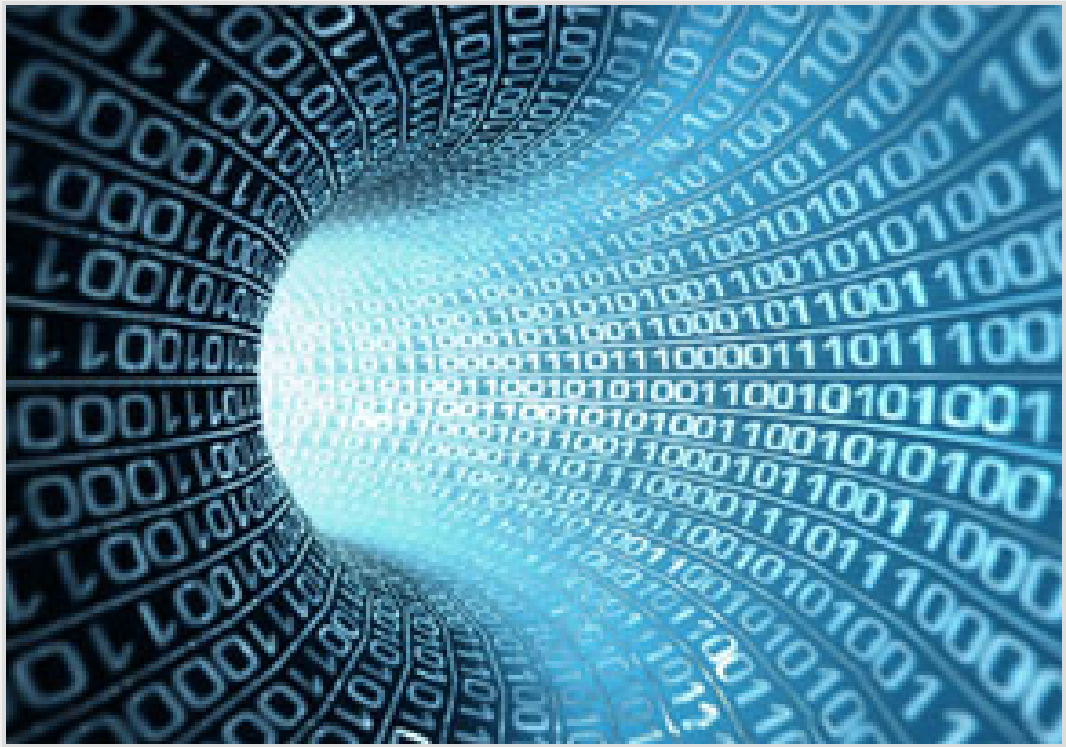
StatLab Workshop

Principal Components Analysis with R

W. D. Brinda

Yale University

01/28/2019



Outline

1 Theory
- Dimension reduction
- The PCA algorithm

Outline

1. Theory
   - Dimension reduction
   - The PCA algorithm

2. Implementation in R
   - Running example code together
   - Exercises to try for yourself

Theory

Dimension reduction

Dimension reduction means applying some procedure to a dataset that results in a new dataset with fewer columns. The new dataset can be considered a lower-dimensional "approximation" of the original one.

Dimension reduction

Dimension reduction means applying some procedure to a dataset that results in a new dataset with fewer columns. The new dataset can be considered a lower-dimensional "approximation" of the original one.

There are a variety of reasons why one might want to do this. For example,

Dimension reduction

Dimension reduction means applying some procedure to a dataset that results in a new dataset with fewer columns. The new dataset can be considered a lower-dimensional "approximation" of the original one.

There are a variety of reasons why one might want to do this. For example,

- Visualization, e.g. reduce many quantitative variables down to a few, then draw scatterplots

Dimension reduction

Dimension reduction means applying some procedure to a dataset that results in a new dataset with fewer columns. The new dataset can be considered a lower-dimensional "approximation" of the original one.

There are a variety of reasons why one might want to do this. For example,

- Visualization, e.g. reduce many quantitative variables down to a few, then draw scatterplots
- Meaningful interpretations of relationships among variables

Dimension reduction

Dimension reduction means applying some procedure to a dataset that results in a new dataset with fewer columns. The new dataset can be considered a lower-dimensional "approximation" of the original one.

There are a variety of reasons why one might want to do this. For example,

- Visualization, e.g. reduce many quantitative variables down to a few, then draw scatterplots
- Meaningful interpretations of relationships among variables
- Decreasing variance of estimation (while likely increasing bias)

Dimension reduction

Dimension reduction means applying some procedure to a dataset that results in a new dataset with fewer columns. The new dataset can be considered a lower-dimensional "approximation" of the original one.

There are a variety of reasons why one might want to do this. For example,

- Visualization, e.g. reduce many quantitative variables down to a few, then draw scatterplots
- Meaningful interpretations of relationships among variables
- Decreasing variance of estimation (while likely increasing bias)
- A technique that you want to use is computationally impossible or intractable in the data's original dimension

Bias-variance trade-off

Suppose $\hat{\theta}$ is an estimator for an unknown parameter vector $\theta \in \mathbb{R}^d$. The expected squared error of $\hat{\theta}$ is equal to its "squared bias" plus its "variance":

Bias-variance trade-off

Suppose $\hat{\theta}$ is an estimator for an unknown parameter vector $\theta \in \mathbb{R}^d$. The expected squared error of $\hat{\theta}$ is equal to its "squared bias" plus its "variance":

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = \|\mathbb{E}\hat{\theta} - \theta\|^2 + \mathbb{E}\|\hat{\theta} - \mathbb{E}\hat{\theta}\|^2.$$

Bias-variance trade-off

Suppose $\hat{\theta}$ is an estimator for an unknown parameter vector $\theta \in \mathbb{R}^d$. The expected squared error of $\hat{\theta}$ is equal to its "squared bias" plus its "variance":

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = \|\mathbb{E}\hat{\theta} - \theta\|^2 + \mathbb{E}\|\hat{\theta} - \mathbb{E}\hat{\theta}\|^2.$$

(A trade-off between bias and variance holds not only for estimation of an unknown parameter or distribution, but also for prediction of new data.)

Bias-variance trade-off

Suppose $\hat{\theta}$ is an estimator for an unknown parameter vector $\theta \in \mathbb{R}^d$. The expected squared error of $\hat{\theta}$ is equal to its "squared bias" plus its "variance":

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = \|\mathbb{E}\hat{\theta} - \theta\|^2 + \mathbb{E}\|\hat{\theta} - \mathbb{E}\hat{\theta}\|^2.$$

(A trade-off between bias and variance holds not only for estimation of an unknown parameter or distribution, but also for prediction of new data.)

