Lecture 7B. Unsupervised Techniques I

DS-GA 1015, Text as Data
Arthur Spirling

March 23, 2021

Where Are We?

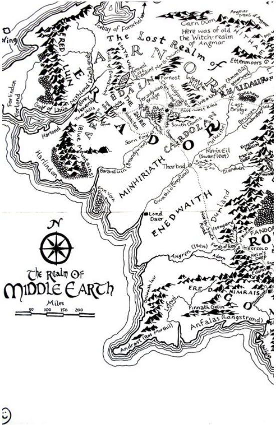
Where Are We?

The Realm of Middle Earth
Miles
50 100 150 200
The Lost Realm of Arnor
Here was of old the Witch-realm of Angmar
Carn Dûm
Angmar
Mtns of Angmar
Icebay of Forochel
Ettenmoors
Eriador
Ered Luin
Forlindon
Forlond
Gulf of Lhûn
Mithlond
(Grey Havens)
Tower Hills
Harlond
Harlindon
(Blue Mountains)
Arthedain
Emyn Uial
Nenuial
(Lake Evendim)
North Downs
Fornost
Weather Hills
Weathertop
(Amon Sûl)
Rhudaur
Trollshaws
Last Bridge
Bruinen
(Loudwater)
The Shire
Hobbiton
White Downs
Brandywine Bridge
Bree
Old Forest
East-West Road
South Downs
Midgewater
Chetwood
Sarn Ford
Baranduin
(Brandywine)
Cardolan
Mitheithel
Eregion
(Hollin)
Moria Gate
Nîn-in-Eilph
(Swanfleet)
Tharbad
Glanduin
Minhiriath
Eryn Vorn
Gwathlo (Greyflood)
Enedwaith
Dunland
Lond Daer
North-South Road
Isengard
Fangorn
Fords of Isen
Angren (Isen)
Adorn
Westfold
Helm's Deep
Edoras
Ered Nimrais
Gondor
Pinnath Gelin
Lefnui
Anfalas (Langstrand)
Andrast (Ras Morthil)

Where Are We?

We've covered supervised learning: the situation in which we have labeled data.

Where Are We?



We've covered supervised learning: the situation in which we have labeled data.

Now begin to think about documents that are not labelled but within which we expect some important, latent structure.

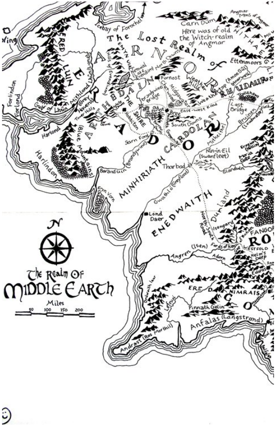
Where Are We?

We've covered supervised learning: the situation in which we have labeled data.

Now begin to think about documents that are not labelled but within which we expect some important, latent structure.

cover some fundamental techniques for accessing that structure

Where Are We?



We've covered supervised learning: the situation in which we have labeled data.

Now begin to think about documents that are not labelled but within which we expect some important, latent structure.

cover some fundamental techniques for accessing that structure

and demonstrate challenges that emerge in interpreting the results.

Terminology

Terminology

Unsupervised techniques:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—

Terminology

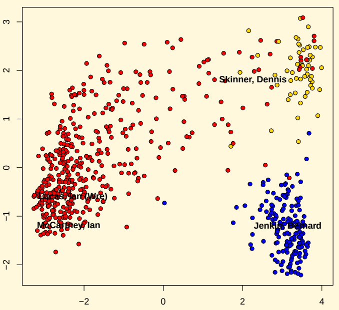
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

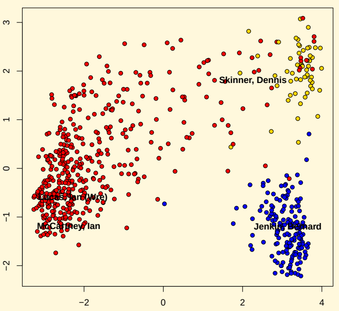
e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

