INTRODUCTION TO MULTIVARIATE STATISTICAL ANALYSIS

HEINRICH MATZINGER
Georgia Tech
E-mail: matzi@math.gatech.edu

May 21, 2015

Contents

1 A supervised learning problem: statistical classification

Assume that we run a fishing boat-factory which is highly automatized: the fish we catch is sorted automatically by a robot. The robot measures the size of the fish and then decides which type of fish it is. To simplify our present discussion we assume at first only two types of fishes: tuna and salmon. Assume that the robot can measure three sizes: small, medium and large. Now, let the code for the sizes be:

$$1 = small, 2 = medium, 3 = large.$$

There is an underling "probability model"also called "probability distribution" or in statistical parlance "the population distribution". The length of the fish will be denoted by X and the "class", that is the type of the fish is denoted by Y. We assume given a joint probability table:

	1	2	3
`tuna`	$P(Y = \texttt{tuna}, \texttt{X} = 1)$	$P(Y = \texttt{tuna}, \texttt{X} = 2)$	$P(Y = \texttt{tune}, \texttt{X} = 3)$
`salmon`	$P(Y = \texttt{salmon}, \texttt{X} = 1)$	$P(Y = \texttt{salmon}, \texttt{X} = 2)$	$P(Y = \texttt{salmon}, \texttt{X} = 3)$

.

The best possible decision rule is based on choosing the class which has highest probability given the size. Let us see an example:

We may have:

	1	2	3
`tuna`	0.1	0.2	0.3
`salmon`	0.2	0.1	0.1

(1.1)

So, we have $P(Y = salmon, X = 1) = 20\%$. This means that in the waters in which we are fishing, 20-percent of the fish are salmons of size 1. Similarly we have $P(Y = tuna, X = 3) = 0.3$. This means that 30-percent of the fish are *tuna* of size 3. If we know the underlying probability distribution, that is we know the table 1.5 to hold, what is the best decision rule for the robot to classify the fish? Again the robot is only given one of the three sizes $\{1, 2, 3\}$ and has to guess based on that information if it is a tuna or a salmon. Say the robot is given a fish of size 2. This fish is then twice as likely to be a tuna:

$$P(Y = tuna|X = 2) = \frac{0.2}{0.3} = \frac{2}{3}$$

and

$$P(Y = salmon|X = 2) = \frac{0.1}{0.3} = \frac{1}{3}$$

So, the best decision rule is that if you catch a fish of size 2 you classify it as tuna. With that rule whenever you catch a salmon of size 2 it gets misclassified. So, this adds to the total missclasification probability 10%. Similarly, we can device the best rule for size 1 fish as well as size 3 fish:
We have

$$P(Y = tuna|X = 1) = \frac{0.1}{0.3} = 33.\bar{3}\%, P(Y = salmon|X = 1) = \frac{0.2}{0.3} = 66.\bar{6}\%$$

So, $66.\bar{6}\%$ of the fish of size 1 are tuna, and hence it makes sense to classify the fish of size 1 as tuna. In this manner every salmon of size 1 will be missclassified adding 10% error into the missclassification percentage.
similarly with size 3 we get:

$$P(Y = tuna|X = 3) = \frac{0.3}{0.4} = \frac{3}{4}, P(Y = salmon|X = 3) = \frac{0.1}{0.4} = \frac{1}{4}$$

leading us to classify fish of size 3 as tuna.

The classification rule $g(.)$ can be viewed as a function from the set $\{1, 2, 3\}$ to the classes set $\{tuna, salmon\}$. In fancy machine learning parlance, classification rules are called *classifiers*. The set $\{1, 2, 3\}$ would be the feature space and *tuna* and *salmon* are the classes. The best rule is denoted by g^* and is called a *Bayes classifier*. So, formally the Bayse classifier is defined by:

$$g^*(x) = tuna \ \texttt{ if and only if } \ \frac{\texttt{P(Y = tuna|X = x)}}{\texttt{P(Y = salmon|X = x)}} \geq 1 \tag{1.2}$$

(when the conditional probabilities are exactly equal, we could classify either way. Here, we chose to assign the fish in case of equal probability to the tuna class).
Recall the formula for conditional probability of A given B. This formula is as follows: $P(A|B) = P(A \cap B)/P(B)$. In the present case, we condition on $X = x$. So, we can also rewrite the formula which defines the Bayes classifier. For this note that

$$P(Y = \texttt{tuna}|\texttt{X} = \texttt{x}) = \frac{\texttt{P(Y = tuna, X = x)}}{\texttt{P(X = x)}}$$

and

$$P(Y = \texttt{salmon}|\texttt{X} = \texttt{x}) = \frac{\texttt{P(Y = salmon, X = x)}}{\texttt{P(X = x)}}$$

. Hence,

$$\frac{P(Y = \texttt{tuna}|\texttt{X} = \texttt{x})}{P(Y = \texttt{salmon}|\texttt{X} = \texttt{x})} = \frac{P(Y = \texttt{tuna}, \texttt{X} = \texttt{x})/\texttt{P(X = x)}}{P(Y = \texttt{salmon}, \texttt{X} = \texttt{x})/\texttt{P(X = x)}} = \frac{P(Y = \texttt{tuna}, \texttt{X} = \texttt{x})}{P(Y = \texttt{salmon}, \texttt{X} = \texttt{x})}.$$

Applying this last equation to 1.2 yields:

$$g^*(x) = tuna \ \texttt{ if and only if } \ \frac{\texttt{P(Y = tuna, X = x)}}{\texttt{P(Y = salmon, X = x)}} \geq 1 \tag{1.3}$$

So, this is the second form of the equation which defines the Bayse classifier.
Finally let $\pi_{\texttt{tuna}}$ denote the probability of a tuna fish and $\pi_{\texttt{salmon}}$ denote the probability of a salmon. Hence,

$$\pi_{\texttt{tuna}} = P(Y = \texttt{tuna}) \ , \ \pi_{\texttt{salmon}} := \texttt{P(Y = salmon)}$$

The third way to rewrite the equations leading to the Bayse classifier, is obtained by using Bayse theorem and is:

$$g^*(x) = tuna \ \ \texttt{if and only if} \ \ \frac{\pi_{\texttt{tuna}} \cdot \texttt{P(X = x|Y = tuna)}}{\pi_{\texttt{salmon}} \cdot \texttt{P(X = x|Y = salmon)}} \geq 1 \tag{1.4}$$

In the present case, the best classifier is given by

$$g^*(1) = salmon, g^*(2) = tuna, g^*(3) = tuna$$

Our optimal decision rule can be represented in our table by the green entries:

	1	2	3
tuna	0.1	0.2	0.3
salmon	0.2	0.1	0.1

(1.5)

whilst the *misclassifiction probability* is given by the red entries:

$$\texttt{misclassification probability of } g^* \ = \ P(g^*(X) \neq Y) = 0.1 + 0.1 + 0.1 = 0.3.$$

This means that on the long run your robot will miscalssify 30% of the fish! And this is the best you can do, if you have no other information than the three sizes $\{1, 2, 3\}$. This number of 30% of course assumes the probabilities to be given in table 1.5 to be the correct probabilities.

Now, there is just one additional idea behind statistical classification: in general the probabilities given in table 1.5 are not exactly known. So, we need to catch some fish, label them manually as salmon or tuna and then estimate the probabilities given in the table 1.5. The fish we catch to figure our what a good classification rule is is called *training sample*.
let us give an example: Say we catch hundred fish which leads to the following frequency table:

	1	2	3
tuna	4	10	45
salmon	15	6	20

(1.6)

So, we have 4 fish which are tuna of size 1. This means 4% of our fish in the training sample are *tuna* of size 1. Hence, we estimate the probability for tuna of size 1 to be

$$\hat{P}(Y = tuna, X = 1) = 0.04$$

Similarly we caught 15 salmon of size 1. This represents 15% of our caught fish, which leads to our estimate:

$$\hat{P}(Y = salmon, X = 1) = 0.15$$

Based on this data, our decision rule is that for a fish of size 1, we classify it as a salmon because

$$\hat{P}(Y = tuna, X = 1) \leq \hat{P}(Y = salmon, X = 1).$$

Note that if we have a very large number of fish which we caught, then the estimated probabilities become indistinguishable close to the true probabilities. In that case our decision rule based on the annotated sample and the estimated probabilities is the same as the best rule that is the Bayesian classifier. So, we have the estimated probabilities given in the table

	1	2	3
`tuna`	0.04	0.10	0.45
`salmon`	0.15	0.06	0.2

(1.7)

where again green is four our decision rule and red represents the classification errors. The classification rule (classifier) which we chose is given by

$$g(1) = salmon, g(2) = tuna, g(3) = tuna$$

Is this the best possible classifier? The answer is we never know for absolutely sure, since we don't know the true probabilities but have only some estimates. However, if we have enough fish in our training sample, then the estimated probabilities will be very close to the true ones. In that case, decision rule based on the estimated probabilities will be the same as the one which would be based on the true probabilities. Hence, with enough data at hand, in the current example we are likely to get the Bayesian classifier.

Consider next an example of detecting counterfeit coins among old historical coins. Say you would investigate coins from the roman time. Back then, the coins where not as precisely minted as nowadays. So, there might be an even bigger fluctuation between weights and other parameters even for the official coins. Assume we are given a training sample of 10 coins. We let a specialist examine them. He will be able to recognize the counterfeit ones from the authentic ones. Then we want to determine a test for the collector to perform at home based on the bias of the coin. We assume that authentic coins tend to have other biases than counterfeit ones and we want to use this to propose a home-test for collectors. Hence, we throw each coin 1000 times and count the number of heads. Assume our training data looks as follows:

$$(y_1, x_1) = (1, 499), (y_2, x_2) = (1, 501), (y_3, x_3) = (1, 506), (y_4, x_4) = (1, 509), (y_5, x_5) = (1, 505)$$

which are the coins which are not counterfeit and the counterfeit ones

$$(Y_6, x_6) = (0, 480), (y_7, x_7) = (0, 505), (y_8, x_8) = (0, 515), (y_9, x_9) = (0, 520), (y_{10}, x_{10}) = (0, 520)$$

So, here $Y = 0$ stands for counterfeit coins and 1 is for authentic mint. These are the two classes. So, for example, the first coin in our training data is authentic, and after throwing it 1000 times we got 499 heads.
Now, recall that when we flip a coin independently n times and count the number of heads, we get a binomial variable with parameter n and p. So, let Z denote the number of heads throwing one specific coin 1000 times. The binomial distribution tells us the probability:

$$P(Z = z) = \binom{n}{p} p^z (1-p)^{n-z}$$

for all $z \leq n$. Here p again designates the probability to get a head when we throw the die once. For a fair coin we would have $p = 0.5$. We can now apply the Bayes approach for determining the best classifier. For this we first assume that the non-counterfeit coins have all a probability of head equal to p_1, whilst the counterfeit ones have their probability of head equal to p_0. To find the best possible classification rule we simple classify a coin as authentic if given the number of heads, the probability to be authentic is bigger than the probability to be counterfeit. In other words, the area $\mathcal{C}_1$, where the Bayes classifier classifies a coin as authentic is defined by the equation

$$\frac{\pi_1 P_1(Z = z|Y = 1)}{\pi_0 P_0(Z = z|Y = 0)} \geq 1$$

which is equivalent to:

$$\frac{\pi_1 \binom{n}{z} p_1^z (1-p_1)^{n-z}}{\pi_0 \binom{n}{z} p_0^z (1-p_0)^{n-z}} = \frac{\pi_1 p_1^z (1-p_1)^{n-z}}{\pi_0 p_0^z (1-p_0)^{n-z}} \geq 1$$

and hence taking the logarithm on both sides of the last inequality above whilst assuming $p_1 > p_0$, we find that the rule which does best at classifying the data classifies as authentic $(Y + 1)$ when:

$$z \geq \left(\ln(\frac{p_1(1-p_0)}{p_0(1-p_1)}) \right)^{-1} \cdot \left((\ln \pi_0 - ln\pi_1) + n \ln \left(\frac{1-p_0}{1-p_1} \right) \right). \tag{1.8}$$

There are now two approaches possible from this point on:

- **GENERATIVE APPROACH:** We can estimate the probabilities p_0 and p_1 and then plug the estimates into equation 1.8 to get an estimated classification boundary. We always denote estimates by putting a hat on the estimated symbol. With our current data we find:

  $$\hat{p}_1 = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{499 + 501 + 506 + 509 + 505}{5000} = 0.504$$

  and similarly

  $$\hat{p}_0 = \frac{x_6 + x_7 + x_8 + x_9 + x_{10}}{5000} = \frac{480 + 505 + 515 + 520 + 520}{5} = 0.508$$

  The estimated probabilities $\hat{\pi}_1$ and $\hat{\pi}_0$ are simply the relative frequencies of counterfeit and authentic in our training data:

  $$\hat{\pi}_1 = 0.5, \hat{\pi}_0 = 0.5$$

  We can now plug in our estimates into the formula for the classification boundary given in 1.8 to obtain the estimated classification boundary $\hat{z}_c$:

  $$\hat{z}_c = \left(\ln(\frac{\hat{p}_1(1-\hat{p}_0)}{\hat{p}_0(1-\hat{p}_1)}) \right)^{-1} \cdot \left((\ln \hat{\pi}_0 - ln\hat{\pi}_1) + n \ln \left(\frac{1-\hat{p}_0}{1-\hat{p}_1} \right) \right) = 506.0001$$

  This would then lead to the rule that when $z > 506.0001$ we classify as authentic.

- **DISSCRIMINATIVE APPROACH** We calculated and found that the best possible classification rule is of the type: $z \geq$ *constant* is classified as authentic. So,

$$\mathcal{C}_1 = \{z \geq \texttt{constant}\}$$

The constant is not known. Above we estimated it. Another approach is simply to look for which such rule which assigns class 1 when $z \geq$ *constant* does best on our data given at hand. In other words, instead of estimating the parameters of the model, we can try for several values of `constant` and look for which value the rule is best on the training data. Then hope is that for other coins which come from a similar sample, the classification rule would do similarly well. So, in our case, assign $Y = 1$ if $z \leq 510$ (or any constant between 509 and 515) this is the classification rule which does best on the sample data: It makes two mistakes out of 10, so we can estimate the classification error with this rule to be 20%. This will tend to not be un unbiased estimate.

WHICH APPROACH SHOULD WE USE NOW? Generative or discriminative? Reality is not all black or all white actually: the rules we use for discriminative approach are usually obtained in the first place from a probability model! So, often we mix: we calculate the best classifier for a given probability distribution which is known up to certain parameters. The best classification rule then also depends on these unknown parameters. In the generative approach, we would then estimate these parameters. In the discriminative approach, we consider the family of decision rules which depend on the parameter. Among, these we chose the one which is best at classifying the training data.

2 The two dimensional covariance matrix

For two random variable X and Y the covariance is defined by

$$COV(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Some properties are given below, where X, Y, X_1, X_2, Y_1, Y_2 are random variables whilst a, b, c, d are non-random constants.

- Covariance is symmetric

$$COV(X, Y) = COV(Y, X)$$

- We have linearity with respect to the first entry:

$$COV(aX_1 + bX_2, Y) = aCOV(X_1, Y) + bCOV(X_2, Y)$$

.

- We also have linearity with respect to the second entry

$$COV(X, aY_1 + bY_2) = aCOV(X, Y_1) + bCOV(, Y_2)$$

.

- The covariance of a variable with itself is the variance

$$COV(X, X) = VAR[X].$$

- Assume that X and Y are independent of each other. Then,

$$COV(X, Y) = 0$$

the reverse is not necessarily true, that is there exists variables with 0 covariance but which are not independent of each other. the reverse is true however when X and Y are jointly normal as we will see in the next section.

The proofs of these properties can be found in matzingers intro to probability lecture notes and have to be known for the next test. In multivariate statistics we consider random vectors. Let us start with a two dimensional random vector:

$$\vec{X} = (X, Y)$$

A typical example of such a two dimensional random vector would be the impact point of a shell in traditional artillery shooting. When we fire every time with the same ammunition and the gun oriented exactly the same way, the shells impact points will non-the-less not be exactly the same. This imprecision leads to the shell impact point being a "natural" random vector. Say, now that $\vec{X}_i = (X_i, Y_i)$ is the impact point of the i-th artillery shell on the ground. We assume that the impacts points are independent of each other and all have the same probability distribution. (The conditions do not change and we shoot with the same artillery gun pointed in exactly the same direction with the same type of ammunition. Weather conditions do not change). So, here we are we have an i.i.d. sequence of random vectors:

$$\vec{X}_1 = (X_1, Y_1), \vec{X}_2 = (X_2, Y_2), \vec{x} = (X_3, Y_3), \ldots$$

For the random vector $\vec{X} = (X, Y)$, we represent the covariances between the different entries of the vector in matrix format. This matrix is then called *covariance matrix of* $\vec{x}$ or simply *covariance of* $\vec{x}$. So, the covariance of $\vec{X}$ is given by:

$$COV[\vec{X}] = \begin{pmatrix} COV(X, X) & COV(X, Y) \\ COV(Y, X) & COV(Y, Y) \end{pmatrix}$$

What is the covariance matrix good for? let us see an example of how it is used. Say X is the value of one dollar of a first stock today in a year from now. Similarly, let Y denote

the value of a second stock in a year from now. Again, we take one dollar worth of stock today and look how much it is worth in a year. The single period portfolio investment problem is now defined as follows:
how do you invest a given amount of money into these two stocks so as to maximize the expected gain and minimize risk. More precisely, we put q_1 cents into the first stock and q_2 cents into the second stock. then at the end of the year we will have a value equal to

$$q_1X + q_2Y.$$

(We assume that during the year we are not allowed to trade this stock. So, we consider a *passive investment policy*). The value of the portfolio at the end of the year is thus $q_1X + q_2Y$ and is a random variable. At the beginning of the year, when we have to make our investment decision and determine q_1 and q_2, the value of the portfolio at the end of the year is of course not yet known.
The risk is represented by the variance:

$$VAR[q_1X+q_2Y] = COV[q_1X+q_2Y, q_1X+q_2Y] = q_1^2COV(X,X)+2q_1q_2COV(X,Y)+q_2^2COV(Y,Y).$$

The covariance above are supposed to be known to the investor, which could have determined them by estimation from previous years. The expected gain which we want to maximize is

$$E[q_1X + q_2Y] = q_1E[X] + q_2E[Y].$$

So, the optimal one period portfolio investment strategy is found by maximizing

$$q_1E[X] + q_2E[Y]$$

under the constrain

$$q_1^2COV(X,X) + 2q_1q_2COV(X,Y) + q_2^2COV(Y,Y) \leq constant_1$$

where the constant $constant_1 > 0$ depends on how much risk the investor is willing to bear. Also, the total amount of money is usually given so that another condition is

$$q_1 + q_2 = contant_2,$$

with the total amount of money to be invested denoted by$constant_2$ and known to us. Finally, we may not be allowed to borrow money, and hence we would have as additional constrain

$$q_1, q_2 \geq 0$$

.
The remarkable thing to realize, is that for solving this one period optimal portfolio investment problem, we do not need to know the exact distribution of $\vec{X}$: we just need the expectation and the covariance matrix! The same holds true when instead of investing only in two stocks we invest into several stocks.

2.0.1 Principal direction of a covariance matrix

When we shoot many times we see the there is much more dispersion (=fluctuation) in the direction of shooting then perpendicular to it. Again we assume that we are shooting with the same artillery gun, with the tube pointed in exactly the same direction and under the same conditions. So, with such a data set of impact points there is a direction in which the coordinates fluctuate maximally, this is usually the direction in which we are shooting. The direction perpendicular to this is the direction in which the impact points fluctuate least. With a plot of the impact points it is usually where simple to see approximately what these directions are. But how could we calculate them based on the covariance matrix of $\vec{X} = (X, Y)$? (Again $\vec{X}$ represents the impact point of a shell.) The answer is simple: the eigenvectors of the covariance matrix $cov(\vec{X})$ represent the direction of maximum resp. minimum spread of the artillery shells impact points. The reason is as follows:
to project on a line passing through the origin and the unit-vector

$$\vec{u} = (u_1, u_2)$$

we simply build the dot product. That is assume the vector (u_1, u_2) has length 1:

$$u_1^2 + u_2^2 = 1.$$

Then, the dot product

$$\vec{u} \cdot \vec{X} = u_1 X + u_2 Y$$

gives us the projection of the vector $\vec{X}$ onto the straight line $\vec{u} \cdot t$. So, to find the direction $\vec{u}$ of maximal dispersion (=fluctuation), we search for $\vec{u}$ which maximizes

$$VAR[\vec{u} \cdot \vec{X}] = VAR[u_1 X + u_2 Y] =$$
$$= COV[u_1 X + u_2 Y, u_1 X + u_2 Y] = u_1^2 COV(X, X) + 2u_1 u_2 COV(X, Y) + u_2^2 COV(Y, Y)$$

under the constrain

$$u_1^2 + u_2^2 = 1.$$

To solve this constrained optimization problem, we find the gradients and set them to point into the same direction. Hence,

$$\vec{grad}(VAR[\vec{u} \cdot \vec{X}]) = \begin{pmatrix} 2u_1 COV(X, X) + 2u_2 COV(X, Y) \\ 2u_1 COV(X, Y) + 2u_2 COV(Y, Y) \end{pmatrix} = 2 \begin{pmatrix} COV[X, X) & COV(X, Y) \\ COV(Y, X) & COV(Y, Y) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

should be colinear with

$$\vec{grad}(u_1^2 + u_2^2) = 2 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

So in other words we look for λ and a vector (u_1, u_2) so that

$$\lambda \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} COV(X, X) & COV(X, Y) \\ COV(X, Y) & COV(Y, Y) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

But the last equation above is the equation for a eigenvector with eigenvalue λ of the covariance matrix $COV[\vec{X}]$! We have just established that the direction of maximal dispersion is given as an eigenvector of the covariance matrix. The same thing holds true for the direction of minimum dispersion. Note that we can find such a direction of maximal dispersion and minimal dispersion for any covarianc matrix, there is no need for artillery shooting. Only that in artillery shooting these direction then have a simple physical interpretation: the direction of maximum dispersion is the direction in which we shoot. And the direction perpendicular to this is the direction of minimum dispersion. Again, the orthogonality of these two directions is not just given with artillery shooting: it holds true for any covariance matrices. The reason is simply that these directions are eigenvectors and for symmetric matrices, the eigenvectors corresponding to different eigenvalues are always perpendicular to each other. To see why consider the following: let A be a symmetric matrix and $\vec{v}_1$ and $\vec{v}_2$ to eigenvectors with corresponding eigenvalues λ_1 and λ_2. If we assume $\lambda_1 \neq \lambda_2$, then the eigenvectors are perpendicular. Indeed, consider the dot product

$$\lambda_1 \vec{v}_2 \cdot \vec{v}_1 = \vec{v}_2^T \lambda_1 \vec{v}_1 = \vec{v}_2^T A \vec{v}_1 = (A\vec{v}_2)^T \vec{v}_1 = \lambda_2 \vec{v}_2 \vec{v}_1 \tag{2.1}$$

If $\vec{v}_2$ and $\vec{v}_1$ would not be perpendicular to each other, then their dot product would not be 0. Hence, in the sequence of equations 2.1, we could divide on the very right and very left by $\vec{v}_2 \cdot \vec{v}_1$ leading to

$$\lambda_1 = \lambda_2$$

which is a contradiction since we assumed $\lambda_1 \neq \lambda_2$. So, if there are different eigenvalues then the corresponding eigenvectors must satisfy $\vec{v} \cdot \vec{v}_2 = 0$ and hence be orthogonal to each other. And this in terms leads to the direction of maximal dispersion and minimal dispersion to be orthogonal to each other. We also get that the covariance of the coordinate in these two directions is 0 as we will see.
Another important fact is that the eigenvalues λ_1 and λ_2 represent the variance of the impact points projected in each of the two eigenvalues directions. To see that this is indeed true, take $\vec{u}$ to be the eigenvector corresponding to the eigenvalue λ_1. We assume $\vec{u} = (u_1, u_2)$ to have unit length: $u_1^2 + u_2^2 = 1$. Then,

$$\begin{aligned}
&VAR[\vec{X} \cdot \vec{u}] = \\
&= COV(Xu_1 + Yu_2, Xu_1 + Yu_2) = u_1^2 COV(X,X) + 2u_1 u_2 COV(X,Y) + u_2^2 COV(Y,Y) = \\
&= (u_1, u_2) \begin{pmatrix} COV(X,X) & COV(X,Y) \\ COV(Y,X) & COV(Y,Y) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = (u_1, u_2) \lambda_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \lambda_1 (u_1^2 + u_2^2) = \lambda_1.
\end{aligned}$$

The same thing can be shown of course for λ_2 and the corresponding eigenvector.

