Matrix methods – Principal component analysis (1)

MATH 2740 – Mathematics of Data Science – Lecture 10

Julien Arino

julien.arino@umanitoba.ca

Department of Mathematics @ University of Manitoba

Fall 202X

The University of Manitoba campuses are located on original lands of Anishinaabeg, Ininew, Anisininew, Dakota and Dene peoples, and on the National Homeland of the Red River Métis. We respect the Treaties that were made on these territories, we acknowledge the harms and mistakes of the past, and we dedicate ourselves to move forward in partnership with Indigenous communities in a spirit of Reconciliation and collaboration.

Outline

A running example: hockey players

Change of basis

Example of change of basis

A crash course on probability

Dimensionality reduction

One of the reasons the SVD is used is for dimensionality reduction. However, SVD has many many other uses

Now we look at another dimensionality reduction technique, PCA

PCA is often used as a blackbox technique, here we take a look at the math behind it

What is PCA?

Linear algebraic technique

Helps reduce a complex dataset to a lower dimensional one

Non-parametric method: does not assume anything about data distribution (distribution from the statistical point of view)

A running example: hockey players

Change of basis

Example of change of basis

A crash course on probability

A 2D example

Dataset (link) of height and weight of some hockey players

```r
# From https://figshare.com/ndownloader/files/5303173
data = read.csv("https://github.com/julien-arino/math-of-data-science/raw/re
dim(data)

## [1] 6292   13
```

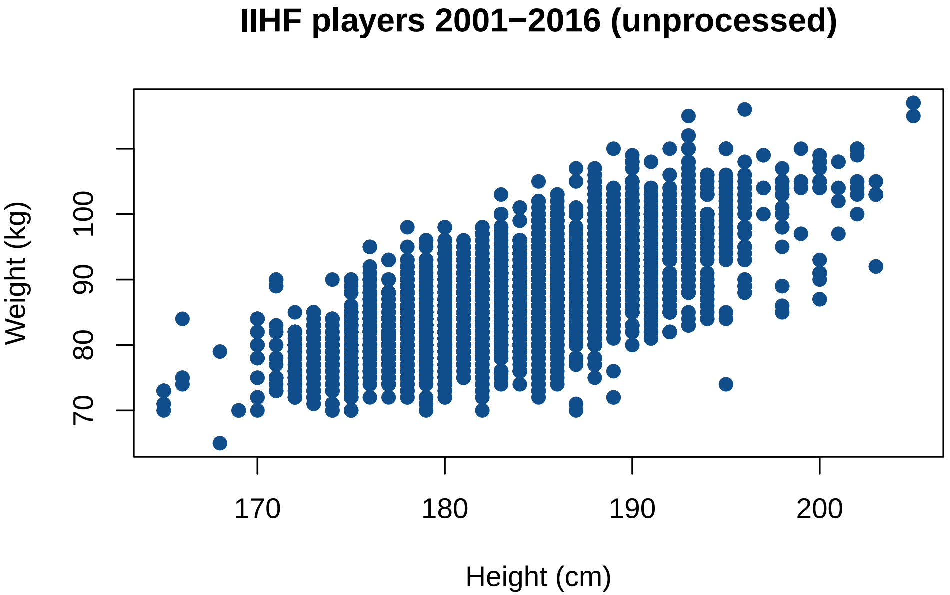
In case you are wondering, this is a database of ice hockey players at IIHF world championships, 2001-2016, assembled by the dataset's author

See some comments here

```
head(data, n=3)

##   year country no                         name position side height weight        bi
## 1 2001     RUS 10    tverdovsky oleg              D    L    185     84 1976-05
## 2 2001     RUS  2  vichnevsky vitali              D    L    188     86 1980-03
## 3 2001     RUS 26 petrochinin evgeni              D    L    182     95 1976-02
##                      club      age cohort      bmi
## 1   anaheim mighty ducks  24.95277   1976 24.54346
## 2   anaheim mighty ducks  21.11978   1980 24.33228
## 3 severstal cherepovetal  25.22930   1976 28.68011
```

