

Lecture 7B. Unsupervised Techniques I

DS-GA 1015, Text as Data
Arthur Spirling

March 25, 2019

Where Are We?

Where Are We?



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

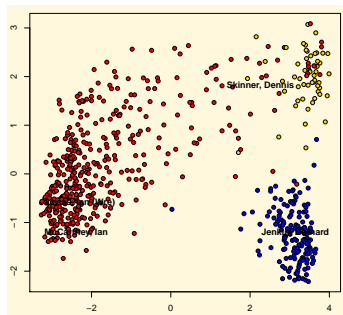
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

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?

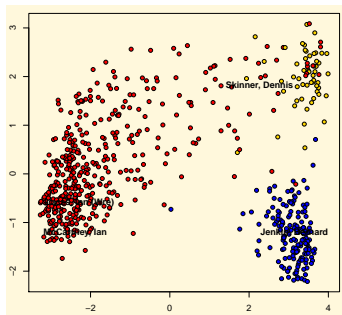


Terminology

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?

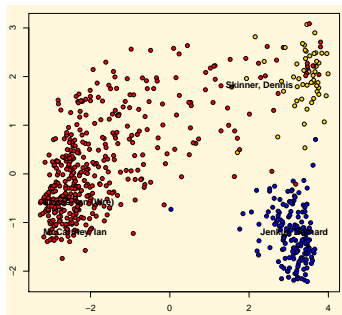
Supervised techniques:



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?

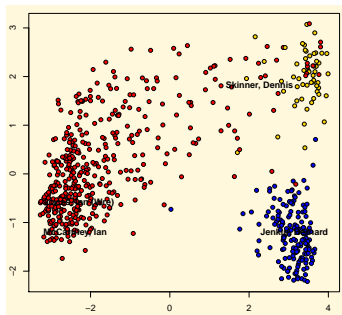


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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?



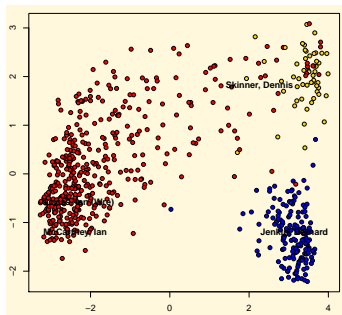
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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?



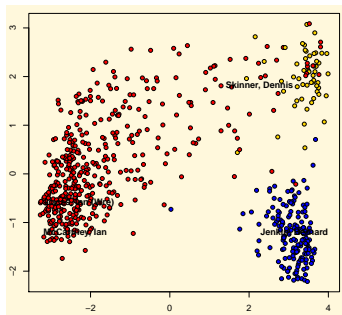
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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 ($y \in \{0, 1\}$)?


CRITIC REVIEWS FOR STAR WARS: EPISODE VII - THE FORCE AWAKENS

All Critics (313) | Top Critics (48) | My Critics | Fresh (293) | Rotten (20)


 The new movie, as an act of pure storytelling, streams by with fluency and zip.


[Full Review...](#) | December 21, 2015

 **Anthony Lane**
New Yorker
★ Top Critic


 While Star Wars: The Force Awakens gets temporarily bogged down taking us back to the world that we left in 1983, it introduces us to the new and exciting torch-bearers of the franchise.


[Full Review...](#) | December 30, 2015

 **Blake Howard**
Graffiti With Punctuation

 At the end The Force Awakens looks more like a nostalgic film that will work as a transition to the new Star Wars' age. [Full Review in Spanish]

[Full Review...](#) | December 29, 2015

 **Salvador Franco Reyes**

 This film is a well-planned product that balances nostalgia with the capacity to attract new generations into the Star Wars universe. [Full Review in Spanish]

[Full Review...](#) | December 29, 2015

Overview: Unsupervised Learning

Overview: Unsupervised Learning

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

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,

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.

→ 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	...
					...

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	...
					...