2.1 Estimation of covariance matrix

Say again that we observe several artillery impact points:

$$\vec{X}_1 = (X_1, Y_1), \vec{X}_2 = (X_2, Y_2), \vec{X}_3 = (X_3, Y_3), \ldots, \vec{X}_n = (X_n, Y_n)$$

With this data set how do we estimate the covariance matrix given by:

$$COV[\vec{X}] = \begin{pmatrix} COV(X,X) & COV(X,Y) \\ COV(Y,X) & COV(Y,Y) \end{pmatrix}?$$

Note that

$$\begin{aligned} COV(X,X) &= E[X^2] - (E[X]^2), \\ COV(X,Y) &= E[XY] - E[X]E[Y], \\ COV(Y,Y) &= E[Y^2] - E[Y]^2 \end{aligned}$$

Hence, the covariance matrix can be written as

$$COV[\vec{X}] = \begin{pmatrix} E[X^2] & E[XY] \\ E[YX] & E[Y^2] \end{pmatrix} - \begin{pmatrix} E[X]^2 & E[X]E[Y] \\ E[Y]E[X] & E[Y]^2 \end{pmatrix} \tag{2.2}$$

The expression on the right side of the last equation above contains only expectations. Expectations are long term averages of the random variables when we repeat the experiment many times independently. (Law of large numbers: when you throw the same die many times independently and calculate the average, you get about the expected value. Provided you throw it many times) So, we are going to simply estimate all the expectations in the least expression above by taking the corresponding averages. the estimates of something is then denoted by putting a hat on that thing. So, we use the estimates:

$$\hat{E}[X] := \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\hat{E}[Y] := \frac{Y_1 + Y_2 + \ldots + Y_n}{n}$$

$$\hat{E}[X^2] := \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{n}$$

$$\hat{E}[Y^2] := \frac{Y_1^2 + Y_2^2 + \ldots + Y_n^2}{n}$$

$$\hat{E}[XY] := \frac{X_1Y_1 + X_2Y_2 + \ldots + X_nY_n}{n}$$

The estimate for the covariance matrix is now obtained by replacing in formula 2.2 the different expectations by their respective estimates. We find as estimate of the covariance matrix:

$$C\hat{O}V[\vec{X}] = \begin{pmatrix} \hat{E}[X^2] & \hat{E}[XY] \\ \hat{E}[YX] & \hat{E}[Y^2] \end{pmatrix} - \begin{pmatrix} \hat{E}[X]^2 & \hat{E}[X]\hat{E}[Y] \\ \hat{E}[Y]\hat{E}[X] & \hat{E}[Y]^2 \end{pmatrix}$$

and hence

$$C\hat{O}V[\vec{X}] = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n X_i^2 & \sum_{i=1}^n X_iY_i \\ \sum_{i=1}^n Y_iX_i & \sum_{i=1}^n Y_i^2 \end{pmatrix} - \begin{pmatrix} \bar{X}^2 & \bar{X} \cdot \bar{Y} \\ \bar{X} \cdot \bar{Y} & \bar{Y}^2 \end{pmatrix}$$

Where $\bar{X}$ and $\bar{Y}$ represent the sample means

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

and

$$\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y_n}{n}$$

2.2 Example

Again assume that we should with an artillery gun. let (X_i, Y_i) denote the impact point on the ground of the i-th shell we shoot. The canon tube points always in the same direction and we shoot under the same circumstances with the same ammunition. Hence,

$$\vec{X}_1 = (X_1, Y_1), \vec{X}_2 = (X_2, Y_2), \ldots, \vec{X}_n = (X_n, Y_n)$$

are ii.d random vectors. Assume that we get the following 10 impact points

X_i	Y_i
1.11	0.47
1.13	1.99
−3.84	−2.82
1.77	0.26
0.28	1.55
−1.46	−2.84
0.52	−0.28
−2.50	−1.70
−3.07	−5.32
0.75	1.84

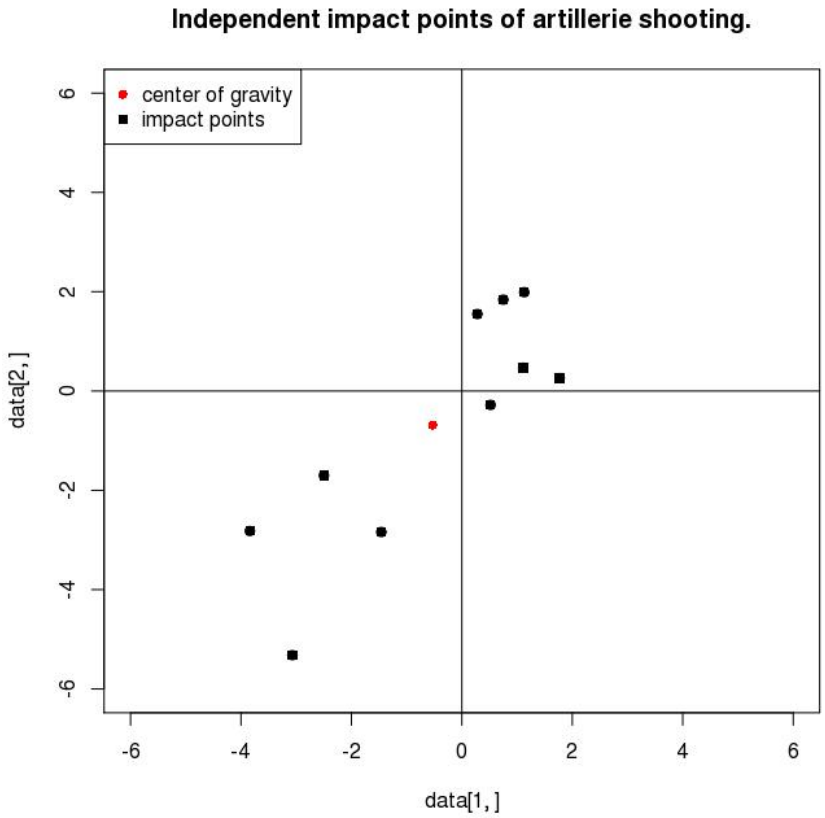
We can represent these impact points on a map. (That is we plot the above points in two dimensions). The result can be seen in figure 5. In that figure we see that the direction of maximal dispersion is approximately $(1, 1)$ and the direction of minimum dispersion is orthogonal given by $(-1, 1)$. Indeed, the covariance matrix we used to simulate this data is

$$COV(\vec{X}) = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \tag{2.3}$$

and one can verify that indeed for this covariance matrix the eigenvectors are $(1, 1)$ and $(-1, 1)$. The corresponding eigenvalues are 9 and 1. So, the standard deviation in the direction of the eigenvectors is 3 and 1 respectively. And indeed when we look at the impact points in figure 5 we see that the average fluctuation in the direction $(1, 1)$ is approximately 3, whilst in the direction of minimum fluctuation it is about 1.
The center of gravity of the impact points is represented by the red square in figure 5. Its value is given by

$$(\bar{X}, \bar{Y}) = (\frac{X_1 + \ldots + X_{10}}{10}, \frac{Y_1 + \ldots + Y_{10}}{10}) = (-0.531, -0.685)$$

Figure 1:



This is also an estimate for the expectation $E[\vec{X}] = (E[X], E[Y])$ so that the estimates value in the present case is

$$(\hat{E}[X], \hat{E}[Y]) = (-0.531, -0.685).$$

In reality we had used $E[X] = 0$ and $E[Y] = 0$ so the estimate is not that far from the true value. With more artillery impact points the precision would be greater. Finally we can estimate the covariance matrix. If X denotes the matrix with two column and 10 rows containing our impact points, then the covariance matrix estimate can be written as:

$$C\hat{O}V[\vec{X}] = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n X_i^2 & \sum_{i=1}^n X_i Y_i \\ \sum_{i=1}^n Y_i X_i & \sum_{i=1}^n Y_i^2 \end{pmatrix} - \begin{pmatrix} \bar{X}^2 & \bar{X}\cdot\bar{Y} \\ \bar{X}\cdot\bar{Y} & \bar{Y}^2 \end{pmatrix} =$$

$$\frac{1}{n}X^{\mathrm{T}}X - \begin{pmatrix} \bar{X}^2 & \bar{X}\bar{Y} \\ \bar{X}\bar{Y} & \bar{Y}^2 \end{pmatrix} =$$

$$\begin{pmatrix} 3.91 & 4.04 \\ 4.04 & 5.73 \end{pmatrix} - \begin{pmatrix} 0.28 & 0.36 \\ 0.36 & 5.2 \end{pmatrix} = \begin{pmatrix} 3.62 & 3.68 \\ 3.68 & 5.26 \end{pmatrix}$$

the last matrix above is our estimate for the covariance matrix given in 2.3. Note that the difference is not too big. Again with more data points, the precision would be better.

3 Precision of estimate of eigenvalues and eigenvectors of covariance matrix in the low dimensional case.

In the artillery shooting example, one might one to know how precisely the eigenvectors can be determined. Recall that the eigenvector with biggest eigenvalue represent the direction in which we are shooting. So, we can use the eigenvector with biggest eigenvalue to find a line on which the enemy artillery gun is located. Of course, one would then like to know the precision with which the eigenvector is determined, since if the precision is bad we might not have enough information about the location of the enemy artillery gun. So, here we assume a three dimensional situation, where the shells explode in the air. So, the explosion point of the i-th shell is denoted by $\vec{X}_i = (X_i, Y_i, Z_i)$. We assume that the same artillery gun shoots many round under the same conditions. Hence, we have a sequence of i.i.d. vectors

$$\vec{X}, \vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$$

corresponding to the different explosion points in the air. (Here $\vec{X} = (X, Y, Z)$ so this is one impact point without index, to later on simplify notations). We will also assume that

$$E[\vec{X}] = E[(X, Y, Z)] = E[\vec{X}_i] = (E[X_i], E[Y_i], E[Z_i]) = (0, 0, 0).$$

We will explain later why in many applications, this assumption is realistic. Furthermore, we assume that the direction in which we are shooting is given by $(1, 0, 0)$. We also assume that X_i, Y_i and Z_i are independent of each other. Hence, the covariance matrix is given by

$$COV[\vec{X}] = \begin{pmatrix} \sigma_X^2 & 0 & 0 \\ 0 & \sigma_Y & 0 \\ 0 & 0 & \sigma_Z^2 \end{pmatrix} = \begin{pmatrix} E[X^2] & E[XY] & E[XZ] \\ E[YX] & E[Y^2] & E[YZ] \\ E[ZX] & E[ZY] & E[Z^2] \end{pmatrix}$$

Now recall the Central Limit Theorem: say we have variables $W_1, W_2, \ldots$ which are i.i.d then we have that for n large enough, the properly re-scaled sum is approximately standard normal:

$$\frac{W_1 + W_2 + \ldots + W_n - nE[W_1]}{\sqrt{n}} \approx \mathcal{N}(0, 1).$$

the goal is to figure out how precise our estimates for the eigenvalues and eigenvectors are. Since the expectation is 0, in our estimate of the covariance matrix we can leave the part which estimates the expectation out. So, we use the following estimate for the covariance matrix:

$$C\hat{O}V[\vec{X}] = \begin{pmatrix} \frac{X_1^2 + \ldots + X_n^2}{n} & \frac{X_1Y_1 + \ldots + X_nY_n}{n} & \frac{X_1Z_1 + \ldots + X_nZ_n}{n} \\ \frac{Y_1X_1 + \ldots + Y_nX_n}{n} & \frac{Y_1^2 + \ldots + Y_n^2}{n} & \frac{Y_1Z_1 + \ldots + Y_nZ_n}{n} \\ \frac{Z_1X_1 + \ldots + Z_nX_n}{n} & \frac{Z_1Y_1 + \ldots + Z_nY_n}{n} & \frac{Z_1^2 + \ldots + Z_n^2}{n} \end{pmatrix}$$

We can now apply the Central Limit Theorem to all the entries of the estimated covariance matrix above. For example take W_i to be equal to $W_i = X_iY_i$. Then,

$$\frac{X_1Y_1 + \ldots + X_nY_n}{n} - E[X_1Y_1] = \frac{1}{\sqrt{n}} \frac{W_1 + W_2 + \ldots + W_n - E[W_1]}{\sqrt{n}} \approx \sigma_{W_1} \frac{\mathcal{N}(0,1)}{\sqrt{n}} = \frac{\sigma_{X_1}\sigma_{Y_1}}{\sqrt{n}} \mathcal{N}(0,1) \tag{3.1}$$

So, take the difference between the estimated covariance matrix and the real one:

$$E = C\hat{O}V[\vec{X}] - COV[\vec{X}] =$$
$$= \begin{pmatrix} \frac{X_1^2+\ldots+X_n^2}{n} - E[X^2] & \frac{X_1Y_1+\ldots+X_nY_n}{n} - E[XY] & \frac{X_1Z_1+\ldots+X_nZ_n}{n} - E[XZ] \\ \frac{Y_1X_1+\ldots+Y_nX_n}{n} - E[YX] & \frac{Y_1^2+\ldots+Y_n^2}{n} - E[Y^2] & \frac{Y_1Z_1+\ldots+Y_nZ_n}{n} - E[YZ] \\ \frac{Z_1X_1+\ldots+Z_nX_n}{n} - E[ZX] & \frac{Z_1Y_1+\ldots+Z_nY_n}{n} - E[ZY] & \frac{Z_1^2+\ldots+Z_n^2}{n} - E[Z^2] \end{pmatrix}$$

With the Central Limit Theorem applied to each of the entries of the last matrix above in the same way as in 3.1, we find

$$C\hat{O}V[\vec{X}] - COV[\vec{X}] \approx \frac{1}{\sqrt{n}} \begin{pmatrix} \sigma_{X^2}\mathcal{N}_{11} & \sigma_X\sigma_Y\mathcal{N}_{12} & \sigma_X\sigma_Z\mathcal{N}_{13} \\ \sigma_Y\sigma_X\mathcal{N}_{21} & \sigma_{Y^2}\mathcal{N}_{22} & \sigma_Y\sigma_Z\mathcal{N}_{23} \\ \sigma_Z\sigma_X\mathcal{N}_{31} & \sigma_Z\sigma_Y\mathcal{N}_{32} & \sigma_{Z^2}\mathcal{N}_{33} \end{pmatrix} \tag{3.2}$$

where $\mathcal{N}_{ij}$ are all standard normal variables and $\mathcal{N}_{ij} = \mathcal{N}_{ji}$ for all $i, j = 1, 2, 3$. Furthermore, the $\mathcal{N}_{ij}$ which are different from each other, are automatically independent of each other. This follows from the following argument:
by the multidimensional Central Limit theorem the "vector-matrix'

$$\begin{pmatrix} \mathcal{N}_{11} & \mathcal{N}_{12} & \mathcal{N}_{13} \\ \mathcal{N}_{21} & \mathcal{N}_{22} & \mathcal{N}_{23} \\ \mathcal{N}_{31} & \mathcal{N}_{32} & \mathcal{N}_{33} \end{pmatrix}$$

is multivariate normal. This means any entries with covariance being 0 are also automatically independent. But for example

$$COV(XY, XZ) = E[XYXZ] - E[XY] \cdot E[XZ] = E[X^2]E[Y]E[Z] - E[X]E[Y]E[X]E[Z] = 0$$

Hence,

$$COV\left(\frac{X_1Y_1 + \ldots + X_nY_n}{n}, \frac{X_1Z_1 + \ldots + X_nZ_n}{n}\right) =$$
$$= \frac{1}{n^2} \sum_{i,j} COV(X_iY_i, X_jZ_j) = \frac{1}{n^2} \sum_i COV(X_iY_i, X_iZ_i) = 0$$

Which implies that $\mathcal{N}_{12}$ and $\mathcal{N}_{13}$ are uncorrelated and hence independent of each other asymptotically since they are jointly normal.
Next we are going to establish the formula for the estimated eigenvalue and eigenvectors of the covariance matrix. Again, here the estimated eigenvalues and eigenvectors are simply

the eigenvectors and eigenvalues of the estimated covariance matrix. We assume here σ_X, σ_Y and σ_z to all have different values from each other. Let A denote the covariance matrix, E again the error-matrix (that is the difference between the estimated and the true covariance matrix). Let $\vec{\mu} = (1, 0, 0)^T$ be the eigenvector with the biggest eigenvalue of $A = COV[\vec{X}]$. Let $\lambda = \sigma_X^2$ denote the biggest eigenvalue of the covariance matrix A and let $\lambda + \Delta\lambda$ denote the biggest eigenvalue of the estimated covariance matrix.
So the estimated covariance matrix is $A + E$, hence the true covariance matrix plus a "perturbation" E. Let $\vec{v} = \vec{\mu} + \Delta\vec{\mu}$ be the eignevector with biggest eigenvalue for the estimated covariance matrix and assume that $\Delta\vec{\mu}$ is orthogonal to μ. Hence $\Delta\vec{\mu} = (0, \Delta\mu_Y, \Delta\mu_X)^T$. With these notations, we have:

$$(A + E)(\vec{\mu} + \Delta\vec{\mu}) = (\lambda + \Delta\lambda)(\vec{\mu} + \Delta\vec{\mu}). \tag{3.3}$$

Also, since $\vec{\mu}$ is an eigenvector of A, we have:

$$A\vec{\mu} = \lambda\vec{\mu} \tag{3.4}$$

Subtracting equation 4.1 from 3.4, we find:

$$(A - I\lambda)\Delta\vec{\mu} = -E\vec{\mu} + \Delta\lambda\vec{\mu} + -E\Delta\vec{\mu} + \Delta\lambda\Delta\vec{\mu}. \tag{3.5}$$

In the last equation above the two terms:

$$-E\Delta\vec{\mu} + \Delta\lambda\Delta\vec{\mu}.$$

is asymptotically smaller than the other terms, since these terms are of order constant over n, whilst the other terms are of order constant over $\sqrt{n}$.
So, we find that

$$(A - I\lambda)\Delta\vec{\mu} \approx -E\vec{\mu} + \Delta\lambda\vec{\mu}$$

and hence

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_Y^2 - \sigma_X^2 & 0 \\ 0 & 0 & \sigma_Z^2 - \sigma_X^2 \end{pmatrix} \begin{pmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} \approx -\frac{1}{\sqrt{n}} \begin{pmatrix} \sigma_{X^2}\mathcal{N}_{11} \\ \sigma_X\sigma_Y\mathcal{N}_{21} \\ \sigma_X\sigma_Z\mathcal{N}_{31} \end{pmatrix} + \begin{pmatrix} \Delta\lambda \\ 0 \\ 0 \end{pmatrix} \tag{3.6}$$

Which implies that

$$\Delta\mu_Y \approx -\frac{\mathcal{N}_{21}}{\sqrt{n}} \frac{\sigma_Y\sigma_X}{(\sigma_Y^2 - \sigma_X^2)}$$

and

$$\Delta\mu_Z \approx -\frac{\mathcal{N}_{21}}{\sqrt{n}} \frac{\sigma_Y\sigma_X}{(\sigma_Z^2 - \sigma_X^2)}$$

and finally

$$\Delta\lambda \approx \sigma_{X^2} \frac{\mathcal{N}_{11}}{\sqrt{n}}$$

We obtain similar estimates for the other eigenvector/eigenvalue pairs.