As usual, it is a good idea to plot this to get a sense of the lay of the land


IIHF players 2001–2016 (unprocessed)
Weight (kg)
70
80
90
100
170
180
190
200
Height (cm)

The author of the study is interested in the evolution of weights, so it is likely that the same person will be in the dataset several times

Let us check this: first check will be `FALSE` if the number of unique names does not match the number of rows in the dataset

```
length(unique(data$name)) == dim(data)[1]

## [1] FALSE

length(unique(data$name))

## [1] 3278
```

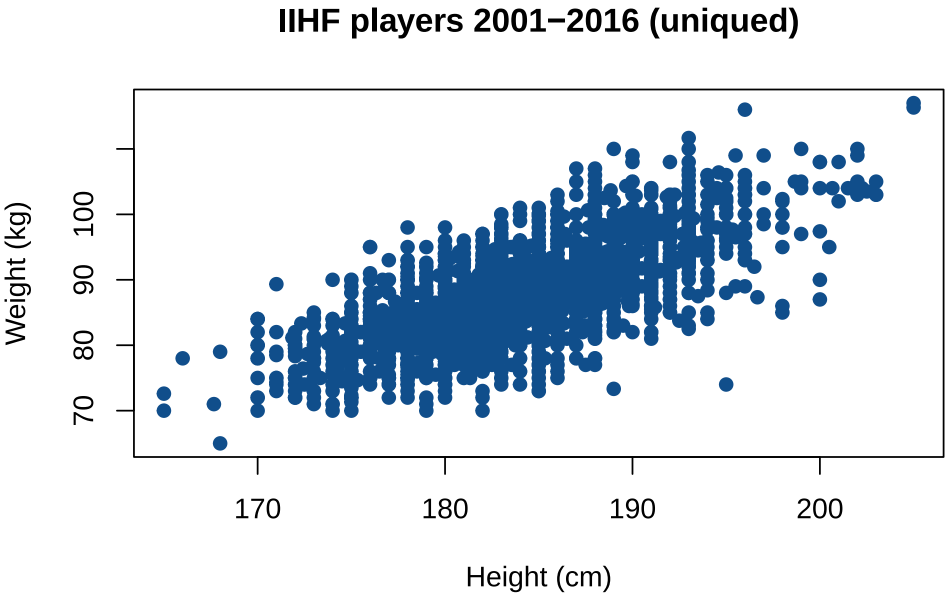
Not interested in the evolution of weights, so simplify: if more than one record for someone, take average of recorded weights and heights

To be extra careful, could check as well that there are no major variations on player height (homonymies?)

```
data_simplified = data.frame(name = unique(data$name))
w = c()
h = c()
for (n in data_simplified$name) {
    tmp = data[which(data$name == n),]
    h = c(h, mean(tmp$height))
    w = c(w, mean(tmp$weight))
}
data_simplified$weight = w
data_simplified$height = h
```

```r
data = data_simplified
head(data_simplified, n = 6)

##                     name weight height
## 1      tverdovsky oleg    84.0  185.0
## 2   vichnevsky vitali    86.0  188.0
## 3 petrochinin evgeni    95.0  182.0
## 4      zhdan alexander    85.5  178.5
## 5      orekhovsky oleg    88.0  175.0
## 6        zhukov sergei    92.5  193.0
```


IIHF players 2001–2016 (uniqued)
Weight (kg)
70
80
90
100
170
180
190
200
Height (cm)

