Danmarks Tekniske Universitet

Trapyng

Multivariate Statistics

Manuel Montoya Catalá - manuwhs@gmail.com - s162706

Created July, 2017. Last modified February 8, 2018

Contents

1 Introduction

In this volume, we will apply linear methods to different time series obtained from the price, to see how this whole branch of knowledge could help us understand the signals and make profit in the market.
Initially we will talk about the analysis of a single time series and then we will see how to combine them.
This part has a big influence from the TSA course at DTU but everyhing was reimplemented in Python and focus towards analysis.
We will start with a MA recap and then explain AR, which can be seen. We will see ARMA, ARIMA, VARMA, how to use them for ex python libraries for them.

1.1 Showing some data

The next graph shows the price evolution of AAPL and GOOGL. These securities should be related up some point since they belong to the same country and financial sector

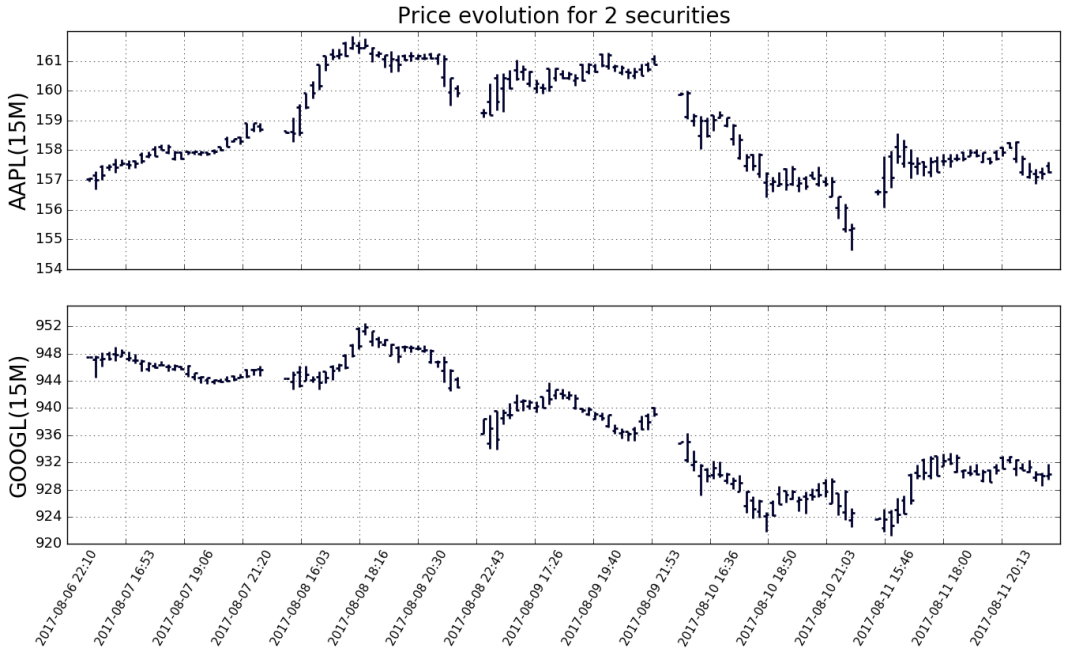


Figure 1: SMA and its window

As we can see, even though point by point they do not exactly match, they kinda go in the same direction.
The following shows the returns for the CLOSE price

