INTRODUCTION TO MULTIVARIATE STATISTICAL ANALYSIS

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

April 1, 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:

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

(when the conditional probabilites are exactly equal, we could classify either way. Here, we chose to asign the fish in case of equal probabiloity to the tuna class).
Recall the formula for conditional probability of A given B. This formula ois 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

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

and

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

. Hence,

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

Applying this last equation to 1.2 yields:

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

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:

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

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:

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

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: $z \geq \textit{constant}$ 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 $z \geq \textit{constant}$ 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

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

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 eignevalue λ_1. We assume $\vec{u} = (u_1, u_2)$ to have unit length: $u_1^2 + u_2^2 = 1$. Then,

$$VAR[\vec{X} \cdot \vec{u}] =$$
$$= COV(Xu_1 + Yu_2, Xu_1 + Yu_2) = u_1^2 COV(X,X) + 2u_1u_2 COV(X,Y) + u_2^2 COV(Y,Y) =$$
$$= (u_1, u_2) \begin{pmatrix} COV(X,X) & COV(X,Y) \\ COV(Y,X) & COV(Y,Y) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = (u_1, u_2)\lambda_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \lambda_1(u_1^2 + u_2^2) = \lambda_1.$$

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 artillerie 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 avearges. the estimates of somthing 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

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

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

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

and

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

3 Example

4 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:

$$f_{(H,R)}(\vec{x}) = \frac{1}{2\pi\sigma_H\cdot\sigma_R}\exp(-0.5(\vec{x}-\vec{\mu})^T\Sigma_{\vec{X}}^{-1}(\vec{x}-\vec{\mu})) \tag{4.1}$$

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 combinations 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

$$Z_a = a_1 H + a_2 B = (a_1, a_2) \cdot \vec{X}$$

and

$$Z_b = b_1 H + b_2 B = (b_1, b_2) \cdot \vec{X},$$

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

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{det(A)} f_{\vec{X}}(A^{-1}\vec{z}) \tag{4.2}$$

Now together 4.2 and 4.1, yield

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{2\pi \det(A)\sigma_H \sigma_R} \exp(-0.5(\vec{z} - \vec{\mu}_z)^T A^{-1T} \Sigma_{\vec{X}}^{-1} A^{-1}(\vec{Z} - \vec{\mu}_z) \tag{4.3}$$

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 4.3 yields

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{2\pi \det(A)\sigma_H \sigma_R} \exp(-0.5(\vec{z} - \mu_{\vec{z}})^T \Sigma_{\vec{z}}^{-1}(\vec{z} - \vec{\mu}_z)). \tag{4.4}$$

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 4.4 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 4.1 *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.

4.1 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 4.1 *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 having covariance 0, they must be independent. But, note that we can write Y as

$$Y = U + aX$$

where a is the constant

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

This leads to the following lemma:

Lemma 4.2 *Assume that $\vec{X} = (X_1, X_2, \ldots, X_n)$ is a normal vector with 0 expectation: $E[X_i] = 0$ for all $i = 1, 2, 3, \ldots, n$. Then there exists independent normal variables*

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

so that $X_1 = U_1$ and for every $i = 2, 3, \ldots, i-1$ we have:
U_{i+1} is independent of $X_1, X_2, \ldots, X_i$ and there exists constants $a_{i1}, a_{i2}, \ldots, a_{ii}$ so that

$$X_{i+1} = U_{i+1} + a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ii}X_i$$

In other words X_{i+1} is obtained from $X_1, X_2, \ldots, X_i$ by taking a linear combination with non-random coefficients and adding an independent normal term.