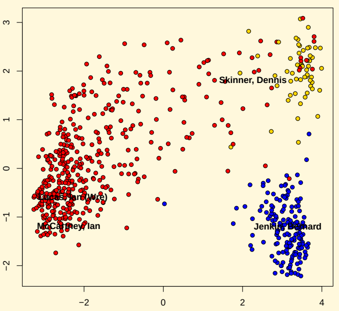


Supervised techniques:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

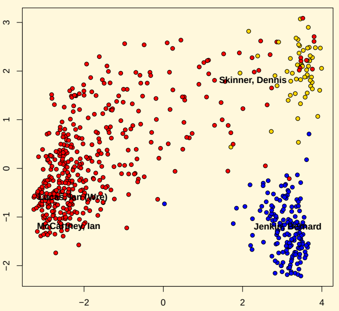


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



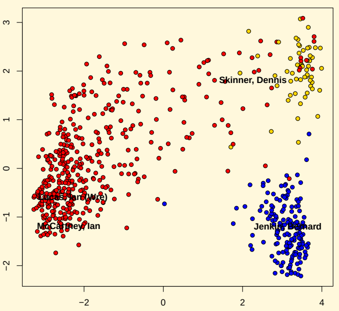
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



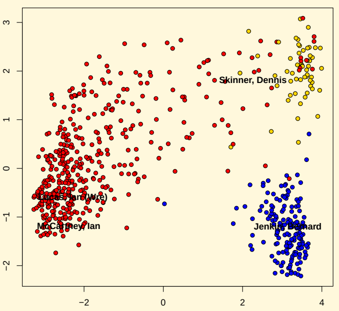
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ($y \in \{0, 1\}$)?

Terminology

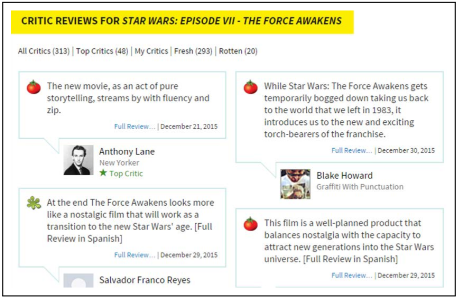
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ($y \in \{0, 1\}$)?



Overview: Unsupervised Learning

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties,

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is,

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers,

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

→ look for (dis)similarities between documents (or observations):

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

$\rightarrow$ look for (dis)similarities between documents (or observations): it will be up to us to *interpret* what the groups/dimensions/concepts represent after the technique has been used.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

→ look for (dis)similarities between documents (or observations): it will be up to us to *interpret* what the groups/dimensions/concepts represent after the technique has been used.

So. . .

So. . .

in contrast to supervised approaches,

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible?

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible? do somewhat similar (e.g. 2, 3, 4 clusters) specifications imply the same thing?

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible? do somewhat similar (e.g. 2, 3, 4 clusters) specifications imply the same thing?

(not "what is the recall/precision/accuracy?")

Motivating Problem

Motivating Problem

Have an $n \times p$ matrix,

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze:

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Name	Party	Vote 1	Vote 2	Vote 3	
Ainsworth, Peter (E S)	Con	NA	1	NA	...
Alexander, Douglas	Lab	NA	0	0	...
Allan, Richard	LD	1	0	1	...
Allen, Graham	Lab	0	0	0	...
Amess, David	Con	1	1	NA	...
					$\ddots$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Text: n speakers, p features in the speeches (often $p > n$ for text problems)

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Text: n speakers, p features in the speeches (often $p > n$ for text problems)

Name	Party	'cost'	'spend'	'tax'	
Ainsworth, Peter (E S)	Con	0.00	0.01	0.30	...
Alexander, Douglas	Lab	0.32	0.20	0.86	...
Allan, Richard	LD	0.99	0.82	0.61	...
Allen, Graham	Lab	0.52	0.86	0.34	...
Amess, David	Con	0.07	0.34	0.33	...
					$\ddots$

Principal Components Analysis

PCA: Introduction

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data

- reduce the data/problem down to this information

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data
- reduce the data/problem down to this information
- simplify data

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data
- reduce the data/problem down to this information
- simplify data
- analyze data in terms of its patterns/groups

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data
- reduce the data/problem down to this information
- simplify data
- analyze data in terms of its patterns/groups

Generally: represent this information as new (and smaller number of) variables known as *principal components*

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data
- reduce the data/problem down to this information
- simplify data
- analyze data in terms of its patterns/groups

Generally: represent this information as new (and smaller number of) variables known as *principal components*

Overview

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other,

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional 'map' of observations in new space:

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional 'map' of observations in new space:

$\rightarrow$ each observation now has a value on each principal component called its (factor) score,

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional 'map' of observations in new space:

$\rightarrow$ each observation now has a value on each principal component called its (factor) score, which are projections of (original) observations onto the PCs

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional 'map' of observations in new space:

$\rightarrow$ each observation now has a value on each principal component called its (factor) score, which are projections of (original) observations onto the PCs

Interpretation of given PC: depends on correlation between component and (original) variable—known as loading

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional 'map' of observations in new space:

$\rightarrow$ each observation now has a value on each principal component called its (factor) score, which are projections of (original) observations onto the PCs

Interpretation of given PC: depends on correlation between component and (original) variable—known as loading

Method: (eigen-) decomposition of cov matrix or singular value decomposition of data matrix

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional 'map' of observations in new space:

$\rightarrow$ each observation now has a value on each principal component called its (factor) score, which are projections of (original) observations onto the PCs

Interpretation of given PC: depends on correlation between component and (original) variable—known as loading

Method: (eigen-) decomposition of cov matrix or singular value decomposition of data matrix

Method

Method

PCA performs a linear transformation

Method

PCA performs a linear transformation on the original variables into new coordinate system,

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points:

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but,

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line,

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance (cf OLS)

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance (cf OLS)

All subsequent components captures (sequentially) less variability

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance (cf OLS)

All subsequent components captures (sequentially) less variability

Assumptions: observations are independent

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance (cf OLS)

All subsequent components captures (sequentially) less variability

Assumptions: observations are independent and X is p-variate normal (may not find highest variance projection if not)

Method

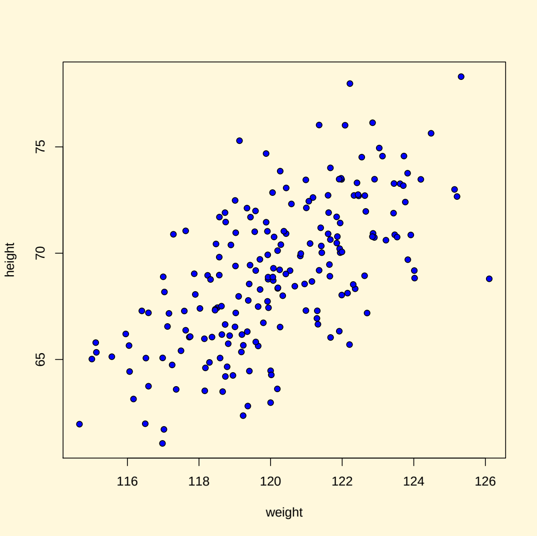
PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance (cf OLS)

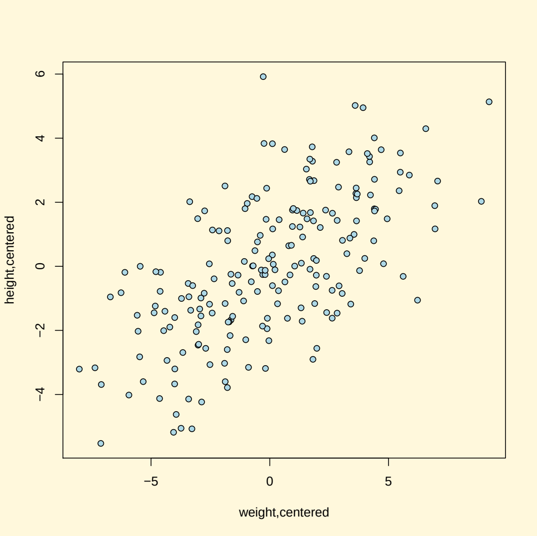
All subsequent components captures (sequentially) less variability

Assumptions: observations are independent and X is p-variate normal (may not find highest variance projection if not)