4 Principal components of covariance matrix and factor analysis

So, far we have seen examples of artillery shooting. Let us consider next the case of a portfolio. Say $\vec{X} = (X, Y, Z)^T$ contains the information about three stocks. Say X is the change in value of the first stock from today to tomorrow. Similarly, Y is the change in value of the second stock and Z is the change in value of the third stock. We assume again $E[\vec{X}] = (E[X], E[Y], E[Z])^T = (0, 0, 0)^T$ this is realistic since the daily change has an expectation part which is negligible compared to the standard deviation. Let us assume that $\vec{X}_i = (X_i, Y_i, Z_i)^T$ is the change from day i to day $i + 1$. We assume that the changes are i.i.d. and hence $\vec{X}, \vec{X}_1, \vec{X}_2, \ldots$ are supposed to be i.i.d. Now assume that there are two sectors of the industry to which these stocks belong. For example, high-tech and energy. Let S be an index for the first sector and T be an index for the second. Often times one models stocks as a linear combination on indexes. So we would assume that we have

$$X = a_X S + b_X T + \epsilon_X$$
$$Y = a_Y S + b_Y T + \epsilon_Y$$
$$Z = a_Z S + b_Z T + \epsilon_Z$$

Here a_X, a_Y, a_z and b_X, b_y, b_z are supposed to be non-random coefficients. Furthermore we assume that

$$\epsilon_X, \epsilon_Y, \epsilon_Z$$

are independent of each other and of S and T and have 0 expectation. To simplify our discussion at first we assume that they also have same standard deviation:

$$\sigma^2 = \sigma^2_{\epsilon_X} = \sigma^2_{\epsilon_Y} = \sigma^2_{\epsilon_Z}.$$

The covariance between X and Y is then given:

$$COV(X, Y) = COV(a_X S + b_X T + \epsilon_X, a_Y S + b_Y T + \epsilon_Y) =$$
$$= a_x a_y COV(S, S) + a_X b_Y COV(S, T) + b_X a_Y COV(T, S) + b_X b_Y COV(T, T)$$

Let us first assume that S and T are independent of each other so that $COV(S, T) = COV(T, S) = 0$. we also assume S and T standardized so that $COV(S, S) = VAR[S] = 1$ and $COV(T, T) = 1$. Assume also at first that X and Y depend only on S and that Z depends only on T. This means $a_z = 0$ and $b_X = b_Y = 0$. In that case:

$$X = a_X S + \epsilon_X, Y = a_Y S + \epsilon_Y, Z = b_Z T + \epsilon_Z$$

The covariance matrix in that case is then given by:

$$COV[\vec{X}] = \begin{pmatrix} COV(X,X) & COV(X,Y) & COV(X,Z) \\ COV(Y,X) & COV(Y,Y) & COV(Y,Z) \\ COV(Z,X) & COV(Z,Y) & COV(Z,Z) \end{pmatrix} =$$

$$= \begin{pmatrix} COV(a_X S, a_X S) & COV(a_X S, a_Y S) & COV(a_1 S, b_Z T) \\ COV(a_Y S, a_X X) & COV(a_Y S, a_y S) & COV(a_Y S, b_Z T) \\ COV(b_Z T, a_X S) & COV(b_Z T, a_Y Y) & COV(b_Z T, b_Z T) \end{pmatrix} + \begin{pmatrix} COV(\epsilon_X, \epsilon_X) & 0 & 0 \\ 0 & COV(\epsilon_Y, \epsilon_Y) & 0 \\ 0 & 0 & COV(\epsilon_Z, \epsilon_Z) \end{pmatrix}$$

$$= \begin{pmatrix} a_x a_x & a_x a_y & 0 \\ a_y a_x & a_y a_y & 0 \\ 0 & 0 & b_Z b_Z \end{pmatrix} + \begin{pmatrix} \sigma_X^2 & 0 & 0 \\ 0 & \sigma_Y^2 & 0 \\ 0 & 0 & \sigma_Z^2 \end{pmatrix}$$

$$= \begin{pmatrix} a_x a_x & a_x a_y & 0 \\ a_y a_x & a_y a_y & 0 \\ 0 & 0 & b_Z b_Z \end{pmatrix} + \sigma^2 I$$

where I stands for the 3×3 identity matrix. Now here we have that the matrix

$$\begin{pmatrix} a_x a_x & a_x a_y & 0 \\ a_y a_x & a_y a_y & 0 \\ 0 & 0 & b_Z b_Z \end{pmatrix}$$

has two eigenvectors with non-zero eigenvalues. These are the vector

$$\vec{u}_1 = (a_X, a_Y, 0)^T$$

with eigenvalue $\lambda_1 = a_X^2 + a_Y^2$ and another eigenvector

$$\vec{u}_2 = (0, 0, 1)$$

with eigenvalue $\lambda_2 = b_Z^2$. Adding $\sigma^2 I$ does not change the eigenvectors. it merely increases the eigenvalues by σ^2. So, the eigenvectors $\vec{u}_1$ and $\vec{u}_2$ are also the eigenvectors of the covariance matrix! Now, we can use these eigenvectors to try to group stocks "belonging to the same sector". For this simply take the non-zero entries of $\vec{u}_1$: the first and second entry are non-zero so we can assume that X and Y belong to a same industry sector. Then, the eigenvector $\vec{u}_2$ has only its third entry non-zero. So, we decide that it represents by itself a sector. Now, why would this be useful? Answer: it is mainly useful when we do not know how to group things. For example, you work with futures and you want to see if there are groups which tend to go together. Now, here in the current example, our covariance matrix has a simple block structure. So in principal we can just look at the correlation matrix and group the stocks which are highly correlated into groups. No, need for eigenvectors there. So, when are eigenvectors most useful? Answer: there could be several such indexes and they need not correspond to sectors but can be overlapping. Think for example of the introducing an additional variable measuring the general state of the economy. Call it M. Now, say that we have $2p$ stocks. The first p depend only on S and M so that for $i = 1, 2, \ldots, p$ we have

$$X_i = a_i S + c_i M + \epsilon_i$$

Then the stocks with indices from $p+1$ to $2p$ belong to a second sector and will depend only on T and M, So that

$$X_i = b_i T + c_i M + \epsilon_i$$

when $i = p+1, \ldots, 2p$. The covariance matrix we find in that case is given by:

$$COV[\vec{X}] =$$
$$\begin{pmatrix} a_1a_1 & a_1a_2 & a_1a_3 & \ldots & a_1a_p & 0 & 0 & 0 & \ldots & 0 \\ a_2a_1 & a_2a_2 & a_2a_3 & \ldots & a_2a_p & 0 & 0 & 0 & \ldots & 0 \\ a_3a_1 & a_3a_2 & a_3a_3 & \ldots & a_3a_p & 0 & 0 & 0 & \ldots & 0 \\ \ldots & & & & & & & & & \\ \ldots & & & & & & & & & \\ a_pa_1 & a_pa_2 & a_pa_3 & \ldots & a_pa_p & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & b_1b_1 & b_1b_2 & b_1b_3 & \ldots & b_1b_p \\ 0 & 0 & 0 & 0 & 0 & b_2b_1 & b_2b_2 & b_2b_3 & \ldots & b_2b_p \\ 0 & 0 & 0 & 0 & 0 & b_3b_1 & b_3b_2 & b_3b_3 & \ldots & b_3b_p \\ \ldots & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & b_pb_1 & b_pb_2 & b_pb_3 & \ldots & b_pb_p \end{pmatrix} +$$

$$+ \begin{pmatrix} c_1c_1 & c_1c_2 & c_1c_3 & \ldots & c_1c_p & c_1c_{p+1} & c_1c_{p+2} & c_1c_{p+3} & \ldots & c_1c_{2p} \\ c_2c_1 & c_2c_2 & c_2c_3 & \ldots & c_2c_p & c_2c_{p+1} & c_2c_{p+2} & c_2c_{p+3} & \ldots & c_2c_{2p} \\ c_3c_1 & c_3c_2 & c_3c_3 & \ldots & c_3c_p & c_3c_{p+1} & c_3c_{p+2} & c_3c_{p+3} & \ldots & c_3c_{2p} \\ \ldots & & & & & & & & & \\ \ldots & & & & & & & & & \\ \ldots & & & & & & & & & \\ \ldots & & & & & & & & & \\ \ldots & & & & & & & & & \\ c_{2p}c_1 & c_{2p}c_2 & c_{2p}c_3 & \ldots & c_{2p}c_p & c_{2p}c_{p+1} & c_{2p}c_{p+2} & c_{2p}c_{p+3} & \ldots & c_{2p}c_{2p} \end{pmatrix} + \sigma^2 I =$$
$$= \vec{a} \cdot \vec{a}^T + \vec{b} \cdot \vec{b}^T + \vec{c} \cdot \vec{c}^T + \sigma^2 I$$

where we have

$$\vec{a} = (a_1, a_2, \ldots, a_p, 0, \ldots, 0)^T, \vec{b} = (0, 0, \ldots, 0, b_1, b_2, \ldots, b_p)^T, \vec{c} = (c_1, c_2, \ldots, \ldots, c_{2p})^T$$

Now, if there would not be the term $\vec{c} \cdot \vec{c}^T$ in the last formula above for the covariance matrix, then, the two eigenvectors with biggest eigenvalues would be $\vec{a}$ and $\vec{b}$. Their eigenvalues would be $\vec{a}^t\vec{a} + \sigma^2$ and $\vec{b}^t\vec{b} + \sigma^2$. All other eigenvectors would have eigenvalues σ^2 which is much smaller. So, the way to find which stocks are in the same industry sector would be to find the two eigenvectors with biggest eigenvalues. The non-zero entries of such an eigenvector shows which stocks belong to the same sector. But now instead, we have added the term $\vec{c}\vec{c}^t$ to the covariance matrix. Then $\vec{a}$ and $\vec{b}$ are no longer eigenvectors. However take the three eigenvectors with biggest eigevalues: $\vec{u}_1, \vec{u}_2, \vec{u}_3$. These vectors are then linear combinations of $\vec{a}$, $\vec{b}$ and $\vec{c}$. This follows from the fact that the matrix $COV[\vec{X}] - \sigma^2 I$ has all its columns being linear combination of those three vectors. Also, we know that an eigenvector must always be in the linear span of the matrix. So, in other words, we have three equations:

$$\vec{a} = r_{11}\vec{u}_1 + r_{12}\vec{u}_2 + r_{13}\vec{u}_3$$
$$\vec{b} = r_{21}\vec{u}_1 + r_{22}\vec{u}_2 + r_{23}\vec{u}_3$$
$$\vec{c} = r_{31}\vec{u}_1 + r_{32}\vec{u}_2 + r_{33}\vec{u}_3$$

The coefficients r_{ij} are not known when we analyse a covariance matrix. Nor will the vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$ be known apriori in general. The one thing known, are the eigenvectors which here are denoted by $\vec{u}_1$, $\vec{u}_2$ and $\vec{u}_3$. The eigenvectors of the covariance matrix are called *principal components*. So, given the principal components, we try to " find back vectors like $\vec{a}$, $\vec{b}$ and $\vec{c}$", which will show us which stocks belong to the same sector. This operation is called *factor analysis*: This consist of finding coefficients r_{ij} for which the resulting $\vec{a}$ and $\vec{b}$ have many close to 0 entries and $\vec{a}$ and $\vec{b}$ tend to be orthogonal to

each other. Unfortunately enough, there is usually not a unique solution. Also, we have to define what criteria we use.

Then, applying this kind of calculation to every entry, we find that the covariance matrix is given by

$$\begin{aligned}
COV[\vec{X}] = & \begin{pmatrix} COV(X,X) & COV(X,Y) & COV(X,Z) \\ COV(Y,X) & COV(Y,Y) & COV(Y,Z) \\ COV(Z,X) & COV(Z,Y) & COV(Z,Z) \end{pmatrix} = \\
= & \ cov_{(S,S)} \begin{pmatrix} a_x^2 & a_x a_y & a_x a_z \\ a_y a_x & a_y^2 & a_y a_z \\ a_z a_x & a_z a_y & a_z a_z \end{pmatrix} + cov_{(S,T)} \begin{pmatrix} a_x b_x & a_x b_Y & a_x b_z \\ a_y b_x & a_y b_Y & a_y b_z \\ a_z b_x & a_z b_y & a_z b_z \end{pmatrix} + \\
& + cov_{(S,T)} \begin{pmatrix} b_x a_x & b_x a_y & b_x a_z \\ b_y a_x & b_y a_y & b_y a_z \\ b_z a_x & b_z a_y & b_z a_z \end{pmatrix} + cov_{(T,T)} \begin{pmatrix} b_x b_x & b_x b_y & b_x b_z \\ b_y b_x & b_y b_y & b_y b_z \\ b_z b_x & b_z b_y & b_z b_z \end{pmatrix} + \sigma^2 I = \\
= & \ COV(S,S)\vec{a}\cdot\vec{a}^T + COV(S,T)\vec{a}\cdot\vec{b}^T + COV(T,S)\vec{b}\cdot\vec{a}^T + COV(T,T)\vec{b}\cdot\vec{b}^T + \sigma^2 I
\end{aligned}$$

where $\vec{a} = (a_X, a_Y, a_Z)^T$ and $\vec{b} = (b_X, b_Y, b_Z)^T$ whilst I represents the identity matrix. Now not that the image space of our covariance matrix without the term $\sigma^2 I$ is two dimensional and is spanned by $\vec{a}^T$ and $\vec{b}^T$. So, without the term $\sigma^2 I$ there are only two non-zero eigenvalues and their corresponding eigenvectors span the linear space generated by $\vec{a}$ and $\vec{b}$. Imagine now a similar situation but with p stock instead of only 3. Again, we assume that all the stocks depend on two indeces S and T through some non-random linear coefficients as before. So, again $\vec{X}, \vec{X}_1, \vec{X}_2, \ldots$ are i.i.d vectors but this time the dimension be p:

$$\vec{X} = (X_1, X_2, \ldots, X_{100})^T$$

and

$$\vec{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip}$$

and

$$X_j = a_j S + b_j T + \epsilon_j$$

where the coefficients a_j and b_j are non random and

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_p$$

have expectation 0 are uncorrelated and are uncorrelated with S and T. The covariance of X_i with X_l is given by

$$\begin{aligned}
COV(X_i, X_l) = & \\
& = Cov(a_i S + b_i T + \epsilon_i, a_l S + b_i T = \epsilon_l) = \\
& = a_i a_l COV(S,S) + a_i b_l COV(S,T) + b_i a_l COV(T,S) + b_i b_l COV(T,T)
\end{aligned}$$

for $i \neq l$. hence the covariance matrix is given by the same formula as before

$$\begin{aligned}
& COV[\vec{X}] = \\
& = COV(S,S)\cdot\vec{a}\cdot\vec{a}^T + COV(S,T)\cdot\vec{a}\cdot\vec{b}^T + COV(T,S)\cdot\vec{b}\cdot veca^t + COV(T,T)\cdot\vec{b}\vec{b} + \sigma^2 I
\end{aligned}$$

So, despite the space having a high dimension p, the covariance matrix when we subtract $\sigma^2 I$ has only rank two. This means only two non-zero eigenvalue for the matrix $COV[\vec{X}] - \sigma^2 I$. Adding the identity matrix time σ^2 moves all eigenvalues up by σ^2. Hence, the covariance matrix $COV[\vec{X}]$ will have two "big "eigenvalues and all others will be equal to σ^2. To understand that the covariance matrix is only of dimension 2 means that there is a very simple structure. When we calculate the spectrum of the matrix, we see that there are two eigenvalues which are separated from the rest and can hence conclude that there is " a simple structure" behind the covariance matrix.

Now, in reality we will never now the exact covariance matrix, but only the covariance matrix up to an "estimation-error-matrix" E. So, the question is if instead of the matrix $COV[\vec{X}]$ we are given

$$COV[\vec{X}] + E$$

where $E = C\hat{O}V[\vec{X}] - COV[\vec{X}]$ will we still be able to recognize that there is a low dimensional structure behind the covariance matrix. Of course, this will depend on "how big" the perturbation-matrix E is. At this stage we have to introduce perturbation results for symmetric matrices. This is the content of the next subsection.

4.1 Perturbation results for symmetric matrices

We are going to look at how much a perturbation can modify the eigenvalues and eigenvectors of a symmetric matrix. Of course, the change will depend on how big the "perturbation" is. Our measure of choice will be the spectral norm for symmetric matrices. For a symmetric matrix A, we will denote the spectral norm of A by $|A|$. (The spectral norm can also be defined for non-symmetric matrices, but here we use only symmetric). The spectral norm is given as the biggest absolute value of an eigenvalue.
Let us see an example:

$$A = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

here the biggest eigenvalue is 4. Hence, the spectral norm is 4:

$$|A| = 4$$

Now, let us add a "small perturbation". For this take E to be equal to

$$E = \begin{pmatrix} 0.02 & 0 & 0 \\ 0 & 0.03 & 0 \\ 0 & 0 & 0.01 \end{pmatrix}$$

The resulting "perturbated matrix" is

$$A + E = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} + \begin{pmatrix} 0.02 & 0 & 0 \\ 0 & 0.03 & 0 \\ 0 & 0 & 0.01 \end{pmatrix} = \begin{pmatrix} 4.02 & 0 & 0 \\ 0 & 3.03 & 0 \\ 0 & 0 & 2.01 \end{pmatrix}$$

So, here the eigenvalues of A are $\lambda_1 = 4$, $\lambda_2 = 3$ and $\lambda_3 = 2$. The eigenvalues of the perturbation matrix $A + E$ are $\lambda_1^* = 4.02$, $\lambda_2^* = 3.03$ and $\lambda_3^* = 2.01$ We see that non of the eigenvalues got changed by more than 0.03. The reason is that the biggest eigenvalue of E is 0.03, so that $|E| = 0.03$. In other words we have

$$|\lambda_i - \lambda_i^*| \leq |E|,$$

for any $i = 1, 2, 3$. In the case, that both matrices A and E are diagonal, the last inequality above is easy to understand. But it also holds, when the matrices A and E are not diagonal but just symmetric as we show below in Theorem 4.1.
Now let us consider the problem of finding the vector $(u_1, u_2, u_3)^T$ of Euclidean norm 1 so that $A\vec{u}$ has maximal Euclidean norm. We will denote by $|\vec{x}|$ the Euclidean norm of a vector $\vec{x}$. so, in the case of the current matrix A we find that

$$\max |A\vec{u}| = \sqrt{4^2u_1^2 + 3^2u_2^2 + 2^2u_3^2}$$

under the constrain $\mu_1^2 + \mu_2^2 + \mu_3^2 = 1$ In the current case, of a diagonal matrix A, it is easy to see that the constraint maximum is given by the biggest eigenvalue 4. This is true in general, not just for diagonal

matrices. It can be shown to hold true in general by using Lagrange multipliers. So, we have in general for any symmetric $n \times n$-matrix A that the spectral norm $|A|$ is equal to the maximum:

$$|A| = \max_{\vec{\mu} \in S^{n-1}} |A\vec{u}|$$

where S^{n-1} denotes the surface of the unit ball centered at the origin in $\mathbb{R}^n$.
A third equivalent way to characterize the spectral norm of a symmetric matrix is the maximum

$$|A| := \max_{\vec{u},\vec{v} \in S^{n-1}} \vec{v} A \vec{u}$$

where the maximum is taken over all pairs of vectors $\vec{u}$ and $\vec{v}$ on S^{n-1}. Another formula for the spectral norm of a diagonal matrix is

$$|A| = \max_{\vec{\mu} \in S^{n-1}} |\vec{\mu} A \vec{\mu}|.$$

Again, the result are simple: they say that if the spectral norm of E is less than $\epsilon > 0$, then the eigenvalues move by less than ϵ. This is the statement of the next theorem:

Theorem 4.1 *Let A and E be two symmetric $n \times n$ matrices and assume that $|E| \leq \epsilon$. Assume that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of A. Let $\lambda_1^* \geq \lambda_2^* \geq \ldots \lambda_n^*$ be the eigenvalues of $A + E$. Then,*

$$|\lambda_i - \lambda_i^*| \leq \epsilon = \texttt{spectral norm of perturbation}.$$

Proof. We first do the proof for the biggest eigenvalue λ_1 of A. Let λ_1^* be the biggest eigenvalue of $A + E$. By our characterization of the spectral norm, we have

$$|(A+E)\vec{u}| = |A\vec{u} + E\vec{u}| \leq |A\vec{u}| + |E\vec{u}| \leq |A| \cdot |\vec{u}| + |E| \cdot |\vec{u}| \tag{4.1}$$

Let $B^n(1)$ denote the unit ball in $\mathcal{R}^n$. We can then apply inequality 4.1 to our characterization of spectral norm:

$$|A+E| = \max_{\vec{u} \in B^n(1)} |(A+E)\vec{u}| \leq \max_{\vec{u} \in B^n(1)} ||A| \cdot |\vec{u}| + |E| \cdot |\vec{u}| \leq |A| + |E|$$

hence, the biggest eigenvalue λ_1^* of $A + E$ must be less than $|A| + |E|$. But $|A|$ is equal to the biggest eigenvalue λ_1 of A. Hence,

$$\lambda_1^* - \lambda_1 \leq |E|$$

Similarly we can prove the converse, that is that $\lambda_1 - \lambda_1^*$ is less or equal to $|E|$. This finishes proving that

$$|\lambda_1 - \lambda_1^*| \leq |E|.$$

■

In factor analysis we saw how important eigenvectors of covariance matrix are: they give us "indexes which allow us to understand the data better". Now, we usually don't know the exact value of the covariance matrix, but only have an estimate. Then, typically we take the eigenvectors of the estimated covariance matrix as our estimate of the eigenvectors. In low dimension there might not be a big difference between the estimated covariance matrix and the true one. But, in high dimension, this difference can entrirely mess up things. So, we want to find a way, to bound the difference between the estimated and the true eigenvectors. Among others, this should tell us how much samples of a random vector we need to be able to estimate the covariance matrix sufficiently precisely for whatever perpupse we have in mind. Next, we present a very simple formula, which is the name of the game. The error for the estimated eigenvector is less than twice the fraction of spectral norm of perturbation divided by spectral gap. This main result is given in theorem 4.2. It is the corner-stone for getting a theoretical formula for bounding the error in the estimated eigenvectors in the high-dimensional case. (In the lower-dimensional case, we had more explicit stuff, where we did not just bound the error but where able to describe explicitly the asymptotic distribution of the coordinates of the error vector in section 3).