In general, "simpler" models result in larger squared bias but smaller variance. An essential part of the practice of statistics and machine learning is choosing a procedure that balances bias and variance such that their sum is nearly as small as possible.

Bias-variance trade-off

Suppose $\hat{\theta}$ is an estimator for an unknown parameter vector $\theta \in \mathbb{R}^d$. The expected squared error of $\hat{\theta}$ is equal to its "squared bias" plus its "variance":

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = \|\mathbb{E}\hat{\theta} - \theta\|^2 + \mathbb{E}\|\hat{\theta} - \mathbb{E}\hat{\theta}\|^2.$$

(A trade-off between bias and variance holds not only for estimation of an unknown parameter or distribution, but also for prediction of new data.)

In general, "simpler" models result in larger squared bias but smaller variance. An essential part of the practice of statistics and machine learning is choosing a procedure that balances bias and variance such that their sum is nearly as small as possible.

Dimension reduction techniques tend to increase squared bias and decrease variance.

Linear transformations for dim reduction

There are two common *linear* transformations for reducing the dimension of a set of quantitative variables.

Linear transformations for dim reduction

There are two common *linear* transformations for reducing the dimension of a set of quantitative variables.

- Principal components analysis (PCA) "spreads out the *points* as much as possible"

Linear transformations for dim reduction

There are two common *linear* transformations for reducing the dimension of a set of quantitative variables.

- Principal components analysis (PCA) "spreads out the *points* as much as possible"
- Linear discriminant analysis (LDA) "spreads out the *groups* as much as possible" (requires a categorical variable)

Linear transformations for dim reduction

There are two common *linear* transformations for reducing the dimension of a set of quantitative variables.

- Principal components analysis (PCA) "spreads out the *points* as much as possible"
- Linear discriminant analysis (LDA) "spreads out the *groups* as much as possible" (requires a categorical variable)

PCA is our focus, but we'll also see LDA in one of the R code examples later.

A "simple" explanation

In the following slides, I've tried to present a thorough explanation of the PCA algorithm *as simply as possible*... that doesn't mean that it's *simple*.

A "simple" explanation

In the following slides, I've tried to present a thorough explanation of the PCA algorithm *as simply as possible*... that doesn't mean that it's *simple*.

To be frank, it's probably more than you can process in a single sitting. Just do your best to follow along and see how much you can make sense of. Then, if you'd like to really understand the details, take your time looking over these notes again later.

Spreading out the data points

Consider a data frame with n observations of d quantitative variables. One could try to imagine the high-dimensional scatterplot of the n observations as data points in $\mathbb{R}^d$. For simplicity, just think of an ordinary scatterplot of two variables.

Spreading out the data points

Consider a data frame with n observations of d quantitative variables. One could try to imagine the high-dimensional scatterplot of the n observations as data points in $\mathbb{R}^d$. For simplicity, just think of an ordinary scatterplot of two variables.

What exactly does it mean to "spread out" the points as much as possible? Can you think of a common statistic for quantifying how spread out a set of numbers is?

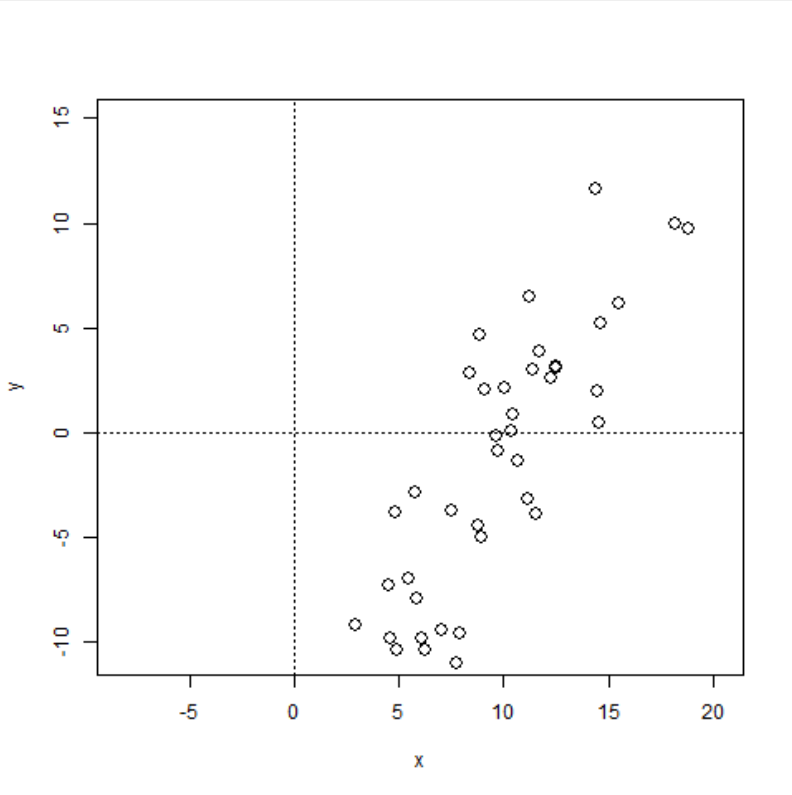
Spreading out the data points

Consider a data frame with n observations of d quantitative variables. One could try to imagine the high-dimensional scatterplot of the n observations as data points in $\mathbb{R}^d$. For simplicity, just think of an ordinary scatterplot of two variables.

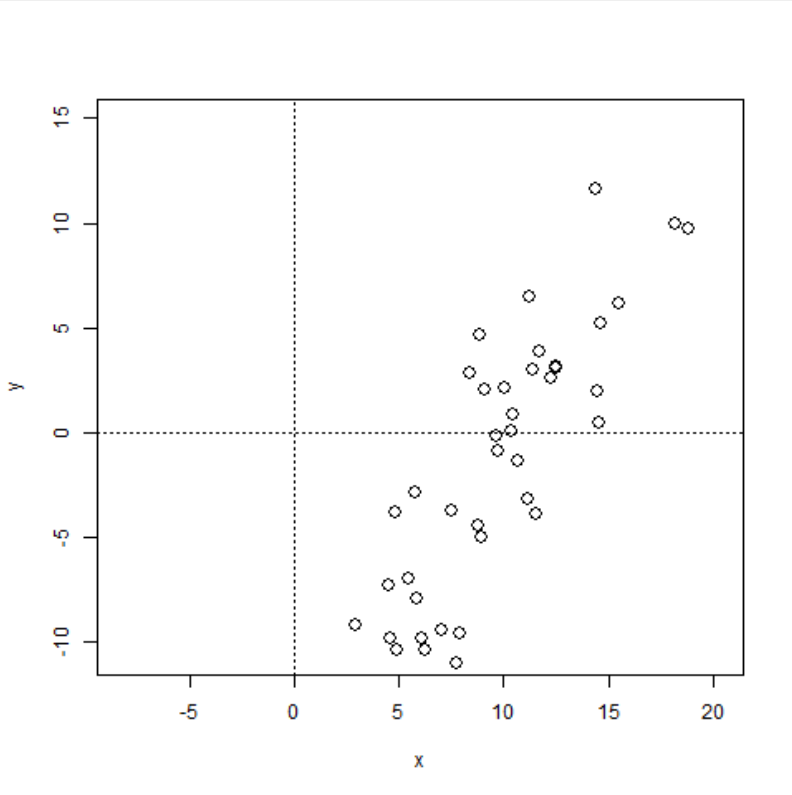
What exactly does it mean to "spread out" the points as much as possible? Can you think of a common statistic for quantifying how spread out a set of numbers is?