Centre the data

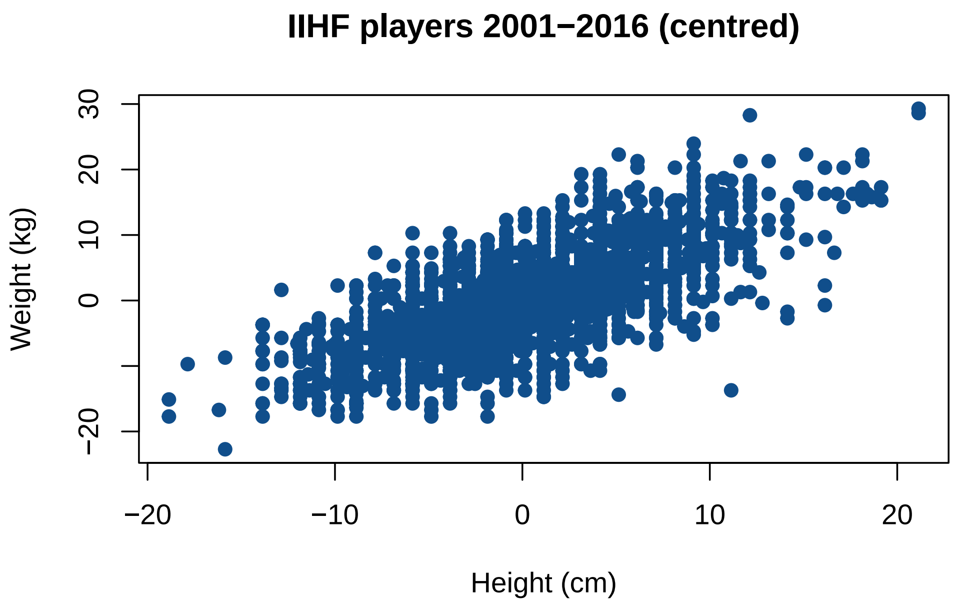
```
mean(data$weight)

## [1] 87.71555

mean(data$height)

## [1] 183.8596

data$weight.c = data$weight-mean(data$weight)
data$height.c = data$height-mean(data$height)
```


IIHF players 2001–2016 (centred)
Weight (kg)
30
20
10
0
−20
−20
−10
0
10
20
Height (cm)

Setting things up

Each player is a row in the matrix (an *observation*), each variable (*height* and *weight*) is a column

After deduplication, we have an 3278×2 matrix (actually, 3278×4 if we consider the uncentred and centred variables, but we will use one or the other, not both uncentred and centred)

We want to find what carries the most information

For this, we are going to project the information in a new basis in which the first "dimension" will carry most information (in a sense we'll define later), the second dimension will carry a little less, etc.

In order to do so, we need to learn how to change bases

A running example: hockey players

Change of basis

Example of change of basis

A crash course on probability

In the following slide,

$$[\boldsymbol{x}]_{\mathcal{B}}$$

denotes the coordinates of $\boldsymbol{x}$ in the basis $\mathcal{B}$

The aim of a change of basis is to express vectors in another coordinate system (another basis)

We do so by finding a matrix allowing to move from one basis to another

Change of basis

Definition 80 (Change of basis matrix)

$\mathcal{B} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ and $\mathcal{C} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ bases of vector space V
The **change of basis matrix** $P_{\mathcal{C}\leftarrow\mathcal{B}} \in \mathcal{M}_n$,

$$P_{\mathcal{C}\leftarrow\mathcal{B}} = [[\boldsymbol{u}_1]_{\mathcal{C}} \cdots [\boldsymbol{u}_n]_{\mathcal{C}}]$$

has columns the coordinate vectors $[\boldsymbol{u}_1]_{\mathcal{C}}, \ldots, [\boldsymbol{u}_n]_{\mathcal{C}}$ of vectors in $\mathcal{B}$ with respect to $\mathcal{C}$

Theorem 81

$\mathcal{B} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ and $\mathcal{C} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ bases of vector space V and $P_{\mathcal{C}\leftarrow\mathcal{B}}$ a change of basis matrix from $\mathcal{B}$ to $\mathcal{C}$

1. $\forall \boldsymbol{x} \in V$, $P_{\mathcal{C}\leftarrow\mathcal{B}}[\boldsymbol{x}]_{\mathcal{B}} = [\boldsymbol{x}]_{\mathcal{C}}$
2. $P_{\mathcal{C}\leftarrow\mathcal{B}}$ *s.t.* $\forall \boldsymbol{x} \in V$, $P_{\mathcal{C}\leftarrow\mathcal{B}}[\boldsymbol{x}]_{\mathcal{B}} = [\boldsymbol{x}]_{\mathcal{C}}$ *is* ***unique***
3. $P_{\mathcal{C}\leftarrow\mathcal{B}}$ *invertible and* $P_{\mathcal{C}\leftarrow\mathcal{B}}^{-1} = P_{\mathcal{B}\leftarrow\mathcal{C}}$