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:

- 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:

- 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:

- 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:

- 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:

- 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

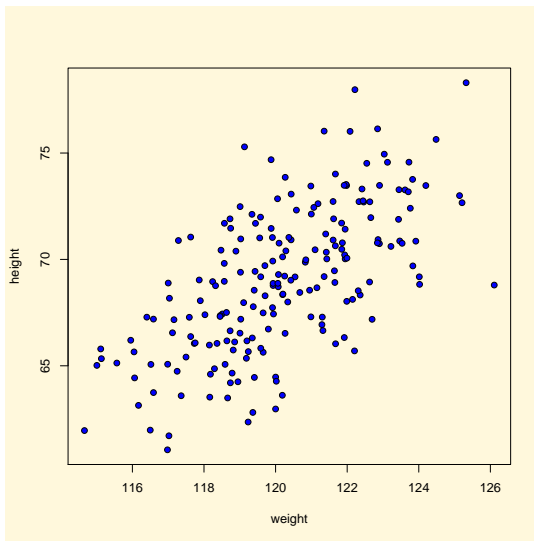
PCA performs a **linear transformation** on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the **most information** about that data

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**)

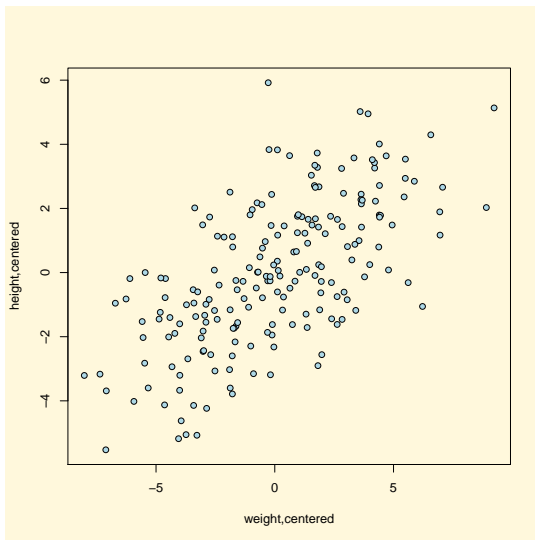
All subsequent components captures (sequentially) less variability