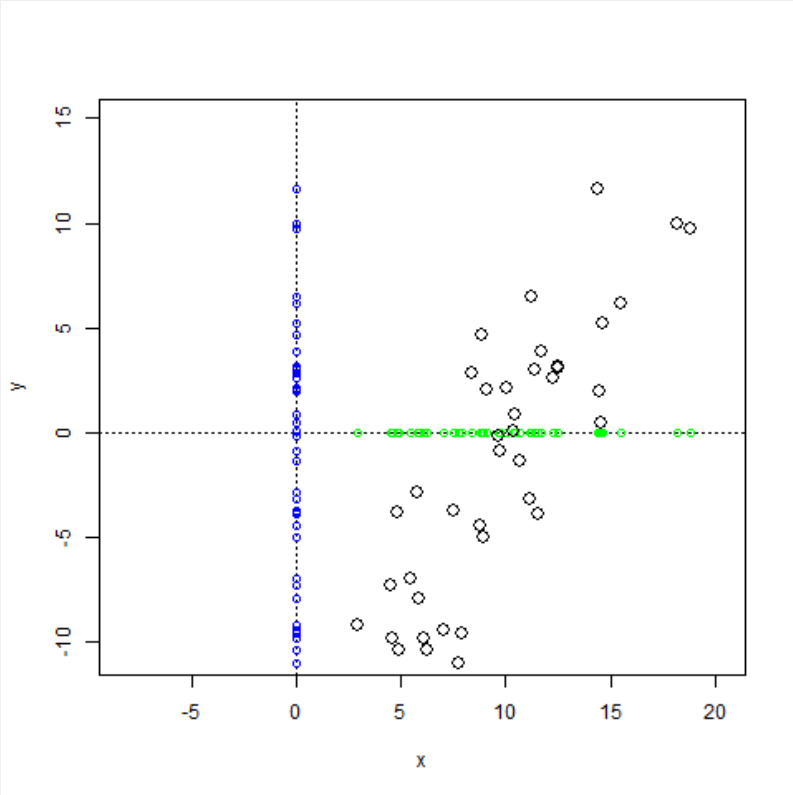
How about *variance*, the average squared difference from the mean. Each variable has a variance, and some variables will have larger variances than others.

A cloud of data point and their coordinates



A cloud of data point and their coordinates





Alternative bases

Consider the first observation $\mathbf{x}_1 = (x_{i,1}, \ldots, x_{i,d})$. It's the sum of its orthogonal projections onto the coordinate axes:

$$\mathbf{x}_1 = x_{1,1}\mathbf{e}_1 + \ldots + x_{1,d}\mathbf{e}_d$$

where $\mathbf{e}_j$ is the jth standard basis vector (it has a 1 as its jth entry and zeros elsewhere).

Alternative bases

Consider the first observation $\mathbf{x}_1 = (x_{i,1}, \ldots, x_{i,d})$. It's the sum of its orthogonal projections onto the coordinate axes:

$$\mathbf{x}_1 = x_{1,1}\mathbf{e}_1 + \ldots + x_{1,d}\mathbf{e}_d$$

where $\mathbf{e}_j$ is the jth standard basis vector (it has a 1 as its jth entry and zeros elsewhere).

However, $\mathbf{x}_1$ is also the sum of its orthogonal projections onto *any* orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_d$. The general formula is simply

$$\mathbf{x}_1 = (\mathbf{x}_1^T\mathbf{u}_1)\mathbf{u}_1 + \ldots + (\mathbf{x}_1^T\mathbf{u}_d)\mathbf{u}_d.$$

Alternative bases

Consider the first observation $\mathbf{x}_1 = (x_{i,1}, \ldots, x_{i,d})$. It's the sum of its orthogonal projections onto the coordinate axes:

$$\mathbf{x}_1 = x_{1,1}\mathbf{e}_1 + \ldots + x_{1,d}\mathbf{e}_d$$

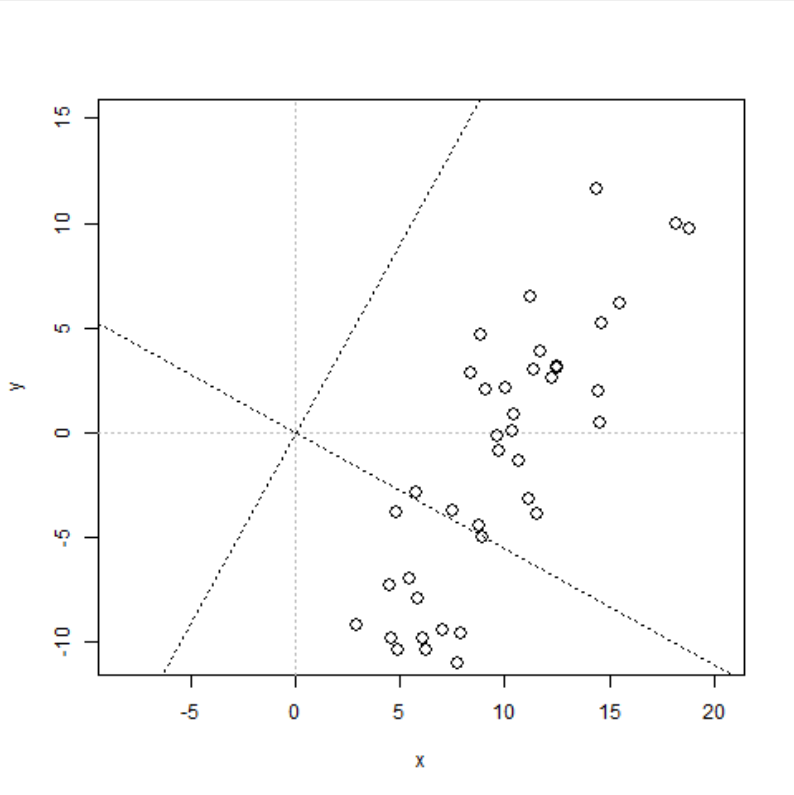
where $\mathbf{e}_j$ is the jth standard basis vector (it has a 1 as its jth entry and zeros elsewhere).

