Statistical learning

Report 1

Maciej Szczutko

2024-03-17

Exercises

Ex1

$$\mathbf{A} = \left[\begin{array}{rr} 3 & -1 \\ -1 & 3 \end{array}\right]$$

1. Is A is symmetric? Yes. It's trivial.
2. Spectral decomposition.

Lets use fact the characteristic polynomial for 2x2 matrix can be written as:

$$\lambda^2 - tr(A)\lambda + det(A).$$

So we get $p_A(\lambda) = \lambda^2 - 6\lambda + 8$. Then we easily find roots $\lambda_1 = 2$ and $\lambda_2 = 4$.

Solving equation

$$Ax = \lambda_i x$$

for $i = 1, 2$ we obtain $e_1 = (1, 1)^T$ and $e_2 = (-1, 1)^T$ for λ_1, λ_2 respectively. We can norm vector to unity or just multiply one of matrix by square of scaling factor.

Then spectal decomposition is:

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix}$$

3. One way of writing the spectral decomposition of $\mathbf{A}$ is

$$\lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T$$

This is just direct sum of subspace spanned by eigen vectors. The first component represent space spanned by λ_1 and second one by another one.

4. Use the spectral decomposition of A given above and find $\sqrt{A}$. Check that the matrix you found satisfies

$$\sqrt{\mathrm{A}}\sqrt{\mathrm{A}} = \mathrm{A}$$

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix}.$$

Because A has representation $P\Lambda P^{-1}$ the $\sqrt{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix}$.

From spectral decomposition we can easily check that

$$\begin{aligned}\sqrt{A}\sqrt{A} &= P\Lambda^{\frac{1}{2}}P^{-1}P\Lambda^{\frac{1}{2}}P^{-1} = \\ &= P\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}P^{-1} = P\Lambda P = A.\end{aligned}$$

Ex2

Consider the spectral decomposition of a positive definite matrix as given in Lecture 1:

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$$

The columns of $\mathbf{P}$ are made of eigenvectors $\mathbf{e}_i, i = 1, \dots, n$ and they are orthonormalized, i.e. their lengths are one and they are orthogonal (peripendicular) one to another. The diagonal matrix $\boldsymbol{\Lambda}$ has the corresponding (positive) eigenvalues on the diagonal. Provide argument for the following

1. $\mathbf{P}^T = \mathbf{P}^{-1}$

Let the $< u, v >$ be a standard scalar product of 2 vectors in R^n. Lets use δ_{ij} for Knocker delta. We can write $P = (C_1, C_2, \dots, C_n)$ where $C_i \in R^n$. Then $P^T = (C_1^T, C_2^T, \dots, C_n^T)$.

From orthogonality we have

$$\langle C_i, C_j \rangle = \delta_{ij}$$

.

$$P^T P = (\langle C_i, C_j \rangle)_{1 \leq i,j \leq n} = I_n$$

2. Determinant of $\boldsymbol{\Lambda}$ is equal to the product of the terms on the diagonal.

This is consequences of fact the determinant is product of eigen values. For diagonal matrices the eigen values are just the element on the diagonals.

3. Dederminant of $\mathbf{A}$ is the same as that of $\boldsymbol{\Lambda}$.

$$\begin{aligned}&\det(A) = \det\left(P\Lambda P^{-1}\right) = \\ &= \det(P)\det(\Lambda)\det\left(P^{-1}\right) = \\ &= \det(PP^{-1})\det(\Lambda) = \det(\Lambda)\end{aligned}$$

4. Find the inverse matrix to $\boldsymbol{\Lambda}$, i.e. $\boldsymbol{\Lambda}^{-1}$.

For diagonal matrix multiplication simplify to Hadamard product (element-wise product). Using this we can easily guess that Λ^{-1} is just

$$\begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_n} \end{bmatrix}.$$

5. A simple way to determine the inverse of a matrix A from its spectral decomposition is through $\mathbf{A}^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}^T$.

Verify that the right hand side of the above indeed define the inverse of $\mathbf{A}$.

$$A \cdot P\Lambda^{-1}P^\top = (P\Lambda P^\top)P\Lambda^{-1}P^\top = P\Lambda(P^\top P)\Lambda^{-1}P^\top = P(\Lambda\Lambda^{-1})P^\top = I$$

If inverse of A exist is must be this matrix.

