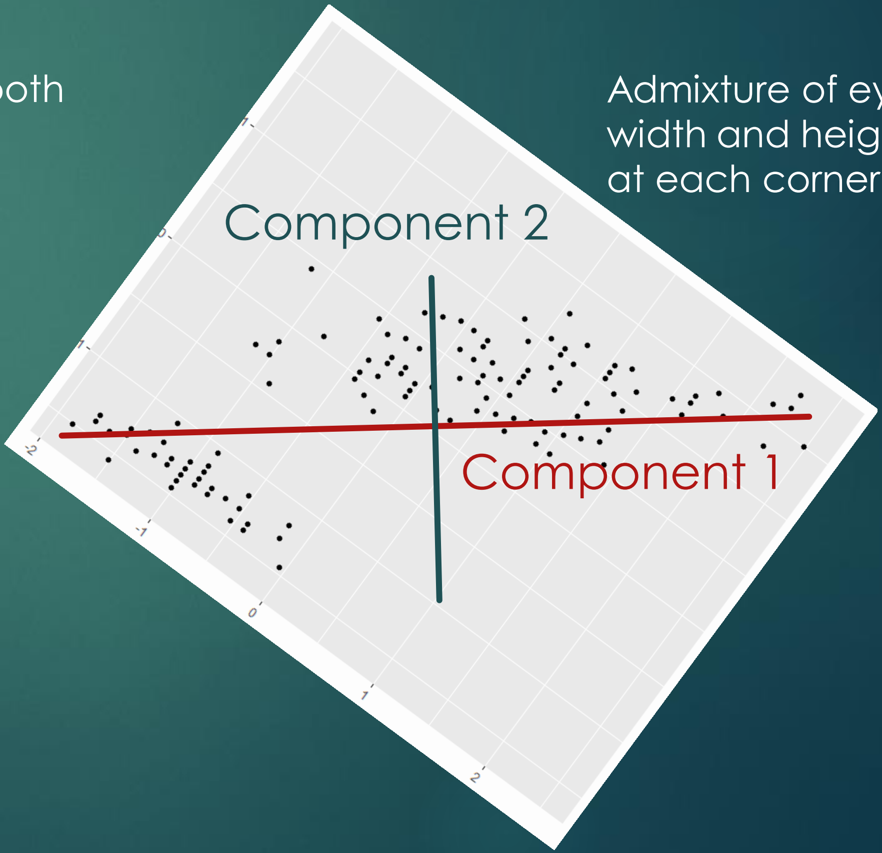
Raw scales could be considered a “component”, but they…

- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
  - A principal component maximises the amount of variation captured.