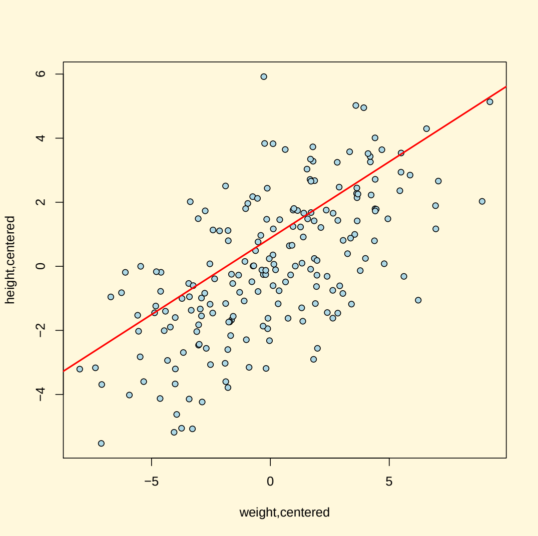
Heights, Weights



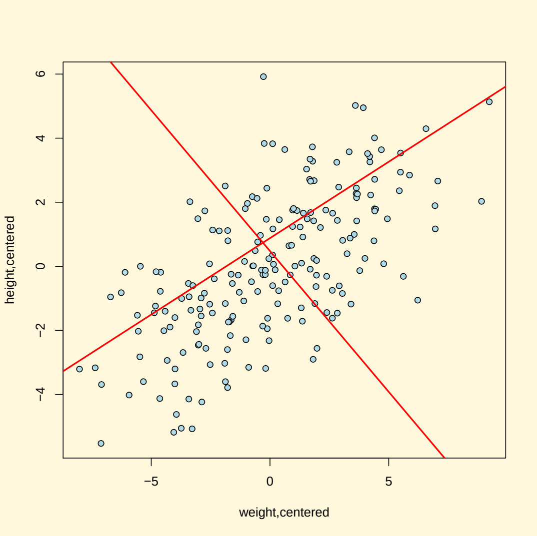
Heights, Weights



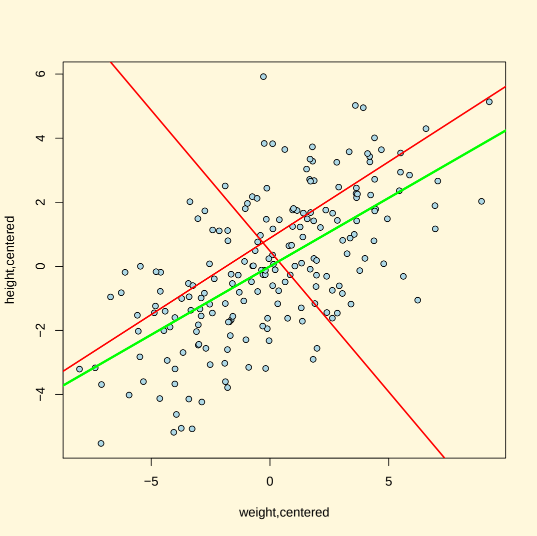
Heights, Weights



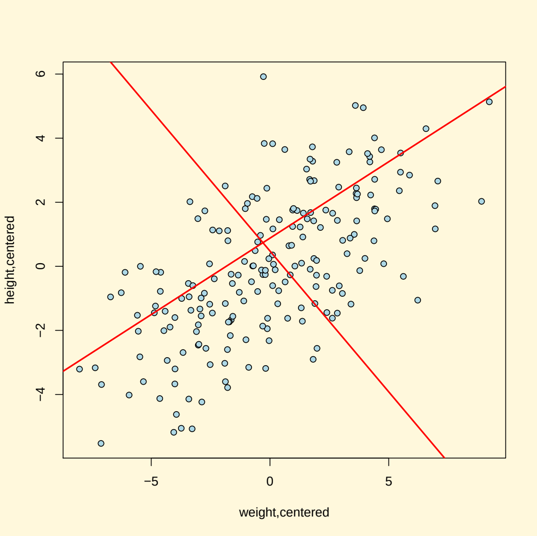
Heights, Weights



Heights, Weights



In new space



In new space

Formally

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1}x_i$

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a}_1'\mathbf{X} = \sum_{i=1}^{p} a_{i1}x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1} x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1} x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2)

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1}x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\mathrm{var}(z_1)$ is maximized

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1}x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\mathrm{var}(z_1)$ is maximized

NB We have to constrain $\sum(a_1^2) = a_{11}^2 + a_{12}^2 + \ldots a_{1p}^2 = 1$ (why?)

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1} x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\mathrm{var}(z_1)$ is maximized

NB We have to constrain $\sum(a_1^2) = a_{11}^2 + a_{12}^2 + \ldots a_{1p}^2 = 1$ (why?)

And we call z_{i1} the principal component score for the ith observation on the 1st dimension

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1}x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\mathrm{var}(z_1)$ is maximized

NB We have to constrain $\sum(a_1^2) = a_{11}^2 + a_{12}^2 + \ldots a_{1p}^2 = 1$ (why?)