However, $\mathbf{x}_1$ is also the sum of its orthogonal projections onto *any* orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_d$. The general formula is simply

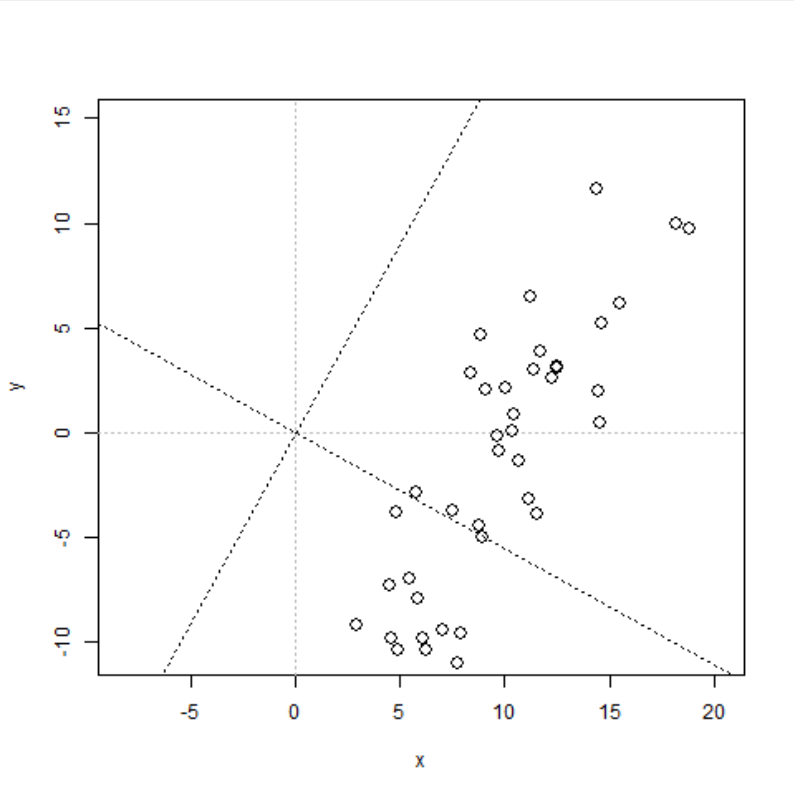
$$\mathbf{x}_1 = (\mathbf{x}_1^T\mathbf{u}_1)\mathbf{u}_1 + \ldots + (\mathbf{x}_1^T\mathbf{u}_d)\mathbf{u}_d.$$

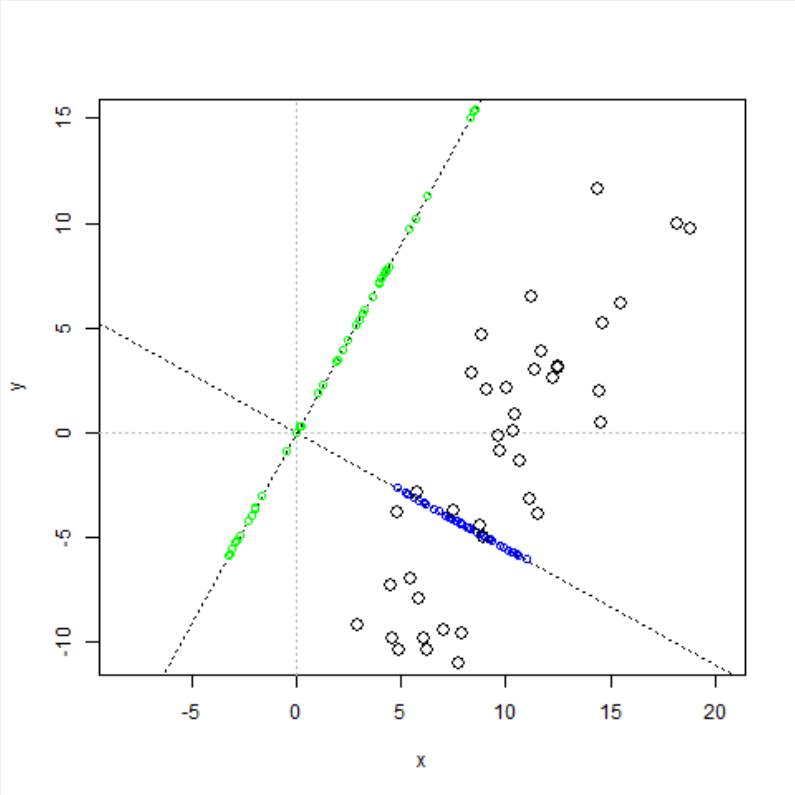
The other observations can be represented likewise. Thus, the *coordinates* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *with respect to* a basis vector $\mathbf{u}$ are $\mathbf{x}_1^T\mathbf{u}, \ldots, \mathbf{x}_n^T\mathbf{u}$.

Alternative axes and coordinates



Alternative axes and coordinates





Preview of PCA argument

Assuming the mean of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the zero vector:

Preview of PCA argument

Assuming the mean of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the zero vector:

- The average of the coordinates with respect to any basis vector $\mathbf{u}$ is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ where $\mathbb{X}$ is the matrix whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Preview of PCA argument

Assuming the mean of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the zero vector:

- The average of the coordinates with respect to any basis vector $\mathbf{u}$ is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ where $\mathbb{X}$ is the matrix whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.

Preview of PCA argument

Assuming the mean of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the zero vector:

- The average of the coordinates with respect to any basis vector $\mathbf{u}$ is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ where $\mathbb{X}$ is the matrix whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.
- Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

Preview of PCA argument

Assuming the mean of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the zero vector:

- The average of the coordinates with respect to any basis vector $\mathbf{u}$ is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ where $\mathbb{X}$ is the matrix whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.
- Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (with respect to any basis vector) is exactly the same as the variance of the coordinates of the *centered* vectors $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$.

Preview of PCA argument

Assuming the mean of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the zero vector:

- The average of the coordinates with respect to any basis vector $\mathbf{u}$ is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ where $\mathbb{X}$ is the matrix whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.
- Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (with respect to any basis vector) is exactly the same as the variance of the coordinates of the *centered* vectors $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$.