High on both



Low on both

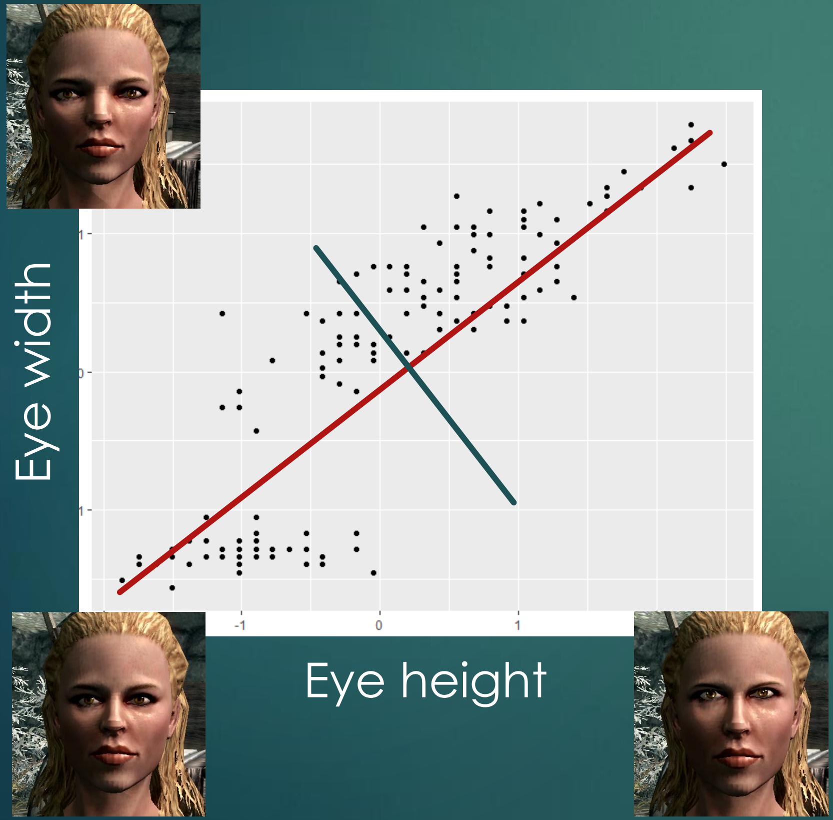


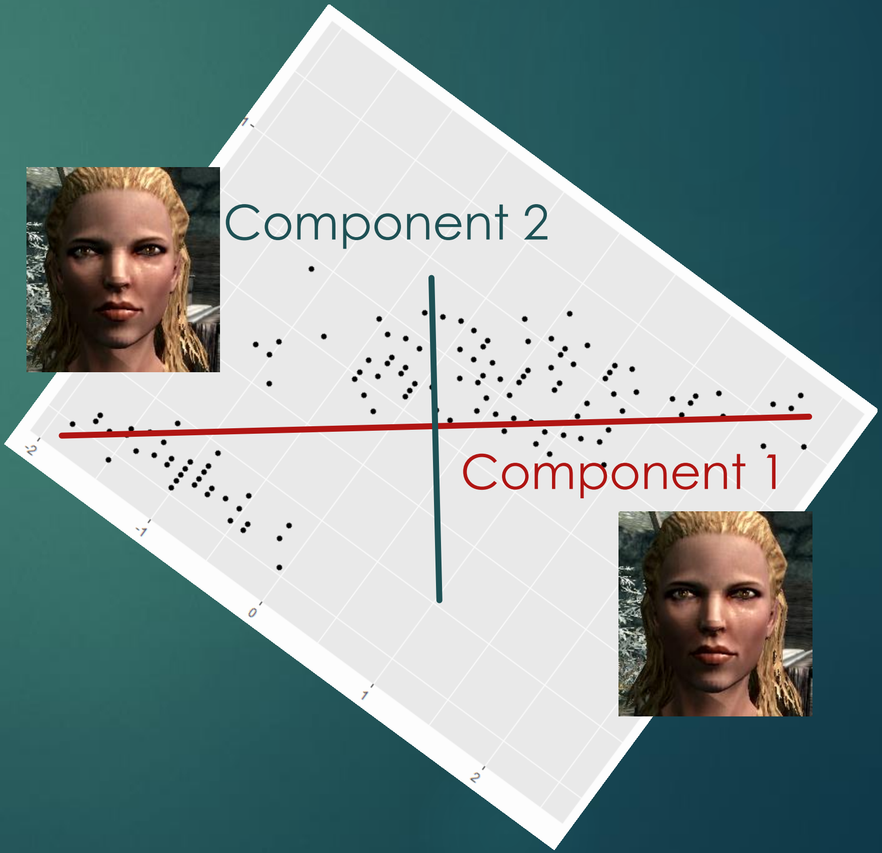
Admixture of eye width and height at each corner

Drawing the line: Raw score to PC

Raw scales could be considered a “component”, but they…

- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
  - A principal component maximises the amount of variation captured.





Drawing the line: Raw score to PC

Raw scales could be considered a “component”, but they…

- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
  - A principal component maximises the amount of variation captured.
  - This is conceptually the same as finding the line of best fit in a linear model.