Theorem 4.2 *Let A and E to symmetric $n \times n$ matrices and assume that $|E| \leq \epsilon$. Assume that all eigenvalues are unique so that $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ be the eigenvalues of A. Let $\lambda_1^* \geq \lambda_2^* \geq \ldots \lambda_n^*$ be the eigenvalues of $A + E$. Let $\vec{\mu}$ be the eigenvector of A with eigenvalue λ_1 and let $\Delta\vec{\mu}$ denote the change in eigenvector. This means that $\vec{\mu} + \Delta\vec{\mu}$ is the eigenvector of $A + E$ corresponding to λ_1^*. We also assume that $\vec{\mu} + \Delta\vec{\mu}$ has length 1 and $\Delta\vec{\mu}$ is perpendicular to $\vec{\mu}$. In this case, provided that*

$$\epsilon \leq \frac{\lambda_1 - \lambda_2}{2},$$

we have

$$|\Delta\vec{\mu}| \leq \frac{2\epsilon}{\lambda_1 - \lambda_2}$$

Proof. From equation 3.5 we find:

$$(A - I\lambda_1)\Delta\vec{\mu} = -E(\vec{\mu} + \Delta\vec{\mu}) + \Delta\lambda(\vec{\mu} + \Delta\vec{\mu}). \tag{4.2}$$

Now we assume that

$$|\vec{\mu} + \Delta\vec{\mu}| = 1.$$

We can do this because eigenvectors are only defined up to a scalar. So, to determinate the amplitude by which they change we need to normalize them. (Otherwise we could just multiply them with a very big number, and then any change if small would appear big in norm). Hence, since the spectral norm of E is less than ϵ we find that

$$|E(\vec{\mu} + \Delta\vec{\mu})| \leq \epsilon \tag{4.3}$$

By theorem 4.1, we know that $|\Delta\lambda| \leq \epsilon$. This then implies that

$$|\Delta\lambda(\vec{\mu} + \Delta\vec{\mu})| \leq \epsilon \tag{4.4}$$

Taking the Euclidean norm on both sides of equation 3.5, yields:

$$|(A - I\lambda_1)\Delta\vec{\mu}| = |E(\vec{\mu} + \Delta\vec{\mu}) \ + \ \Delta\lambda(\vec{\mu} + \Delta\vec{\mu})| \leq |E(\vec{\mu} + \Delta\vec{\mu})| \ + \ |\Delta\lambda(\vec{\mu} + \Delta\vec{\mu})| \leq \epsilon + \epsilon = 2\epsilon \tag{4.5}$$

where for the last inequality above we used inequalities 4.3 and 4.4. Now, we took $\Delta\vec{\mu}$ to be in the plane orthogonal to $\vec{mu}$. Recall when we work in the basis given by the eigenvectors of A, we find that $(A - I\lambda_1)$ is a diagonal matrix given by

$$A - \lambda_1 I = \begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 - \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 - \lambda_1 & 0 & \ldots & 0 \\ \ldots & & & & & \\ 0 & 0 & 0 & 0 & \ldots & \lambda_n - \lambda_1 \end{pmatrix}$$

and the vector

$$\Delta\vec{\mu} = (0, \Delta\mu_2, \Delta\mu_3, \ldots, \Delta\mu_n)^T.$$

Now, the eigenvalues where taken in decreasing order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. It follows that in the matrix $A - \lambda_1 I$ the smallest absolute value of a non-zero entry in the diagonal is $\lambda_1 - \lambda_2$. The vector $\Delta\vec{\mu}$ is not affected by the non-zero entry, since its first entry is 0. And hence

$$(\lambda_1 - \lambda_2)|\Delta\vec{\mu}| \leq |(A - \lambda_1 I)\Delta\vec{\mu}|$$

Applying the last inequality above to 4.5, finally yields the desired result:

$$|\Delta\vec{\mu}| \leq \frac{2\epsilon}{\lambda_1 - \lambda_2}.$$

■

4.2 Bounds for the spectral norm of the perturbation matrix : simplified case of independent entries above the diagonal.

The next question is how to bound the spectral norm of the symmetric error matrix E in our applications, here to covariance matrix estimation. We assume that all the entries of E have expectation 0.
Recall that in our low-dimensional case we had found that asymptotically $E = C\hat{O}V[\vec{X} - COV[\vec{X}]$ behaves asymptotically like

$$E = C\hat{O}V[\vec{X}] - COV[\vec{X}] \approx \frac{1}{\sqrt{n}} \begin{pmatrix} \sigma_{X^2}\mathcal{N}_{11} & \sigma_X\sigma_Y\mathcal{N}_{12} & \sigma_X\sigma_Z\mathcal{N}_{13} \\ \sigma_Y\sigma_X\mathcal{N}_{21} & \sigma_{Y^2}\mathcal{N}_{22} & \sigma_Y\sigma_Z\mathcal{N}_{23} \\ \sigma_Z\sigma_X\mathcal{N}_{31} & \sigma_Z\sigma_Y\mathcal{N}_{32} & \sigma_{Z^2}\mathcal{N}_{33} \end{pmatrix} \tag{4.6}$$

where the matrix on the right side of the last equation above has independent normal entries above the diagonal and $\mathcal{N}_{ij}$ are i.i.d. for $i \leq j \leq p$ standard normal. Let us first assume that all the σ's are equal to 1. And let us assume that instead of three dimensions we have p dimension. Then, for such a symmetric matrix of i.i.d. standard normal entries above the diagonal the spectral norm can be bound in a simple way. Now, the actual estimated covariance matrix is slightly more complicated than a matrix of independent entries above the diagonal, so we will treat that case only in the next subsection. But, for understanding the method of bounding the spectral norm of the perturbation it is best to start with a symmetric perturbation matrix which consists of independent entries above the diagonal. Later, in the next section, we will see that to view the estimated covariance matrix as a matrix of independent normal entries above the diagonal, can be a good approximation when the number of samples is several times bigger than the dimension of the vectors considered. This approximation leads however to a complete misunderstanding of what is going on when the number of samples is strictly less than the dimension of the vectors. In that case, the estimated covariance matrix is defective, whilst a matrix with independent normal entries above the diagonal has always full rank!
So, in this current subsection we consider symmetric matrices with independent normal entries above the diagonal. The technique to bound the perturbation's spectral norm is less opaque for symmetric matrices with i.i.d. entries above the diagonal, than if we use righatway the true estimated covariance matrix.
So, we have such a symmetric matrix E with all entries above the diagonal independent normal. Let us assume to start with that all the entries are standard normal. Take now $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ to be a vector of Euclidean length 1:

$$|\vec{\mu}| = \mu_1^2 + \mu_2^2 + \ldots + \mu_p^2 = 1$$

Let us consider

$$\vec{\mu}^T \cdot E \cdot \vec{\mu} = \sum_{ij} E_{ij}\mu_i\mu_j.$$

This expression has expectation 0. Let us calculate the variance then:

$$\begin{aligned} &VAR[\vec{\mu}^T \cdot E \cdot \vec{\mu}] = VAR[\sum_{ij} E_{ij}\mu_i\mu_j] = \\ &= VAR[\sum_{i>j} 2E_{ij}\mu_i\mu_j + \sum_i E_{ii}\mu_i^2] = \sum_{i>j} 4VAR[E_{ij}]\mu_i^2\mu_j^2 + \sum_i VAR[E_{ii}]\mu_i^4 = \\ &= \sum_{i,j} 2\mu_i^2\mu_j^2 - \sum_i 2\mu_i^4 + \sum_i \mu_i^4 = \\ &= 2\sum_i \mu_i^2 \sum_j \mu_j^2 - \sum_i \mu_i^4 = 2 - \sum_i \mu_i^4 \leq 2. \end{aligned}$$

So, we know now that $\vec{\mu}^T E \vec{\mu}$ is a normal with expectation 0 and standard deviation less than 2 for $\vec{\mu} \in S^{p-1}$. Our goal here is to bound the spectral norm of E, which is the maximum value of $\vec{\mu}^T E \vec{\mu}$ when $\vec{\mu}$ ranges over S^{p-1}. The good news is that to find the order of magnitude of that maximum in the present case we need only find the maximum over a ϵ-net of S^{p-1}. A set C is called an ϵ-*net* of a set B

in a metric space if for each point x of B there is a point y of C so that $d(x,y) \leq \epsilon$. Let us next give the lemma which shows how we use ϵ-nets to find the order of magnitude of the maximum we are interested in:

Lemma 4.1 *Let $\epsilon > 0$ and let E be a symmetric $p \times p$-matrix. Then, $\mathcal{N}_\epsilon \subset S^{p-1}$ be a ϵ-net of S^{p-1}. Then,*

$$|E| \leq \frac{1}{1-2\epsilon} \max_{\vec{\mu} \in \mathcal{N}_\epsilon} \vec{\mu}^T E \vec{\mu}$$

Proof. Let $\vec{x}$ be the unit vector so that

$$\vec{x}^T E \vec{x} = |E|$$

In other words $\vec{x}$ is the eigenvector with biggest eigenvalue. Let $\vec{y}$ be a vector from the ϵ-net N_ϵ which is not futher from $\vec{y}$ than a Euclidian distance ϵ. Let us denote by $\vec{\epsilon}$ the difference:

$$\vec{\epsilon} = \vec{x} - \vec{y}.$$

Then by definition the Euclidian norm of $\vec{\epsilon}$ is less or equal to ϵ. Now, we have

$$|E| = (\vec{y} + \vec{\epsilon})^T \cdot E(\vec{y} + \vec{\epsilon}) = \vec{y}^T E \vec{y} + \vec{\epsilon} E \vec{x} + \vec{y} E \vec{\epsilon} \leq \vec{y}^T E \vec{y} + 2|E|\epsilon$$

Taking the maximum for $\vec{y}$ over $\mathcal{N}_\epsilon$ of the last equation above yields:

$$|E| \leq \sup_{\vec{y} \in \mathcal{N}_\epsilon} \vec{y}^T E \vec{y} + 2|E|\epsilon$$

and hence

$$|E| \leq \frac{1}{1-2\epsilon} \sup_{\vec{y} \in N_\epsilon} \vec{y}^T E \vec{y}.$$

This finishes this proof. ■

So, now we know that in order to figure out the order of magnitude of the spectral norm of E, it is enough to determine the maximum of $\vec{u}^T E \vec{u}$ over a ϵ-net of the unit sphere S^{p-1}. Now, how big such a maximum can get may depend on the criminality of the ϵ-net. So, we are interested in the minimal possible criminality of a ϵ-net of S^{p-1}. That minimal criminality will be called the covering number of S^{p-1}.

Lemma 4.2 *Let $\mathcal{N}_\epsilon^p$ be a ϵ-net with minimal criminality. So, $|\mathcal{N}_\epsilon^p|$ is the covering number of S^{p-1}. We have the following upper bound*

$$\texttt{covering number of } \mathtt{S}^{\mathtt{p}-1} = |\mathcal{N}_\epsilon^{\mathtt{p}}| \leq \left(1 + \frac{2}{\epsilon}\right)^{\mathtt{p}} \tag{4.7}$$

Proof. Let us cover the surface of the unity sphere by adding ball of radius $\epsilon > 0$ one after another. We place the center of the next ball in a place which is not covered. This way the centers of the balls are all away from each other by strictly more than ϵ. So, if we reduce the size of each of these balls by a factor 0.5 then they do not longer intersect. Let $\mathcal{N}$ denote the cardinality of the covering obtained by successively adding balls with centers located in the non-covered area. Clearly $\mathcal{N}_\epsilon^n \leq \mathcal{N}$. But, the reduced balls of the covering, are all contained in the ball centered at the origin and with diameter $1 + \frac{\epsilon}{2}$. Since they don't intersect, we find the the volume they cover must be less than the volume of the ball of radius $1 + \frac{\epsilon}{2}$. This leads to

$$\mathcal{N} \cdot B^p(\frac{\epsilon}{2}) \leq B^p(1 + \frac{\epsilon}{2})$$

and hence

$$\mathcal{N} \cdot (\frac{\epsilon}{2})^p B(1) \leq (1 + \frac{\epsilon}{2})^p B(1)$$

which implies

$$\mathcal{N} \leq \left(1 + \frac{2}{\epsilon}\right)^p.$$

Hence, since $\mathcal{N}$ is bigger or equal to $\mathcal{N}_\epsilon^p$, this implies inequality 4.7. ■

So finding the order of magnitude of the spectral norm of E boils down, to figuring out the maximum of the expression

$$\vec{\mu}^T E \vec{\mu} \tag{4.8}$$

where $\vec{\mu}$ ranges over a ϵ-net of S^{p-1}. Expression 4.8 is a normal random variable with expectation 0 and standard deviation less or equal to 2. Such a variable hence most of the times takes values between -4 and 4. However the maximum we consider will be of a much bigger order. The reason is that we consider many such variables, and hence, with many variables, the odds are that at least some of them will be exceptionally big. To quantify this phenomena we need to determine the order of the probability for a normal random variable be big. This is the content of the next lemma:

Lemma 4.3 *Let $\mathcal{N}(0,1)$ be a standard normal. Let $s > 0$. Then, we have*

$$P(\mathcal{N}(0,1) \geq s) \leq 0.5 \cdot \exp(-s^2/2).$$

Proof. Let $s > 0$. We have

$$P(\mathcal{N}(0,1) \geq s) = \int_s^\infty \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx = \int_0^\infty \frac{\exp(-(s+y)^2/2)}{\sqrt{2\pi}} dy \tag{4.9}$$

where to obtain the last equation above we operated the change of variable $x = s + y$. Now, for $s, y > 0$ we have

$$-(s+y)^2 \leq -s^2 - y^2$$

This implies that the right most side of the chain of equations 4.9, is less than

$$\int_0^\infty \frac{\exp(-(s+y)^2/2)}{\sqrt{2\pi}} dy \leq \exp(-s^2/2) \int_0^\infty \frac{\exp(-y^2/2}{\sqrt{2\pi}} dy = \exp(-s^2/2) \cdot P(\mathcal{N}(0,1) \geq 0) = 0.5 \exp(-s^2/2)$$

The last equation together with 4.9 finishes to prove that

$$P(\mathcal{N}(0,1) \geq s) \leq 0.5 \exp(-s^2/2).$$

■

We are now ready to explain why a symmetric $p \times p$-matrix E with standard normal entries which are independent above the diagonal has spectral norm of order $\sqrt{p}$ times constant. The idea is very simple: to get the order of magnitude of the spectral norm of E it is enough to find the order of magnitude of the maximum of $\vec{\mu}^T E \vec{\mu}$ where $\vec{\mu}$ ranges over a $\epsilon > 0$-net $\mathcal{N}_\epsilon$. We also need $\epsilon < 0.5$, for guaranteeing that the maximum over the ϵ-net of S^{p-1} is the same order as the maximum over all of S^{p-1}.
Now, this is the same principle as when you are climbing in the mountains: if you climb often the risk of an accident becomes bigger. so, for example if every time you climb the risk of an accident is $1/1000$ and you climb 10 times, then the overall risk of an accident happening during one of your climbs is going to be $10/1000 = 0.01$. Here, the vectors $\vec{\mu}$ in our ϵ-net $\mathcal{N}_\epsilon^p$ are going to be the climbs. The accident will be that $\vec{\mu}^T E \vec{\mu}$ is going to be exceptionally big. So, using the same formula as for the climbs:

P(An accident happens during one of the climbs) $\leq$ number of climbs$\times P$(Accident happens during one climb).

So, instead of number of climbs we will have number of elements in the ϵ-net, hence the covering number. So, applying this we find:

$$P(\text{There is a } \vec{\mu} \text{ in the } \epsilon\text{-net so that } \vec{\mu}^T E \vec{\mu} \geq t) \leq |\mathcal{N}_\epsilon^p| \cdot P(\vec{\mu}^T E \vec{\mu}) \leq 0.5(1 + \frac{2}{\epsilon})^p \cdot \exp(-t^2/8)$$

where we used our upper bound 4.7 for the covering number and the following:

$$P(\vec{\mu}^T E\vec{\mu} \geq t) \leq P(2\mathcal{N}(0,1) \geq t) \leq 0.5\exp(-t^2/8). \tag{4.10}$$

The last inequality was derived from Lemma 4.3 and the fact that $\vec{\mu}^T E\vec{\mu}$ is normal with expectation 0 and standard deviation less or equal than 2.
Now, take for t the value

$$t_0 := \sqrt{p}\cdot\sqrt{8\ln(1+\frac{2}{\epsilon})}.$$

With that value t_0 for t, the bound on the right side if inequality 4.10 is equal to 0.5. So, when you take t to be even bigger than t_0 by a quantity s, then the bound on the right side of inequality 4.10 becomes less than $0.5\exp(-s^2/2)$. This is the content of the next theorem:

Theorem 4.3 *Let E_{ij} for $i,j \leq p$ be standard normal variable so that $E_{ij} = E_{ji}$ for all $i,j \leq p$ and so that E_{ij} for $i \leq j \leq p$ is a collection of i.i.d. standard normal variables. Let E denote the $p \times p$-matrix, given by*

$$E = (E_{ij})$$

Let ϵ be a constant so that $0 < \epsilon < 0.5$. Then, for every $s \geq 0$ we have:

$$P(|E| \geq \mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s}) \leq \mathtt{0.5}\exp(-\mathtt{s}^2/\mathtt{2}) \tag{4.11}$$

where

$$\mathtt{constant}_\epsilon = \frac{\sqrt{\mathtt{8}\cdot\ln(\mathtt{1}+\frac{\mathtt{2}}{\epsilon})}}{\mathtt{1}-\mathtt{2}\epsilon}.$$

We also have a lower bound. That is there exists $c_0, c_1 > 0$ not depending on p so that

$$P(|E| \leq c_0\cdot\sqrt{p}) \leq \exp(-c_1 p)$$

Proof. ■

Now, the above lemma is for random matrices where all entries are standard normal. In reality, this is seldom the case for our covariance-matrix-estimation-error. Rather, asymptotically, we get an expression like in 4.6 where there are coefficients $\sigma_i\sigma_j$ multiplying the standard normal variables. So, let us assume that we have a matrix E with normal entries with 0 expectation and so that the entries above the diagonal are independent of each other. Furthermore, we assume

$$E_{ij} = \mathcal{N}(0, \frac{\sigma_i\sigma_j}{\sqrt{n}})$$

for all $i,j \in \{1,2,\ldots,p\}$. So, in other words, we have a sequence of non-random values: $\sigma_1 > \sigma_2 > \ldots > \sigma_p$ and the entry E_{ij} can be viewed as a standard normal time the constant $\sigma_i\sigma_j/\sqrt{n}$. So we have

$$E = \frac{1}{\sqrt{n}}\begin{pmatrix} \sigma_1 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma_3 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & \sigma_p \end{pmatrix}\cdot\begin{pmatrix} \mathcal{N}_{11} & \mathcal{N}_{12} & \mathcal{N}_{13} & \ldots & \mathcal{N}_{1p} \\ \mathcal{N}_{21} & \mathcal{N}_{22} & \mathcal{N}_{23} & \ldots & \mathcal{N}_{2p} \\ \mathcal{N}_{31} & \mathcal{N}_{32} & \mathcal{N}_{33} & \ldots & \mathcal{N}_{3p} \\ & & & \ldots & \\ \mathcal{N}_{p1} & \mathcal{N}_{p2} & \mathcal{N}_{p3} & \ldots & \mathcal{N}_{pp} \end{pmatrix}\cdot\begin{pmatrix} \sigma_1 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma_3 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & \sigma_p \end{pmatrix}$$

Where N_{ij} with $i \leq j \leq p$ are standard normal independent of each other and $N_{ij} = N_{ji}$ for all $i,j \leq p$. Let D_σ denote the following diagonal matrix:

$$D_\sigma := \begin{pmatrix} \sigma_1 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma_3 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & \sigma_p \end{pmatrix}$$

Let E^* denote the symmetric matrix with the normalized normal entries:

$$E^* = \begin{pmatrix} \mathcal{N}_{11} & \mathcal{N}_{12} & \mathcal{N}_{13} & \dots & \mathcal{N}_{1p} \\ \mathcal{N}_{21} & \mathcal{N}_{22} & \mathcal{N}_{23} & \dots & \mathcal{N}_{2p} \\ \mathcal{N}_{31} & \mathcal{N}_{32} & \mathcal{N}_{33} & \dots & \mathcal{N}_{3p} \\ & & & \dots & \\ \mathcal{N}_{p1} & \mathcal{N}_{p2} & \mathcal{N}_{p3} & \dots & \mathcal{N}_{pp} \end{pmatrix}$$

Then, we have for any $\vec{\mu} \in S^{p-1}$, that

$$|\vec{\mu}^T E \vec{\mu}| = |\vec{\mu}^T D E^* D \vec{\mu}| = |\vec{\mu} D|^2 \cdot \left| \frac{\vec{\mu}^T}{|\vec{\mu} D|} \cdot E^* \cdot \frac{\vec{\mu}}{|\vec{\mu} D|} \right| \leq |\vec{\mu} D|^2 \cdot |E^*| \tag{4.12}$$

where we used the fact that the renormalized vector

$$\vec{y} := \frac{\vec{\mu} D}{|\vec{\mu} D|}$$

has norm 1, and hence $\vec{y}^T E^* \vec{y}$ is less than the spectral norm $|E^*|$ of E^*. But, clearly, since $\vec{\mu}$ has norm 1, we get that $|\vec{\mu} D| \leq \sigma_1$. Applying the last inequality to 4.12, yields,

$$|\vec{\mu}^T E \vec{\mu}| \leq \sigma_1^2 |E^*|$$

Taking the supremum of the last inequality for $\vec{\mu}$ ranging over S^{p-1} yields:

$$|E| \leq |E^*| \sigma_1^2.$$

Hence,

$$|E^*| \leq \mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s}$$

implies

$$|E| \leq \sigma_1^2 (\mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s}).$$

In terms of the probabilities we thus have

$$P(|E| \geq \sigma_1^2 (\mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s})) \;\; \leq \;\; \mathtt{P}(|\mathtt{E}^*| \geq \mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s}) \tag{4.13}$$

We can now use the bound 4.11, but apply it to the matrix E^* instead of E, since now the matrix E^* is the one with the standard normal entries. Together with the inequality 4.13 we find

$$P(P(|E| \geq \sigma_1^2 (\mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s})) \;\leq\; \exp(-\mathtt{s}^2/2).$$

The same approach leads to a similar lower bound. So, this gives us the next theorem:

Theorem 4.4 *Let E be a $p \times p$ symmetric matrix with zero expectation and normal entries which are independent of each other above the diagonal. Assume also that $\sigma_1 > \sigma_2 > \ldots > \sigma_p \geq 0$ are non-random numbers and assume that E_{ij} has standard deviation equal to $\frac{\sigma_i \sigma_j}{\sqrt{n}}$ for all $i, j \in \{1, 2, \ldots, p\}$. Then we have*

$$P(|E| \geq \frac{\sigma_1^2}{\sqrt{n}} (\mathtt{constant}_\epsilon \times \sqrt{\mathtt{p}} + \mathtt{s})) \;\; \leq \;\; \exp(-\mathtt{s}^2/2)$$

and

$$P(|E| \leq \frac{\sigma_p^2 c_0}{\sqrt{n}} \cdot \sqrt{p}) \leq \exp(-c_1 p)$$

The above corrolarry gives the correct order of the spectral norm of $|E|$ up to a constant when σ_1 and σ_p are of the same order. (Equal to each other up to a constant). Now, in many situations this will not be the case: we saw in our simple models to explain factor analysis that there will be a few eigenvalues of much bigger order then the other eigenvalues. Non-the-less, the above corollary will be very useful to calculating the exact order in the realistic case where there are eigenvalues of different order. This will be seen in subsection 4.5.