And we call z_{i1} the principal component score for the ith observation on the 1st dimension

Subsequent PCs are found subject to the constraint that

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1}x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\mathrm{var}(z_1)$ is maximized

NB We have to constrain $\sum(a_1^2) = a_{11}^2 + a_{12}^2 + \ldots a_{1p}^2 = 1$ (why?)

And we call z_{i1} the principal component score for the ith observation on the 1st dimension

Subsequent PCs are found subject to the constraint that $\mathrm{cov}(z_k, z_l) = 0$ and that $a_k' a_k = 1$.

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a_1'X} = \sum_{i=1}^{p} a_{i1} x_i$

where $\mathbf{X}$ is the $p \times n$ data matrix (here, 2×200), $\mathbf{a_1} = (a_1, a_2, \ldots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\mathrm{var}(z_1)$ is maximized

NB We have to constrain $\sum(a_1^2) = a_{11}^2 + a_{12}^2 + \ldots a_{1p}^2 = 1$ (why?)

And we call z_{i1} the principal component score for the ith observation on the 1st dimension

Subsequent PCs are found subject to the constraint that $\mathrm{cov}(z_k, z_l) = 0$ and that $a_k' a_k = 1$.

btw Presumably, we wouldn't fit two components to two variables (why?)

So. . .

So. . .

a_k obtained via an singular value decomposition of `x` via `prcomp()`

So...

a_k obtained via an singular value decomposition of `x` via `prcomp()`

Alternative:

So. . .

a_k obtained via an singular value decomposition of `x` via `prcomp()`

Alternative: obtain a_k via eigenvectors of original data's var-cov matrix

So. . .

a_k obtained via an singular value decomposition of `x` via `prcomp()`

Alternative: obtain a_k via eigenvectors of original data's var-cov matrix (via `princomp()`)

So...

a_k obtained via an singular value decomposition of `x` via `prcomp()`

Alternative: obtain a_k via eigenvectors of original data's var-cov matrix (via `princomp()`)

Can usually be done very quickly,

So. . .

a_k obtained via an singular value decomposition of `x` via `prcomp()`

Alternative: obtain a_k via eigenvectors of original data's var-cov matrix (via `princomp()`)

Can usually be done very quickly, though may require no missingness.

Notes: Origin of the PCs

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation.

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

When we center all the variables, they by definition have means of zero.

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

When we center all the variables, they by definition have means of zero. But the PCs are just linear combinations of these variables, so they must *also* have means of zero.

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

When we center all the variables, they by definition have means of zero. But the PCs are just linear combinations of these variables, so they must *also* have means of zero. e.g. what is mean of $z = 2x_1 + 3x_2$ when mean of x_1 and x_2 are zero?

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

When we center all the variables, they by definition have means of zero. But the PCs are just linear combinations of these variables, so they must *also* have means of zero. e.g. what is mean of $z = 2x_1 + 3x_2$ when mean of x_1 and x_2 are zero?

So those PCs must go through the "origin" of the new data.

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

When we center all the variables, they by definition have means of zero. But the PCs are just linear combinations of these variables, so they must *also* have means of zero. e.g. what is mean of $z = 2x_1 + 3x_2$ when mean of x_1 and x_2 are zero?

So those PCs must go through the "origin" of the new data. BTW, we have to center the data to make sure the PCs find highest variance projections (rather than finding the mean of the data).

Notes: Origin of the PCs

We observed that the principal components must go through the (new mean-centered) mean of the observations. Note that this mean need not correspond to an actual observation. Why?

When we center all the variables, they by definition have means of zero. But the PCs are just linear combinations of these variables, so they must *also* have means of zero. e.g. what is mean of $z = 2x_1 + 3x_2$ when mean of x_1 and x_2 are zero?

So those PCs must go through the "origin" of the new data. BTW, we have to center the data to make sure the PCs find highest variance projections (rather than finding the mean of the data). Suppose not, and one variable has much larger magnitude than another—PC1 will be that variable.

Aside: Singular Value Decomposition

Aside: Singular Value Decomposition

We have a document term matrix $\mathbf{X}$ of dimensions $D \times T$, which we have centered (subtracting column means).

Aside: Singular Value Decomposition

We have a document term matrix $\mathbf{X}$ of dimensions $D \times T$, which we have centered (subtracting column means). SVD is a way to factorize this matrix (make a product of other matrices).

Aside: Singular Value Decomposition

We have a document term matrix **X** of dimensions $D \times T$, which we have centered (subtracting column means). SVD is a way to factorize this matrix (make a product of other matrices).