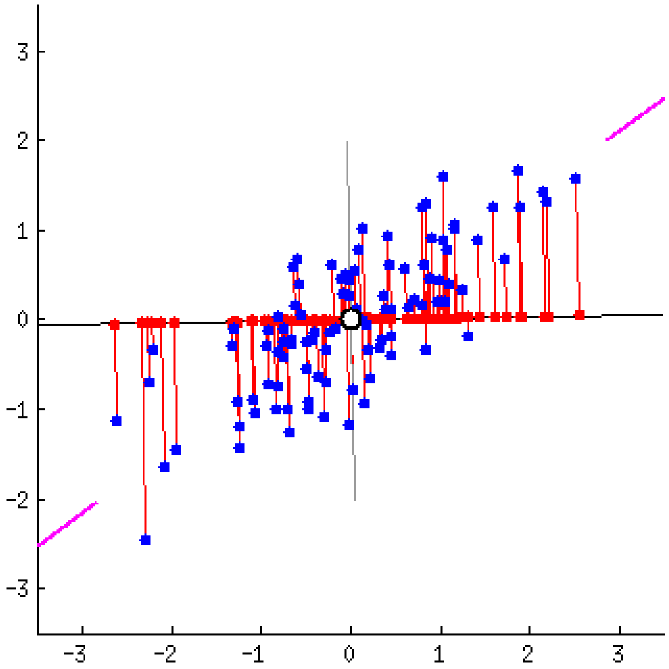
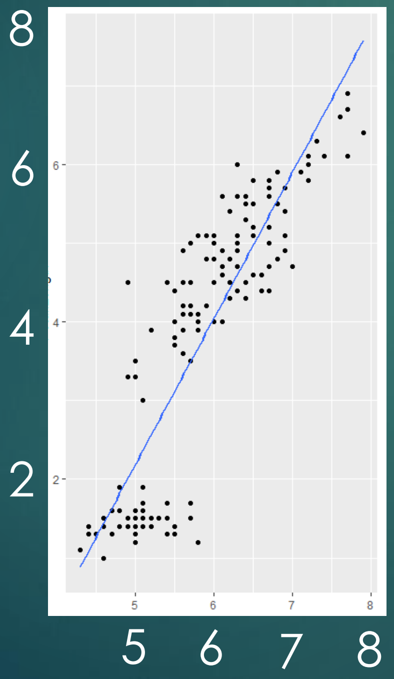


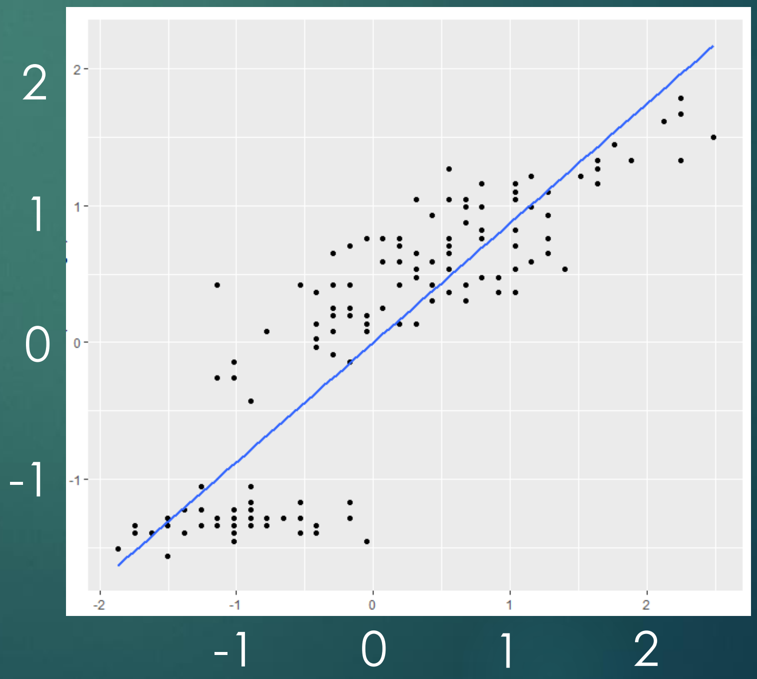
Image source: https://towardsdatascience.com/