4.3 Precise proof for bounds for the spectral norm of the perturbation matrix when estimating the covariance matrix

Let us assume that $\vec{X} = (X_1, X_2, \ldots, X_p)$ is a normal vector with independent components and 0 expectation

$$E[X_1] = E[X_2] = \ldots = E[X_p] = 0.$$

. We can always make a change of coordinates to get independent components. Now, we assume that we have many independent copies of the vector $\vec{X}$ Hence,

$$\vec{X}, \vec{X}_1, \vec{X}_2, \ldots$$

is an i.i.d. sequence of normal vectors where

$$\vec{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip}).$$

Let X denotes the $n \times p$ random matrix having its i-th row equal to $\vec{X}_i$. So, we have a sample of size n of random vectors of dimension p. We use this sample to estimate the covariance matrix. Since, the expectation is 0, we have

$$C0V[X_i, X_j] = E[X_i X_j]$$

for all $i, j \leq p$. Hence, we can use as estimate of the covariance of X_i with X_j, the following:

$$\hat{COV}(X_i, X_j) = \frac{X_{1i}X_{1j} + X_{2i}X_{2j} + \ldots + X_{ni}X_{nj}}{n}.$$

for any pair $i, j \leq p$. So, estimating each entry of the covariance matrix of $\vec{X}$, gives our estimate of the covariance matrix which can also be expressed in terms of the matrix X as follows:

$$\hat{COV}[\vec{X}] = (\hat{COV}(X_i, X_j))_{i,j} = \frac{1}{n} X^t X.$$

In the previous section on the "small dimensional case" (Section 3), we took p fixed and let n go to infinity. Here, we will be in the case, where we let n and p both go to infinity at the same time. Typically we assume that there is a constant C so that $n = Cp$. Now, if $C < 1$ that is $n < p$ then the estimated covariance matrix does not have full rank. Thus, if for example we would be in the portfolio problem, and we would use the estimated covariance matrix instead of the real one with $n < p$ we would find erornously an investement oportunity with 0 estimated variance.....
Now, in the Section 3 on the "low dimensional case", we saw that the approximation error of the random matrix behaves roughly like an symmetric matrix with independent normal entries above the diagonal. Such a matrix has full rank, and hence "behaves entirely different" from the estimated covariance matrix in the case when $n < p$. But, for a constant C not depending on n which is quite a bit bigger than 1, it is our understanding that if we put $n = Cp$, there is almost no difference for practical puposes between the symmetric matrix with with independent normal entries like given in formula 3.2. and reality. So for practical purposes, if $n = Cp$ and $C >> 1$, then (in our understanding) most of the times you can assume that the estimated covariance matrix is obtrained from the original one by adding a symmetric matrix with i.i.d. normal entries above the diagonal. But, this is not a proof, but so far a heuristic argument. So, in this current section we are going to prove a formula like we obtained in the previous subsection bounding the spectral norm of the estimation error for the covariance matrix. The idea is very similar, but slightly more complicated though the approach is the same. We start with assuming that all the entries of the random vector have same standard deviation equal to 1. (Until recently for the case when the entries had different standard deviation, the order of magnitude of the spectral norm of the covariance matrix approximation error was not known.....and still to this day, there is a forumal by Koltschinksii and Lounici which is only up to a constant the constant being not known. We are currently working on this.) So, let us start with the case where all the components have standard deviation 1:

Theorem 4.5 *Let X be a random $n \times p$-matrix with i.i.d. standard normal entries. Let E be the estimation error of the covariance matrix:*

$$E := \frac{1}{n} X^t \cdot X - I$$

where I is the $p \times p$ identity matrix. Then there exists a constant $c_3 > 0$ not depending on p so that if $n \geq c_3 p$, we have that, with high probability, the order of the spectral norm of the error-matrix is no more than $\sqrt{p}/\sqrt{n}$ times a constant. let ϵ be a constant such that $0 < \epsilon < 0.5$.
Let us formulate our theorem more precisely:
there exists $c_3 > 0$ not depending on p, so that if $n \geq c_3 p$ we have:

$$P\left(|E| \geq \frac{c_3}{1-2\epsilon} \cdot \frac{\sqrt{p}}{\sqrt{n}} \right) \leq \exp(-p)$$

Proof. We are going to bound the spectral norm of the matrix

$$n \cdot E = X^t X - nI \tag{4.14}$$

where I represents the $p \times p$ identity matrix. So we want to show that the matrix E has typically spectral norm no bigger than order $\sqrt{p}/\sqrt{n}$ times a constant. So, this means we want to prove for the symmetric matrix $n \cdot E$ that typically the spectral norm is not bigger than $\sqrt{p}\sqrt{n}$. As, we did in the previous section, we use the fact that for a symmetric matrix, to find the order of the spectral norm, we work with a ϵ-net $\mathcal{N}_\epsilon^p$ of minimum cardinality. So, let $\mathcal{N}_\epsilon^p \subset S^{p-1}$ be such a set. Thus we have according to formula 4.7 in Lemma 4.2, that

$$|\mathcal{N}_\epsilon^p| \leq \left(1 + \frac{2}{\epsilon}\right)^p \tag{4.15}$$

Let now $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ be a vector of $\mathcal{N}_\epsilon^p$. Hence, $\vec{\mu}$ is a Euclidian unit vector. So, the spectral norm of nE is given by

$$|nE| = \max_{\vec{\mu} \in S^{p-1}} |\vec{\mu}^T (nE) \vec{\mu}| = \max_{\vec{\mu} \in S^{p-1}} |\vec{\mu}^T \left(X^T X - nI\right) \vec{\mu}| \tag{4.16}$$

Again, according to Lemma 4.1, when we take the maximum above for $\vec{\mu}$ ranging over the ϵ-net $\mathcal{N}_\epsilon^p$, instead of all of S^{p-1} we get the same order of magnitude up to a constant. So, we are going to find a likely bound for the maximum like in **??**, but where $\vec{\mu}$ ranges over the ϵ-net $\mathcal{N}_\epsilon^p$ instead of S^{p-1}. So, for one $\vec{\mu} \in S^{p-1}$, we need to bound the probability that

$$\vec{\mu}^T (X^T X - nI) \vec{\mu} = \vec{\mu}^T X^T X \vec{\mu} - n \vec{\mu}^T \vec{\mu} \tag{4.17}$$

is big. Now, note that

$$E[X^t X] = nI$$

and hence

$$E[\vec{\mu}^T (X^T X) \vec{\mu}] = n \vec{\mu}^T \vec{\mu}$$

So, the expression 4.17 is obtained from taking the random number

$$\vec{\mu}^T X^T X \vec{\mu} \tag{4.18}$$

and subtracting its expectation from it. Now, expression 4.18 is symply the Euclidian norm squared for the random vector $X\vec{\mu}$. But, this random vector consists of n independent standard normal entries so we can write:

$$X\vec{\mu} = (\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_n)^T$$

where $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_n$ are i.i.d. standard normal entries. So expression 4.18 is a chi-square variable with n degrees of freedom:

$$\vec{\mu}^T X^T X \vec{\mu} = \mathcal{N}_1^2 + \mathcal{N}_2^2 + \ldots + \mathcal{N}_n^2$$

subtracting its expectation amounts to the same as subtracting from each $\mathcal{N}_1^2$ one unit. So, we find:

$$\vec{\mu}^T X^T X \vec{\mu} - n\vec{\mu}^T X \vec{\mu} = \vec{\mu}^T X^T X \vec{\mu} - E[\vec{\mu}^T X^T X \vec{\mu}] = (\mathcal{N}_1^2 - 1) + (\mathcal{N}_2^2 - 1) + \ldots + (\mathcal{N}_n^2 - 1). \quad (4.19)$$

The variables $\mathcal{N}_i^2$ don't behave like normal in terms of their tail. Instead, they are merely subexponential. So, we can use the large deviation result below in Lemma 4.4 for summs of centered subexponential variables. We can apply thus Lemma 4.4 with $K = 4$. Let $Y_i = (X_i^2 - 1)$ With the help of equation 4.19, we find:

$$P(\vec{\mu}^T E \vec{\mu} \geq c_3 \frac{\sqrt{p}}{\sqrt{n}}) = P(\vec{\mu}^T X^T X \vec{\mu} - n\vec{\mu}^T \vec{\mu} \geq c_3 \sqrt{n}\sqrt{p}) = \quad (4.20)$$

$$P(Y_1 + Y_2 + \ldots + Y_n \geq c_3 n \cdot \frac{\sqrt{p})}{\sqrt{n}}) \leq \quad (4.21)$$

$$\leq 2\exp(-c \cdot c_3 (n(\frac{\sqrt{p}}{4\sqrt{n}})^2) = 2\exp(-c \cdot c_3 \frac{p}{16}) \quad (4.22)$$

where the very last inequality above was obtained with the help of Lemma 4.4 takine ϵ to be

$$\epsilon := \frac{\sqrt{p}}{\sqrt{n}}.$$

Note that according to Lemma 4.4, the last inequality above only holds when $\epsilon^2/K^2 \leq 1$. (Note that when $\epsilon/K > 1$, then the formula for the bound from Lemma 4.4 changes and is no longer useful for us.) So, in other words we need $n \geq K^2 p$ for our inequality to hold. This is to say that we take n larger than a certain constant number of times the dimension p of the space.....
Now, to find the probability upper bound for that

$$\vec{\mu}^T E \vec{\mu} \leq c_2 \sqrt{n}\sqrt{p}$$

holds for all $\vec{\mu}$ in our ϵ-net $\mathcal{N}_\epsilon^p$ we simply need to multiply the probability bound for one such $\vec{\mu}$ by the upper bound for the number of elements in the ϵ-net. So, we get:

$$P(\exists \vec{\mu} \in \mathcal{N}_\epsilon^p \ , \ |\vec{\mu}^T E \vec{\mu}| \geq c_3 \sqrt{n}\sqrt{p})) \leq |\mathcal{N}_\epsilon^p| \cdot \exp(-c \cdot c_3 \frac{p}{16}) \leq (1 + \frac{2}{\epsilon})^p \exp(-\frac{c \cdot c_3}{16} \cdot p)$$

the bound on the right side of the last equation is less than $\exp(-p)$ as soon as c_3 is big enough. More precisely, just take c_3 so that it satisfies

$$c_3 > \frac{1}{c}\left(16\ln(1 + \frac{2}{\epsilon}) + 16\right)$$

and then the bound becomes less or equal to $2\exp(-p)$. By lemma 4.1 when the maximum of $|\vec{\mu}^T E \vec{\mu}|$ is bound for $\vec{\mu}$ ranging over the ϵ-net $\mathcal{N}_\epsilon^p$, then we have to multiply that bound by $\frac{1}{1-2\epsilon}$ in order to find a bound for the maximum with $\vec{\mu}$ ranging over all of S^{p-1}, This finishes the proof. then ■

heini

Lemma 4.4 *Let $Y_1, Y_2, \ldots$ be i.i.d. sub-exponential variables with parameter K and expectation 0, we have: for every $\epsilon > 0$, we have:*

$$P(\sum_{i=1}^{n} Y_i| \geq \epsilon n) \leq 2\exp(-c \min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right) \cdot n).$$

Proof. ■
compare the last lemma above to Azuma-Hoeffding for a sum of i.i.d.:

Lemma 4.5 *let $Y_1, Y_2, \ldots$ be i.i.d. random variables with 0 expecation which are bounded by the number a:*

$$P(|Y_i| \leq a) = 1.$$

Then, we have

$$P(|Y_1 + Y_2 + \ldots + Y_n| \geq \epsilon n) \leq 2\exp(-\frac{n\epsilon^2}{a^2}).$$

Now above we have found the order for approximation error of the covariance matrix to be $\sqrt{p}/\sqrt{n}$ times constant. yBut this bound was only valid, when we deal with normal random vectors with standard deviation one. Until recently, there was no tide bound for the order of magnitude of the spectral norm of the approximation error when the standard deviations are of different orders. Recently, Koltschinksii and klounici where able to obtain a fomula using a Talagrand-measure concentration inequality. However, they do not provide a way to determine, a constant in front of their order, but merely give the fact that such a universal constant whic does not depend on p exists. So, let us mention what there order is: For this we still have

$$\vec{X}, \vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$$

is an i.i.d. sequence of normal random vectors with expectation 0. As before X is a $n \times p$-matrix with independent normal entries so that the i-th row of X is equal to $\vec{X}_i$. But, this time the entries of X are independent but not i.i.d. They are merely i.i.d. in each column. More precisely we assume that we have $\sigma_1 > \sigma_2 > \ldots > \sigma_p$ are constants and the j-th entry of the random vector $\vec{X}_i$ has standard deviation σ_j. In other words, all entries in column j of the random matrix X are i.i.d. normal with standard deviation equal to σ_j. for that situation, the order found by Koltschinsky and Klounici for the spectral norm of the approximation error matrix of the covariance matrix is given by the formula:

$$\sigma_1\sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2}$$

which is the order up to a constant. But, again they do not provide a way to determine a constant to put in front of that order, but merely provide an existence proof. Let us state their result in a theorem:

Theorem 4.6

4.4 Precise coordinate-wise understanding of the error made in estimating the eigenvectors of covariance matrix in high-dimensional case

Equation 3.5 was used to find an asymptotic expression for the error in estimating the eigenvectors of the covariance matrix in the low dimensional case. That meant that we have p fixed and let n go to infinity. In this subsection we are finding means of obtaining the same asymptotic formula but when both p and n go to infinity at the same time. More precisely, we are interested in situations where there is a constant $C > 0$ and $n = Cp$. When the dimension is not hold fixed as n goes to infinity, but rather grows at the same time, then things become different. That is the approximation 3.6 which was obtained from equation 3.5 by leaving out two smaller order terms is not apriori valid: small terms can not necessarily be left out when their number grows to infinity. So, the matrix term which are left out because each of the entry goes to 0 faster than $1/\sqrt{n}$ can no longer be left out apriori. (since they are part of a matrix whos dimension goes to infinity. And a matrix with small entries but high dimension could potentially have a large spectral norm and can hence not be automatically discarded)
So in the case the dimension p is not fixed, we can not leave out apriori the two "smaller order" terms of equation 3.5. They might be small only in the low dimensional case. So, instead we are taking all of equation 3.5 at first, without leaving out anything. So, we assume a three dimensional random vector with independent normal entries $\vec{X} = (X, Y, Z)$ having each expectation 0. We take a three dimensional vector

$\vec{X}$ to simplify notation, but the formulas obtained will remain valid for a p dimensional random vector with independent components. We assume that the standard deviations are decreasing: $\sigma_X > \sigma_Y > \sigma_Z$. The covariance matrix we consider is hence

$$COV[\vec{X}] = \begin{pmatrix} \sigma_X^2 & 0 & 0 \\ 0 & \sigma_Y^2 & 0 \\ 0 & 0 & \sigma_Z^2 \end{pmatrix} \tag{4.23}$$

which corresponds to the matrix A in formula 3.5. Now, we want at first an exact formula and not an approximation. So, instead of $\mathcal{N}_{ij}$ we will have the exact values denoted by $\hat{\mathcal{N}}_{ij} := \frac{1}{\sigma_i \sigma_j} E_{ij}$ for $i \neq j$. And $\hat{\mathcal{N}}_{ii} = \frac{1}{\sigma_{X_i^2}} E_{ii}$ where $X_1 = X, X_2 = Y, X_3 = Z$. Again, as before here E denotes the noise matrix

$$E = C\hat{O}V[\vec{X}] - COV[\vec{X}],$$

and E_{ij} is the ij-th entry of that covariance-estimation-error-matrix. So, now we can write equation 3.5 with A being the covariance matrix 4.23. We consider the eigenvector $\vec{\mu} = (1, 0, 0)$ of A with eigenvalue σ_X^2. So, without leaving out terms, we find instead of the approximation 3.6, the following exact equation:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_Y^2 - \sigma_X^2 - \Delta\lambda & 0 \\ 0 & 0 & \sigma_Z^2 - \sigma_X^2 - \Delta\lambda \end{pmatrix} \begin{pmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} =$$
$$= \frac{-1}{\sqrt{n}} \begin{pmatrix} \sigma_{X^2}\hat{\mathcal{N}}_{11} \\ \sigma_X\sigma_Y\hat{\mathcal{N}}_{21} \\ \sigma_X\sigma_Z\hat{\mathcal{N}}_{31} \end{pmatrix} + \begin{pmatrix} \Delta\lambda \\ 0 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{n}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_{Y^2}\hat{\mathcal{N}}_{22} & \sigma_Y\sigma_Z\hat{\mathcal{N}}_{23} \\ 0 & \sigma_Y\sigma_Z\hat{\mathcal{N}}_{32} & \sigma_{Z^2}\hat{\mathcal{N}}_{33} \end{pmatrix} \begin{pmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} +$$
$$- \frac{1}{\sqrt{n}} \begin{pmatrix} 0 & \sigma_X\sigma_Y\hat{\mathcal{N}}_{12} & \sigma_X\sigma_Z\hat{\mathcal{N}}_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix}$$

the above equation for matrices can be "separated into two parts". First the single equation for $\Delta\lambda$:

$$\Delta\lambda = \frac{1}{\sqrt{n}}\sigma_{X^2}\hat{\mathcal{N}}_{11} + \frac{\sigma_X}{\sqrt{n}}(\sigma_Y\hat{\mathcal{N}}_{12}\Delta\mu_Y + \sigma_Z\hat{\mathcal{N}}_{13}\Delta\mu_Z)$$

and then the $p-1$ dimensional equation for $\Delta\vec{\mu}$ given as follows:

$$\begin{pmatrix} \sigma_Y^2 - \sigma_X^2 - \Delta\lambda & 0 \\ 0 & \sigma_Z^2 - \sigma_X^2 - \Delta\lambda \end{pmatrix} \begin{pmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} =$$
$$= \frac{-1}{\sqrt{n}} \begin{pmatrix} \sigma_X\sigma_Y\hat{\mathcal{N}}_{21} \\ \sigma_X\sigma_Z\hat{\mathcal{N}}_{31} \end{pmatrix} - \frac{1}{\sqrt{n}} \begin{pmatrix} \sigma_{Y^2}\hat{\mathcal{N}}_{22} & \sigma_Y\sigma_Z\hat{\mathcal{N}}_{23} \\ \sigma_Y\sigma_Z\hat{\mathcal{N}}_{32} & \sigma_{Z^2}\hat{\mathcal{N}}_{33} \end{pmatrix} \begin{pmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix}$$

If $\Delta\lambda$ is given we can solve the above equation for $\Delta\vec{\mu} = (\Delta\mu_Y, \Delta\mu_Z)$ to find:

$$\Delta\vec{\mu} = \begin{pmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} =$$
$$\left(I - \frac{-1}{\sqrt{n}} \begin{pmatrix} \frac{1}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda} & 0 \\ 0 & \frac{1}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{Y^2}\hat{\mathcal{N}}_{22} & \sigma_Y\sigma_Z\hat{\mathcal{N}}_{23} \\ \sigma_Y\sigma_Z\hat{\mathcal{N}}_{32} & \sigma_{Z^2}\hat{\mathcal{N}}_{33} \end{pmatrix} \right)^{-1} \cdot \frac{-1}{\sqrt{n}} \begin{pmatrix} \frac{\sigma_X\sigma_Y}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda}\hat{\mathcal{N}}_{21} \\ \frac{\sigma_X\sigma_Z}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda}\hat{\mathcal{N}}_{31} \end{pmatrix}$$

where I is the identity matrix. Now, let D be the matrix

$$D := \frac{-1}{\sqrt{n}} \begin{pmatrix} \frac{1}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda} & 0 \\ 0 & \frac{1}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{Y^2}\hat{\mathcal{N}}_{22} & \sigma_Y\sigma_Z\hat{\mathcal{N}}_{23} \\ \sigma_Y\sigma_Z\hat{\mathcal{N}}_{32} & \sigma_{Z^2}\hat{\mathcal{N}}_{33} \end{pmatrix}$$

We thus have

$$\Delta\vec{\mu} = \begin{pmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} = -\frac{1}{\sqrt{n}}(I-D)^{-1}\begin{pmatrix} \frac{\sigma_X\sigma_Y}{\sigma_Y^2-\sigma_X^2-\Delta\lambda}\hat{\mathcal{N}}_{21} \\ \frac{\sigma_X\sigma_Z}{\sigma_Z^2-\sigma_X^2-\Delta\lambda}\hat{\mathcal{N}}_{31} \end{pmatrix}$$

Now when the spectral norm of D is less than 1, then we get the formual:

$$(I-D)^{-1} = I + D + D^2 + D^3 + \dots$$

In that case, $(I-D)^{-1}$ can be approximated by I and we find

$$\Delta\vec{\mu} = \begin{pmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{pmatrix} \approx -\frac{1}{\sqrt{n}}\begin{pmatrix} \frac{\sigma_X\sigma_Y}{\sigma_Y^2-\sigma_X^2-\Delta\lambda}\hat{\mathcal{N}}_{21} \\ \frac{\sigma_X\sigma_Z}{\sigma_Z^2-\sigma_X^2-\Delta\lambda}\hat{\mathcal{N}}_{31} \end{pmatrix}$$

with the relative error in that approximation being less than $\frac{|D|}{1-|D|}$. Note that this is our formula (with additional $\Delta\lambda$) from the finite dimensional approximation). We can now rewrite the above formula for a general p dimensional random vector: $\vec{X} = (X_1, X_2, \dots, X_p)$ with independent normal entries all having expectation 0. Let σ_i denote the standard deviation of X_i and we assume $\sigma_1 > \sigma_2 > \dots > \sigma_p$. Let E be the $p \times p$ covariance-matrix-estimation-error-matrix:

$$E = C\hat{O}V[\vec{X}] - COV[\vec{X}].$$

Hence, the covariance matrix is the $p \times p$-matrix, given by:

$$COV[\vec{X}] = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ & & & \dots & \\ 0 & 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

and $\vec{\mu}$ is the eigenvector of the above covariance matrix $\vec{\mu} = (1, 0, 0, 0, \dots, 0)$ with eigenvalue σ_1^2. Let E_1 denote the $(p-1) \times (p-1)$-matrix obtained from E by deleting the first row and the first column. We can now write the formula for the change in the eigenvector $\vec{\mu}$ when instead of taking the covariance matrix, we take its estimate:

$$\Delta\vec{\mu} = \begin{pmatrix} \Delta\mu_2 \\ \Delta\mu_3 \\ \dots \\ \Delta\mu_p \end{pmatrix} = -\frac{1}{\sqrt{n}}(I-D)^{-1}\begin{pmatrix} \frac{\sigma_1\sigma_2}{\sigma_2^2-\sigma_1^2-\Delta\lambda}\hat{\mathcal{N}}_{21} \\ \frac{\sigma_1\sigma_3}{\sigma_3^2-\sigma_1^2-\Delta\lambda}\hat{\mathcal{N}}_{31} \\ \dots \\ \frac{\sigma_1\sigma_p}{\sigma_p^2-\sigma_1^2-\Delta\lambda}\hat{\mathcal{N}}_{p1} \end{pmatrix}$$

where $\hat{\mathcal{N}}_{ij} := \sqrt{n}\frac{E_{ij}}{\sigma_i\sigma_j}$ for all $i \neq j$ with $i, j \leq p$ and

$$D := -\begin{pmatrix} \frac{1}{\sigma_2^2-\sigma_1^2-\Delta\lambda} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_3^2-\sigma_1^2-\Delta\lambda} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sigma_4^2-\sigma_1^2-\Delta\lambda} & \dots & 0 \\ & & & \dots & \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_p^2-\sigma_1^2-\Delta\lambda} \end{pmatrix} \cdot E_1$$

So, we get the same formula for the approximation of the eigenvector change as soon as the spectral norm of D is quite a bit less than 1. This is the case (sufficient condition, but not necessarily necessary condition) when the number of sample is of order large constant times:

$$(\sigma_1^2 - \sigma_2^2)^2 \cdot \sigma_2^2 \cdot (\sigma_2^2 + \sigma_3^2 + \dots + \sigma_p^2)$$

4.5 Bounds for the spectral norm of the estimation error in the covariance matrix for models with eigenvalues of different orders

5 Multivariate normal distribution

We will study the multivariate normal distribution. Assume for example H be the height of a human being and R the ratio between the high and the hip width. The two variables might well be independent of each other. Furthermore, if we believe that the height of an individual is due to a sum of little independent contributions (food habits, genetics, illnesses,...) then according to the central limit theorem, H should be approximately normal. Same thing for R. Let μ_H, resp. μ_R be the expectation of H and or R respectively. Let σ_H and σ_R be the respective standard deviation. Then, the probability density function of H is given by

$$f_H(x) = \frac{1}{\sqrt{2\pi}\sigma_H} \exp(-(x-\mu_H)^2/2\sigma_H^2)$$

whilst the probability density function of R is given by

$$f_R(x) = \frac{1}{\sqrt{2\pi}\sigma_R} \exp(-(x-\mu_R)^2/2\sigma_R^2)$$

The joint density function of two variables which are independent of each other is given by their product. Hence,

$$f_{(H,R)}(x_1,x_2) = f_H(x_1)\cdot f_R(x_2) = \frac{1}{2\pi\sigma_H\cdot\sigma_R}\exp(-0.5(\frac{(x_1-\mu_H)^2}{\sigma_H^2}+\frac{(x_2-\mu_R)^2}{\sigma_R^2}))$$