Row-reduction method for changing bases

Theorem 82

$\mathcal{B} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ and $\mathcal{C} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ bases of vector space V. Let $\mathcal{E}$ be any basis for V,

$$B = [[\boldsymbol{u}_1]_{\mathcal{E}}, \ldots, [\boldsymbol{u}_n]_{\mathcal{E}}] \ and \ C = [[\boldsymbol{v}_1]_{\mathcal{E}}, \ldots, [\boldsymbol{v}_n]_{\mathcal{E}}]$$

and let $[C|B]$ be the augmented matrix constructed using C and B. Then

$$RREF\left([C|B]\right) = [\mathbb{I}|P_{\mathcal{C}\leftarrow\mathcal{B}}]$$

If working in $\mathbb{R}^n$, this is quite useful with $\mathcal{E}$ the standard basis of $\mathbb{R}^n$ (it does not matter if $\mathcal{B} = \mathcal{E}$)

So the question now becomes

How do we find what new basis to look at our data in?

(Changing the basis does not change the data, just the view you have of it)

(Think of what happens when you do a headstand.. your up becomes down, your right and left switch, but the world does not change, just your view of it)

(Changes of bases are *fundamental* operations in Science)

A running example: hockey players

Change of basis

Example of change of basis

A crash course on probability

Worked example: change of basis

Problem: Find the change of basis matrix from basis $\mathcal{B}$ to basis $\mathcal{C}$ in $\mathbb{R}^2$, where

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{C} = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}$$

Then use this to find the coordinates of $\boldsymbol{x} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ in basis $\mathcal{C}$

Step 1 – Set up the matrices

- $\mathcal{B}$ is the standard basis of $\mathbb{R}^2$, so $[\boldsymbol{u}_1]_{\mathcal{E}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $[\boldsymbol{u}_2]_{\mathcal{E}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
- For $\mathcal{C}$: $[\boldsymbol{v}_1]_{\mathcal{E}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $[\boldsymbol{v}_2]_{\mathcal{E}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

Step 2 – Row reduce

Using Theorem 82, we form the augmented matrix $[C|B]$:

$$[C|B] = \left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{array}\right]$$

Row reduce to RREF:

$$\left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{array}\right] \xrightarrow{R_2 \leftarrow R_2 - R_1} \left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \end{array}\right]$$

$$\xrightarrow{R_2 \leftarrow -\frac{1}{2} R_2} \left[\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 0 & 1 & \frac{1}{2} & -\frac{1}{2} \end{array}\right]$$

$$\xrightarrow{R_1 \leftarrow R_1 - R_2} \left[\begin{array}{cc|cc} 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & -\frac{1}{2} \end{array}\right]$$

Step 3 – Extract the change of basis matrix

From the RREF form $[\mathbb{I}|P_{\mathcal{C}\leftarrow\mathcal{B}}]$, we get:

$$P_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}$$

Verification: Let's check that this matrix works correctly.

- $[\boldsymbol{u}_1]_{\mathcal{C}} = P_{\mathcal{C}\leftarrow\mathcal{B}}[\boldsymbol{u}_1]_{\mathcal{B}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$
- Check: $\frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Step 4 – Coordinates of $\boldsymbol{x}$ in basis $\mathcal{C}$

Now we find $[\boldsymbol{x}]_{\mathcal{C}}$ for $\boldsymbol{x} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$:

$$[\boldsymbol{x}]_{\mathcal{C}} = P_{\mathcal{C}\leftarrow\mathcal{B}}[\boldsymbol{x}]_{\mathcal{B}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{1}{2} \end{pmatrix}$$