6. All satisfies :)

Ex 3

L - length distribution W - weight distribution

$\mu_L = 49.8 \quad \sigma_L^2 = 2.5^2 = 6.25 \quad \mu_W = 3343 \quad \sigma_W^2 = 528^2 = 278784$

1. Write explicitly the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. $\begin{bmatrix}3343 & 49.8\end{bmatrix}$

$$\Sigma = \begin{pmatrix} \sigma_W^2 & corr(W,L)\sigma_W\sigma_L \\ corr(L,W)\sigma_W\sigma_L & \sigma_L^2 \end{pmatrix} =$$

$\Sigma = \begin{bmatrix} 278784 & 990 \\ 990 & 6.25 \end{bmatrix}$

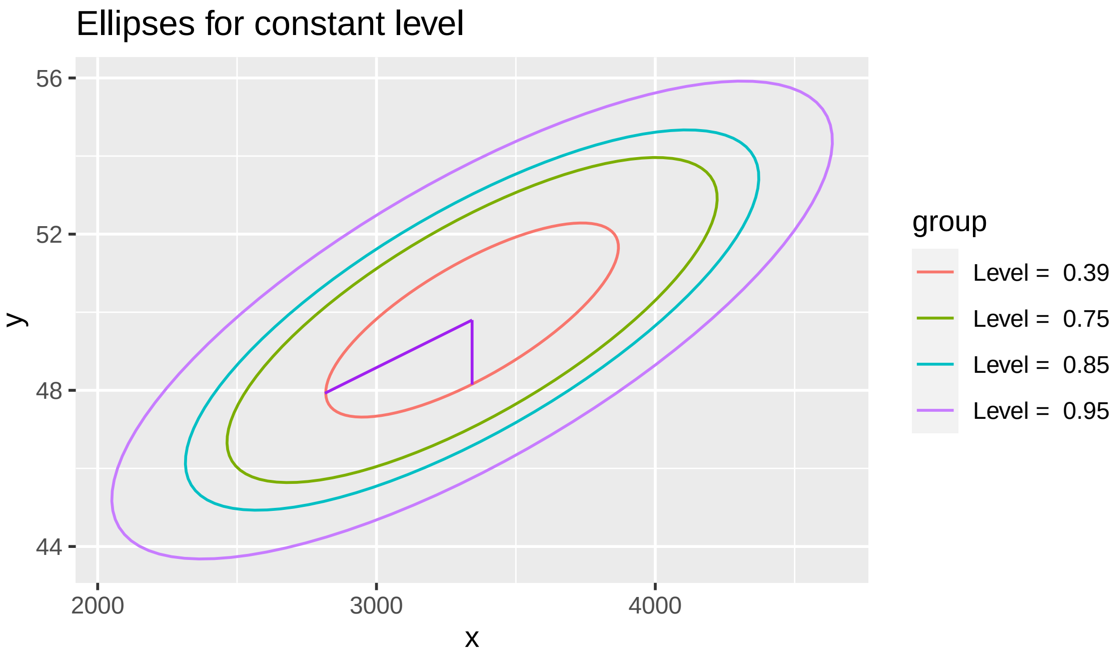
2. Density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$f(\mathbf{x}) = \frac{1}{(2\pi)873.098} \exp\left(-\frac{1}{2}(\mathbf{x} - \begin{pmatrix} 3343 \\ 49.8 \end{pmatrix})^T \begin{pmatrix} 0 & -0.001 \\ -0.001 & 0.366 \end{pmatrix} (\mathbf{x} - \begin{pmatrix} 3343 \\ 49.8 \end{pmatrix})\right)$$

3. Find eigenvalues and eigenvectors of the covariance matrix

$\Sigma = PVP^T$ where $V = \begin{bmatrix} 278787.5157 & 0 \\ 0 & 2.7343 \end{bmatrix}$ and $P = \begin{bmatrix} -1 & 0.0036 \\ -0.0036 & -1 \end{bmatrix}$.



I don't know how to better represent this vectors. They look for not orthogonal due graph proportion. But we know the scaling factor is $\chi^2_{df}(\alpha)$ so we can select such level that this factor will be 1. Then we clearly see that eigenvector point axes direction.

4. How many parameters characterize a bivariate normal distribution?

The p-dimensional normal distribution is characterized by $2p + \binom{p}{2}$. p means, p variances and correlations.

5. This one dimensional normal distribution. $L \sim N(49.8, 6.25)$.
6. The best guess will be 49.8 ± 7.5[cm] by three σ rule.

Ex 4 (conditional distribution)

1. What is the distribution of L given this additional information? Give its name and parameters. It is conditional distribution $L|W$. From lecture we know it will be also normal distribution. For bivariate normal distribution the mean and variance are given as below:

$$E(X \mid Y=y) = \mu_X + \sigma_X \frac{y-\mu_Y}{\sigma_Y}$$
$$\operatorname{Var}(X \mid Y=y) = \sigma_X^2 \left(1-\rho^2\right)$$

$$L|W \sim N(52.221875, 2.734375)$$

2. Improve your previous guess and provide with accuracy limits.

The best guess will be 52.221875 ± 4.962[cm] by three σ rule.

3. Compare the answers from this and previous problems and comment how additional information affected the prediction value and accuracy.