Let $\widetilde{\mathbb{X}}$ be the matrix whose rows are $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$, and let $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_1$. Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

Average of coordinates

Suppose that that the mean $\frac{1}{n}\sum_i \mathbf{x}_i$ is equal to the zero vector $\mathbf{o}$, that is, the vector with zero as every entry.

Average of coordinates

Suppose that that the mean $\frac{1}{n}\sum_i \mathbf{x}_i$ is equal to the zero vector $\mathbf{o}$, that is, the vector with zero as every entry.

Then the average of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to a unit vector $\mathbf{u}$ is

$$\frac{1}{n}(\mathbf{x}_1^T\mathbf{u} + \ldots + \mathbf{x}_n^T\mathbf{u}) = \underbrace{\frac{1}{n}(\mathbf{x}_1 + \ldots + \mathbf{x}_n)^T}_{\mathbf{o}}\mathbf{u}$$
$$= 0$$

Variance of coordinates

If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{o}$, then the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to a unit vector $\mathbf{u}$ is simply the average of the squared coordinates

$$\frac{1}{n}[(\mathbf{x}_1^T\mathbf{u})^2 + \ldots + (\mathbf{x}_n^T\mathbf{u})^2].$$

Variance of coordinates

If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{0}$, then the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to a unit vector $\mathbf{u}$ is simply the average of the squared coordinates

$$\frac{1}{n}[(\mathbf{x}_1^T\mathbf{u})^2 + \ldots + (\mathbf{x}_n^T\mathbf{u})^2].$$

Let's derive a more useful representation of this quantity in terms of the matrix $\mathbb{X}$ whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The coordinates $\mathbf{x}_1^T\mathbf{u}, \ldots, \mathbf{x}_n^T\mathbf{u}$ are the entries of the vector

$$\mathbb{X}\mathbf{u} = \begin{bmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{u} \\ | \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{u} \\ \vdots \\ \mathbf{x}_n^T\mathbf{u} \end{bmatrix}.$$

Variance of coordinates

If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{0}$, then the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to a unit vector $\mathbf{u}$ is simply the average of the squared coordinates

$$\frac{1}{n}[(\mathbf{x}_1^T\mathbf{u})^2 + \ldots + (\mathbf{x}_n^T\mathbf{u})^2].$$

Let's derive a more useful representation of this quantity in terms of the matrix $\mathbb{X}$ whose rows are $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The coordinates $\mathbf{x}_1^T\mathbf{u}, \ldots, \mathbf{x}_n^T\mathbf{u}$ are the entries of the vector

$$\mathbb{X}\mathbf{u} = \begin{bmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{u} \\ | \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{u} \\ \vdots \\ \mathbf{x}_n^T\mathbf{u} \end{bmatrix}.$$

Therefore, the variance of the coordinates can be represented

$$\frac{1}{n}\sum_i (\mathbf{x}_1^T\mathbf{u})^2 = \frac{1}{n}\|\mathbb{X}\mathbf{u}\|^2 = \frac{1}{n}(\mathbb{X}\mathbf{u})^T(\mathbb{X}\mathbf{u}) = \mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}.$$

Maximizing variance

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{o}$, we've realized that the variance of their coordinates with respect to any unit vector $\mathbf{u}$ equals the *quadratic form* $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.

Maximizing variance

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{o}$, we've realized that the variance of their coordinates with respect to any unit vector $\mathbf{u}$ equals the *quadratic form* $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.

It turns out that this is easy to maximize via a *spectral decomposition*. The symmetric matrix $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ can be rewritten in terms of its eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ and orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$:

$$\frac{1}{n}\mathbb{X}^T\mathbb{X} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_d \mathbf{q}_d \mathbf{q}_d^T.$$

Maximizing variance

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{0}$, we've realized that the variance of their coordinates with respect to any unit vector $\mathbf{u}$ equals the *quadratic form* $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.

It turns out that this is easy to maximize via a *spectral decomposition*. The symmetric matrix $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ can be rewritten in terms of its eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ and orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$:

$$\frac{1}{n}\mathbb{X}^T\mathbb{X} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_d \mathbf{q}_d \mathbf{q}_d^T.$$

Therefore

$$\begin{aligned}\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u} &= \mathbf{u}^T(\lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_d \mathbf{q}_d^T)\mathbf{u} \\ &= \lambda_1(\mathbf{u}^T\mathbf{q}_1)^2 + \ldots + \lambda_d(\mathbf{u}^T\mathbf{q}_d)^2.\end{aligned}$$

The numbers $\mathbf{u}^T\mathbf{q}_1, \ldots, \mathbf{u}^T\mathbf{q}_n$ are exactly the coordinates of $\mathbf{u}$ with respect to the orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_d$. Because $\mathbf{u}$ is a unit vector, their squares must sum to 1 (Parseval's identity). Therefore, $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is a weighted average of the eigenvalues $\lambda_1, \ldots, \lambda_d$. Its maximum possible value of λ_1 is achieved if $\mathbf{u}$ equals $\mathbf{q}_1$.

Maximizing variance

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{0}$, we've realized that the variance of their coordinates with respect to any unit vector $\mathbf{u}$ equals the *quadratic form* $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.

It turns out that this is easy to maximize via a *spectral decomposition*. The symmetric matrix $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ can be rewritten in terms of its eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ and orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$:

$$\frac{1}{n}\mathbb{X}^T\mathbb{X} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_d \mathbf{q}_d \mathbf{q}_d^T.$$

Therefore

$$\begin{aligned}\mathbf{u}^T(\tfrac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u} &= \mathbf{u}^T(\lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_d \mathbf{q}_d^T)\mathbf{u} \\ &= \lambda_1(\mathbf{u}^T\mathbf{q}_1)^2 + \ldots + \lambda_d(\mathbf{u}^T\mathbf{q}_d)^2.\end{aligned}$$

The numbers $\mathbf{u}^T\mathbf{q}_1, \ldots, \mathbf{u}^T\mathbf{q}_n$ are exactly the coordinates of $\mathbf{u}$ with respect to the orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_d$. Because $\mathbf{u}$ is a unit vector, their squares must sum to 1 (Parseval's identity). Therefore, $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is a weighted average of the eigenvalues $\lambda_1, \ldots, \lambda_d$. Its maximum possible value of λ_1 is achieved if $\mathbf{u}$ equals $\mathbf{q}_1$.

By similar reasoning, among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

Centering

Define the *mean vector* $\bar{\mathbf{x}}$ to be the vector whose first entry is the average of the first entries of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and so on. The *centered* points $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$ have two key properties.