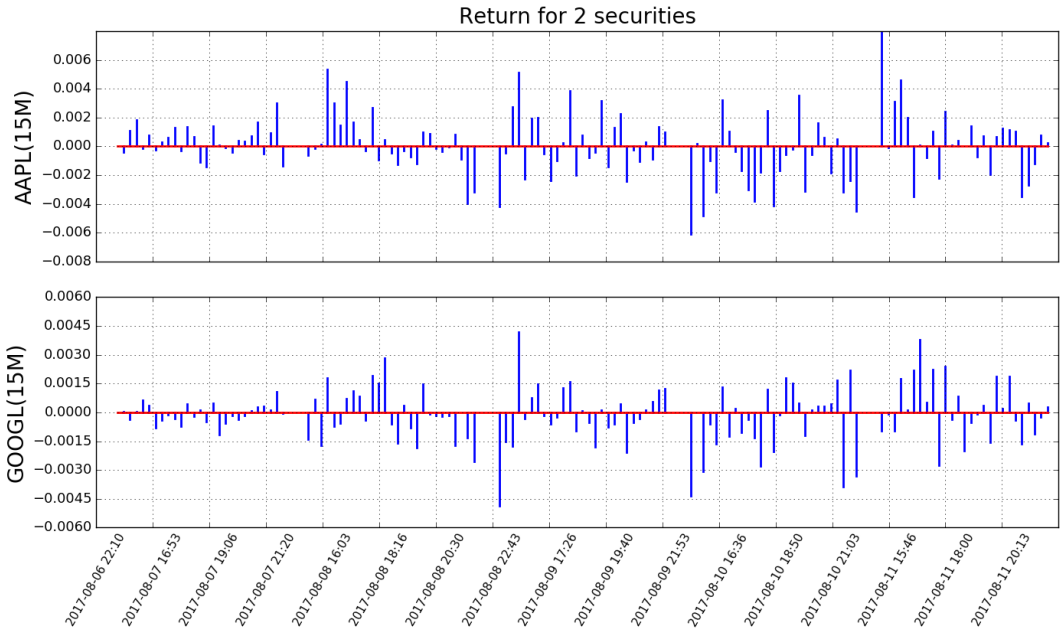


Figure 2: SMA and its window

The returns doe not seem random, they are pretty correlated and autocorrelated, we can see clusters of mroe return. Also, due to the gaps, there are huge returns, this are more magnified among weeks. What is the best time frame to see correlation ?

Since the return from one day to the other is way bigger, more time has passes so the distribution is another one, we remove them. They can also offer valuable info, but for the simplistic puporse we will dischard them for now.

The next graph shows the scatter plot return between both securities. Each of the return vectors contains 130 data points, to which we have fitted a gaussian distribution by means of using the unbiased estimators of the mean and variance. In the figure we can appreciate:

- The histograms for the returns of both assets do not seem specially gausssian. The probability around the mean greater, more spiky than that of a gaussian, so these distribution have a big central component. Having values close to mean is more likely than in the gaussian distribution.
- The distribution also shouws heavier tails than those of a Gaussian. The heavier tails mean that "unlike" events are more likely in this distribution.
- We can see the estimated gaussian distribution fitted to the individual assets. Comparing it with the histogram, we observe that the estimation is half way the high central peaks and the lower tails.
- The correlation is 0.54 which seems pretty high.
- We can see many outliers in the graph. Points with return significantly bigger than the rest, most likely due to the release of important news.