From the conditional variance formula we can see that the width of interval will stayed the same iff the variables are uncorrelated. Otherwise we improve the "prediction". This is beacuse we lower the variance ($p \in (0,1)$) and the width of interval getting shorter.

Ex 5

Problem 5 Let $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$ be independent $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vectors of a dimension p.

1. Find the distribution of each of the following vectors:

$$\mathbf{V}_1 = \frac{1}{4}\mathbf{X}_1 - \frac{1}{2}\mathbf{X}_2 + \frac{1}{4}\mathbf{X}_3$$
$$\mathbf{V}_2 = \frac{1}{4}\mathbf{X}_1 - \frac{1}{2}\mathbf{X}_2 - \frac{1}{4}\mathbf{X}_3$$

Let use the fact if $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ then $X+Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$ and $-Y \sim \mathcal{N}(-\mu_Y, \Sigma_Y)$. Also we can use how the formula for linear transformation for normal.

$$Z = AX + b \sim \mathcal{N}\left(A\mu + b, A\Sigma A^{\mathrm{T}}\right) (*)$$

Using this we obtain

$$\mu_1 = \frac{1}{4}\mu - \frac{1}{2}\mu + \frac{1}{4}\mu = 0 \text{ and } \mu_2 = \frac{1}{4}\mu - \frac{1}{2}\mu - \frac{1}{4}\mu = -\frac{1}{2}\mu$$

Multiplying by scalar scale covariance matrix by square $(*)$.

$$\Sigma_1 = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{3}{8}\Sigma \text{ and } \Sigma_1 = \frac{1}{16}\Sigma + \frac{1}{4}\Sigma + \frac{1}{16}\Sigma = \frac{3}{8}\Sigma$$

2. Find the joint distribution of the above vectors.

Let's write those vectors as single one of dimension $3p$.

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

And define matrix

$$A = \begin{pmatrix} \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{pmatrix}$$

Then we can easily calculate μ and Σ from $(*)$.

The mean will be simply

$$\begin{pmatrix} 0 \\ -\frac{1}{2}\mu \end{pmatrix}$$

$$B\Sigma_{3p}B^T = \begin{pmatrix} \frac{3}{8}\Sigma_p & \frac{1}{4}\Sigma_p \\ \frac{1}{4}\Sigma_p & \frac{3}{8}\Sigma_p \end{pmatrix}$$

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\mu \end{pmatrix}, \begin{pmatrix} \frac{3}{8}\Sigma_p & \frac{1}{4}\Sigma_p \\ \frac{1}{4}\Sigma_p & \frac{3}{8}\Sigma_p \end{pmatrix}\right)$$

Project part 1

1 Sample statistic

i	μ_i	Var	ρ
L	49.2376358695652	4.44330270260278	915.295510869565
W	3233.54510869565	220276.657663117	915.295510869565

2 Normality

I will verify normality through histogram and qq-plots inspections.