Drawing the line: Raw score to PC

Raw scales could be considered a “component”, but they…

- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
  - A principal component maximises the amount of variation captured.
  - This is conceptually the same as finding the line of best fit in a linear model.
  - Standardize all variables first

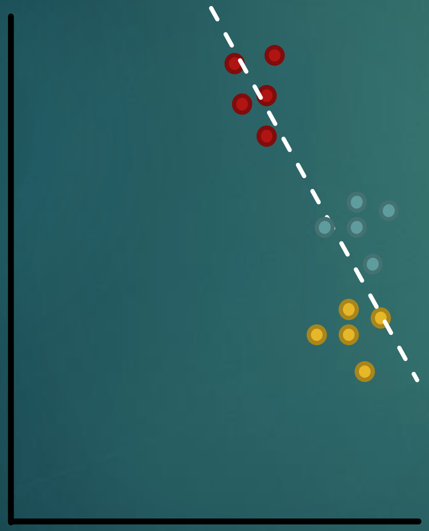




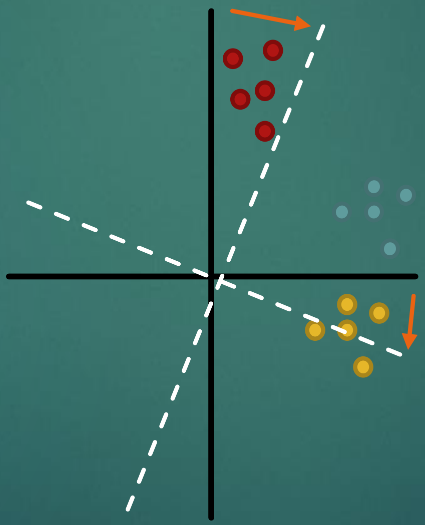
Drawing the line: Raw score to PC

Raw scales could be considered a “component”, but they…

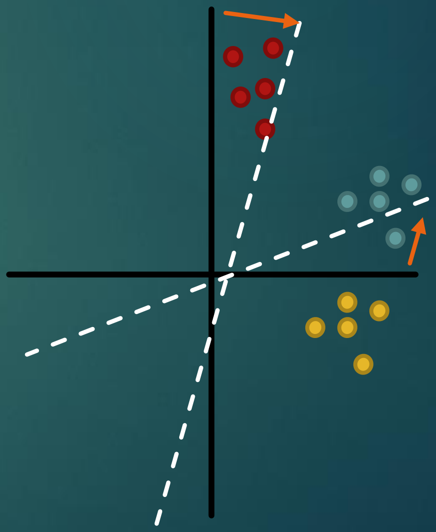
- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
- Have differing relationships to one another (covariance)
  - Typically Principal Components just take one of two highly correlated items
  - Side note: rotation (technically no longer PCA if you rotate)