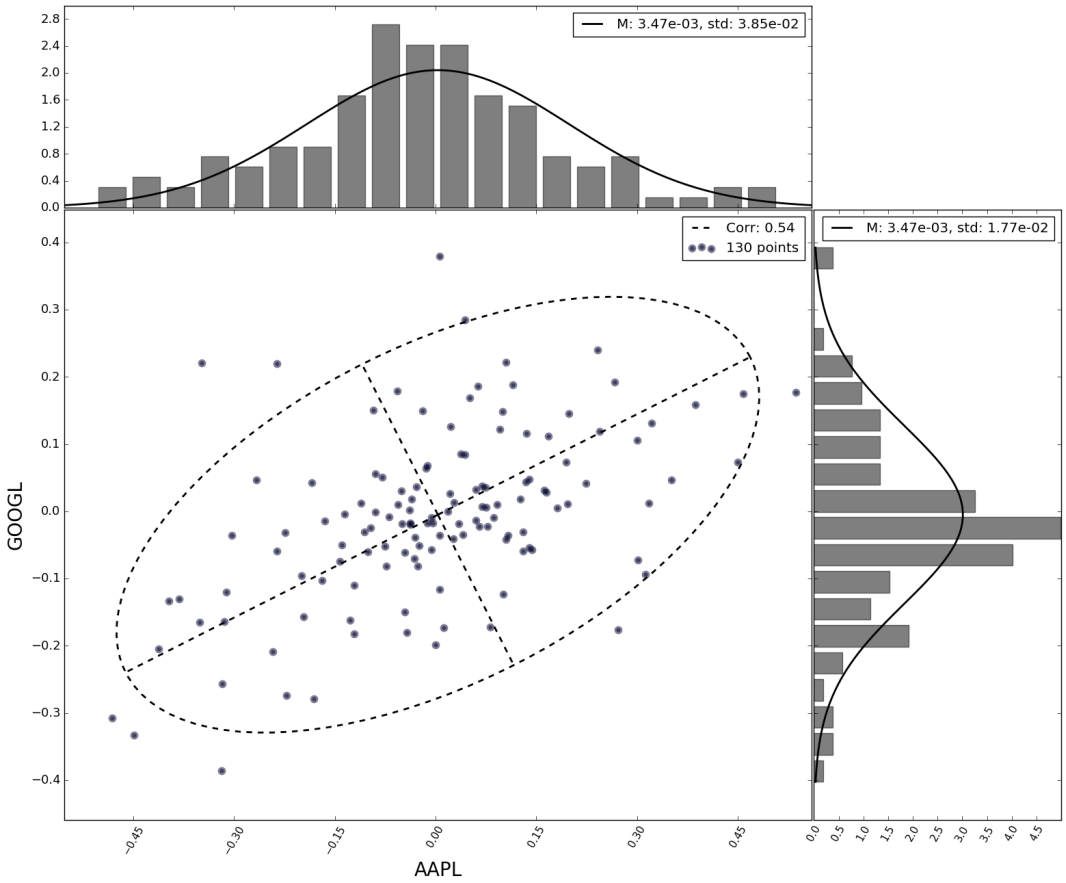


Figure 3: SMA and its window

Looking at the graph, we see that we only have 130 samples, since we only have 5 days in intervals of 15 min during 15:30 to 22:00 so we have 6h30min of market activity, which is 26 samples per day. I DONT KNOW WHY WE GET 27, we should not get the first 1 !! The estimated distributions also do not take into account possible time-relations between the returns or the changing of the distribution over time. Overall we could make the following questions.

- Are 130 enough samples for the estimated parameters ? How sure are we that they are the real ones ?
- Should we use other price instead of the CLOSE price ? LOW ? HIGH ? AVERAGE ?
- Should we do any transformation of the data to make it more Gaussian ?
- How can we check mathmatically how "Gaussian" something is ?
- Are there other distributions that are handy to use (exponential family i.e.) to model these returns ?
- Can we capture the time relation between the returns ?
- How much will the statistical properties of this return will change over time ?
- Can we capture the overall changing dynamics of the return ?
- How is the correlation at this timeframe related with the correlation at other timeframe ?
- What other measures of relationship can we model between the 2 variables ?

Well, as can be seen there are a lot of quesitons now ! Lets try to answer them as best as possible ! We will need pretty much dedicated Secitons for them. This was the introduciton of the data !

THIS WAS IN P-VALUE, it needs to be changed. In many ocasions we estimate parameters of our model. Due to noise of the samples, we could have got another value for the parameter, it depends just on the data we obtained.
So we can make the following quetions:

- How sure are we of our parameter is the proper one ?
- Can we tell a range of that ?

We have seen how we have already fitted the Gaussian distribution to this. We might not want to make assumptions on the distribution .We can fit kernel modes.
TODO: Missalingment between plot and histogram barchart
The following graph shows the estimation of KDE when fitting a gaussian to every point.