Centering

Define the *mean vector* $\bar{\mathbf{x}}$ to be the vector whose first entry is the average of the first entries of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and so on. The *centered* points $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$ have two key properties.

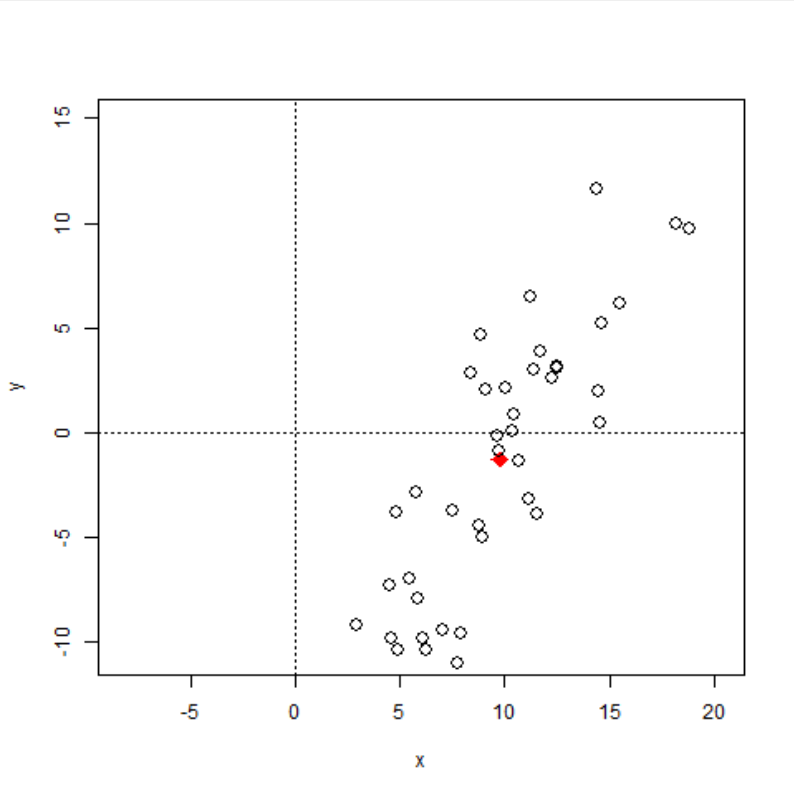
- The coordinates of the centered points with respect to any basis vector have the same variance as the coordinates of the original points with respect to that basis vector. (Variances don't change when you translate the whole cloud of points by same vector.)

Centering

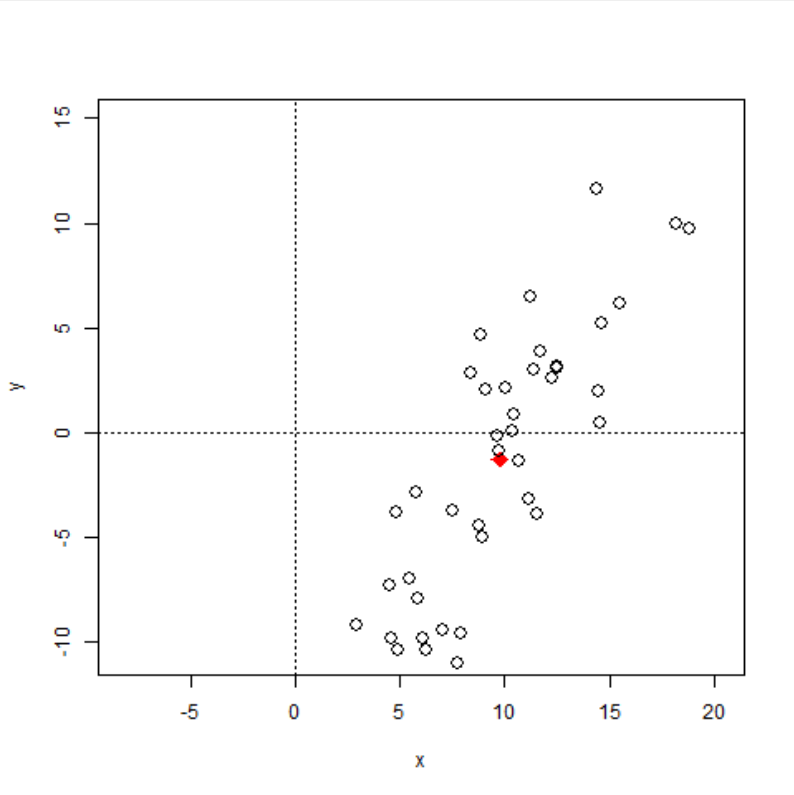
Define the *mean vector* $\bar{\mathbf{x}}$ to be the vector whose first entry is the average of the first entries of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and so on. The *centered* points $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$ have two key properties.

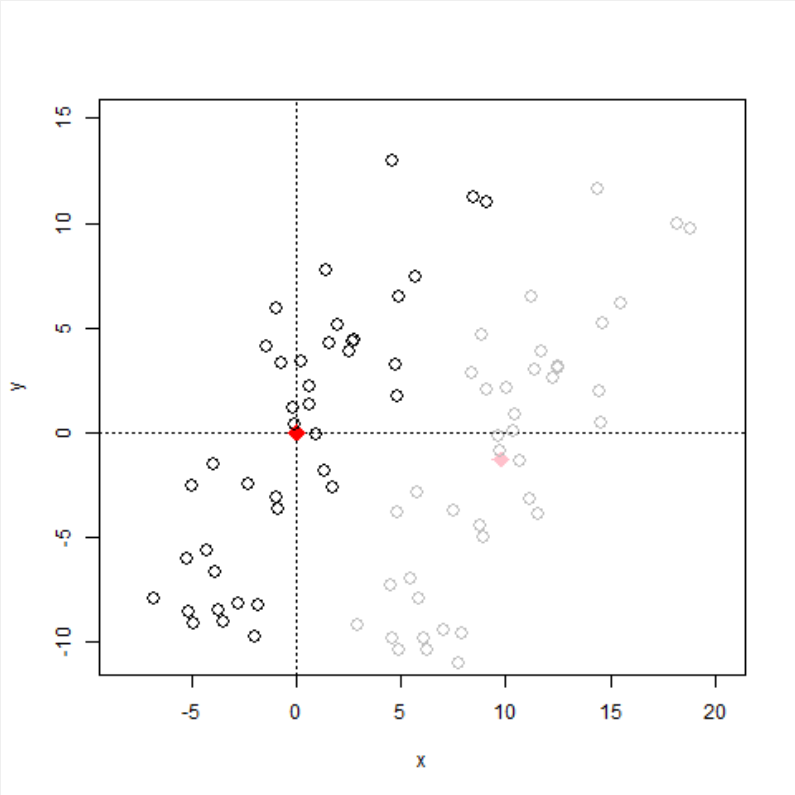
- The coordinates of the centered points with respect to any basis vector have the same variance as the coordinates of the original points with respect to that basis vector. (Variances don't change when you translate the whole cloud of points by same vector.)
- The centered points have mean $\mathbf{o}$. $(\frac{1}{n}\sum_i(\mathbf{x}_i - \bar{\mathbf{x}}) = (\frac{1}{n}\sum_i \mathbf{x}_i) - \bar{\mathbf{x}} = \mathbf{o}.)$

Centering the data points



Centering the data points





PCA algorithm

The variance-maximizing basis vector for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the same as the variance-maximizing basis vector for the centered points, which we know how to find since they have mean $\mathbf{o}$.