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

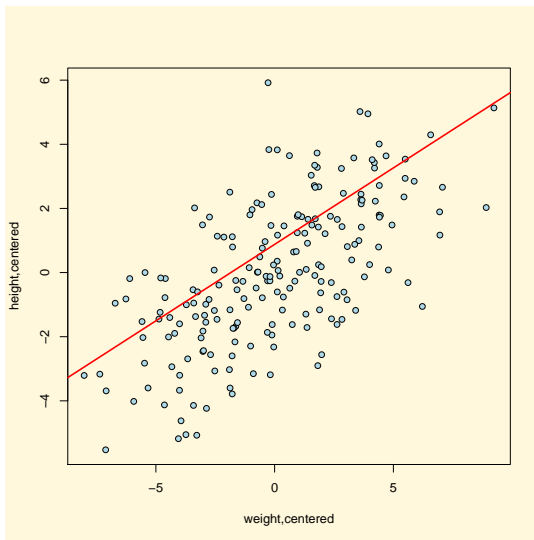
Heights, Weights



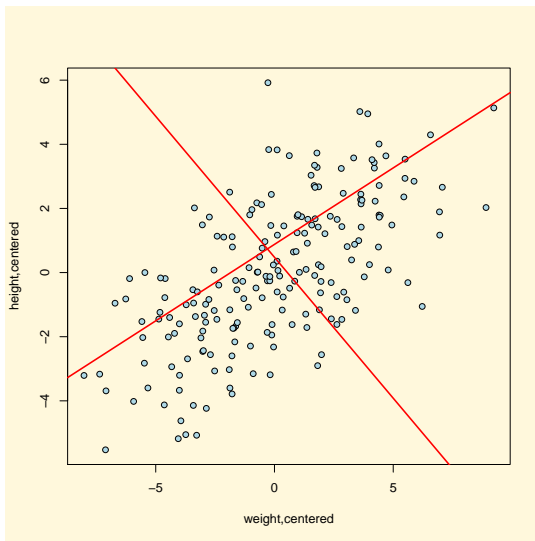
Heights, Weights



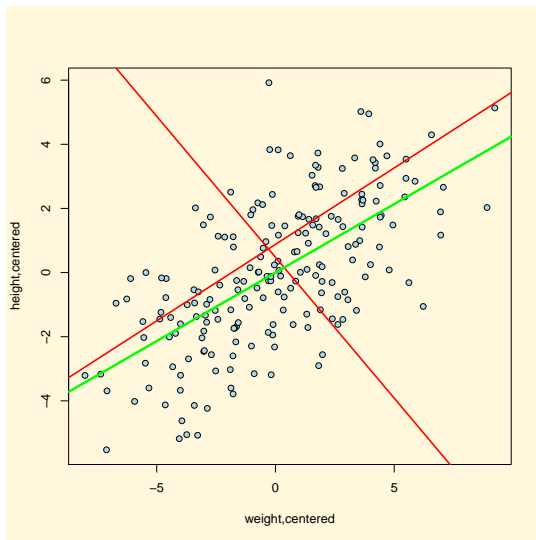
Heights, Weights



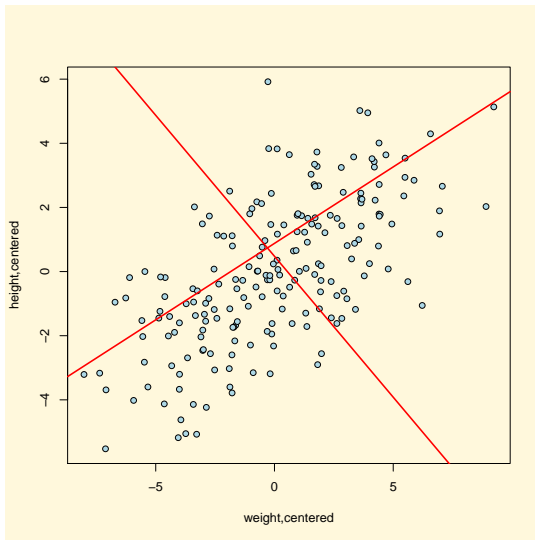
Heights, Weights



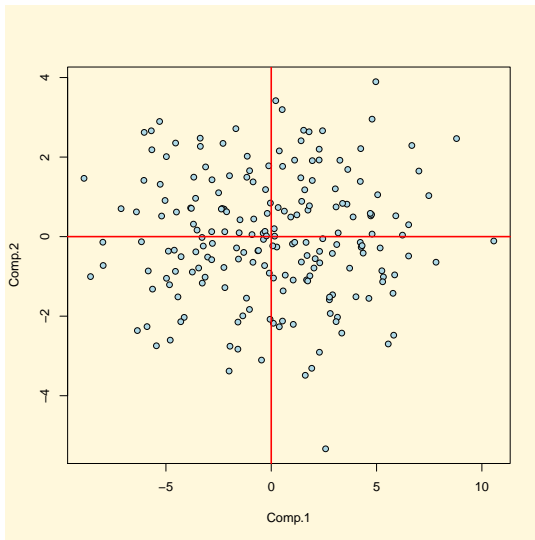
Heights, Weights



In new space



In new space



Formally

Formally

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

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 $z_1 \equiv \mathbf{a}'_1 \mathbf{X} = \sum_{i=1}^p a_{i1} x_i$

where \mathbf{X} is the $p \times n$ data matrix (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$ vector

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 (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$ vector (here 1×2)

Formally

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

where \mathbf{X} is the $p \times n$ data matrix (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\text{var}(z_1)$ is maximized

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 (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\text{var}(z_1)$ is maximized

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

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 (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\text{var}(z_1)$ is maximized

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

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

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 (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$ vector (here 1×2) picked such that $\text{var}(z_1)$ is maximized

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

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

Subsequent PCs are found subject to the constraint that

Formally

Obtain the first PC as the solution of $z_1 \equiv \mathbf{a}'_1 \mathbf{X} = \sum_{i=1}^p a_{i1} x_i$
where \mathbf{X} is the $p \times n$ data matrix (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$
vector (here 1×2) picked such that $\text{var}(z_1)$ is maximized

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

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

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

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 (here, 2×200), $\mathbf{a}_1 = (a_1, a_2, \dots, a_p)$ is $1 \times p$
vector (here 1×2) picked such that $\text{var}(z_1)$ is maximized

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

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

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

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

So...

So...

a_k obtained via an **singular value decomposition** of \mathbf{x} via `prcomp()`

So...

a_k obtained via an **singular value decomposition** of \mathbf{x} via `prcomp()`

Alternative:

So...

a_k obtained via an **singular value decomposition** of \mathbf{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 \mathbf{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 \mathbf{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 \mathbf{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.