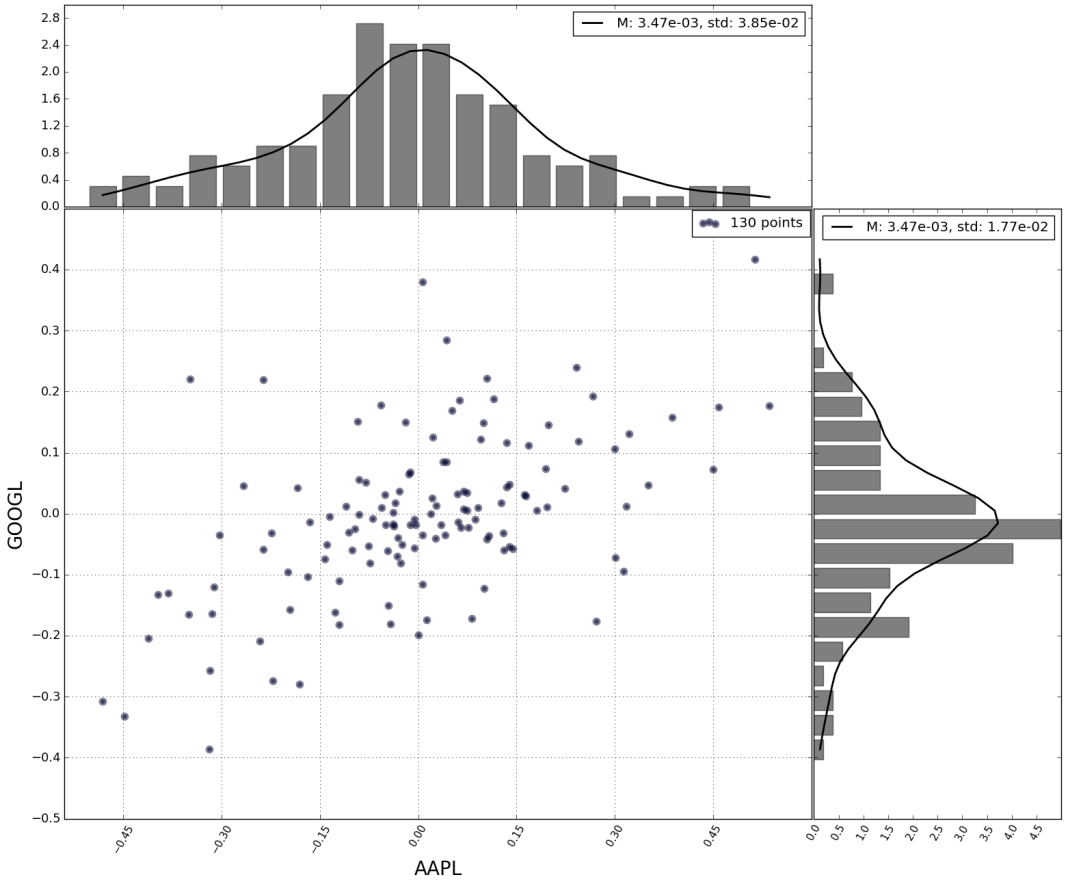


Figure 4: SMA and its window

Now in 3D we can see the joint distribution. Notice how P(x,y) = P(x)P(y) and since P(x) and P(y) ¿ 0, then Pxy is bigger.