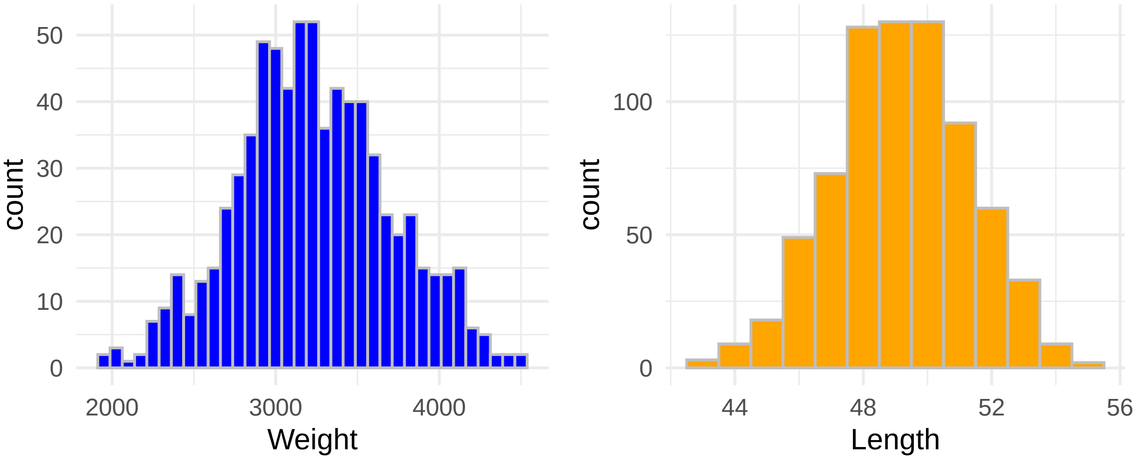
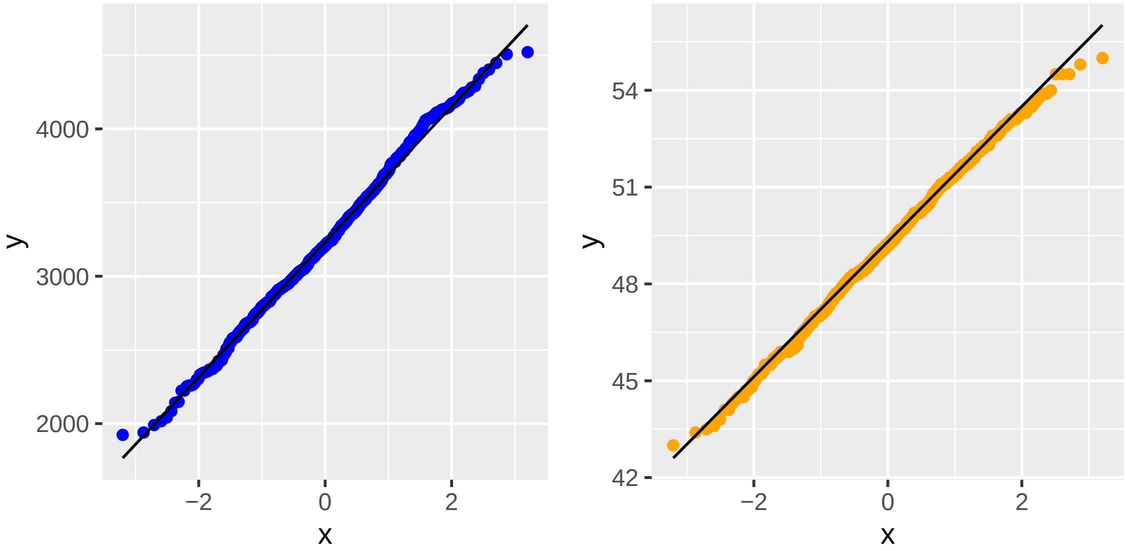


Figure 1: Historams of marignals

Histograms for *Weight* and *Length* seems to be symmetric without heavy tails. Let's take a further look at the qq-plots



The sample quantiles are very close to the theoretical ones. We can assume the marginals from sample comes from normal distribution.

3 Find the ellipsoids that would serve classification regions for scores as described above.

According to method description we need to plot ellipses for two levels (0.95 and 0.75). On the plot the area for given score are marked with colour. Each sample inside ellipse represented level 0.75 got score 2. If it outside 0.75, but inside 0.95 – got score 1. Otherwise we assign 0 score.

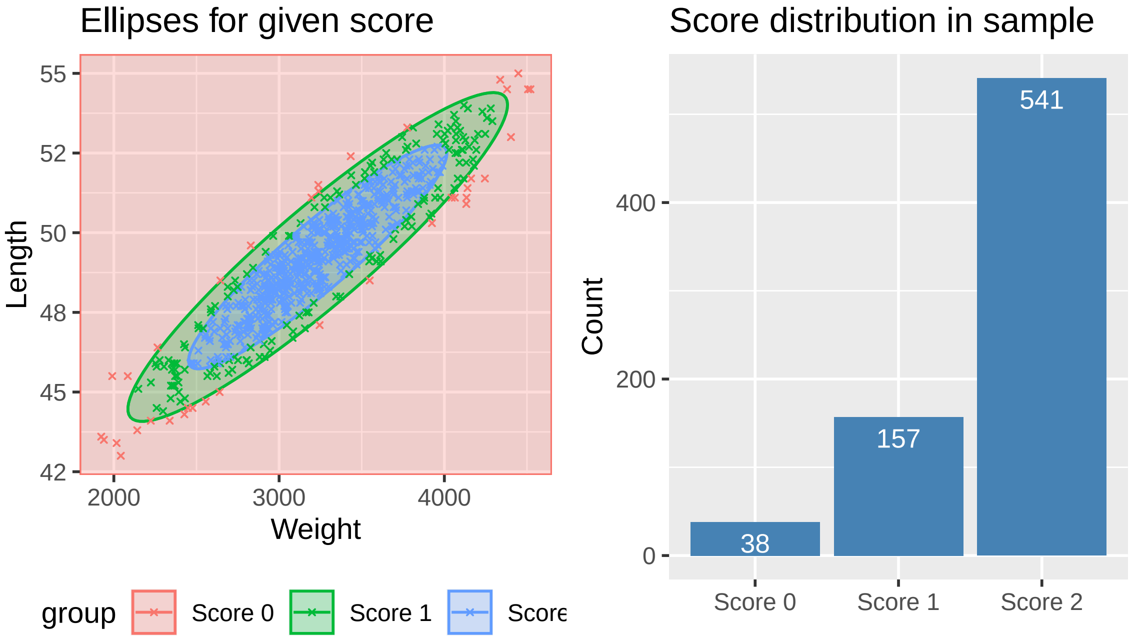


Figure 2: Score graphs

4 How many children would score zero, one, and two, respectively? Illustrate this classification on the graphs.

The group distribution is presented on the graph. Indeed the shape remind the χ^2_2 distribution. This would be seen better if we classified into more than 3 classes (but using same methodology i.e. 10 classes and each class is defined by 9 elipses for levels $0.1, 0.2, \ldots, 0.9$).