Verification: Check that this gives us back the original vector:

$$\frac{5}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} + \frac{1}{2} \\ \frac{5}{2} - \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Answer: The coordinates of $\boldsymbol{x} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ in basis $\mathcal{C}$ are $[\boldsymbol{x}]_{\mathcal{C}} = \begin{pmatrix} \frac{5}{2} \\ \frac{1}{2} \end{pmatrix}$

Alternative method – Direct calculation

We could also solve this directly by setting up the system:

$$\boldsymbol{x} = c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2$$

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

This gives us the system:

$$\begin{aligned} c_1 + c_2 &= 3 \\ c_1 - c_2 &= 2 \end{aligned}$$

Solving: $c_1 = \frac{5}{2}$, $c_2 = \frac{1}{2}$

This confirms our result: $[\boldsymbol{x}]_{\mathcal{C}} = \begin{pmatrix} \frac{5}{2} \\ \frac{1}{2} \end{pmatrix}$

A running example: hockey players

Change of basis

Example of change of basis

A crash course on probability

Why probability?

We said earlier that we would look for a basis in which the first dimension carries most information

But how do we define *information*?

We use concepts from probability and statistics to do so

A good measure of information is *variance* (how much data varies around the mean)

Brief “review” of some probability concepts

Proper definition of *probability* requires to use *measure theory*.. will not get into details here

A **random variable** X is a *measurable* function $X : \Omega \to E$, where Ω is a set of outcomes (*sample space*) and E is a measurable space

$\mathbb{P}(X \in S \subseteq E) = \mathbb{P}(\omega \in \Omega | X(\omega) \in S)$

Distribution function of a r.v., $F(x) = \mathbb{P}(X \leq x)$, describes the distribution of a r.v.

R.v. can be discrete or continuous or .. other things.

Definition 83 (Variance)

Let X be a random variable. The **variance** of X is given by

$$\operatorname{Var} X = E\left[(X - E(X))^2\right]$$

where E is the expected value

Definition 84 (Covariance)

Let X, Y be jointly distributed random variables. The **covariance** of X and Y is given by

$$\operatorname{cov}(X, Y) = E\left[(X - E(X))\,(Y - E(Y))\right]$$

Note that $\operatorname{cov}(X, X) = E\left[(X - E(X))^2\right] = \operatorname{Var} X$

In practice: “true law” versus “observation”

In statistics: we reason on the *true law* of distributions, but we usually have only access to a sample

We then use **estimators** to .. estimate the value of a parameter, e.g., the mean, variance and covariance

Definition 85 (Unbiased estimators of the mean and variance)

Let $x_1, \dots, x_n$ be data points (the *sample*) and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

be the **mean** of the data. An unbiased estimator of the variance of the sample is

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Definition 86 (Unbiased estimator of the covariance)

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be data points,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

be the means of the data. An estimator of the covariance of the sample is

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

What does covariance do?

Variance explains how data disperses around the mean, in a 1-D case

Covariance measures the relationship between two dimensions. E.g., height and weight

More than the exact value, the sign is important:

- $\text{cov}(X, Y) > 0$: both dimensions change in the same "direction"; e.g., larger height usually means higher weight
- $\text{cov}(X, Y) < 0$: both dimensions change in reverse directions; e.g., time spent on social media and performance in this class
- $\text{cov}(X, Y) = 0$: the dimensions are independent from one another; e.g., sex/gender and "intelligence"

The covariance matrix (we usually have more than 2 variables)

Definition 87

Suppose p random variables $X_1, \ldots, X_p$. Then the covariance matrix is the symmetric matrix

$$\begin{pmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_p) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ \operatorname{cov}(X_p, X_1) & \operatorname{cov}(X_p, X_2) & \cdots & \operatorname{cov}(X_p, X_p) \end{pmatrix}$$

i.e., using the properties of covariance,

$$\begin{pmatrix} \operatorname{Var} X_1 & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_p) \\ \operatorname{cov}(X_1, X_2) & \operatorname{Var} X_2 & \cdots & \operatorname{cov}(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ \operatorname{cov}(X_1, X_p) & \operatorname{cov}(X_2, X_p) & \cdots & \operatorname{Var} X_p \end{pmatrix}$$