PCA algorithm

The variance-maximizing basis vector for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the same as the variance-maximizing basis vector for the centered points, which we know how to find since they have mean $\mathbf{o}$.

Let $\widetilde{\mathbb{X}}$ be the matrix whose rows are $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$, and let $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_1$. Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

PCA algorithm

The variance-maximizing basis vector for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the same as the variance-maximizing basis vector for the centered points, which we know how to find since they have mean $\mathbf{o}$.

Let $\widetilde{\mathbb{X}}$ be the matrix whose rows are $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$, and let $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_1$. Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

Notice that $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ is the covariance matrix of the data points.

Recap of PCA argument

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are centered:

Recap of PCA argument

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are centered:

- The average of the coordinates is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.

Recap of PCA argument

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are centered:

- The average of the coordinates is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.

Recap of PCA argument

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are centered:

- The average of the coordinates is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.
- Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

Recap of PCA argument

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are centered:

- The average of the coordinates is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.
- Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (with respect to any basis vector) is exactly the same as the variance of the coordinates of the *centered* vectors $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$.

Recap of PCA argument

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are centered:

- The average of the coordinates is zero, so their variance is equal to the average of the squared coordinates which can be represented $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$.
- Let $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. Spectral decomposition shows that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is maximized if $\mathbf{u}$ equals $\mathbf{q}_1$.
- Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (with respect to any basis vector) is exactly the same as the variance of the coordinates of the *centered* vectors $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$.

Let $\widetilde{\mathbb{X}}$ be the matrix whose rows are $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$, and let $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ have eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ with orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$. The variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_1$. Among all unit vectors that are orthogonal to $\mathbf{q}_1$, the variance of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is maximized by $\mathbf{q}_2$, and so on.

The eigenvector basis coordinates

The spectral decomposition of the covariance matrix $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ is usually expressed as $\mathbb{Q}\mathbb{\Lambda}\mathbb{Q}^T$ where $\mathbb{\Lambda}$ is a diagonal matrix of eigenvalues and the columns of $\mathbb{Q}$ are orthonormal unit eigenvectors.

The eigenvector basis coordinates

The spectral decomposition of the covariance matrix $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ is usually expressed as $\mathbb{Q}\mathbb{\Lambda}\mathbb{Q}^T$ where $\mathbb{\Lambda}$ is a diagonal matrix of eigenvalues and the columns of $\mathbb{Q}$ are orthonormal unit eigenvectors.

The coordinates of $\mathbf{x}_1$ with respect to the eigenvectors of the covariance matrix are precisely the entries of

$$\mathbb{Q}^T\mathbf{x}_1 = \begin{bmatrix} - & \mathbf{q}_1 & - \\ & \vdots & \\ - & \mathbf{q}_d & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{x}_1 \\ | \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^T\mathbf{x}_1 \\ \vdots \\ \mathbf{q}_d^T\mathbf{x}_1 \end{bmatrix}.$$

The eigenvector basis coordinates

The spectral decomposition of the covariance matrix $\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}$ is usually expressed as $\mathbb{Q}\Lambda\mathbb{Q}^T$ where Λ is a diagonal matrix of eigenvalues and the columns of $\mathbb{Q}$ are orthonormal unit eigenvectors.

The coordinates of $\mathbf{x}_1$ with respect to the eigenvectors of the covariance matrix are precisely the entries of

$$\mathbb{Q}^T\mathbf{x}_1 = \begin{bmatrix} - & \mathbf{q}_1 & - \\ & \vdots & \\ - & \mathbf{q}_d & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{x}_1 \\ | \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^T\mathbf{x}_1 \\ \vdots \\ \mathbf{q}_d^T\mathbf{x}_1 \end{bmatrix}.$$

Indeed, one can easily compute the coordinates of all the observations at once:

$$\mathbb{X}\mathbb{Q} = \begin{bmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{q}_1 & \cdots & \mathbf{q}_d \\ | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{q}_1 & \cdots & \mathbf{x}_1^T\mathbf{q}_d \\ \vdots & \ddots & \vdots \\ \mathbf{x}_n^T\mathbf{q}_1 & \cdots & \mathbf{x}_n^T\mathbf{q}_d \end{bmatrix}.$$

The ith row contains the coordinates of $\mathbf{x}_i$ with respect to $\mathbf{q}_1, \ldots, \mathbf{q}_n$.

The principal components

An often more convenient matrix is $\widetilde{\mathbb{X}}\mathbb{Q}$ which contains the coordinates of the *centered* observations with respect to $\mathbf{q}_1, \ldots, \mathbf{q}_n$. In that case, the columns are orthogonal.

$$\begin{aligned}(\widetilde{\mathbb{X}}\mathbb{Q})^T(\widetilde{\mathbb{X}}\mathbb{Q}) &= \mathbb{Q}^T\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}\mathbb{Q} \\ &= \mathbb{Q}^T(n\mathbb{Q}\Lambda\mathbb{Q}^T)\mathbb{Q} \\ &= n\Lambda\end{aligned}$$

The principal components

An often more convenient matrix is $\widetilde{\mathbb{X}}\mathbb{Q}$ which contains the coordinates of the *centered* observations with respect to $\mathbf{q}_1, \ldots, \mathbf{q}_n$. In that case, the columns are orthogonal.

$$\begin{aligned}(\widetilde{\mathbb{X}}\mathbb{Q})^T(\widetilde{\mathbb{X}}\mathbb{Q}) &= \mathbb{Q}^T\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}\mathbb{Q} \\ &= \mathbb{Q}^T(n\mathbb{Q}\Lambda\mathbb{Q}^T)\mathbb{Q} \\ &= n\Lambda\end{aligned}$$

These columns that were called *principal components* of the dataset, though the terminology is used inconsistently. You may see the columns of $\mathbb{X}\mathbb{Q}$ called the principal components. Or you may see $\mathbf{q}_1, \ldots, \mathbf{q}_n$ called the principal components.