PCA: find linear combination of best fit for all points



Varimax = orthogonal, does not allow them to be correlated



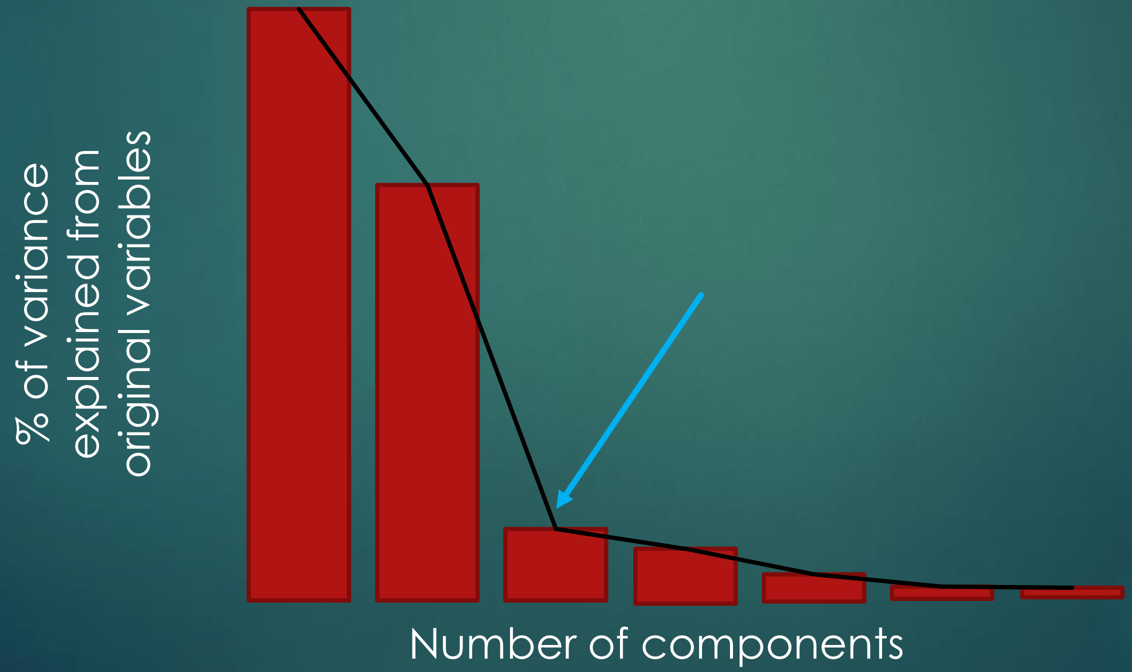
Oblimin = oblique, allows them to be correlated

Drawing the line: Raw score to PC

Raw scales could be considered a "component", but they…

- May not relate directly to the dimension of interest
- Have different degrees of variability (on different scales)
- Have differing relationships to one another (covariance)
- Give no guidance as to what is or is not a useful.

Principal components are fit iteratively, so first combinations explain the most variability. Less explained with each new component, so you can pick just a few based on a cut off (e.g. "elbow" of a scree plot)



Some terms…

- Dimension: any axis of possibly meaningful variation in your data
- Component: a variable summarising the degree to which individuals in a sample vary on a dimension
- Principal component: variables constructed from linear combinations or mixtures of the initial variables.
- Principal Component Analysis: dimension reduction method that transforming a large set of variables into a smaller set (principal components) that still contains most of the information (variability) in the original set.

Image source: https://www.lovelifedrawing.com/

PCA in general

1. Standardization (conversion to z scores):
   - Homogenizes range so they contribute equally to regression line of best fit
2. Covariance matrix computation
   - Computes associations between input variables, to detect what is redundant/can be dropped
3. Find the eigenvectors
   - The directions of the axes where there is the most variance. These are our principal components!

   … and the eigenvalues
   - The amount of variance each principal component explains. This is the values on the scree plot
4. Find the feature vector
   - Arrange our eigenvectors by eigenvalues to see which to keep
5. Recast the data to this new shape space

- Gives you the information you're after
  - Eigenvectors: how much each of the original variables contributes to each principal component
    - Similar to concept of "loading" in FA, though here 'loadings' are orthonormal Eigenvectors
  - Position: where each individual falls on each principal component
    - This is what you'll use in later analyses if you were using PCA for dimension reduction

PCA in R

- Standardization
- Covariance matrix computation
- Fitting the PCA
- Evaluating the PCA
  - Eigenvalues
  - Eigenvectors
  - Position on components

We'll run through an example together, then you try!

Image source: Daniel Gonzalez on unsplash.com