Agenda

1. Housekeeping
   - Homework due Friday
   - Good job Colin!
   - Final presentation times posted
   - Final guidelines posted
   - What the presentations will be like
2. Principal Component Analysis
3. Thursday: final quickfire!

Introduction to Principal Component Analysis

Exploratory Data Analysis

Dr. Cassy Dorff

Principal Component Analysis

- Unsupervised learning
- PCA is both intuitive and technical
- Basic situation
  - you have a high # of variables and you want to use them to:
    a) get a sense of patterns in the data
    b) predict (model) an outcome
    c) reduce dimensions of your data

Principal Component Analysis

You might ask the question:

How do I take all of the variables I've collected and focus on only a few of them?

In technical terms, you want to:

reduce the dimension of your feature space

Dimensionality Reduction

feature elimination: remove features (variables)

- pro: simple
- con: we get no information from these data
- con: requires strong assumptions about the data

Dimensionality Reduction

feature extraction: create new features from combinations of current variables, importance = captures variation

- pro: maintain data and use a principled way of deciding what is ‘important’
- con: a bit more complicated under the hood
- this is what PCA does

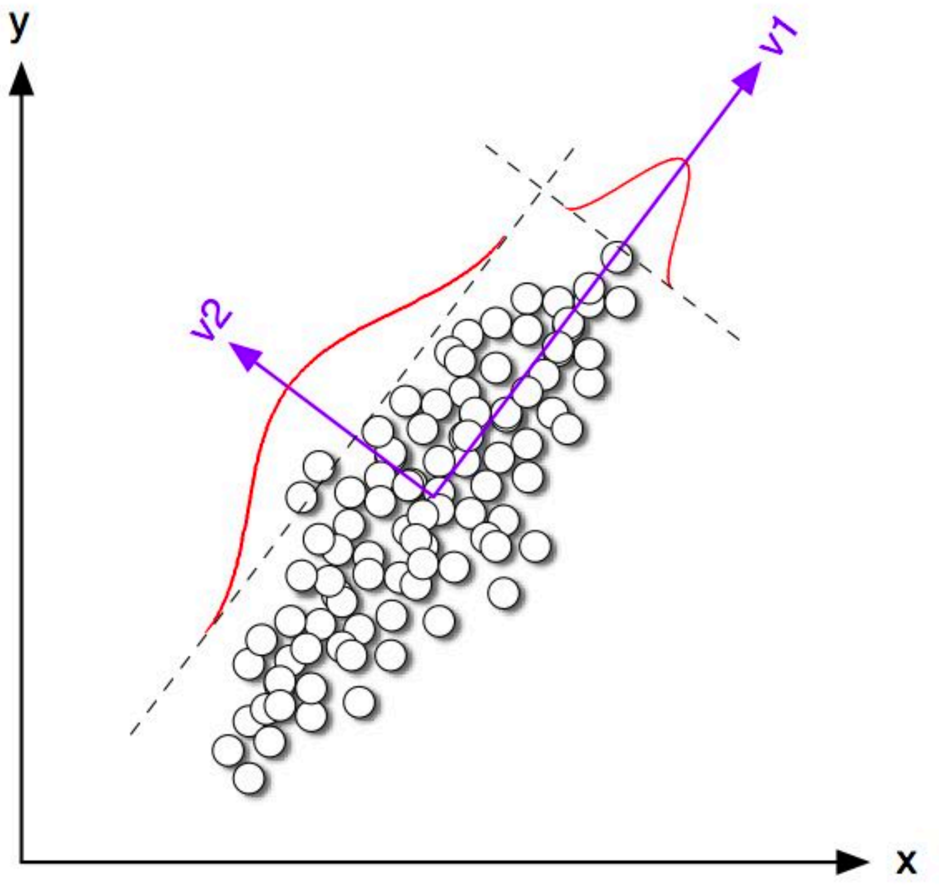
Principal Component Analysis

- When faced with a large set of variables, principal components allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set.

How does it work?