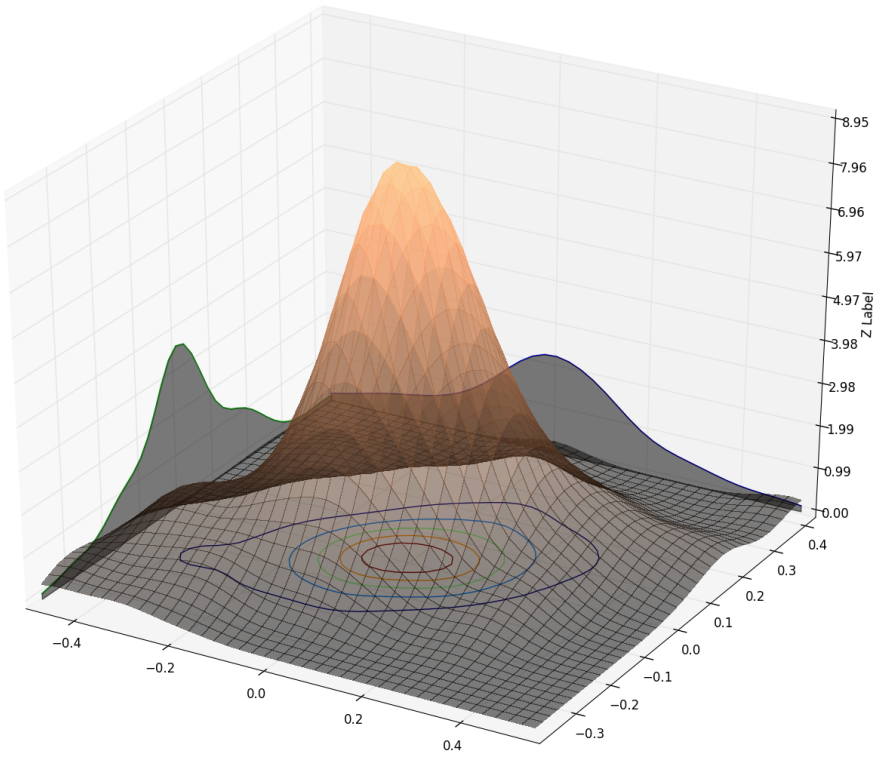


Figure 5: SMA and its window

In the previos graphs we have completely ignored the temporal relationship between the samples, we have implicitly assumed that they are independent. But this could not be the case, there is a set of tests we could perform in order to check a possible relationship between the samples. As an example, we could see the Autocorrelation and Partial Autocorrelation...
XXXXX
In this Section we introduced a graphical analysis of the properties of the data and a set of questions that we could have while looking at it. During this document, we will model and quantize the data, building useful systems that could be able to tell us something relevant about our data. Hopefully this information can have some predictive capability that would help us earn some money.

2 Gaussian Distribution

This Section is dedicated to the Gaussian distribution, it is meant to be both an intuitive and mathematical tutorial in order to undertand and justify many assumptions done in the models. In many scenarios we will assume that the distribution followed by the data is Gaussian, this is done for several reasons.

- It is mathematically simple to work with. The math usually can be reduced. It is symetric, linear, convex...
- Many natural phenomena obey it.
- Central Limit Theorem. The sum of a lot of independent variables follow it, more on this later.

During this tutorial we will start viewing the univariate Gaussian shape of the distribution, along with its intuition and properties than can be easily vidualized. Then we will see its equation, explaining its components and linking it to the intuition and graphs. Once we have a basic understanding of the distribution, then we will see other sets of properties as linearity and the CLT which are not straight forward to see from the graph and equation. Once we understand that, we will see the Multivariate Gaussian distribution, making enphasis on its derivation, conditional distributions and geometrical point of view, giving special attention to the covariance matrix. We will close the section talking about the estimators for the parameters of the distribution.

2.1 Univariate Gaussian

First of all, lets get the simple things right, we will dedicate a big proportion of this document to the analysis of the Gaussian distribution in 1D, 2D and 3D to give an intuition and insight of the distribution that can be generalized later to any number of dimensions. When dimensionality grows, our spatial intuitions fail miserably but we will see that at least the equations remain in the same general form, regardless of later intuitions that we might have.

Next image shows the probability density function, pdf, of 3 independent 1D gaussian distributions, X_1, X_2, X_3, along with 100 samples drawn from each of the distributions. Each of the 3 random belongs to a distirbution $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ where their parameters, $\theta_i = \{\mu_i, \sigma_i^2\}$, can be seen in the graph. From this graphical representation we can make the following observations about the distributions:

- The mean μ is the central point of the distribution, it maximizes the pdf and it is simetric around it. This also implies that the mean and median are the same.
- The distribution is convex, that is, it has a unique maximum point.
- The more variance, the more uniform they are. The bigger the variance, the more uncertainty we have about the possible values that the random variable will take.