Let us define the vector: $\vec{x} = (x_1,x_2)^T$ where ^{T} is the symbol for transpose. Furthermore, let $\vec{X}$ be the random vector equal to $(R,H)^T/$ Since, we assume H and R to be independent of each other, we find that the covariance matrix of $\vec{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ is

$$\Sigma_{\vec{X}} := \begin{pmatrix} COV(H,H) & COV(H,R) \\ COV(R,H) & COV(R,R) \end{pmatrix} = \begin{pmatrix} \sigma_H^2 & 0 \\ 0 & \sigma_R^2 \end{pmatrix}$$

and hence the joint density of (H,R) can also be written in matrix/vector notation as:

$$f_{(H,R)}(\vec{x}) = \frac{1}{2\pi\sigma_H\cdot\sigma_R}\exp(-0.5(\vec{x}-\vec{\mu})^T\Sigma_{\vec{X}}^{-1}(\vec{x}-\vec{\mu})) \tag{5.1}$$

where

$$\vec{\mu} = (\vec{\mu}_H,\vec{\mu}_R)^T$$

and $\Sigma_{\vec{X}}^{-1}$ designates the inverse of the covariance matrix of the random vector $(H,R)^T$.
So far we have considered the case of two variables which are independent and each of them is normal. Often times, like in discriminant analysis we will consider all linear combination of two variables. For example we may have $\vec{X} = (H,B)^T$, but consider different linear combinations of the entries of $\vec{X}$. We could have

$$Z_a = a_1H + a_2B = (a_1,a_2)\cdot\vec{X}$$

and

$$Z_b = b_1H + b_2B = (b_1,b_2)\cdot\vec{X},$$

where the coefficients a_1,a_2,b_1,b_2 are fixed non-random. Let $\vec{Z}$ denote the random vector:

$$\vec{Z} = \begin{pmatrix} Z_a \\ Z_b \end{pmatrix}$$

Then, in matrix notation, we have

$$\vec{Z} = A\vec{X}$$

where

$$A = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}.$$

Equivalently we can write

$$\vec{X} = A^{-1}\vec{Z},$$

where A^{-1} is the inverse of the matrix A. Now we can apply the rule for finding the probability density of a random vector $\vec{Z}$ given the density function of a random vector $\vec{X}$, where $\vec{Z}$ is a linear transform of $\vec{X}$. This rule says that the probability density of $\vec{Z}$ can be obtained from the probability density of $\vec{X}$. For, this we just take the density function of $\vec{X}$ and replace $\vec{x}$ by $A^{-1}\vec{z}$. We also have to divide by the determinant of A. This then yields

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{det(A)} f_{\vec{X}}(A^{-1}\vec{z}) \tag{5.2}$$

Now together 5.2 and 5.1, yield

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{2\pi \det(A)\sigma_H\sigma_R} \exp(-0.5(\vec{z}-\vec{\mu}_z)^T A^{-1T}\Sigma_{\vec{X}}^{-1}A^{-1}(\vec{Z}-\vec{\mu}_z) \tag{5.3}$$

where

$$\vec{\mu}_Z = E[\vec{Z}] = E[A\vec{X}] = AE[\vec{X}] = A\vec{\mu}.$$

Note that the covariance matrix of $\vec{Z}$ is given by

$$\Sigma_{\vec{Z}} = COV[\vec{Z}] = E[(\vec{Z}-E[\vec{Z}])(\vec{Z}-E[\vec{Z}])^T] = E[A(\vec{X}-E[\vec{X}])(\vec{X}-E[\vec{X}])^T A^T] =$$
$$= AE[(\vec{X}-E[\vec{X}])(\vec{X}-E[\vec{X}])^T]A^T = A\Sigma_{\vec{X}}A^T$$

and since $(AB)^{-1} = B^{-1}A^{-1}$, we find

$$\Sigma_{\vec{Z}}^{-1} = (A\Sigma_{\vec{X}}A^T)^{-1} = A^{-1T}\Sigma_{\vec{X}}^{-1}A^{-1}.$$

The last equation applied to 5.3 yields

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{2\pi \det(A)\sigma_H\sigma_R} \exp(-0.5(\vec{z}-\mu_{\vec{z}})^T\Sigma_{\vec{z}}^{-1}(\vec{z}-\vec{\mu}_z)). \tag{5.4}$$

Now note that $det(\Sigma_{\vec{X}}) = \sigma_R^2 \cdot \sigma_H^2$. Furthermore the determinant of a product is the product of the determinants and the transpose does not change the determinant:

$$\det(\Sigma_{\vec{Z}}) = \det(A\Sigma_{\vec{X}}A^T) = \det(A)\det(\Sigma_{\vec{X}})\det(A^T) = \det(A)^2\sigma_H^2\sigma_R^2.$$

The last equation can be used in 5.4 to find the final formula for the probability density of $\vec{Z}$:

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{2\pi\sqrt{det(\Sigma_{\vec{Z}})}} \exp(-0.5(\vec{z}-\mu_{\vec{z}})^T\Sigma_{\vec{z}}^{-1}(\vec{z}-\vec{\mu}_z)).$$

this shows that the probability density of the vector $\vec{Z}$ depends only on the covariance matrix and the expectation and nothing else! The same formula for the density would hold if instead of a vector with 2 entires, the vector $\vec{Z}$ would have n entries. This means, that if a vector is a linear transform of a vector with independent normal entries, then the distribution depends only on the covariance matrix and the expectation. Such linear transform are called multivariate normal vectors. Let us give a precise definition:

Definition 5.1 *Let $\vec{Z} = (Z_1, Z_2, \ldots, Z_n)$ a random vector. Then $\vec{Z}$ is said to be normally or Gaussian distributed if there exists a random vector $(X_1, X_2, \ldots, X_n)$ having independent normal entries and such that there exists a $n \times n$-matrix which is non-random such that $\vec{Z} = A\vec{X}$.*

Immediate consequences are:

- Coefficients of a normal vector are normally distributed. This follows from the fact that the linear combination of independent normals is again normal.
- Linear combinations of the components of a normal vector are normal again.
- The probability density of a normal vector depends only its covariance matrix and expectation.

5.1 Simple structure of conditional probability of normal vector

For any random variables we have that if X and Y are independent, then $COV(X, Y) = 0$. But, in general the reverse implication is not true: there are variables with covariance 0 which are not independent. However, for normal variables when the covariance is 0, they must also be independence. This is the content of the next lemmma:

Lemma 5.1 *Let X and Y be jointly normal. Then if $COV(X, Y) = 0$ we have that X and Y are independent.*

Proof. Assume that X and Y are jointly normal. Then X and Y both have a normal distribution so that $\vec{X} = (X, Y)$ is a normal vector. Let us simulate two independent normals $\mathcal{N}_1$ and $\mathcal{N}_2$. We take the standard deviation and expectation of these two normals so that:

$$\sigma_{\mathcal{N}_1} = \sigma_X, E[X] = E[\mathcal{N}_1]$$

and

$$\sigma_{\mathcal{N}_2} = \sigma_Y, E[Y] = E[\mathcal{N}_2]$$

Note that then $\mathcal{N}_1$ and $\mathcal{N}_\in$ are jointly normal and hence

$$\vec{\mathcal{N}} = (\mathcal{N}_1, \mathcal{N}_2)$$

is a normal vector. The covariance matrix of $\vec{\mathcal{N}}$ is given by

$$COV[\vec{\mathcal{N}}] = \begin{pmatrix} VAR[\mathcal{N}_1] & 0 \\ 0 & VAR[\mathcal{N}_2] \end{pmatrix} = \begin{pmatrix} VAR[X & 0] \\ 0 & VAR[Y] \end{pmatrix} = COV[\vec{X}]$$

where we used the fact that the covariance of $\mathcal{N}_1$ and $\mathcal{N}_2$ must be 0 since they are independent of each other. So, $\vec{N}$ and $\vec{X}$ have the same covariance matrix. They also have the same expectation. Since they are both normal vectors they must have the same distribution. Indeed for normal vectors the distribution only depends on the expectation and the covariance matrix. But, $\mathcal{N}_1$ and $\mathcal{N}_2$ are independent of each other. Since, X and Y have the same joint distribution than $\mathcal{N}_1$ and $\mathcal{N}_2$, we find That X and Y must also be independent of each other. ■

The above lemma allows to decompose a normal vector into independent parts. For this say $\vec{X} = (X, Y)$ is a normal vector with 0 expectation. Let U be equal to

$$U = Y - X\frac{COV(X, Y)}{COV(X, X)}.$$

Then we can see that U and X are uncorrelated:

$$COV(X, U) = COV(X, Y - X\frac{COV(X, Y)}{COV(X, X)}) = COV(X, Y) - COV(X, X)\frac{COV(X, Y)}{COV(X, X)} = 0$$

Hence, X and U have covariance 0, they must be independent. But, note that we can now write Y as

$$Y = U + aX$$

where a is the constant

$$a := \frac{COV(X,Y)}{COV(X,X)}.$$

So basically this is to say, that Y is obtained from X, by multiplying X by a constant and adding an independent normal part. This is a very simple "joint probability structure". In non-normal joint distributions, there can be way more complicated dependencies where the distribution of Y depends in a very complicated manner on X. But for joint normal, the conditional distribution of Y given X is simple. Thus, we have that there exists a non-random constant a which does not depend on x, so that:

$$\mathcal{L}(Y|X = x) = ax + \mathcal{L}(\mathcal{N}(0, \sigma^2))$$

where $\mathcal{L}(Y|X = x)$ stands for the conditional distribution of Y given $X = x$ and $\mathcal{N}(0, \sigma^2)$ is a normal random variable with expectation 0 and variance σ^2. If we want to simulate the random vector (X, Y), we can thus first simulate X. Once we have the value x of X, we then simulate Y by simulating an independent normal with expectation 0 and adding it to ax.
Let us think of an example: say X is the height of a human being and Y is the breadth of his/her hip. Then, a would represent "a ratio between hip-breadth and height". Then you could figure out the expected hip-breadth of an individual if you know his/her height being x: the expected hip breadth is then ax, and his actual hip breadth is obtained from there by adding a random error term U with 0 expectation. If there are different "types of body structures" in the population then obviously this model does not apply. There could for example be five body types corresponding to five such ratios: a_1, a_2, a_3, a_4, a_5. To model the hip-breadth, given the height, we would first throw a five sided die to determine the coefficient a_i. Then we would add the random error term U, so as to get $Y = a_I x + U$, where $I \in \{1, 2, 3, 4, 5\}$ is the random variable corresponding to selecting the body type. **When we also select the coefficient a at random, then we do not have a joint normal distribution! Instead we would have a mixture of normals. That means that given the coefficient a_i, the conditional distribution of the hip-breadth is normal. Hence, once the coefficient a_i is determined, then the hip breadth conditional on the height is normal.**

In principal, we are going to decompose a sequence of jointly variables into independent parts. This is the same idea as Graham Schmidt decomposition for vectors: say you have three (non-random) vectors $\vec{x}, \vec{y}, \vec{z}$. Then one can find orthogonal unit vectors $\vec{u}_1, \vec{u}_2, \vec{u}_3$ so that:
I) $\vec{x}$ and $\vec{u}_1$ are co-linear: there exists a coefficient b_{11} so that $\vec{x} = b_{11}\vec{u}_1$. Indeed, since $\vec{u}_1$ is a unit vector,

$$|\vec{x}| = |b_{11}\vec{u}_1| = |b_{11}| \cdot |\vec{u}_1| = |b_{11}|.$$

Hence, b_{11} is the renormalization coefficient for $\vec{x}$.
II) There exist coefficients b_{21} and b_{22} so that

$$\vec{y} = b_{21}\vec{u}_1 + b_{22}\vec{u}_2. \tag{5.5}$$

In other words, $\vec{u}_2$ corresponds to the direction of orthogonally projecting $\vec{y}$ onto the line defined by $\vec{x}$. The coefficient b_{21} is determined by multiplying equation 5.5 by $\vec{u_1}$ yielding

$$\vec{y} \cdot \vec{u}_1 = b_{21}\vec{u}_1 \cdot \vec{u}_1 = b_{21}.$$

From there we determine $\vec{u}_2$ by first calculating

$$b_{22}\vec{u}_2 = \vec{y} - b_{21}\vec{u}_1$$

and then remormalising the right side of the last equation above. This then yields $\vec{u}_2$. (Recall that renormalizing simply means to divide by the norm of a vector. This corresponds to taking the vector of

same direction but with length 1.)
III) There exists coefficients b_{31}, b_{32}, b_{33} so that

$$\vec{z} = b_{31}\vec{u}_1 + b_{32}\vec{u}_2 + b_{33}\vec{u}_3. \tag{5.6}$$

These coefficients b_{31} can be determined by multiplying (dot product) equation 5.6 by $\vec{u}_1$:

$$\vec{z} \cdot \vec{u}_1 = b_{31}\vec{u}_1 \cdot \vec{u_1} + b_{32}\vec{u}_2 \cdot \vec{u}_1 + b_{33}\vec{u}_1 \cdot \vec{u}_3 = b_{31}\vec{u}_1 \cdot \vec{u}_1 = b_{31},$$

where we used that $\vec{u}_1$ is orthogonal to $\vec{u}_2$ and $\vec{u}_3$. Similarly $b_{32} = \vec{z} \cdot \vec{u}_2$. Once one has determined the coefficients b_{31} and b_{32}, then we get the direction of projecting $\vec{z}$ onto the plan spanned by $\vec{x}$ and $\vec{y}$. This direction is given as:

$$b_{33}\vec{u}_3 = \vec{z} - b_{31}\vec{u}_1 + b_{32}\vec{y}.$$

The coefficient b_{33} is then the norm of the expression on the right side of the last equation above and $\vec{u}_3$ is obtained by renormalizing $\vec{z} - b_{31}\vec{u}_1 + b_{32}\vec{y}$ i.e. dividing that expression by its norm.

In matrix notation this leads to

$$\begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \\ \vec{u}_3 \end{pmatrix} = \begin{pmatrix} \vec{x} \\ \vec{y} \\ \vec{z} \end{pmatrix}$$

Next we will apply the same approach to our jointly normal random variables.

The covariance behaves like the dot product: it is bilinear, and positive. So, we can view it like a dot product: uncorrelated random variables can be seen as orthogonal. Hence, we apply the same scheme as with vectors to our random variables. So, let X, Y and Z be three jointly normal random variables with expectation 0. Then, there exists three uncorrelated (and hence independent) standard normal variables U_1, U_2, U_3 so that there exists a lower triangular matrix B of non-random coefficients

$$B = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \tag{5.7}$$

so that

$$\begin{array}{ll} b_{11}U_1 & = X \\ b_{21}U_1 + b_{22}U_2 & = Y \\ b_{31}U_1 + b_{32}U_2 + b_{33}U_3 & = Z \end{array} \tag{5.8}$$

To determine these coefficients proceed inductively. First b_{11} is simply equal to σ_X. (In other words,

$$U_1 := X/\sigma_X \tag{5.9}$$

. Note that dividing any random variable by its finite standard deviation, makes the standard deviation equal to 1).
Then, take the covariance of the second equation with U_1 yielding

$$b_{21}COV(U_1, U_1) + b_{22}COV(U_2, U_1) = COV(Y, U_1)$$

and hence $b_{21} = \frac{COV(Y,U_1}{COV(U_1,U_1)} = COV(Y, U_1)$. Having determine b_{21}, we get that $b_{22}U_2$ can be determined by the following equation:

$$b_{22}U_2 = Y - b_{21}U_1$$

From there U_2 is determine by renormalizing $b_{22}U_2$, that is dividing by its standard deviation:

$$U_2 = \frac{Y - b_{21}U_1}{\sqrt{VAR[Y - b_{21}U_1]}}$$

(Note that the variance is the covariance of the variable with itself. So, using equation 5.9 and the fact that the variance is the covariance of the variable with itself, we get:

$$VAR[Y - b_{21}U_1] = VAR[Y - \frac{b_{21}}{\sigma_X}X] = COV(Y - \frac{b_{21}}{\sigma_X}X, Y - \frac{b_{21}}{\sigma_X}X) = COV(Y,Y) - 2\frac{b_{21}}{\sigma_X}COV(X,Y) + b_{21}.)$$

The next step, is to determine b_{31} and b_{32} by taking the covariance with U_1 resp. U_2 of the equation $b_{31}U_1 + b_{32}U_2 + b_{33}U_3 = Z$. With the fact that the U_i's are uncorrelated to each other, we get

$$b_{31} = COV(Z,X) \ , \ b_{32} = COV(Y,U_2).$$

Then $b_{33}U_3$ is determined by

$$b_{33}U_3 = Z - b_{31}U_1 + b_{32}U_2 \tag{5.10}$$

and U_3 is determine from there by renormalising:

$$U_{33} = \frac{Z - b_{31}U_1 + b_{32}U_2}{\sqrt{VAR[Z - b_{31}U_1 + b_{32}U_2]}}$$

whilst b_{33} is the standard deviation of the right side of 5.10. So, we have in matrix notation:

$$B\vec{U} = \vec{X} \tag{5.11}$$

where $\vec{U} = (U_1, U_2, U_3)^T$ and $\vec{X} = (X, Y, Z)^T$. Also, the matrix B is given in 5.7. We can multiply both sides of equation 5.11 by B^{-1} and obtain:

$$\vec{U} = A\vec{X} \tag{5.12}$$

where A designates the inverse B^{-1} of the matrix B. Again, A is lower-triangular. So, in other words we have coefficients a_{ij} which are non random so that

$$\begin{aligned} U_1 &= a_{11}X \\ U_2 &= a_{21}X + a_{22}Y \\ U_3 &= a_{31}X + a_{32}Y + a_{33}Z \end{aligned}$$

here

$$A = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

This means that if we know for example the values of X and Y, then Z is obtained by taking a linear combination of these values and adding a normal error term with 0 expectation. This follows from the equation:

$$Z = \frac{1}{a_{33}}U_3 - \frac{a_{31}}{a_{33}}X - \frac{a_{32}}{a_{33}}Y.$$

so, if we know that X took as value x and Y took y, then Z is obtained by adding the independent normal term $\frac{U_3}{a_{33}}$ to the linear combination

$$-\frac{a_{31}}{a_{33}}x - \frac{a_{32}}{a_{33}}y. \tag{5.13}$$

This means that to simulate Z once we have simulated X and Y and obtained the values x and y for them, we proceed as follows:
we simulate independently on what the values x and y are a normal with expectation 0 and standard deviation $1/a_{33}$. Then we add this term to expression 5.13 in order to obtain a value for Z.

The same approach for normal random vectors with more than three entries. this is the the result of the next lemma:

Lemma 5.2 *Assume that $\vec{X} = (X_1, X_2, \ldots, X_n)^T$ is a normal vector with 0 expectation: $E[X_i] = 0$ for all $i = 1, 2, 3, \ldots, n$. Then there exists independent standard normal variables*

$$U_1, U_2, \ldots, U_n$$

and a lower triangular $n \times n$ matrix $A = (a_{ij})$ with non-random coefficients so that we have:

$$\vec{U} = A\vec{X}$$

where $\vec{U} = (U_1, U_2, \ldots, U_n)^T$. Hence U_i is independent of $X_1, X_2, \ldots, X_{i-1}$ and X_{i+1} is obtained from $X_1, X_2, \ldots, X_i$ by taking a linear combination with non-random coefficients:

$$X_i = \left(\sum_{j=1}^{i-1} X_j \frac{a_{ij}}{a_{ii}} \right) + \frac{U_i}{a_{ii}}.$$

So, the conditional distribution of X_{i+1} given $X_1, X_2, \ldots, X_{i-1}$ is normal with an expectation being a linear combination of the previous X_j's:

$$\mathcal{L}(X_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) = \mathcal{L}\left(\mathcal{N}(\mu = \sum_{j=1}^{i-1} x_j \frac{a_{ij}}{a_{ii}}, \sigma^2 = \frac{1}{a_{ii}^2} \right).$$

Proof. ■

6 Linear discriminant analysis

Assume that we find a skeleton from a person that has been murdered. We are only given the height and the breadth of the hip to guess if it was a man or a women. the height say was 171.4 and the hip measurement for the diseased was 26. To solve our mystery, we are given the measurements of ten people in the same community and whether they are man or women. This ten people are our training data. Here is the data:

Hip	Height	Gender
26.5	171.5	1
28.6	173.0	1
29.3	176.0	1
27.5	176.0	1
28.0	180.5	1
28.7	167.6	0
25.9	154.9	0
31.5	175.3	0
27.5	171.4	0
26.8	157.5	0

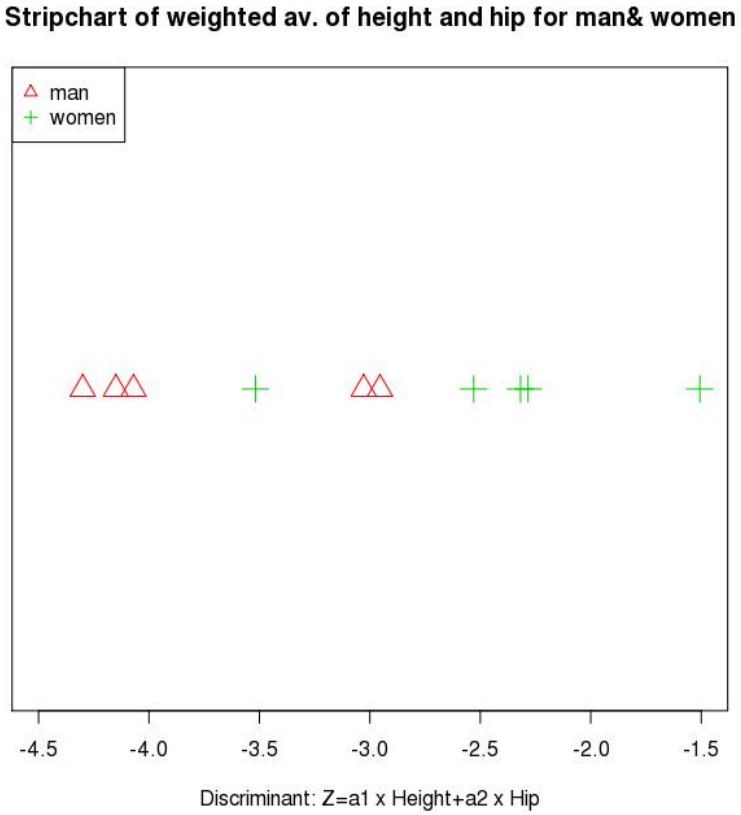
Here 1 stands for man, and 0 for women. We could try to guess if your skeleton is a man or a women just based on the height. But we may feel that it would be safer to use both the hip measurements and the height. How do we proceed? Let $Hight$ be the height and Hip be the breath of the hip. Typically a taller person is likely to be a man. Whilst a broader hip tends to be associated with women. We want to find a linear function of the type $Z = a_1 Hip + a_2 Height$ where a_1 and a_2 are constant, so that base on this "Z-score" we can distinguish well between men and women. We take $a_1 > 0$ and $a_2 < 0$, because large Hip tends to imply a women, whilst on the opposite a large *height* tends to imply that it is a man. Also, a_1 and a_2 should bring hip and height to a similar scale since otherwise if one of them is much bigger, then the other wouldn't work. Indeed, if one is much smaller, then it would not have a lot of

effect, and in that case it would be like just using only one of the variables. So, let us take for example $a_1 = 0.62$ and $a_2 = -0.12$. This then leads to the following table:

a_1Hip + a_2Hight	Hip	Height	Gender
−4.15	26.5	171.5	1
−3.02	28.6	173.0	1
−2.95	29.3	176.0	1
−4.07	27.5	176.0	1
−4.3	28.0	180.5	1
−2.31	28.7	167.6	0
−2.53	25.9	154.9	0
−1.50	31.5	175.3	0
−3.51	27.5	171.4	0
−2.28	26.8	157.5	0

Now, we can look at a strip chart of Z to see if it separates women from men well. This can be seen in figure 2. Indeed it seems that men and women are well separated by Z. Take the rule $z < -2.95$ gives