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 \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:

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:

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

\mathbf{U} has columns which are **left singular vectors** of \mathbf{X}

$\mathbf{\Sigma}$ is diagonal matrix of **singular values** (related to eigenvalues of covariance matrix).

→ $\mathbf{U}\mathbf{\Sigma} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'\mathbf{V} = \mathbf{X}\mathbf{V}$ 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 $\mathbf{\Sigma}$ 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...

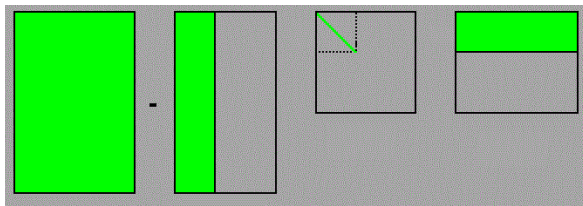
Selecting first k columns of \mathbf{U} and first $k \times k$ part of $\mathbf{\Sigma}$ yields a $D \times k$ matrix of principal components (in terms of their scores), starting with the one which explains most variation.

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/>)

(Partner) Exercise

(Partner) Exercise

Consider the following quiz:

`http://psychcentral.com/quizzes/narcissistic.htm`

googling psychcentral npi seems to get there.

Think about how you would respond to the questions,

(Partner) Exercise

Consider the following quiz:

`http://psychcentral.com/quizzes/narcissistic.htm`

`googling psychcentral npi` seems to get there.

Think about how you would respond to the questions, and fill them in privately if you wish!

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

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

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?

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?

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

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

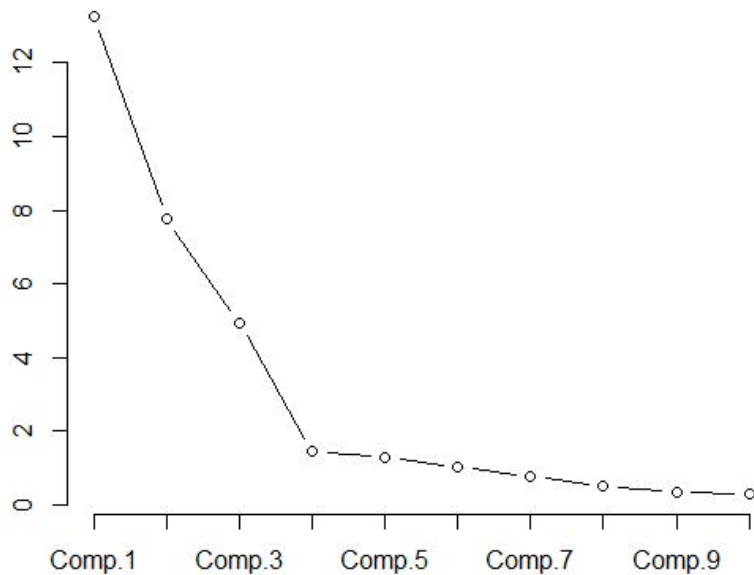
use perhaps (scree) plot, and see when 'next' PC offers 'little' extra variance explained?

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

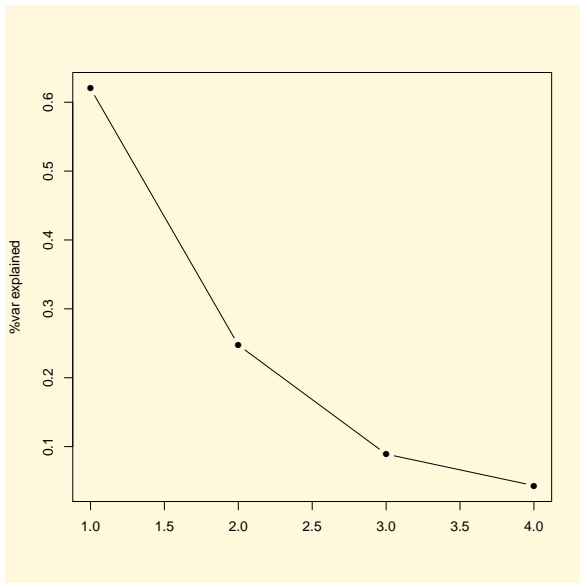
and drop last k PCs whose variances are roughly equal

btw generally like to see an 'elbow'

Good



Bad



Ugly