5. Find the spectral decomposition of the estimated covariance matrix.

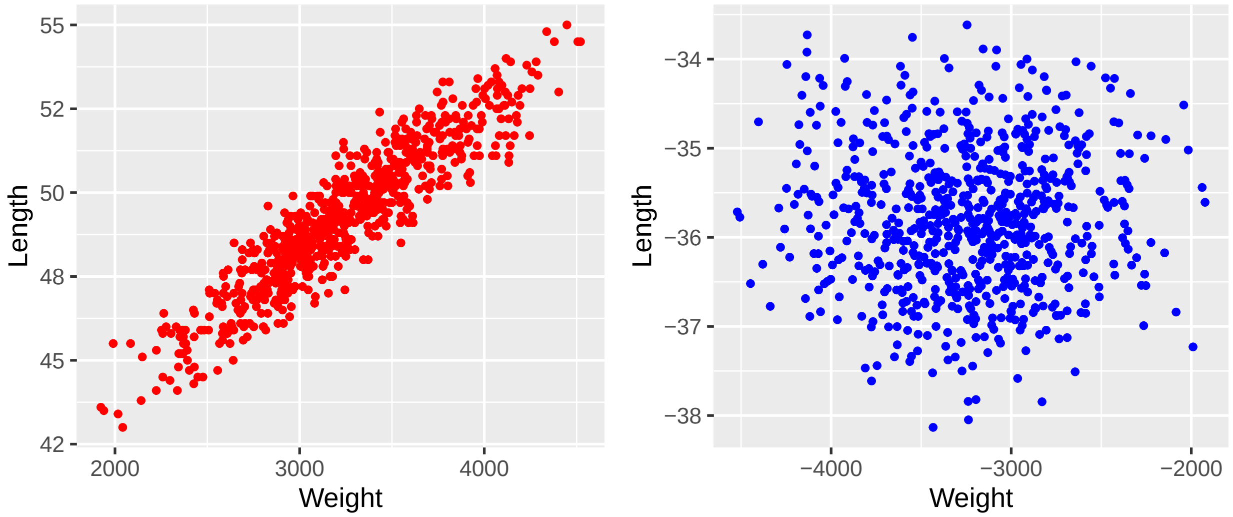
Estimated covariance matrix $\Sigma = \begin{bmatrix} 220276.6577 & 915.2955 \\ 915.2955 & 4.4433 \end{bmatrix}$. Instead of calculating spectral decomposition by hand we will use *eigen* function from base package.

$\Sigma = PVP^T$ where $V = \begin{bmatrix} 220280.4609 & 0 \\ 0 & 0.64 \end{bmatrix}$ and $P = \begin{bmatrix} -1 & 0.0042 \\ -0.0042 & -1 \end{bmatrix}$.

Note I display result with precision to 4 decimal places.

6. $\mathbf{P}^T\mathbf{X}$ vs X

Now we will compare plots for original data and transformed by $\mathbf{P}^T$.



We can see the original data are focus in ellipse and thus variable are dependent. After transformation there is no evident clusters or pattern in data. So we end up with new, independent, random variables. This confirm fact from lecture.

> The distribution of $\mathbf{P}^T(\mathbf{X}-\mu)$ is the same as the $\left(\sqrt{\lambda_1}Z_1,\ldots,\sqrt{\lambda_p}Z_p\right)$, where Z_i 's are independent standard normal random variables.

Here we don't centerize data before transformation, so it won't be standard normal (mean might be nonzero). But the independence property will be preserved.

Project part 2

Basic statistic for ParentsWeightLength.txt dataset.

Table 2: Means for features

Weight	Length	FatherHeight	MotherHeight
3234	49	177	167

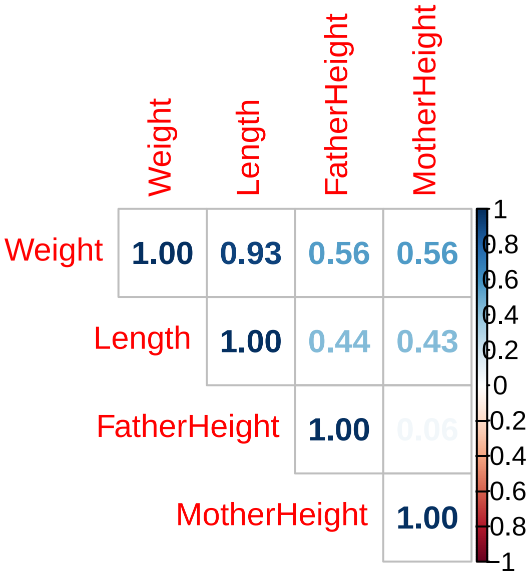


Figure 3: Correlation between features

We can see the correlation between parents's height and newborn weight, length isn't as high as correlation between weight and length. But for sure can have direct impact on this variables. Additional I would assume that parents's height are independent.