Figure 2:



man, and with that rule you classify all but one point correctly in our training data. So, we can now apply this rule to the skeleton which was found. the hip was 26 and the height 171.4. This leads to a score of

$$a_1 \cdot 26 + a_2 \cdot 171.4 = -4.448$$

this value is clearly below -2.95, so we classify the skeleton as having belonged to a man.
Now, a training sample of ten is not enough in reality to have a good estimate for the misclassification probability. The procedure we showed here would work well provided we have enough training data.
In reality we will work with more points in the training data set. Also, we will "optimize" the coefficients a_1 and a_2 so that they separate the man from the women optimally in the following sence:
we calculate the values for constants a_1 and a_2 that are best in terms of separating man and women average of Z whilst maintain the intergroupe variance bounded.. That is we want the mean to be far away but the standard deviation for each group to be small. We assume at first that the covariance matrix for men and women is the same. So, this leads to the following optimization problem:
find constants a_1 and a_2 so as to maximize

$$E[a_1H + a_2B|Y = male] - E[a_1H + a_2B|Y = female]$$

under the constrain

$$VAR[a_1H + a_2B|Y = male] \leq constant$$

. Now,

$$VAR[a_1H+a_2B|Y = male] = a_1^2VAR[H|Y = male]+2a_1a_2COV(H, B|Y = male)+a_2^2VAR[B|Y = male]$$

Let the difference between the expected values of the two groups be designated by

$$\Delta\vec{\mu} = (\mu_1, \mu_2)$$

where

$$\mu_1 = E[H|Y = male] - E[H|Y = female]$$

and

$$\mu_2 = E[B|Y = male] - E[B|Y = female]$$

So, in other words, we want to find a_1, a_2 to maximize

$$h(a_1, a_2) = (a_1, a_2) \cdot \Delta\vec{\mu}$$

under the constrain

$$g(a_1, a_2) = a_1^2VAR[H|Y = male] + 2a_1a_2COV(H, B|Y = male) + a_2^2VAR[B|Y = male]$$

is constant. We are going to solve this problem by using Lagrange multipliers. For this, we have to calculate the gradient of h and the gradient of g and set them to be collinear. So, we find the gradient of h to be equal to:

$$\vec{\mathtt{grad}}\ h = (\mu_1, \mu_2)$$

whilst

$$\vec{\mathtt{grad}}\ g = (a_1, a_2), \begin{pmatrix} COV(H,H) & COV(H,B) \\ COV(B,H & COV(B,B) \end{pmatrix}$$

So, setting the two gradients to point in the same direction, yields

$$\vec{\mathtt{grad}}\ g = \lambda\vec{\mathtt{grad}}\ h$$

for a constant λ. This yields

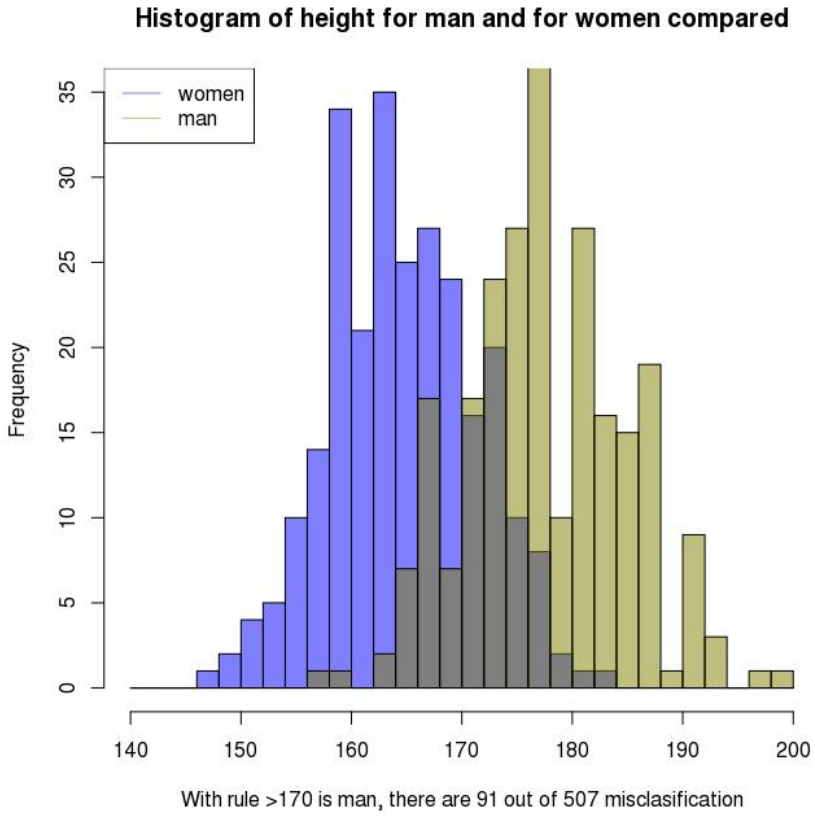
$$(a_1, a_2) = (\mu_1, \mu_2) \begin{pmatrix} COV(H,H) & COV(H,B) \\ COV(B,H) & COV(B,B) \end{pmatrix}^{-1} \tag{6.1}$$

where we only need to determine the vector (a_1, a_2) up to a constant factor. Now, the covariance and the difference in expectations are not exactly known. So, instead we will take their estimates and put them into formula 6.1:

$$(\hat{a}_1, \hat{a}_2) = (\mu_1, \mu_2) \begin{pmatrix} C\hat{O}V(H,H) & C\hat{O}V(H,B) \\ C\hat{O}V(B,H) & C\hat{O}V(B,B) \end{pmatrix}^{-1}$$

Now assume that instead of two measurements like hip width and height with have a whole collection of them. We can measure many things from cranial dimensions, to wrist. Say we would have a rather big data set with maybe about 500 women and men. Then, we could try to discriminate using height and a hip parameter. But, typically we would expect that as we add more of the measurements, the separation between women and men becomes better and better until it is close to 100%. The reason is that after all when we add enough information it should become possible to tell if we are dealing with a man or a women. so, first we take as discriminant function only the height. The result can be seen in figure 3. In figure 1, we have that 91 out of 507 are classified's. That gives a percentage of about 17%. Thus, if we use only height to discriminate between women and men, we estimate that the classification probability is about 17%. Then, we use all the 24 variables available to us. Now, with two variables it is often possible to find which linear combination makes sense for discrimination without big math formula. But with 24 variables, we need 24 four coefficients. So, it is better to use our "official" formula based on the inverse of the covariance matrix. We did it and found an almost perfect separation between men and women with less than 1 percent error. This can be seen in figure 4.

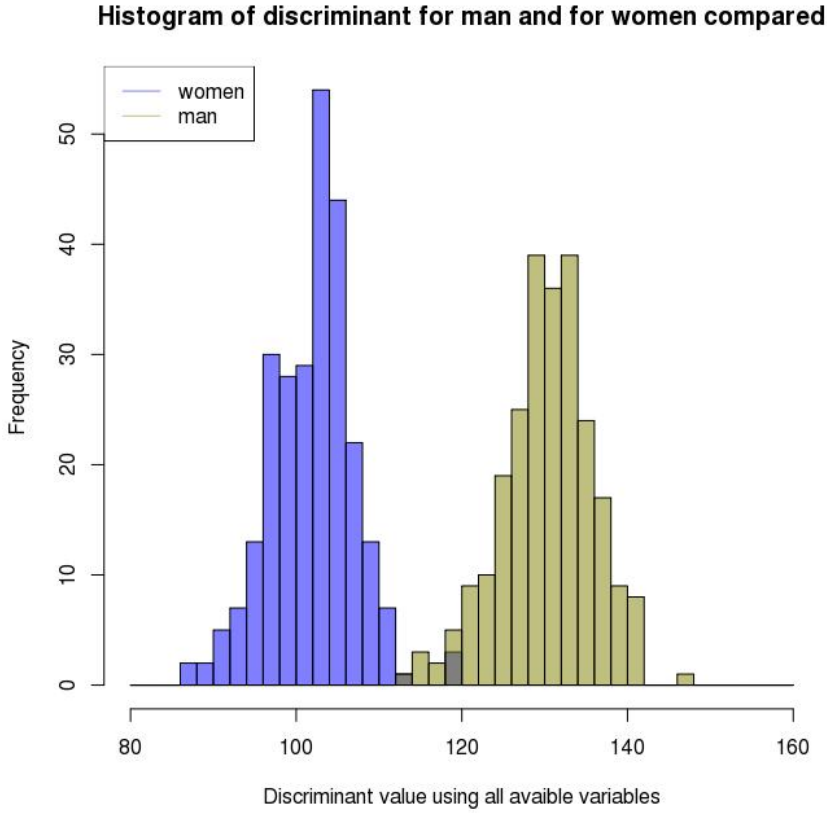
Figure 3:



then, we use the linear discriminant for all the 23 variables available. This is then much more powerful as can be seen in 4

Consider the case of a data of measurements of about 500 man and women. Different parameters where measured. When we run a linear discriminant with all the available measurement we find:

Figure 4:



7 A first application of the spectral method: neighborhood detection

Assume that we record the time people spend with each other on the phone. So, if we have n people we will record that information in a $n \times n$-matrix. We consider the problem where there are subgroups in the community which are not known to us. By looking at the matrix with the phone call time should allow us to tell if there are such subgroups which are closer to each other. Assume there are 10 people the police investigates. Say the expected time during a week they spend on the phone with each other is

given in the following matrix

$$\Sigma = \begin{pmatrix} 9 & 9 & 9 & 9 & 9 & 0 & 0 & 0 & 0 & 0 \\ 9 & 9 & 9 & 9 & 9 & 0 & 0 & 0 & 0 & 0 \\ 9 & 9 & 9 & 9 & 9 & 0 & 0 & 0 & 0 & 0 \\ 9 & 9 & 9 & 9 & 9 & 0 & 0 & 0 & 0 & 0 \\ 9 & 9 & 9 & 9 & 9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 16 & 16 & 16 & 16 & 16 \\ 0 & 0 & 0 & 0 & 0 & 16 & 16 & 16 & 16 & 16 \\ 0 & 0 & 0 & 0 & 0 & 16 & 16 & 16 & 16 & 16 \\ 0 & 0 & 0 & 0 & 0 & 16 & 16 & 16 & 16 & 16 \\ 0 & 0 & 0 & 0 & 0 & 16 & 16 & 16 & 16 & 16 \end{pmatrix}$$

So we see that the first five people communicate with each other 9 minutes on average and the people 6 to 10 communicate with each other is 16 minutes. Between these two groups there is 0 expected communication time. Now, when we consider the matrix Σ there are only two eigenvectors with non-zero eigenvalues. These eigenvectors are given by

$$\vec{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \vec{x}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

with corresponding eigenvalues $\lambda_1 = 45$ and $\lambda_2 = 80$. Now clearly the two eigenvectors $\vec{x}_1$ and $\vec{x}_2$ corresponds each to a group of people who communicate with each other a lot in our model. So, if we would be given the eigenvectors $\vec{x}_1$ and $\vec{x}_2$ we could from there determine which people communicate with each other a lot. But, why would that be needed? Indeed one could just look at the matrix Σ to see which group of people communicate a lot with each other. But, here is the deal: in general we do not directly observe the matrix Σ, which is the matrix of expected times people spend speaking to each other. So, the actual time can fluctuate. And hence in general we will have that what we observe is the actual time people speak to each other given by

$$\Sigma + E$$

where in the present case E is a symmetric matrix with independent entries above the diagonal with 0 expectation so that

$$E[\Sigma + A] = E[\Sigma] + E[E] = \Sigma.$$

So, we started with simulation a "noise" matrix with entries that are independent of each other in the triangle above the diagonal and the entries have 0 expectation. The matrix we got is as follows:

$$E = \begin{pmatrix} 4 & -2 & 2 & 0 & -6 & 2 & 6 & -8 & 0 & 3 \\ -2 & -8 & 1 & -2 & -3 & -1 & 7 & -3 & 2 & -4 \\ 2 & 1 & 11 & -2 & -4 & 4 & -2 & -5 & 13 & 6 \\ 0 & -2 & -2 & 4 & -5 & -3 & -2 & 3 & -4 & -2 \\ -6 & -3 & -4 & -5 & 0 & -4 & 3 & 4 & -10 & -3 \\ 2 & -1 & 4 & -3 & -4 & -8 & 2 & -5 & -2 & -4 \\ 6 & 7 & -2 & -2 & 3 & 2 & 3 & -3 & 1 & -3 \\ -8 & -3 & -5 & 3 & 4 & -5 & -3 & 0 & 8 & 4 \\ 0 & 2 & 13 & -4 & -10 & -2 & 1 & 8 & 12 & 4 \\ 3 & -4 & 6 & -2 & -3 & -4 & -3 & 4 & 4 & 0 \end{pmatrix}$$

Then we add to the original matrix of expected phone time the noise matrix and get

$$\Sigma + E = \begin{pmatrix} 13 & 7 & 11 & 9 & 3 & 2 & 6 & -8 & 0 & 3 \\ 7 & 1 & 10 & 7 & 6 & -1 & 7 & -3 & 2 & -4 \\ 11 & 10 & 20 & 7 & 5 & 4 & -2 & -5 & 13 & 6 \\ 9 & 7 & 7 & 13 & 4 & -3 & -2 & 3 & -4 & -2 \\ 3 & 6 & 5 & 4 & 9 & -4 & 3 & 4 & -10 & -3 \\ 2 & -1 & 4 & -3 & -4 & 8 & 18 & 11 & 14 & 12 \\ 6 & 7 & -2 & -2 & 3 & 18 & 19 & 13 & 17 & 13 \\ -8 & -3 & -5 & 3 & 4 & 11 & 13 & 16 & 24 & 20 \\ 0 & 2 & 13 & -4 & -10 & 14 & 17 & 24 & 28 & 20 \\ 3 & -4 & 6 & -2 & -3 & 12 & 13 & 20 & 20 & 16 \end{pmatrix}$$

the two eigenvectors corresponding to the biggest eigenvalues are

$$\hat{\vec{x}}_1 = \begin{pmatrix} -0.51 \\ -0.35 \\ -0.61 \\ -0.38 \\ -0.22 \\ 0.01 \\ -0.06 \\ 0.21 \\ -0.02 \\ 0.03 \end{pmatrix}, \hat{\vec{x}}_2 = \begin{pmatrix} -0.03 \\ -0.02 \\ -0.13 \\ 0.04 \\ 0.06 \\ -0.34 \\ -0.40 \\ -0.43 \\ -0.57 \\ -0.44 \end{pmatrix}$$

We put the hat on the eigenvector because the eigenvectors of the perturbated matrix can be viewed as estimates of the eigenvectors of the non-perturbated matrix.
Eigenvectors are defined only up to multiplication by a constant. This means that when we multiply an eigenvactor by a non-zero scalar, we get again an eigenvector with the same eigenvalue. Now note that the eigenvectors of the precturbated matrix $\Sigma + E$ are close to the eigenvectors of the unperturbed matrix Σ. But, we could also go into the matrix $\Sigma + E$ and take a column: these columns are the columns of Σ with added noise. And the columns of Σ in the present case are the eigenvectors. So, what is better: taking the eigenvectors of $\Sigma + E$ or the columns of Σ in order to figure out the original eigenvectors? (Again the original eigenvectors tell us which group of people talk to each other a lot). In our case, to compare the, two we are going to multiply each by a factor so as to get them close to the corresponding eigenvector. Otherwise they would not be comparable. So, let us do this with the second eigenvector:

the unperturbated eigenvector is

$$\vec{x}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

We multiply the eigenvector $\hat{\vec{x}}_2$ by the coefficient -2.1748 (found by linear regression) to get a comparable vector:

$$-2.1742 \cdot \hat{\vec{x}}_2 = \begin{pmatrix} 0.06 \\ 0.04 \\ wh0.28 \\ -0.08 \\ -0.13 \\ 0.73 \\ 0.86 \\ 0.93 \\ 1.23 \\ 0.95 \end{pmatrix}$$

Instead, as mentioned, we could also have taken any of the column 6 to 10 in the matrix $\Sigma + E$. Let us take for example the 9-th column $C9$:

$$C9 = \begin{pmatrix} 0 \\ 2 \\ 13 \\ -4 \\ -10 \\ 14 \\ 17 \\ 24 \\ 28 \\ 20 \end{pmatrix}$$

and we multiply $C9$ by the factor 0.0409 (which we found by linear regression) in order to approximate the eigenvector $\vec{x}_2$. This yields:

$$0.0409 \cdot C9 = \begin{pmatrix} 0 \\ 0.08 \\ 0.52 \\ -0.16 \\ -0.4 \\ 0.56 \\ 0.69 \\ 0.97 \\ 1.13 \\ 0.81 \end{pmatrix}$$

We can now compare which one of the two $C9 \cdot 0.0409$ or $-2.1742 \cdot \vec{x}_2$ comes closer to the eigenvector $\vec{x}_2$. We compute the standard deviation of the entries of the difference between each of them and the

original unperturbed eigenvector $\vec{x}_2$. We find:

$$sd(0.0406 \cdot C9 - \vec{x}_2) = 0.29, sd(-2.1748 \cdot \hat{\vec{x}}_2 - \vec{x}_2) = 0.16$$

We see the eigenvector of the perturbated matrix is almost twice closer to the origina eigenvector $\vec{x}_2$. We will see that **in general with a finite structure and a random noise with independent entries and 0 expectation, the precision is improved by a factor of order constant times $\sqrt{n}$. Here $\sqrt{n}$ denotes the size of the matrix**. We will have to define what we mean by finite structure and precision gets improved by a factor $\sqrt{n}$. The eigenvector of the perturbed matrix is better than a column of $\sigma + E$ for recovering the eigenvectors. Another way to see this in our example is if we round of to the closest integer the entries in our vectors which are supposed to approximate $\vec{x}_2$. We find:

$$\texttt{round}(0.0409 \cdot \texttt{C9}) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad , \quad \texttt{round}(-2.1742 \cdot \hat{\texttt{x}}_2) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

We see that rounding of $-2.1742 \cdot \hat{\vec{x}}_2$ we recover $\vec{x}_2$ exactly, whilst with the column $C9$ times 0.0409 we still get an error. In general, with bigger matrices this effect will be even more dramatic: from the column we will not be able to recover the eigenvectors at all, whilst with the eigenvalues we will.
Let us next see an example with a somewhat bigger matrix:

7.1 An example with a bigger matrix

Let us assume that there are 24 people whos phone call we record with two groups of twelve which communicate with each other a lot. The matrix of the expected time people communicate with each

other be given by

$$
\Sigma = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2
\end{pmatrix}
$$

Now we are going to add to Σ a noise matrix E. The noise matrix is symmetric and has i.i.d entries with expectation 0 above the diagonal. The eigenvectors are

$$\vec{x}_1 = (1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0)^T$$

and

$$\vec{x}_2 = (0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1)^T$$

with corresponding eigenvalues $\lambda_1 = 12$ and $\lambda_2 = 24$. Now from the theory of symmetric matrices we know that we can represent σ in terms of its rescaled eigenvectors and eigenvalues. We get

$$\Sigma = \frac{\lambda_1}{\vec{x}_1^2}\vec{x}_1 \cdot \vec{x}_1^T + \frac{\lambda_2}{\vec{x}_2^2}\vec{x}_2 \cdot \vec{X}_2^T$$

So, to reconstitute σ if we are only given $\Sigma + E$ we take the eigenvectors of $\Sigma + E$ and rewrite the above formula. This gives us something which is closer to Σ than $\Sigma + E$. In other words, we take as estimate for Σ the following

$$\hat{\Sigma} = \frac{\hat{\lambda}_1}{\hat{\vec{x}}_1^2}\hat{\vec{x}}_1 \cdot \hat{\vec{x}}_1^T + \frac{\hat{\lambda}_2}{\hat{\vec{x}}_2^2}\hat{\vec{x}}_2 \cdot \hat{\vec{x}}_2^T$$

where $\hat{\vec{x}}_1$ and $\hat{\vec{x}}_2$ are the two eigenvectors with biggest eigenvalues of $\Sigma + E$ and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the non-zero eigenvalues of $\Sigma + E$.

7.2 The basic theory which makes it all work

7.3 Does it make sense to use spectral methods at all

8 Closest neighbor classification

9 The Multivariate T-test

Assume that we shoot with an artillery gun again. Say as usual $\vec{X} = (X, Y)$ is the impact point of a shell. To simplify notation let us assume that we shoot in the direction of the vector $(1, 0)$ and hence (X, Y) are independent of each other. We also assume that both X and Y are normal. Say to start with that $\sigma_X = \sigma_Y = 1$. Summarizing:

Again, as before we assume a situation where the artillery gun shoots many rounds whilst staying in the exact same position the tube oriented the same way. Furthermore, the meteo conditions remain unchanged, so basically the rounds we shoot are i.i.d. Hence, $\vec{X}, \vec{X}_1 =, \vec{X}_2, \ldots$ ae i.i.d. where (X_i, Y_i) is the i-th impact point. Now assume that we are like in world war one: a fixed battlefield where the fronts to not move for weeks and the artillery guns remain in the same position for long times. Say, suddenly we observe an impact point $x = 3, y = 3$. We can use our rule of thumb that variables take value most of the times no further than 2 standard deviation from their expectation. here, both x and y are three standard deviation away from the expected value. (If it was our artillery gun who had been shooting). So, it is unlikely that it is our artillery gun which had been shooting, and we may designate that impact point as an outlier. Next consider a point which is closer to the average impact point $E[(X, Y)] = (0, 0)$. Say, $(x, y) = (1.8, 1.8)$. For this point each coordinate is closer than 2 standard deviation from the expected value. So, if we only look at each coordinate separately we would not classify the point as outlier. But, then again: both coordinates are pretty big, not exactly two, but not very far either. So, to have two pretty big value at the same time might have a relatively small over all probability. And this in terms, might allow us to classify this point as outlier maybe. More precisely, we are going to look at the points which are too far from the center to be likely to be shoot by our gun, so it must have been another one. So, we take the square distance of the impact point from $(E[X], E[Y])$. In our case, this is $X^2 + Y^2$. With X and Y being independent standard normal, the distribution of that squared distance is called chi-square distribution with 2 degrees of freedom. In general we define:

Definition 9.1 *Let $\mathcal{N}_1, \mathcal{N}_2, \ldots$ be a sequence of i.i.d. standard normals. Then,*

$$\mathcal{N}_1^2 + \ldots + \mathcal{N}_n^2$$

is called a Chi-square with n degree of freedoms. It represents the Euclidean distance squared of the random vector $(\mathcal{N}_1, \ldots, \mathcal{N}_n)$ to the origin.