- Covariance matrix: a measure of how each variable is associated with one another
- Eigenvectors: represent directions (imagine a multi-dimensional scatterplot)
- Eigenvalues: represent the magnitude or importance of these directions
- Assumption: more variability in a given direction means higher importance for prediction

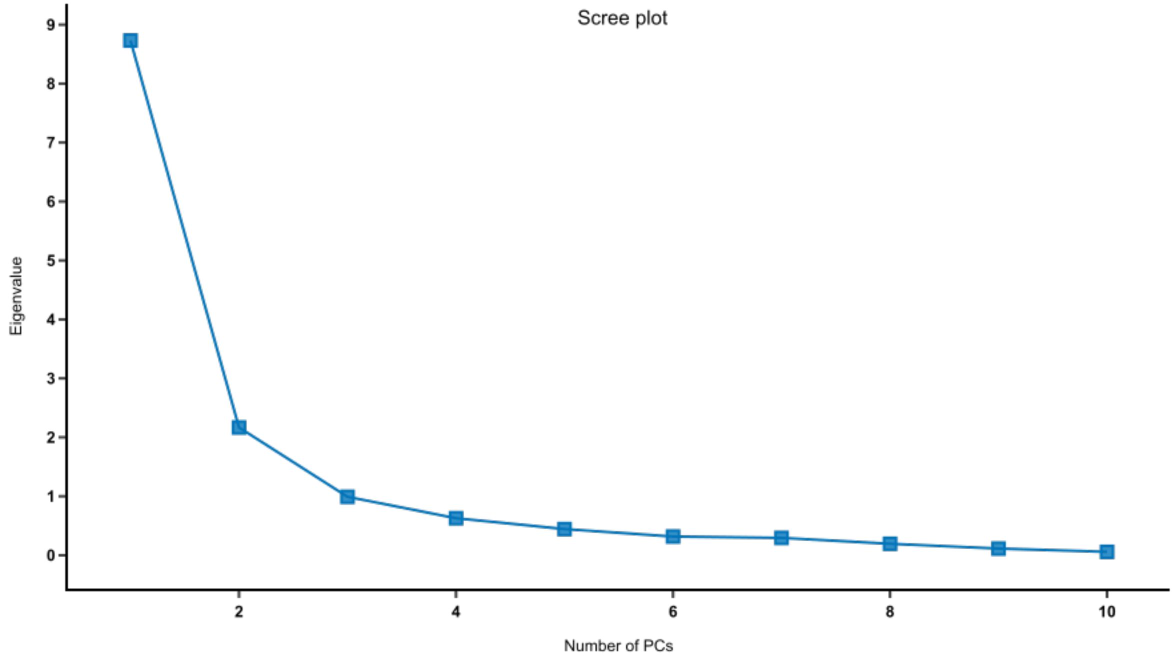
- PCA’s goal is to rotate the coordinate axes of a p-dimensions linear space to new ‘natural’ positions (aka principal axes) such that:
- coordinate axes are ordered: principal axis 1 corresponds to the highest variance in the data, principal axis 2 has the next highest variance, and so on
- the covariance among each pair of principal axes is zero, i.e. the components are uncorrelated



https://stats.stackexchange.com/questions/183236/what-is-the-relation-between-k-means-clustering-and-pca