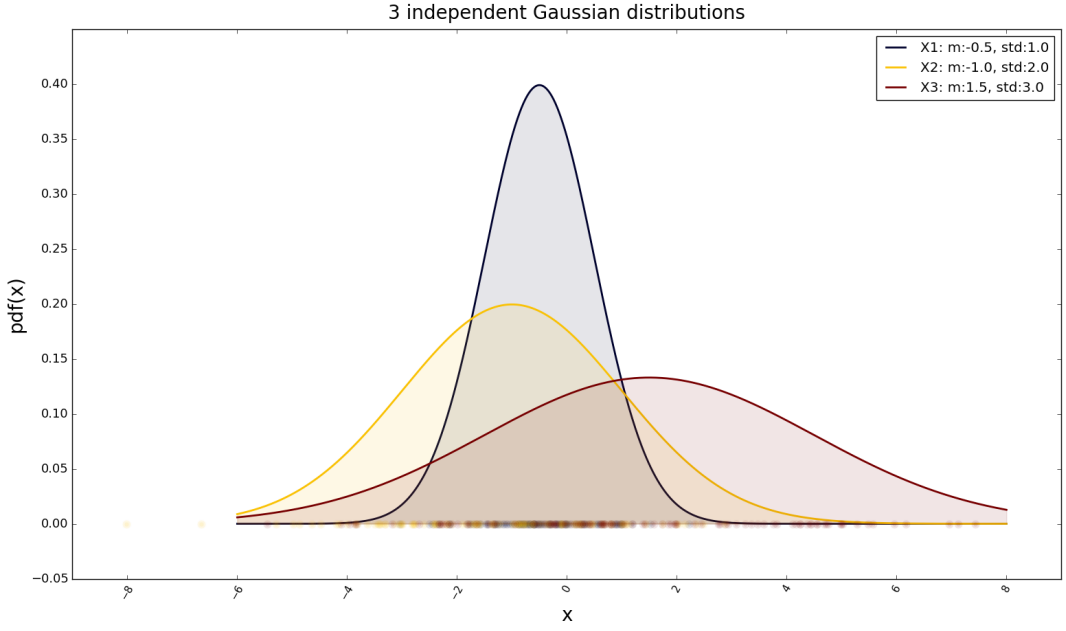


Figure 6: SMA and its window

Now that have an intial idea of what this distribution look like, lets us see the equation of its probability density function. Given that we know the parameters of the distribution $\theta = \mu, \sigma$, the equation is as follows:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Some further observations can be made from this equation:

- The domain of X is the entire real domain R, from inf to inf.
- The first term is the normalization constant and it only depends on the variance.
- The second term is a negative exponential, with maximum value when $x = \mu$ in which case the argument of exponentila is 0. In this situation, the value of the pdf is the normalization constant.
- We can also see in the equation that the pdf is symetric respect to μ due to the squared distance $(x - \mu)$.
- The peak pdf can be bigger than 1, since this is a pdf, not a probability
- The pdf decreases at a speed of a squared negative exponential as x goes away from the mean.
- The exponential decreased is divided by the variance, so the bigger the variace, the slowlier the pdf will decrease. And also, the smaller the normalization constant.

These are some of the straight forward observations that we can make from taking a first sight to the equation and graph of the 1D Gaussian variable. As we will see later, it also has a lot of more properties that are not easy to see at first sight but that makes this distribution the most used in mathematics.

2.2 Properties of the Univariate Gaussian Distribution

2.2.1 Linear Properties

Multiplication for a constant. Sum of gaussians.
Explain that a Gaussian with mean is just the normal one with added a value, the one with bigger variance is just the original one multiplied by a constant, since we have increased the volumen of space then we need to normalize.
Then explain the sum of gaussians is also gaussian.

$$\begin{aligned} X + Y &\sim N(\mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y}) \\ aX + bY &\sim N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{X,Y}) \end{aligned} \tag{1}$$

2.2.2 Central Limit Theorem

2.2.3 Entropy

Exponential family, maximization of entropy given finite free momentums order 1 and 2. Finally, to gain a little more intutition of this.

2.2.4 Estimators

ML unbiased Estimator for the mean and variance.

2.3 Independent Multivariate Gaussian

Now we will present the multidimensional version of the Gaussian, that is, we want to find a joint statistical distribution $pdf(X_1, X_2, X_3....)$ whose marginalized components $X_1, X_2, X_3...$ are Gaussian univariate random variables.

Probably the most straight forward approach is to combine all these random variables into a joint distribution is just to assume that they are independent. In the case in which we have 3 independent random variables, the global distribution is:

$$f(X_1, X_2, X_3) = f(X_1)f(X_2)f(X_3) = \prod_{i=1}^{D=3}\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) exp\left[-\sum_{i=1}^{D=3}\frac{(X_i - \mu_i)^2}{2\sigma_i^2}\right]$$