Now, in our example, the distance square to the origin is

$$d^2 = x^2 + y^2 = 1.8^2 + 1.8^2 = 6.48$$

We can now go into a chi-square table with 2-degrees of freedom and find that with 95% probability the distance square to the origin is less than 5.991465:

$$P(\chi_2 \geq 5.991465) \leq 0.05$$

and hence

$$P(X^2 + Y^2 \geq 5.991465) \leq 0.05$$

So, the chance to get a distance square bigger than 6.48 is less than 5%. Hence it is unlikely that the point $(1.8, 1.8)$ was shot by our artillery gun. If we want to use formal statistical parlance, we would say

that we reject the hypothesis on the 5% significance level that our point $x = 1.8$ and $y = 1.8$ comes from our artillery gun. (Assuming of course, that $\sigma_X = \sigma_Y = 1$, $COV(X, Y) = 0$ and $E[X] = E[Y] = 0$. Also assuming normality). In other words, we are 95%-confident that it was not our artillery gun which shot this shell with impact point $(1.8, 1.8)$. We could thus classify, this point as an outlier, despite each single coordinate by itself not being big enough to qualify as an outlier.
Now, most of the time in artillery shooting we have $\sigma_X \neq \sigma_Y$. What would we do in that case. So assume again that X and Y are independent normal with expectation 0. But this time we have the covariance matrix:

$$COV[\vec{X}] = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}, (E[X], E[Y]) = (0, 0) \tag{9.1}$$

Say for example that $\sigma_X >> \sigma_Y$. Then the impact points will tend to lie in an ellipse and not in a circle anymore: the spread in the x-direction will be much bigger than the spread in the y direction. So, to check for an outlier it would not make sense to take the points outside a certain circle. Instead, we will take ellipses. here is the idea. when we take a normal with 0 expectation and divide it by its standard deviation we get a standard normal. In other words

$$(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y})$$

is a random vector with two independent standard normal entries. For such a random vector the distance square to the origin is a chi-square variable. So,

$$\frac{X^2}{\sigma_X^2} + \frac{Y^2}{\sigma_Y^2}$$

is a chi-square variable with two degrees of freedom. so, we have

$$P(\frac{X^2}{\sigma_X^2} + \frac{Y^2}{\sigma_Y^2} \leq 5.991465) = 0.95.$$

But, not that the set of points (x, y) satisfying

$$\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} \leq 5.991465 \tag{9.2}$$

constitutes an ellipse centered at the origin with principal direction $(1, 0)$ and $(0, 1)$. The ellipse cuts the x coordinate line at $\sigma_X \cdot \sqrt{5.6}$ and the y-coordinate axis at $\sigma_Y \cdot \sqrt{5.991465} \approx \sigma_Y \cdot 2.44$ so basically it is a blow up by $\sqrt{5.991465} \approx 2.44$ of the ellipse centered in 0 having the x and y axis as coordinates and having $2\sigma_X$ as maximum breadth and $2\sigma_Y$ as maximal height. Now note that the inequality 9.2 can also be written in matrix notation as

$$(X, Y)\Sigma^{-1} \begin{pmatrix} X \\ Y \end{pmatrix} \leq 5.991465$$

where Σ^{-1} designate the inverse of the covariance matrix. So, here the test statistic has a chi-square distribution with degree of freedom equal to the dimension of the vector. This formula remains valid, even in other coordinate system, then the coordinate system of the principal components.
We can now define formula a test: say we have a normal random vector $\vec{X} = (X_1, X_2, \ldots, X_p)^T$ with known covariance matrix Σ. Let $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ be a non-random vector (known to us). Assume we want to test the hypothesis H_0: $E[\vec{X}] = \vec{\mu}$ at the significance level $\alpha \in (0, 1)$.
then we simply calculate

$$\vec{X}^T \Sigma^{-1} \vec{X}. \tag{9.3}$$

We compare that value with the α-upperquantile of a chi-square with p degrees of freedom. If the value is bigger than the quantil, then we reject the hypothesis H_0.

The precious case is when we know the covariance matrix. In artillery shooting you have tables with the dispersion in shooting direction and perpendicular to it depending on the distance you shoot. So, if it is about your own artillery gun you may be able to indeed know the covariance matrix. In many cases, however you have to estimate the covariance matrix. In that case, you simply use the estimated covariance matrix instead of the true one in the formula 9.3. So, you calculate the value

$$\vec{X}^T \hat{\Sigma}^{-1} \vec{X}, \tag{9.4}$$

instead of the test statistic given in 9.3. Now to point to an outlier, this test statistic needs to be even a bit higher than for the case with the known covariance matrix. The reason is simple: if the estimated covariance matrix is very wrongly estimated, then this could lead to a high value of the test statistic 9.4 not because the impact point lies far away, but because $\hat{\sigma}$ is wrong. So, to manage to be highly confident that the impact point is indeed far from the center, we need to raise the bar a little. This means that for example in 2 dimensions when we want to have a significance level of 5% we can not just ask for 9.4 to be bigger than 5.991465. We will use another table called the T^2-table instead. With many data point there is however almost no difference between the true covariance matrix and the estimated one. So, in that case of many data-points, even with an estimated covariance matrix, we could use the chi-square distribution and for practical purposes, we would get the same result.
Now, if we observe only one impact point, of course we can not estimate the covariance matrix. So, in stead imagine the following situation: we are shooting at an enemy bunker with coordinate $\vec{\mu} = (\mu_X, \mu_y)^T = (200, 400)^T$ Assume that we do not know the covariance matrix $COV[\vec{X}]$ where $\vec{X} = (X, Y)^T$. We shoot at the point $\vec{\mu}$. Again, (X, Y) is the impact point. yesterday, we had the gun perfectly adjusted in the correct direction:

$$(E[X], E[Y]) = \vec{\mu}^T = (200, 400) \tag{9.5}$$

That was yesterday. We did not move the direction of the gun, but today in the morning when we want to start shooting, we don't know if 9.5 still holds. So, we may have to adjust the direction in which the gun is pointing. Why? Mainly because the meteorological conditions may change from one day to the next. So, you shoot several rounds and then based on that you may adjust your artillery gun or not. Hence, in the morning at dawn, first thing we shoot n rounds and get n impact points $(X_1, Y_1), \ldots, (X_n, Y_n)$. We then take the center of gravity of our round of shooting:

$$\texttt{Center of gravity} = (\bar{\texttt{X}}, \bar{\texttt{Y}})$$

where $\bar{X} = (X_1 + \ldots + X_n)/n$ and $\bar{Y} = (Y_1 + \ldots + Y_n)/n$. If the center of gravity is close to the target $(200, 400)$ we don't readjust. If it is far we need to readjust the artillery gun. but how far is far? In principle, what we are going to do is a multivariate T-test: we can for example take the 5%-confidence level. so, we test that $E[\bar{X}] = 200$ and $E[\bar{Y}] = 400$ on the 5% level. If we reject the hypothesis then we have to readjust the gun. Otherwise we leave it. Why does one do in real artillery life such a test? Because, due to the imprecision (dispersion) at least with classical artillery you rarely get exactly on target. Most of the time you will be a little off. This does however not mean, that your mean trajectory is off: this may just be due to the little impresistion called dispersion in artillery shooting. So, if every time your shells fall a small distance away from the target, you would readjust, then this would not be good: you would continuously readjust which is a lot of work. And you would readjust even when your expected impact point $(E[X], E[Y])$ are right on target. In other words, you readjust only when after shooting a first round there is significant evidence that the gun is badly adjusted, meaning

$$(E[X], E[Y]) \neq (200, 400).$$

Now, how do we do the test? Note that the average impact point of our round of shooting $(\bar{X}, \bar{Y})$ has same expectation as (X, Y):

$$E[\frac{X_1 + \ldots + X_n}{n} = \frac{E[X_1 + \ldots + X_n]}{n} = \frac{nE[X_1]}{n} = E[X_1]$$

and similarly $E[\bar{Y}] = E[Y]$ The covariance matrix gets divided by a factor n when we consider $(\bar{X}, \bar{Y})$ instead of (X, Y). So, basically we can just do our test using the center of gravity $(\bar{X}, \bar{Y})$ instead of (X, Y). The only difference will be that the covariance matrix gets divided by a factor n. And an advantage is now that in case we do not know the covariance matrix, we can still do the test by using the estimated covariance matrix. (With one point alone we can not estimated the covariance matrix). When we estimate the covariance matrix we need to take instead of the chi-square distribution with p degrees of freedom, a multivariate T^2-distribution, with n and p being he parameters of freedom. When n is very large we could just use the chi-square distribution with p-degrees of freedom and it shouldn't't make much of a difference. So, anyhow we calculate the following test statistic, called multivariate T^2-statistic:

$$T_0^2 := n \cdot (\bar{X} - 200, \bar{Y} - 400)\hat{\Sigma}^{-1} \begin{pmatrix} \bar{X} - 200 \\ \bar{Y} - 400 \end{pmatrix} \tag{9.6}$$

where $\hat{\Sigma}$ denotes the estimated covariance matrix. For n very large and p small, we can just use the critical value for a chi-square table with p degrees of freedom. Otherwise we need to use a multivariate T^2-table. This table is rare to find, because up to a linear transformation the T^2variable is equivalent to an F variable. But we have

$$T_0^2 = \frac{(n-1)p}{n-p} F_{p,n-p}$$

In other words you if you want the 95% quantil you go into the F table with p and $n-p$ degree of freedom. There you find that quantil which you multiply by the factor $(n-1)p/(n-p)$ to get the critical value for the multivariate T^2 statistic. That is this will be the number which you compare to the test statistic given in 9.6. If the test statistic exceeds the value, you reject the hypothesis that $E[X] = 200$ and $E[Y] = 400$ and you readjust. Otherwise, you don't need to readjust, because you don't have significant evidence that your artillery gun is not well adjusted (at the 5%-significance level.

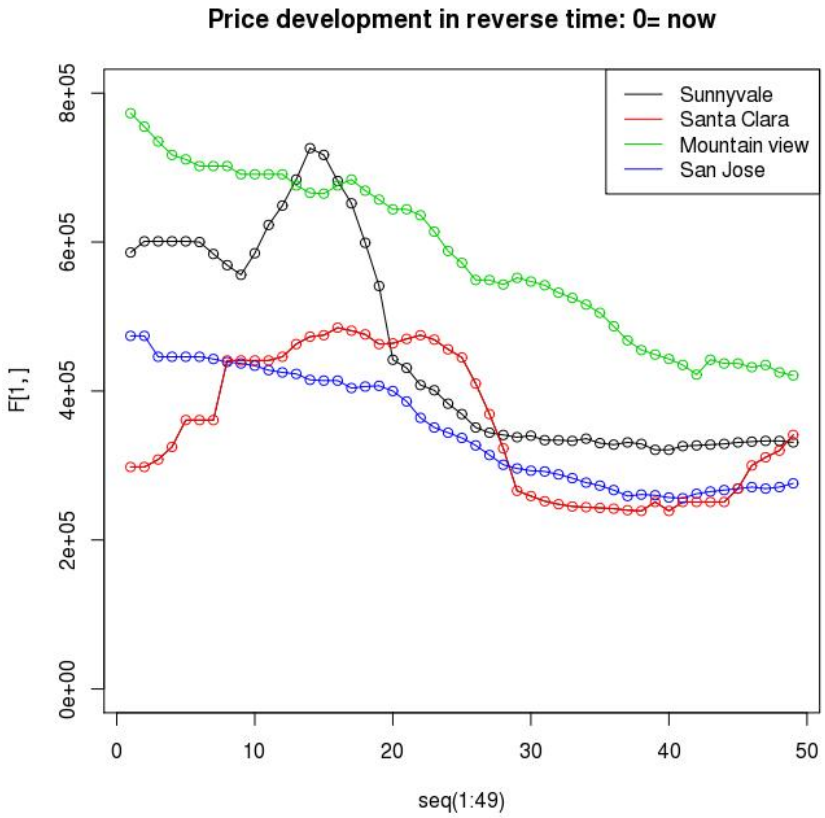
10 Singular value decomposition

In principal component analysis we consider the eigenvectors of the covariance matrix. So, we have a matrix with data X, then we had seen that $\frac{1}{n}XX^T$ is our estimate for the covariance matrix of the rows. For this we assumed the rows of X to be i.i.d. and have expectation $\vec{0}$. So, the estimated principal components of the covariance matrix are then the eigenvectors of XX^t. There are other situations however than covariance matrix estimation where these eigenvectors can be of great importance. Let us consider an example where X is a $n \times p$ matrix recording the house values in different zip-codes. Say we have one value for each month for each zip-code. Each row of X corresponds to a neighborhood and each row to one specific month. We can see a plot of this situation below in figure 5. In that figure basically San Jose and Mountain view follow pretty much the same trend, except Mountain view being more expensive. 5 So, for this take 1.72 times San Jose and you get almost the same plot as for Mountain view. This can be seen in figure **??** below. 6 in the current case San jose and Moutain view are almost the same up to a multiplicative coefficient: they follow the same trend, but one of this neighborhoods is more expensive then the other. So, there is basically one function $i \mapsto f(i)$ where $f(i)$ denotes the average housing price say between these two neighborhoods, so that both the price evolution for San Jose and moutain view can be obtained approximately from the function $f(.)$ by multiplying by a constant. So, there is a one-dimensional function behind both of these markets. Sometimes there could be more than one. for example, $f(.)$ could represent the general price trend in the US and $g(.)$ could represent employment in Silicon valley. Then, maybe all the price evolutions in these different neighborhoods could be approximated by linear combinations of $f(.)$ and $g(.)$. In, other words we would have eight coefficients c_1, c_2, c_3, c_4 and d_1, d_2, d_3, d_4 so that

$$X_{ij} \approx c_j f(i) + d_j g(i). \tag{10.1}$$

So, hence on the i month in the nighborhood j the avearge price is approximately $c_j f(i) + d_j g(i)$. The coefficents c_j and d_j vary from one neighborhood to the other, because these neighborhoods may depend

Figure 5:



to different degrees for example on the general market. Say at first to simplify a little bit, that the approximation 10.1 is not an approximation but holds exactly. Then we would have that:

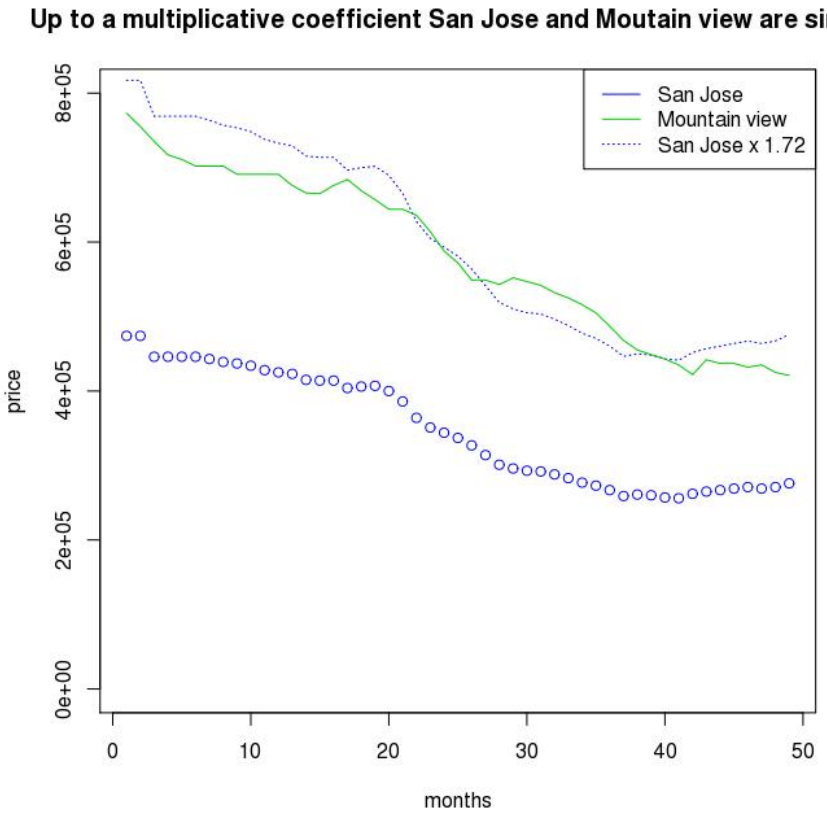
$$\vec{X}_j = c_j\vec{f} + d_j\vec{g} \tag{10.2}$$

where the vector $\vec{X}_j$ represents the j column of our matrix X. hence, $\vec{X}_j$ give the time evolution of the house prices in the neighborhood i. Furthermore, $\vec{f}$ is the column vector of length n with i-th entry equal to $f(i)$. Similarly, $\vec{g}$ is the column vector of length n with i-th entry equal to $g(i)$. So basically what we have is that the columns of X are located (approximately) in a two dimensional linear space spanned by $\vec{f}$ and $\vec{g}$. This implies then that the image space of $X \cdot X^T$ is generated by $\vec{f}$ and $\vec{g}$. hence the eigenvectors with non-zero eigenvalue of $X \cdot X^T$ are in the span of $\vec{f}$ and $\vec{g}$. So to try to find the functions $\vec{f}$ and $\vec{g}$, if we are only given X we can go for the eigenvectors $\vec{U}_1$ and $\vec{U}_2$ with non-zero eigenvalue of $X \cdot X^T$. We then have that $\vec{f}$ and $\vec{g}$ are both linear combination of $\vec{U}_1$ and $\vec{U}_2$. This is for the case that equation 10.2 holds exactly. If instead of equation 10.2, we have only an approximation, $\vec{X}_j \approx c_j\vec{f} + d_j\vec{g}$, then we will merely be able to approximate $\vec{f}$ and $\vec{g}$ by linear combinations of the eigenvectors.
Now let us consider the *Singular Value Decomposition* of X. This is how it is defined. Recall first that the matrix X has dimension $n \times p$, where $n \geq p$. One can always write the matrix X as a product of three matrices:

$$X = UDV^T$$

where D is diagonal matrix of dimension $p \times p$, U is of dimension $n \times p$ and V is of dimension $p \times p$. We also ask that the columns of U are orthonormal and same for V. That is we ask $U^T \cdot U = I$ where I

Figure 6:



denotes the $p \times p$ identity matrix. Similarly we request that V columns be an orthonormal basis. Then such a decomposition of X into a product of three matrices two of which are orthogonal, is called a Singular Value Decompositin (SVD). Now, we have

$$X \cdot X^T = UDV^T \cdot (UDV^T)^T = UDV^T VDU^T = UD^2U^T$$

In other words, if $\vec{U}_i$ denotes the i-th column of the matrix U we get that

$$X \cdot X^t = \sum d_{ii}^2 \vec{U}_i \cdot \vec{U}_i^T$$

this is the same formula as when we write the symmetric matrix $X \cdot X^T$ in terms of its eigenvectors and eigenvalues. Hence, we find that the columns of U are the normalized eigenvectors of $X \cdot X^T$. (There is always still a question about the sign in front of each eigenvector). Furthermore, d_{ii}^2 is then an eigenvalue of $X \cdot X^T$. Hence, the diagonal matrix D is obtained by putting the square roots of $X \cdot X^T$ into the diagonal. Usually we write the eigenvalues in decreasing order, and put the eigenvector also in the same corresponding order.
Now a similar argument shows that the columns of V are eigenvectors of $X^T X$:

$$X^T \cdot X = (UDV^T)^T (UDV^T) = VDU^T UDV^T = VD^2V^T = \sum_{i=1}^{p} d_{ii}^2 \vec{V}_i \tag{10.3}$$

where the vector $\vec{V}_i$ denotes the i-th colum of the matrix V. Clearly, given the right most expression of 10.3, we get that the vectors $\vec{V}_i$ are eigenvectors of $X^T \cdot X$.

so, now say that we are back in the previous example with equation 10.2 holding approximately for every colum j. Assume also that $\vec{f}$ and $\vec{g}$ are the eigenvectors. Then, the matrix U first two and second colum are equal to $\vec{f}$ and $\vec{g}$. The other columns will be eigenvectors which correspond to much smaller eigenvalue, that is eigenvalue which are just noise and which we can leave out in a first approximation. Now, when you consider the equation

$$X = UDV^T = U(DV^T)$$

you see that the matrix DV^T gives us the coefficients c_j and d_j for each column j. So, if we want to for example keep only two eigenvalues, then we compute instead of X the approximation matrix:

$$X_{II} = UD_{II}V^T$$

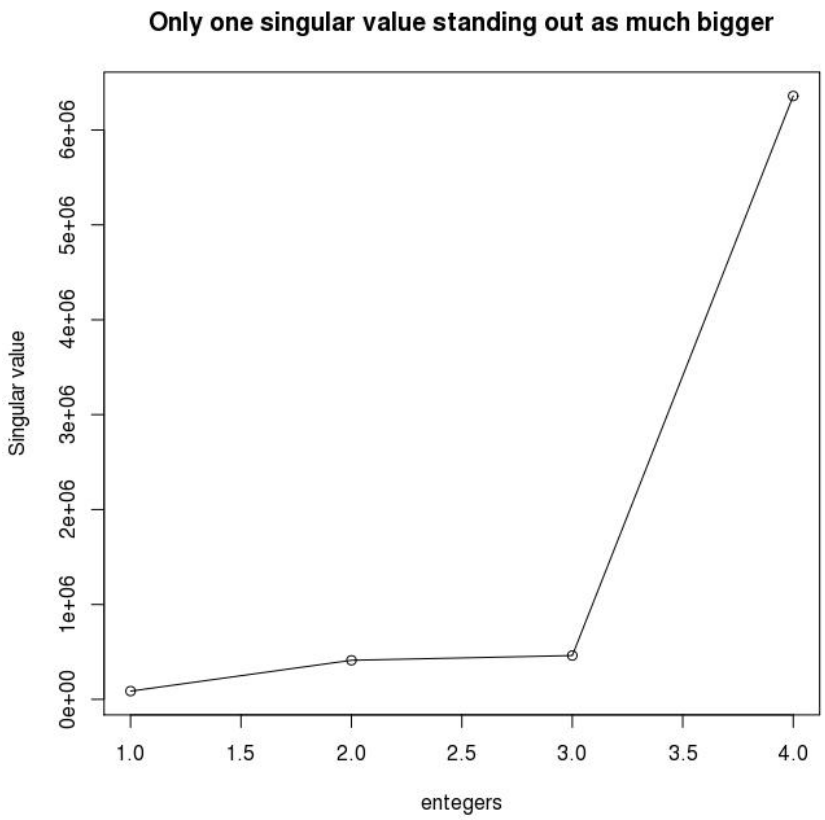
where D_{II} is the diagonal matrix which is obtained from D by keeping only the two biggest eigenvalues. So, in this situation the columns of X_{II} will be approximation of teh columns in X. So for each neighborhood we will have such an approximation. Why would that be good for? Usually we approximate something we don't know and which would be difficult to calculate. But here we are given the matrix X. Well here is the reason: say in one neighbor hood in month there were only three houses sold. And in that month by chance these three houses were all foreclosure which makes the market price in that neighboor hood look smaller than what it would be if we would buy another house. So, to some extend, the hope is that X_{II} can be more accurate than X and to some degree eliminate such a dip which is not corresponding to the true enighborhood price, but which is just due to a momentarily noise. To know how many such underlying functions like $f(.)$ and $g(.)$, we look and plot the singular values of D. (these are the square roots fo the eigenvalues of $X \cdot X^t$). Then we see how many singular values there are which are much bigger than the others. In our present case in figures **??** there is only one which stands out. (Ellbow of polygone). **??**

So, we decide to go for an approxiamtion leaving only one eigenvalue in D. Hence, we go for D_I which is obtained from D by leaving only the biggest eigenvalue. Then, our "denoised" X denoted by X_I is given by the formula:

$$X_I = UD_IV^T$$

. In figure **??** below, we can compare the original X and the one obtained by using only the one biggest eigenvalue for the approximation. (I do not yet add this figure because it is the project which is part fo the final for extra credit).

Figure 7:


Only one singular value standing out as much bigger
Singular value
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
1.0
1.5
2.0
2.5
3.0
3.5
4.0
entegers