An SVD of **X** produces three matrices, and forces the vectors of **X** into a new space:

Aside: Singular Value Decomposition

We have a document term matrix $\mathbf{X}$ of dimensions $D \times T$, which we have centered (subtracting column means). SVD is a way to factorize this matrix (make a product of other matrices).

An SVD of $\mathbf{X}$ produces three matrices, and forces the vectors of $\mathbf{X}$ into a new space:

$$\mathbf{X}_{D\times T} = \mathbf{U}_{D\times D}\mathbf{\Sigma}_{D\times T}\mathbf{V}'_{T\times T}$$

Aside: Singular Value Decomposition

We have a document term matrix $\mathbf{X}$ of dimensions $D \times T$, which we have centered (subtracting column means). SVD is a way to factorize this matrix (make a product of other matrices).

An SVD of $\mathbf{X}$ produces three matrices, and forces the vectors of $\mathbf{X}$ into a new space:

$$\mathbf{X}_{D\times T} = \mathbf{U}_{D\times D}\boldsymbol{\Sigma}_{D\times T}\mathbf{V'}_{T\times T}$$

- $\mathbf{U}$ has columns which are left singular vectors of $\mathbf{X}$
- $\boldsymbol{\Sigma}$ is diagonal matrix of singular values (related to eigenvalues of covariance matrix).
- $\rightarrow$ $\mathbf{U}\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'\mathbf{V} = \mathbf{X}\mathbf{V}$ are the principal component scores.
- $\mathbf{V}$ has columns which are the right singular vectors of $\mathbf{X}$

So. . .

So. . .

Selecting first k columns of $\mathbf{U}$ and first $k \times k$ part of $\mathbf{\Sigma}$ yields a $D \times k$ matrix of principal components

So...

Selecting first k columns of $\mathbf{U}$ and first $k \times k$ part of $\boldsymbol{\Sigma}$ yields a $D \times k$ matrix of principal components (in terms of their scores), starting with the one which explains most variation.

So. . .

Selecting first k columns of $\mathbf{U}$ and first $k \times k$ part of $\mathbf{\Sigma}$ yields a $D \times k$ matrix of principal components (in terms of their scores), starting with the one which explains most variation.

If we then multiple that $D \times k$ matrix by the relevant part of $\mathbf{V}'$

So...

Selecting first k columns of $\mathbf{U}$ and first $k \times k$ part of $\mathbf{\Sigma}$ yields a $D \times k$ matrix of principal components (in terms of their scores), starting with the one which explains most variation.

If we then multiple that $D \times k$ matrix by the relevant part of $\mathbf{V}'$ we get back $\mathbf{X}$ which is $D \times T$ in dimension,

So. . .

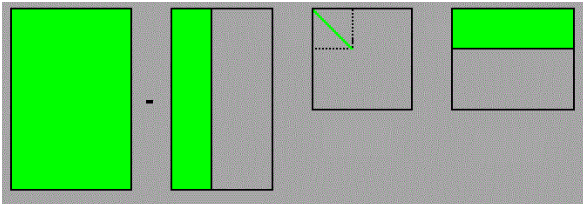
Selecting first k columns of $\mathbf{U}$ and first $k \times k$ part of $\mathbf{\Sigma}$ yields a $D \times k$ matrix of principal components (in terms of their scores), starting with the one which explains most variation.

If we then multiple that $D \times k$ matrix by the relevant part of $\mathbf{V}'$ we get back $\mathbf{X}$ which is $D \times T$ in dimension, but has reduced rank (i.e. is now in a smaller space).

So. . .

Selecting first k columns of $\mathbf{U}$ and first $k \times k$ part of $\mathbf{\Sigma}$ yields a $D \times k$ matrix of principal components (in terms of their scores), starting with the one which explains most variation.

If we then multiple that $D \times k$ matrix by the relevant part of $\mathbf{V}'$ we get back $\mathbf{X}$ which is $D \times T$ in dimension, but has reduced rank (i.e. is now in a smaller space).



(from `http://web.eecs.utk.edu/~mberry/`)

Narcissistic Personality Disorder

Narcissistic Personality Disorder

Narcissistic Personality Disorder is a psychological problem in which people's functioning and relationships with others are damaged by excessive self-importance, lack of empathy etc

Narcissistic Personality Disorder

Narcissistic Personality Disorder is a psychological problem in which people's functioning and relationships with others are damaged by excessive self-importance, lack of empathy etc

Assessed by Narcissistic Personality Inventory: 40 'forced choice' questions.