Normality of marginal distributions

We limit this part to variables *FatherHeight* and *MotherHeight* as the *Weight* and *Length* are the same as in part one of the project.

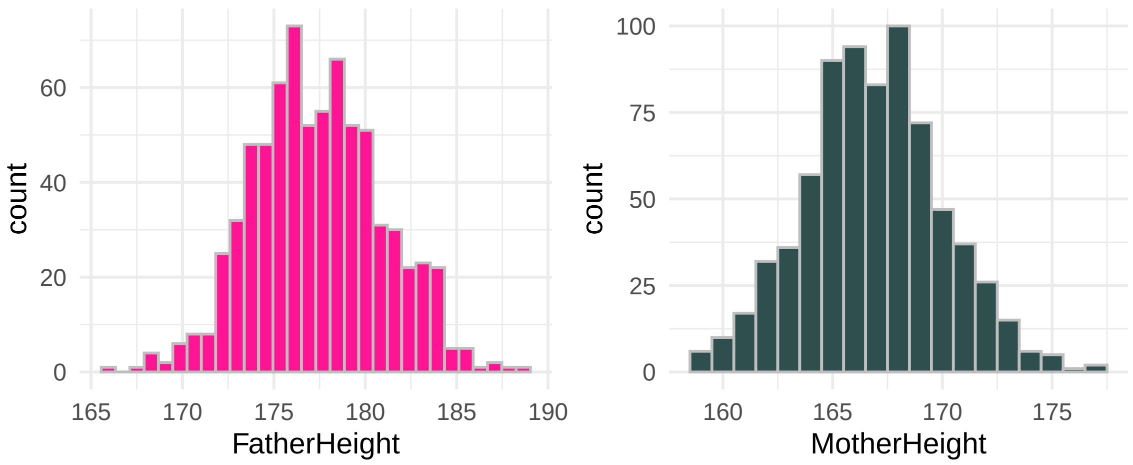


Figure 4: Histogram of marginals (parents)

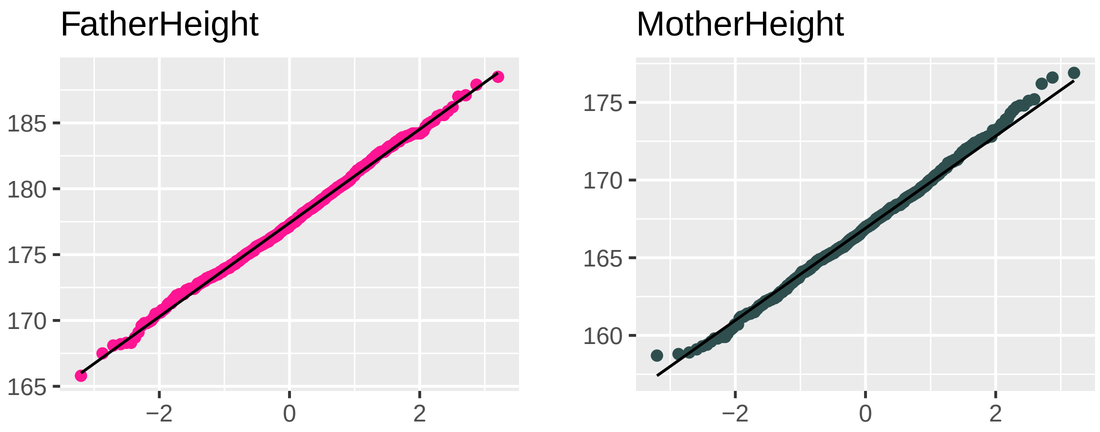


Figure 5: Theoretical vs empirical quantiles

Based on above plots, again, we can assume the data comes from normal distribution.

Contional distribution of Weight and Length

Now we want to derive distribution of *Weight* and *Length* conditioned by distribution of parents's height. We use direct formula from lecture to derive parameters for this distribution.

$$\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{N}_q \left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \right)$$

From the formula we can see that covariance matrix is determined only by covariance matrix of source distribution, but the mean is linear function of random vector.

$$\begin{pmatrix} W \\ L \end{pmatrix} \Bigg| \begin{pmatrix} W_F \\ W_M \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 3234 \\ 49 \end{pmatrix} + \begin{pmatrix} 69.88 & 80.15 \\ 0.25 & 0.28 \end{pmatrix} \begin{pmatrix} W_F - 177 \\ W_M - 167 \end{pmatrix}, \begin{pmatrix} 88858 & 456.8 \\ 456.8 & 2.8 \end{pmatrix} \right)$$

We observe that for each pair covariance is reduced. I'm not sure we can generalize previous findings (exercise 4). I think we do, if we assume that all variables are correlated (there is no pair with 0 correlation).