Singular value decomposition

The eigenvectors and eigenvalues of the covariance matrix can also be obtained directly from a *singular value decomposition* of $\frac{1}{\sqrt{n}}\widetilde{\mathbb{X}}$:

$$\frac{1}{\sqrt{n}}\widetilde{\mathbb{X}} = \mathbb{U}\mathbb{D}\mathbb{V}^T$$

where both $\mathbb{U}$ and $\mathbb{V}$ have orthonormal columns and $\mathbb{D}$ is diagonal with non-negative entries.

Singular value decomposition

The eigenvectors and eigenvalues of the covariance matrix can also be obtained directly from a *singular value decomposition* of $\frac{1}{\sqrt{n}}\widetilde{\mathbb{X}}$:

$$\frac{1}{\sqrt{n}}\widetilde{\mathbb{X}} = \mathbb{U}\mathbb{D}\mathbb{V}^T$$

where both $\mathbb{U}$ and $\mathbb{V}$ have orthonormal columns and $\mathbb{D}$ is diagonal with non-negative entries.

How does this relate to the spectral decomposition of the covariance matrix?

$$\begin{aligned}\frac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}} &= (\frac{1}{\sqrt{n}}\widetilde{\mathbb{X}})^T(\frac{1}{\sqrt{n}}\widetilde{\mathbb{X}}) \\ &= (\mathbb{U}\mathbb{D}\mathbb{V}^T)^T(\mathbb{U}\mathbb{D}\mathbb{V}^T) \\ &= \mathbb{V}\mathbb{D}\mathbb{U}^T\mathbb{U}\mathbb{D}\mathbb{V}^T \\ &= \mathbb{V}\mathbb{D}^2\mathbb{V}^T\end{aligned}$$

reveals a spectral decomposition of the covariance matrix. The columns of $\mathbb{V}$ are unit eigenvectors while the squares of the diagonals in $\mathbb{D}$ are the corresponding eigenvalues.

Variances equal to eigenvalues

The variance of the coordinates with respect to $\mathbf{q}_1$ is

$$\begin{aligned}
\mathbf{q}_1^T\left(\tfrac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}\right)\mathbf{q}_1 &= \mathbf{q}_1^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_d\mathbf{q}_d\mathbf{q}_d^T)\mathbf{q}_1 \\
&= \lambda_1(\underbrace{\mathbf{q}_1^T\mathbf{q}_1}_{1})^2 + \ldots + \lambda_d(\underbrace{\mathbf{q}_1^T\mathbf{q}_d}_{0})^2 \\
&= \lambda_1.
\end{aligned}$$

Variances equal to eigenvalues

The variance of the coordinates with respect to $\mathbf{q}_1$ is

$$\begin{aligned}\mathbf{q}_1^T(\tfrac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}})\mathbf{q}_1 &= \mathbf{q}_1^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_d\mathbf{q}_d\mathbf{q}_d^T)\mathbf{q}_1 \\ &= \lambda_1(\underbrace{\mathbf{q}_1^T\mathbf{q}_1}_{1})^2 + \ldots + \lambda_d(\underbrace{\mathbf{q}_1^T\mathbf{q}_d}_{0})^2 \\ &= \lambda_1.\end{aligned}$$

Likewise the variance of the coordinates with respect to any eigenvector $\mathbf{q}_j$ is the corresponding eigenvalue λ_j. If only the first $k \leq d$ principal components are retained, then the sum of the variances in the lower-dimensional approximation is $\lambda_1 + \ldots + \lambda_k$.

Variance decomposition

The sum of the variances of the eigenvector basis coordinates equals the sum of the variances of the original coordinates.

Variance decomposition

The sum of the variances of the eigenvector basis coordinates equals the sum of the variances of the original coordinates.

A short proof is provided here for your future reference. It uses the fact that the squared norm of a vector is the sum of its squared coordinates with respect to any orthonormal basis (Parseval's identity). Without loss of generality, pretend once again that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{o}$.

$$\sum_{j=1}^{d} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{q}_j^T \mathbf{x}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} (\mathbf{q}_j^T \mathbf{x}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} x_{i,j}^2 = \sum_{j=1}^{d} \frac{1}{n} \sum_{i=1}^{n} x_{i,j}^2$$

Variance decomposition

The sum of the variances of the eigenvector basis coordinates equals the sum of the variances of the original coordinates.

A short proof is provided here for your future reference. It uses the fact that the squared norm of a vector is the sum of its squared coordinates with respect to any orthonormal basis (Parseval's identity). Without loss of generality, pretend once again that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have mean $\mathbf{o}$.

$$\sum_{j=1}^{d} \tfrac{1}{n} \sum_{i=1}^{n} (\mathbf{q}_j^T \mathbf{x}_i)^2 = \tfrac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} (\mathbf{q}_j^T \mathbf{x}_i)^2 = \tfrac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 = \tfrac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} x_{i,j}^2 = \sum_{j=1}^{d} \tfrac{1}{n} \sum_{i=1}^{n} x_{i,j}^2$$

Let v equal the sum of the variances of the original coordinates. If only the first $k \leq d$ principal components are retained, then the proportion of the variance that is preserved is $\frac{\lambda_1 + \ldots + \lambda_k}{v}$.

Scale-dependence

Although variance doesn't depend on location, it does depend on *scale*. If a set of heights were measured in inches, they will have a much larger variances than they would if measured in feet. Likewise, the choice of units of measurement also affect which basis vectors are returned by the PCA algorithm.

Scale-dependence

Although variance doesn't depend on location, it does depend on *scale*. If a set of heights were measured in inches, they will have a much larger variances than they would if measured in feet. Likewise, the choice of units of measurement also affect which basis vectors are returned by the PCA algorithm.

To avoid having PCA depend on the units of measurement, the original variables can be rescaled to each have variance 1 before starting the algorithm.

Scale-dependence

Although variance doesn't depend on location, it does depend on *scale*. If a set of heights were measured in inches, they will have a much larger variances than they would if measured in feet. Likewise, the choice of units of measurement also affect which basis vectors are returned by the PCA algorithm.

To avoid having PCA depend on the units of measurement, the original variables can be rescaled to each have variance 1 before starting the algorithm.

However, if all of the original variables are measured in the same units, you may want to preserve their differing variances rather than rescaling.

Implementation in R

Running example code together

See *PCA-examples.R* Exercises 1 - 3.

Exercises to try for yourself

Work through *PCA-examples.R* Exercise 4.