Narcissistic Personality Disorder

Narcissistic Personality Disorder is a psychological problem in which people's functioning and relationships with others are damaged by excessive self-importance, lack of empathy etc

Assessed by Narcissistic Personality Inventory: 40 'forced choice' questions.

	A	B
1.	I have a natural talent for influencing people.	I am not good at influencing people.
2.	Modesty doesn't become me.	I am essentially a modest person.
3.	I would do almost anything on a dare.	I tend to be a fairly cautious person.
4.	When people compliment me I sometimes get embarrassed.	I know that I am good because everybody keeps telling me so.
5.	The thought of ruling the world frightens the hell out of me.	If I ruled the world it would be a better place.
6.	I can usually talk my way out of anything.	I try to accept the consequences of my behavior.
7.	I prefer to blend in with the crowd.	I like to be the center of attention.
8.	I will be a success.	I am not too concerned about success.
9.	I am no better or worse than most people.	I think I am a special person.
10.	I am not sure if I would make a good leader.	I see myself as a good leader.
11.	I am assertive.	I wish I were more assertive.
12.	I like to have authority over other people.	I don't mind following orders.
13.	I find it easy to manipulate people.	I don't like it when I find myself manipulating people.
14.	I insist upon getting the respect that is due me.	I usually get the respect that I deserve.

Loadings: Raskin et al, 1988

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test,

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

→ Authority, Self-Sufficiency, Superiority, Exhibitionism, Exploitativeness, Vanity, Entitlement.

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

→ Authority, Self-Sufficiency, Superiority, Exhibitionism, Exploitativeness, Vanity, Entitlement.

NB these labels are not in original data:

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

→ Authority, Self-Sufficiency, Superiority, Exhibitionism, Exploitativeness, Vanity, Entitlement.

NB these labels are not in original data: imposed by researcher after studying!

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

→ Authority, Self-Sufficiency, Superiority, Exhibitionism, Exploitativeness, Vanity, Entitlement.

NB these labels are not in original data: imposed by researcher after studying!

Then looked at factor *loadings*:

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

→ Authority, Self-Sufficiency, Superiority, Exhibitionism, Exploitativeness, Vanity, Entitlement.

NB these labels are not in original data: imposed by researcher after studying!

Then looked at factor *loadings*: correlation between items and factors.

Loadings: Raskin et al, 1988

Wanted to understand underpinnings of the (standard) NPI

Had 1000 students take test, then fit 7 principal components (explained 52% of response variance)

→ Authority, Self-Sufficiency, Superiority, Exhibitionism, Exploitativeness, Vanity, Entitlement.

NB these labels are not in original data: imposed by researcher after studying!

Then looked at factor *loadings*: correlation between items and factors.

and Squared factor loading is percent of variance in that variable explained by the factor

Narcissistic Personality Inventory Items and Principal-Component Loadings

Items	Loadings 1	2	3	4	5	6	7
47. I would prefer to be a leader.	.83	.00	−.07	.04	−.12	.07	.22
15. I see myself as a good leader.	.83	.16	.09	−.12	.06	.03	−.14
13. I will be a success.	.67	.00	−.09	−.14	−.14	.17	.26
46. People always seem to recognize my authority.	.66	.02	.06	−.06	.06	.00	.20
2. I have a natural talent for influencing people.	.66	−.15	.02	−.02	.29	.03	−.24
16. I am assertive.	.56	.18	−.02	.22	−.02	−.03	−.27
17. I like to have authority over other people.	.56	.08	−.08	.18	.08	.05	.24
50. I am a born leader.	.35	.20	.22	.00	.09	−.14	−.01
30. I rarely depend on anyone else to get things done.	.02	.61	−.17	.04	.04	.10	−.11
23. I like to take responsibility for making decisions.	.28	.59	−.23	.23	−.12	.00	.02
53. I am more capable than other people.	−.19	.57	.16	.07	.11	.01	.20
45. I can live my life in any way I want to.	−.13	.46	.29	−.02	.05	.05	−.03
29. I always know what I am doing.	.15	.46	−.14	−.03	.30	.01	−.09
48. I am going to be a great person.	.05	.43	.39	.04	−.03	−.05	.00
54. I am an extraordinary person.	.06	.22	.69	−.07	−.06	.01	.06
7. I know that I am good because everybody keeps telling me so.	−.18	.01	.69	.00	.21	.01	.15
36. I like to be complimented.	.00	−.28	.67	.06	.00	.11	−.17
14. I think I am a special person.	.08	.16	.64	−.02	−.09	.17	−.01
51. I wish somebody would someday write my biography.	−.06	−.01	.57	.06	−.22	.09	.00
28. I am apt to show off if I get the chance.	−.04	−.02	.04	.71	−.03	.06	.06
3. Modesty doesn't become me.	−.01	.19	−.01	.69	−.16	−.06	.14
52. I get upset when people don't notice how I look when I go out in public.	−.16	.04	.10	.51	.09	.25	.17