$\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ remind quadratic formula (but is not). I was trying to connect this with formula for special case in 1d ($\sigma_X^2\left(1-\rho^2\right)$) but I can't prove it formally.

Elipsoid for average height parents.

As the conditional mean depends on value of condition I will use mean of the $\begin{pmatrix} W_F \\ W_M \end{pmatrix}$, then conditional distribution has the same mean as non-conditional one. So the only difference is reduced covariance. I expect that now we will be more more restrictive than before.

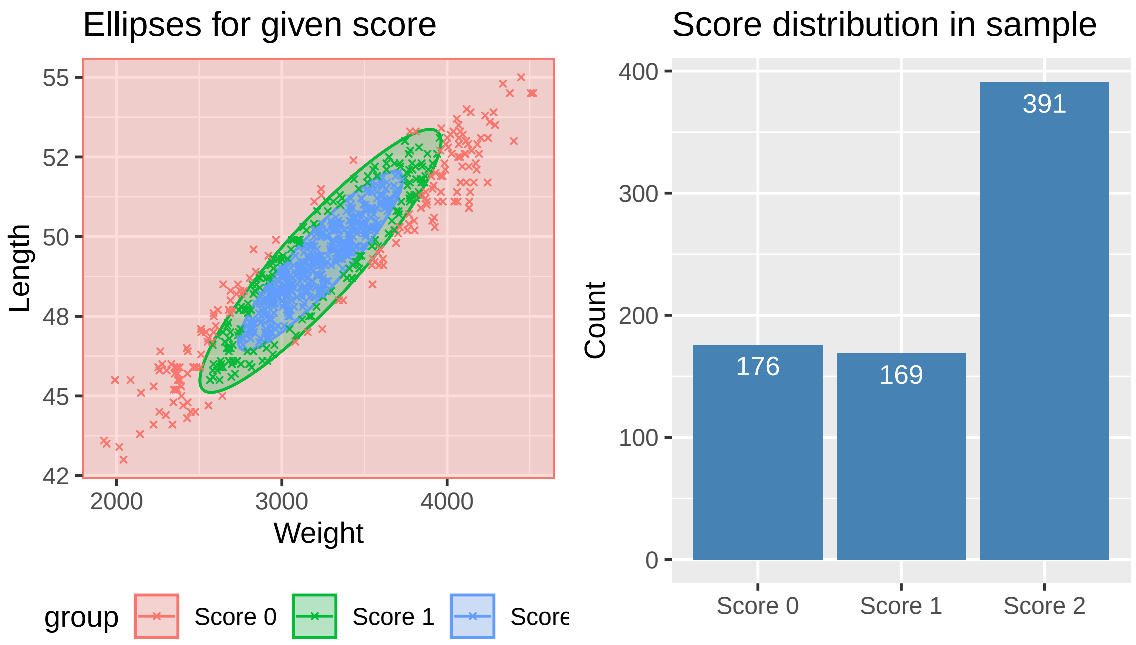


Figure 6: Score graphs

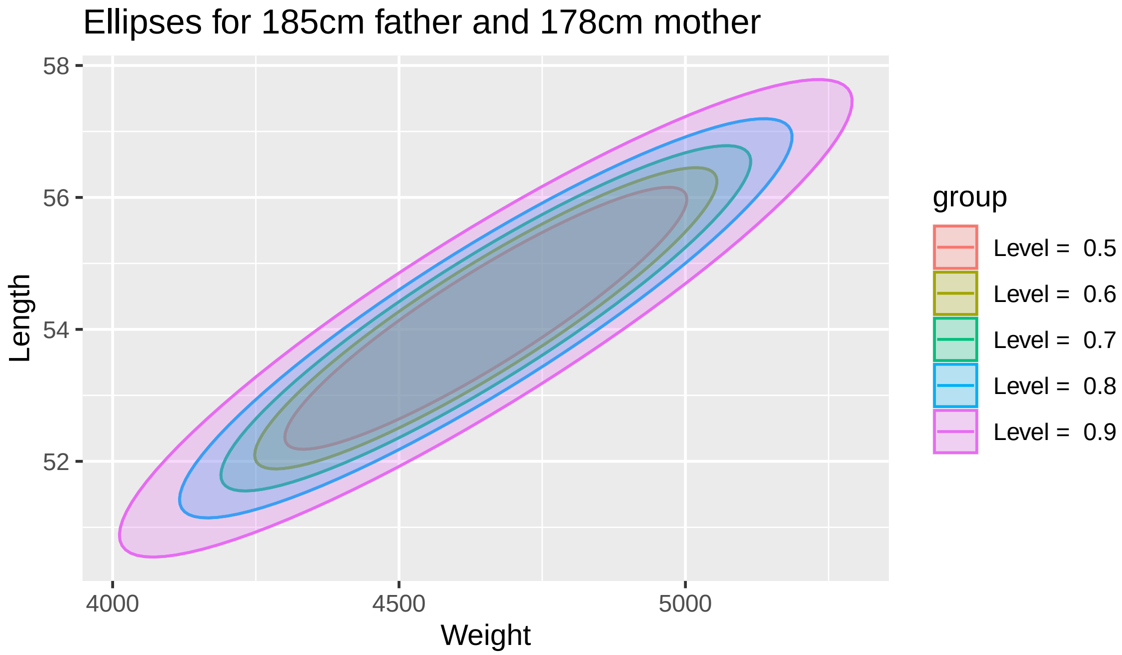
Scores of children

As we can see on the plots, the distribution of scores change if we add extra information about parents AND we assume each new born has average height parents. For me this example illustrate that now the classification rule become to restrictive.

Classification for quite tall parents.

Now we try to provide elipsses for child whose father has 185cm and mother 178cm.