Proof. ■

5 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

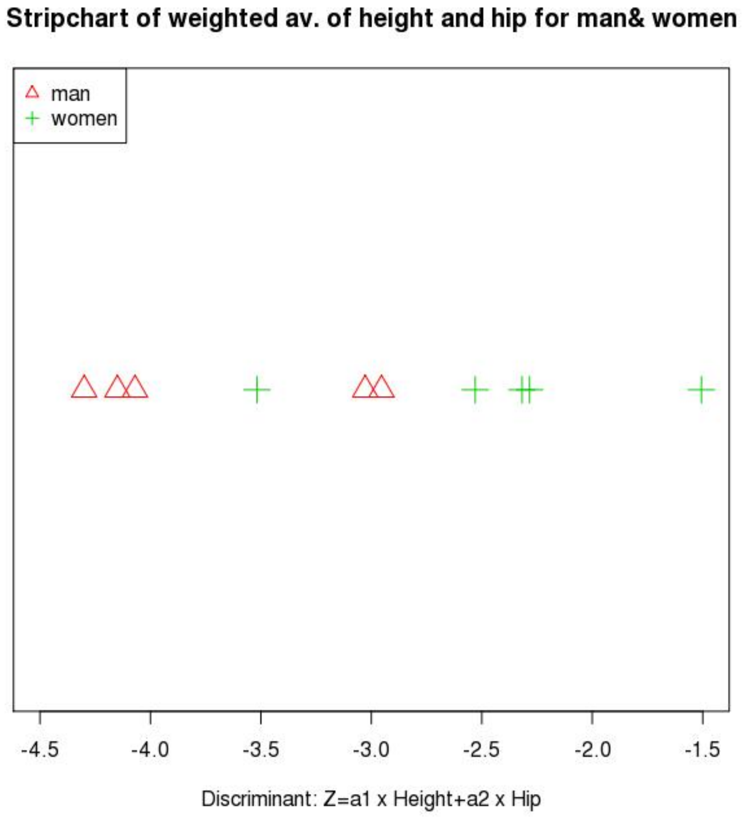
Here 1 stands for man, and 0 for women. We could try to guess if your skeleton is a man or a women just based on the height. But we may feel that it would be safer to use both the hip measurements and the height. How do we proceed? Let $Hight$ be the height and Hip be the breath of the hip. Typically 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 1. Indeed it seems that men and women are well separated by Z.

Figure 1:



Take the rule $z < -2.95$ gives 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:

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

whilst

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

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

$$\vec{\texttt{grad}}\, g = \lambda \vec{\texttt{grad}}\, h$$

for a constant λ. This yields

$$(a_1, a_2) = (\mu_1, \mu_2) \begin{pmatrix} COV(H,H) & COV(H,B) \\ COV(B,H) & COV(B,B) \end{pmatrix}^{-1} \tag{5.1}$$

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 5.1:

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

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 2. In figure 1, we have that 91 out of 507 are misclassified. 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 3.

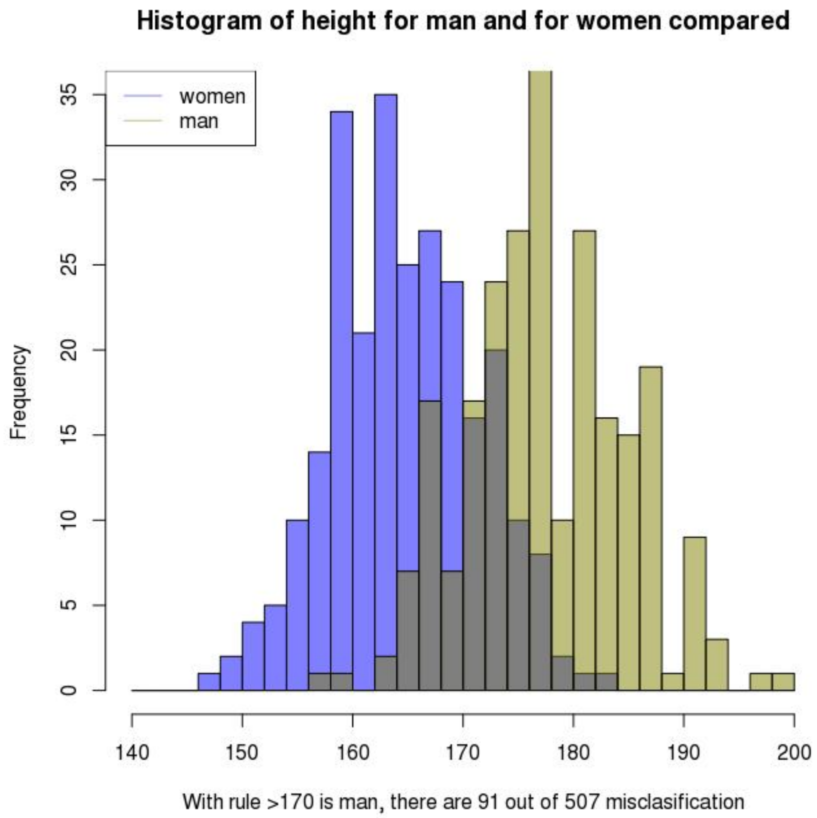
then, we use the linear discriminant for all the 23 variables available. This is then much more powerful as can be seen in 3