Fit

Fit

We have variance explained by each PC: can divide out by total variance explained by all PCs.

Fit

We have variance explained by each PC: can divide out by total variance explained by all PCs.

use perhaps (scree) plot, and see when ‘next’ PC offers ‘little’ extra variance explained?

Fit

We have variance explained by each PC: can divide out by total variance explained by all PCs.

use perhaps (scree) plot, and see when 'next' PC offers 'little' extra variance explained?

or include all PCs up to e.g. 90% of variance explained?

Fit

We have variance explained by each PC: can divide out by total variance explained by all PCs.

use perhaps (scree) plot, and see when 'next' PC offers 'little' extra variance explained?

or include all PCs up to e.g. 90% of variance explained?

and drop last k PCs whose variances are roughly equal

Fit

We have variance explained by each PC: can divide out by total variance explained by all PCs.

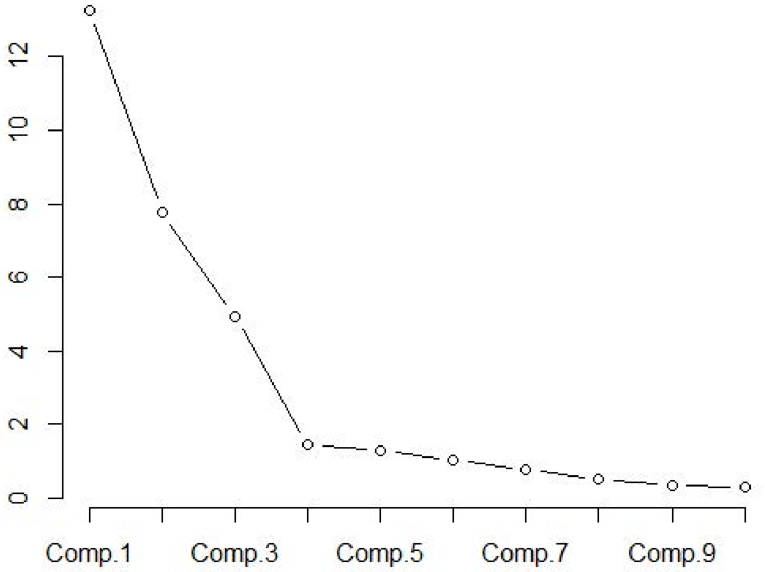
use perhaps (scree) plot, and see when 'next' PC offers 'little' extra variance explained?

or include all PCs up to e.g. 90% of variance explained?

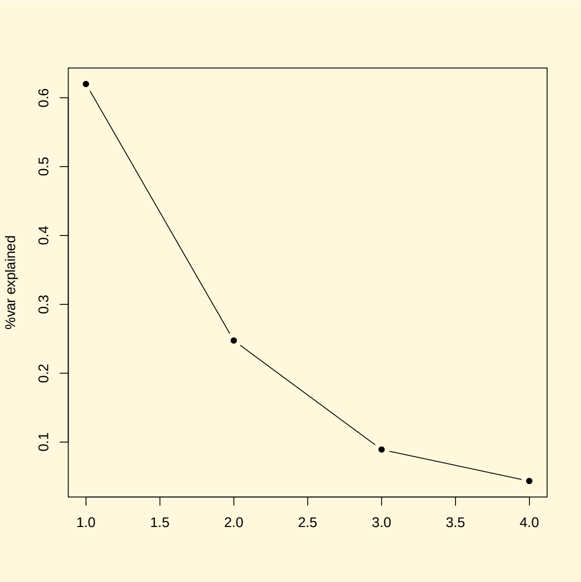
and drop last k PCs whose variances are roughly equal

btw generally like to see an 'elbow'

Good



Bad



Ugly