The mean for conditional distribution is $\begin{bmatrix}4651.5373 & 54.1683\end{bmatrix}$.



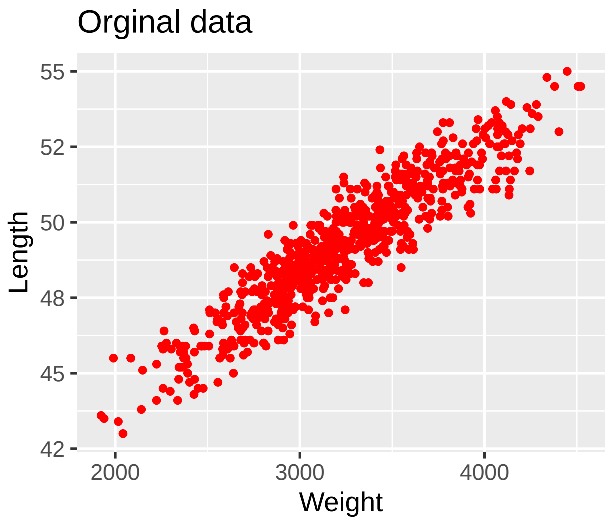
Comparing with previous graph we can observe that area of corresponding ellipses stay the same as the length of axes depends on λ_i, α level and dimension. But the all ellipses are shifted to to higher values as the centrer of ellipses is determined by mean of distribution which is “greater” now.

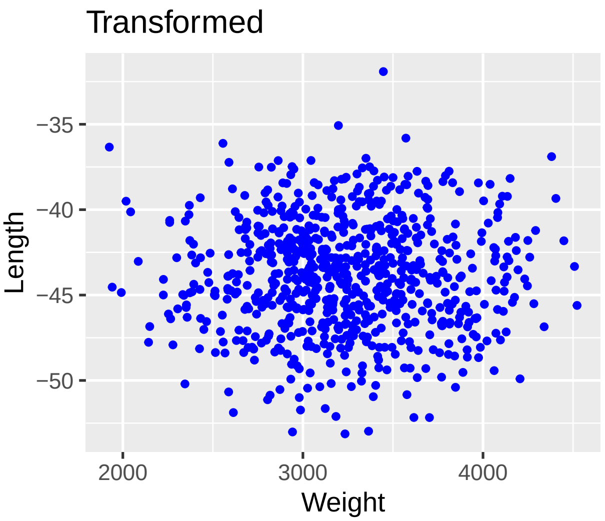
Spectral decomposition

$$\Sigma = PVP^T \text{ where } V = \begin{bmatrix} 220287.5102 & 0 & 0 & 0 \\ 0 & 10.7082 & 0 & 0 \\ 0 & 0 & 4.7884 & 0 \\ 0 & 0 & 0 & 0.4784 \end{bmatrix} \text{ and } P = \begin{bmatrix} 1 & 0.0012 & 0.0046 & 0.0052 \\ 0.0042 & 0.0137 & 0.1926 & -0.9812 \\ 0.0042 & -0.8164 & -0.5644 & -0.1221 \\ 0.0038 & 0.5773 & -0.8027 & -0.1495 \end{bmatrix}.$$

Transformed data

I can’t plot 4d data so I plot only relation between *Weight* and *Length*





The conclusion is the same as in part one. We have other basis now, so the coordinates differ but the independence still holds.