The goal of principal component analysis is to identify the

most meaningful basis to re-express a data set. The hope is

that this new basis will filter out the noise and reveal hidden

structure.

•To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component

**Principal Component Analysis (PCA)**is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

What is an orthonormal basis for (xA;yA)?

A naive choice would be f(1;0); (0;1)g

From linear

algebra we know that all measurement vectors form a linear

combination of this set of unit length basis vectors.

With this assumption

PCA is now limited to re-expressing the data as a linear

combination of its basis vectors.

covariance o f A and B \_ s2

AB =

1

nå

i

aibi

CX \_

1

n

XXT :

CX captures the covariance between all possible pairs of measurements.

The covariance values reflect the noise and redundancy

in our measurements.

\_ In the diagonal terms, by assumption, large values correspond

to interesting structure.

\_ In the off-diagonal terms large magnitudes correspond

to high redundancy.

**Factor loadings:** Correlation between the original variables and the factors, and

the key to understanding the underlying nature of a particular factor. Squared

factor loadings indicate what percentage of the variance in an original variable is

explained by a factor.

* 1. Loadings1that are:more than .5 are typically considered strong

**Eigenvalues:** Also called characteristic roots. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor.

**Orthogonal factor rotation:**Factor rotation in which the factors are extracted so that their axes are maintained at 90 degrees. Each factor is independent of, or orthogonal to, all other factors. The correlation between tehfactors is determined to be zero.

**VARIMAX:**One of the most popular orthogonal factor rotation methods.

1. An orthogonal rotation is just a shift to a new set of coordinate axes in the same space spanned by the principal components

You might be thinking that you could group some of the above variables (manifest variables) above together to represent a particular aspect of patient satisfaction with their GP such as personality, knowledge and treatment.

Two other things to note; firstly often the observable variables are questions in a questionnaire and can be thought of as **items** and consequently each subset of items represents a **scale**

All factor analysis techniques try to clump subgroups of variables together based upon their correlations.

The MSA does not produce a P value but we are aiming for a value over 0.8

PCA in R1

* 1. Package nameFunction name
  2. baseprincomp
  3. psychprincipal
  4. VSS
  5. pcaMethodsAs the name implies this package is all about PCA, and from a modern approach. Will automatically estimate missing values (via traditional, robust, or Bayesian methods) and is useful just for that for any analysis.
  6. pca
  7. Q2 for cross validation
  8. FactoMiner R-commander plugin

• The KMO and Bartlett’ test evaluate all available data

together. A KMO value over 0.6 and a significance level

for the Bartlett’s test below .05 indicate there is a

reasonable amount of correlation in the data

• MSA values how strongly a single variable is correlated

with all the other variables combined. Values above .5

1. are generally considered acceptable

Toothpaste Data

30 respondents were asked to rate the following statements on a 7-point Likert Scale:

1. 1. It is important to buy a toothpaste that prevents cavities.
2. 2. I like a toothpaste that gives shiny teeth.
3. 3. A toothpaste should strengthen gums.
4. 4. I prefer a toothpaste that freshens breath.
5. 5. Prevention of tooth decay is not an important benefit offered
6. by a toothpaste. (reverse coded)
7. 6. An important consideration in buying a toothpaste is
8. attractive teeth.
9. 1 = strongly disagree; 7 = strongly agree

• A further criteria that is often used to determine the number

of factors is the scree-plot. This is a plot of the eigenvalues

against the number of factors, in order of extraction. Often a

break or ‘elbow’ is visible in the plot. Factors up to and

including this elbow are then selected for further analysis,

assuming they all have eigenvalues above 1

vnijs.rady.ucsd.edu/site\_media/materials/tutorials/**R**.**factoranalysis**.pd