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:

6 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

Figure 2:

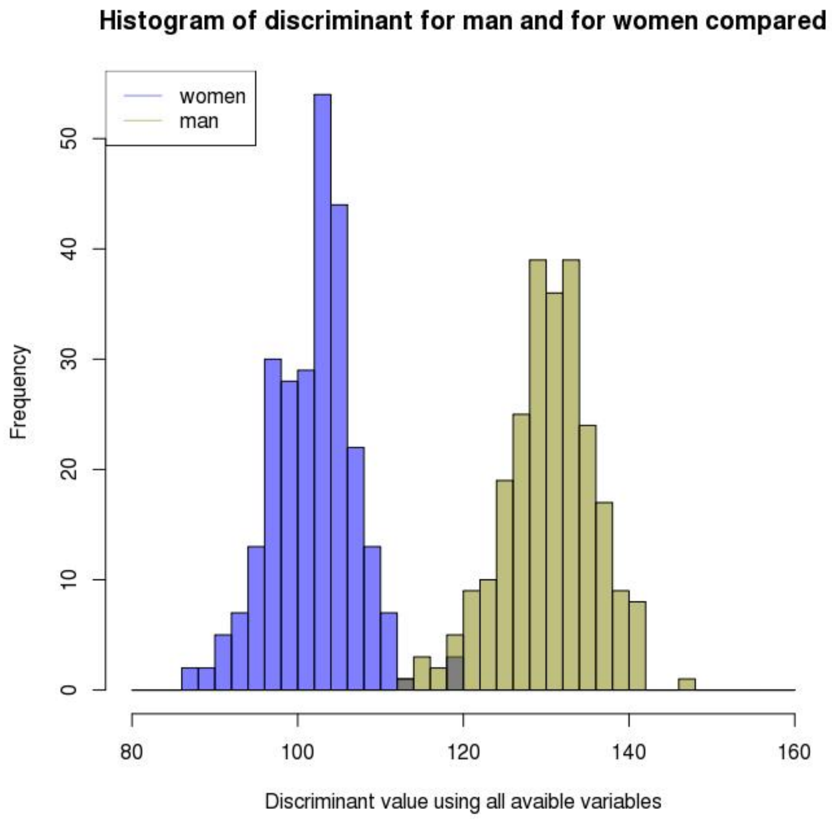


where there are subgroups in the community which are not known to us. By looking at the matrix with the phone call time should allow us to tell if there are such subgroups which are closer to each other. Assume there are 10 people the police investigates. Say the expected time during a week they spend 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

Figure 3:



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:

$$-2.1742 \cdot \hat{\vec{x}}_2 = \begin{pmatrix} 0.06 \\ 0.04 \\ 0.28 \\ -0.08 \\ -0.13 \\ 0.73 \\ 0.86 \\ 0.93 \\ 1.23 \\ 0.95 \end{pmatrix}$$

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:

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

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:

6.1 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 $\Sigma + E$.

6.2 The basic theory which makes it all work

6.3 Does it make sense to use spectral methods at all

7 Closest neighbor classification

8 The Multivariate T-test