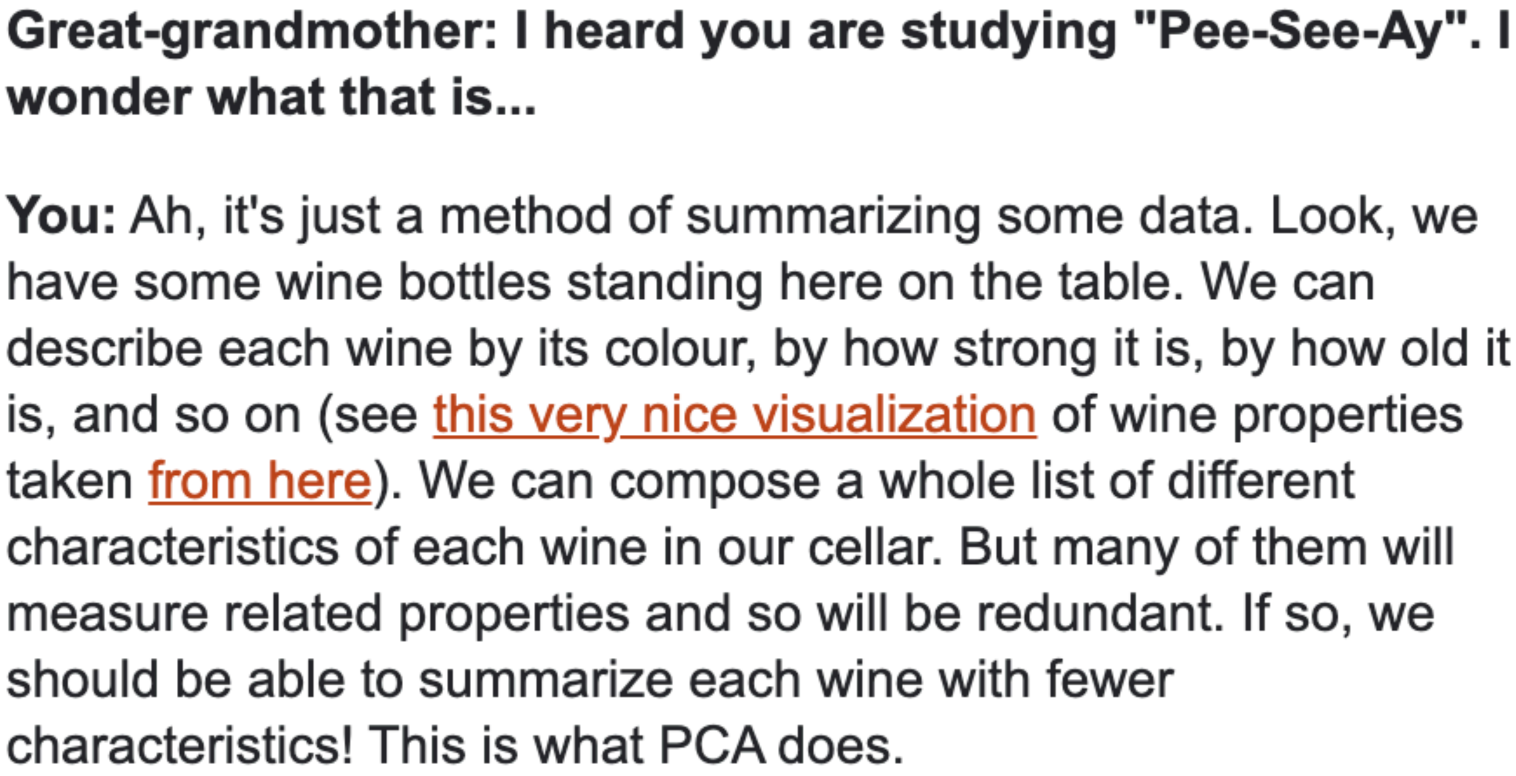
Choosing # of Principal Components

1. Simplicity and convenience (2 dimensions)
2. Scree plot
   - proportion of variance explained
   - (can add cumulative proportion of variance explained as you keep more features)



Tons of information out there:

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

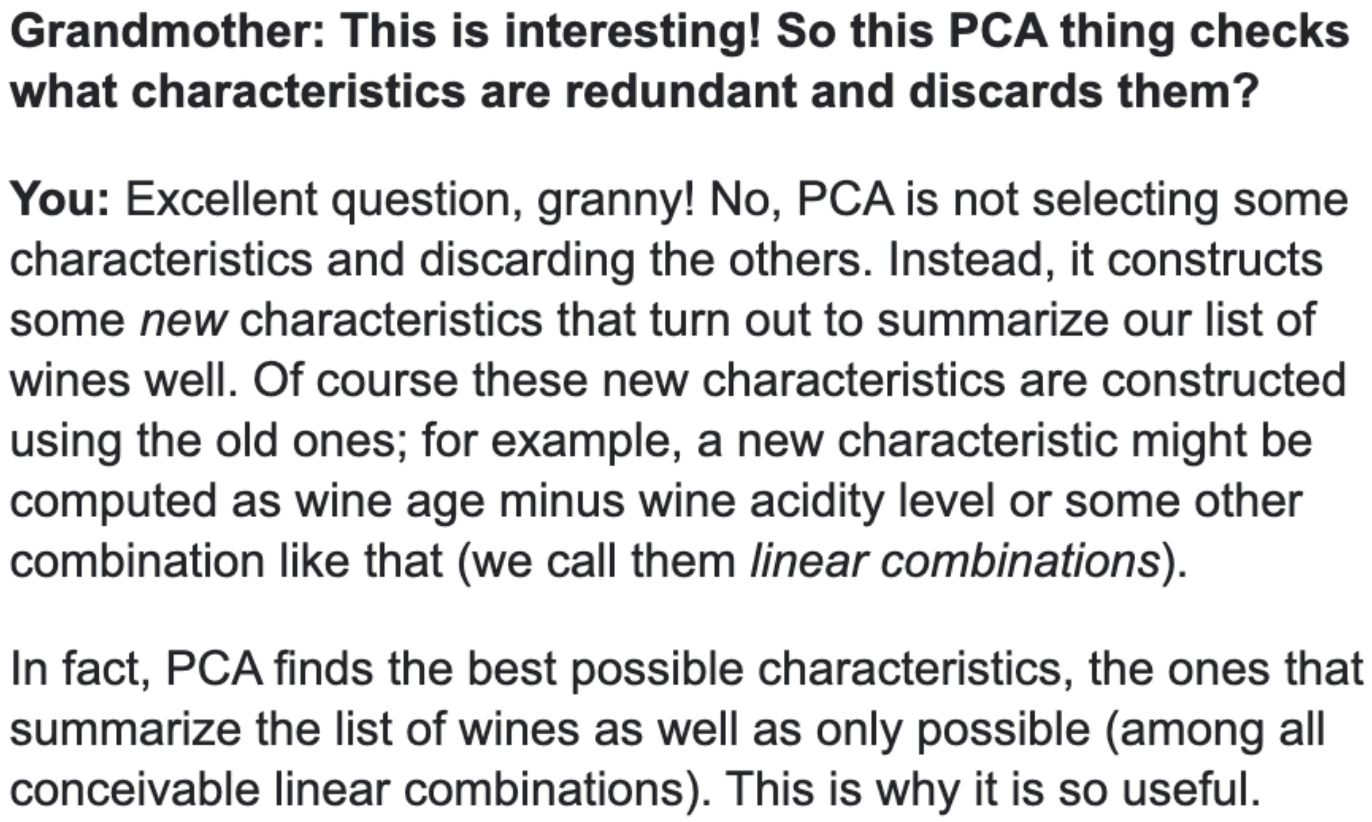


Great-grandmother: I heard you are studying "Pee-See-Ay". I wonder what that is...

You: Ah, it's just a method of summarizing some data. Look, we have some wine bottles standing here on the table. We can describe each wine by its colour, by how strong it is, by how old it is, and so on (see this very nice visualization of wine properties taken from here). We can compose a whole list of different characteristics of each wine in our cellar. But many of them will measure related properties and so will be redundant. If so, we should be able to summarize each wine with fewer characteristics! This is what PCA does.

Tons of information out there:

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues



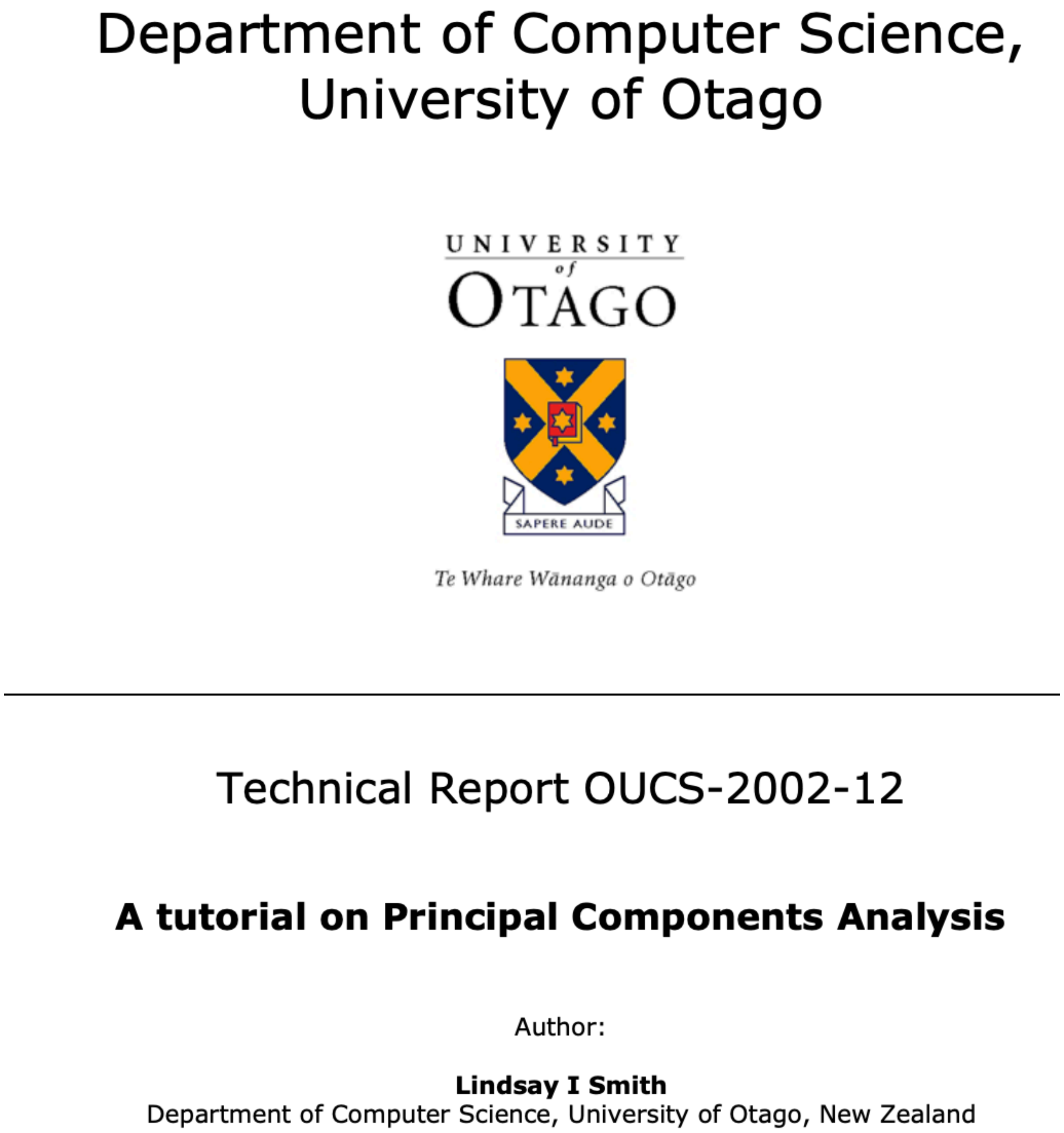
Grandmother: This is interesting! So this PCA thing checks what characteristics are redundant and discards them?

You: Excellent question, granny! No, PCA is not selecting some characteristics and discarding the others. Instead, it constructs some *new* characteristics that turn out to summarize our list of wines well. Of course these new characteristics are constructed using the old ones; for example, a new characteristic might be computed as wine age minus wine acidity level or some other combination like that (we call them *linear combinations*).

In fact, PCA finds the best possible characteristics, the ones that summarize the list of wines as well as only possible (among all conceivable linear combinations). This is why it is so useful.

Recommended Reading

- https://ourarchive.otago.ac.nz/bitstream/handle/10523/7534/OUCS-2002-12.pdf?sequence=1&isAllowed=y



Department of Computer Science, University of Otago

UNIVERSITY of OTAGO

SAPERE AUDE

Te Whare Wānanga o Otāgo

Technical Report OUCS-2002-12

A tutorial on Principal Components Analysis

Author:

Lindsay I Smith
Department of Computer Science, University of Otago, New Zealand

How might you use PCA?

- I might collect data on hundreds of registered horses' physical attributes to model success at the races. Perhaps PCA would pick up on size and breeding location?
- Social scientists use it to predict behavior and generate forecasting models (interpretation of IV is less important than prediction accuracy)
- Equity portfolios: to reduce portfolio risk, allocation strategies are applied to the 'principal portfolios' instead of underlying stocks