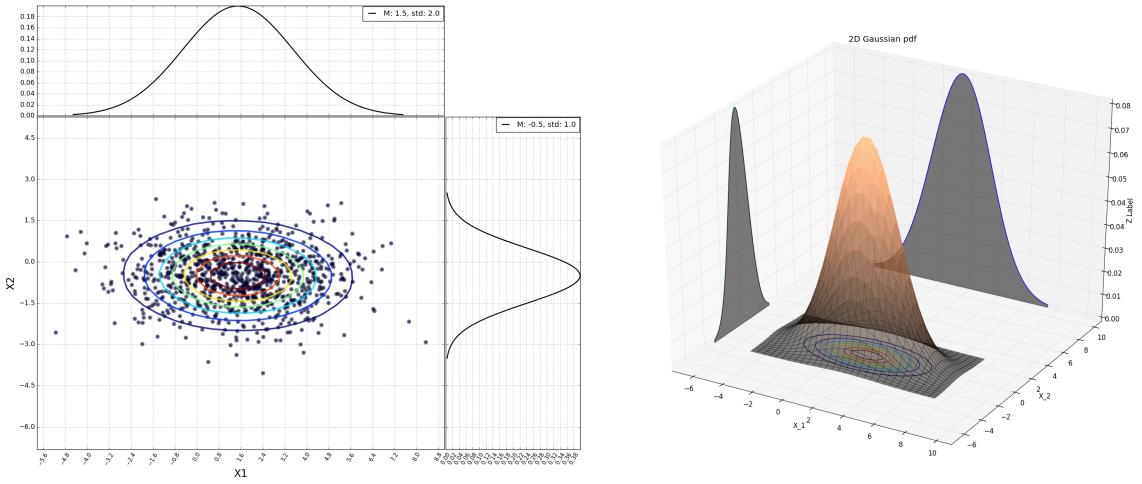
As we can see, this join distribution translates into:

- The product of the normalization constant
- The sum of the exponents of the different individual components

Notice that since the variables are independent, any momentum applied to the distribution can be computed independently to each dimension, without the influence of the others. Lets start using some vectorial notation already. If we define the column vector of independent random variables $X = [X_1, X_2, ...X_D]^T$, with mean vector $E[X] = \mu = [\mu_1, \mu_2, ..., \mu_D]$ and vector of variances $VAR[X] = \sigma^2 = [\sigma_1^2, \sigma_2^2, ..., \sigma_D^2]$, then we have the equation:

$$f(X) = \prod_{i=1}^{D} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \right) exp \left[-\sum_{i=1}^{D} \frac{(X_i - \mu_i)^2}{2\sigma_i^2} \right]$$

This is not the most informative approach to make a multivariate Gaussian distribution since you can use this procedure to create the joint distribution of any set of independent random variables. Nevertheless, lets see how this distribution looks like with 2 random variables so that we build a better intuition. In the next Figure we have 2 visualizations of the joint distribution $f(X_1, X_2)$ and the marginal distributions $f(X_1)$ and $f(X_2)$.



(a) sizeInches = [2,2]

(b) sizeInches = [10,6]

Figure 7: Effect of the figure size in the saved figures

In the image to the left we have a cloud of point drawn from the joint distribution in the main axes, along with the contour lines of the joint pdf. These contour lines show paths of equal probability in the surface that conforms the joint pdf. The other 2 axes contain the marginal distributions of X_1 and X_2. Some observations that can be made:

- The shape of the contour lines are ellipses. This equation can be obtained from the argument of the exponent in the equation.
- The ellipse has no angle with the X axes. As we will see later, this indicates the random variables are independent.

In the image to the right we can see the 2D surface of the joint pdf $f(X_1, X_2)$. Along with the projection of the marginalized distributions at each extreme of the axis, $f(X_1), f(X_2)$. Notice that the marginalized distributions shown have been scaled so that their maximum value is equal to the maximum value of the joint distribution. How did we do this ? and why did we need to do it ?

Notice that $f(X_1, X_2) = f(X_1)f(X_2|X_1)$, and since we know the variables are independent we can assume directly that $f(X_1, X_2) = f(X_1)f(X_2)$. Notice that $f(X_1)$ and $f(X_2)$ are density functions, so their values can be bigger than one, $f(X_i) \in [0, \text{inf}]$ so the joint probability could be bigger or smaller than the marginal probabilities. We can obtain slices of the 2D joint distribution by giving a value to one of the parameters, for example, a vertical slice will correspond to the equation:

$$f(X_1 = x_1, X_2) = f_1(x_1) \cdot f(X_2|X_1 = x_1) = \left[\frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} exp \left[-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} \right] \right] exp \left[-\frac{(X_2 - \mu_2)^2}{2\sigma_2^2} \right]$$

Notice how the first term is a constant since we set $X_1 = x_1$ and the second term is the negative squares exponential that changes with X_2, the vertical random variable.

Notice that in general this will not a probability distribution, since it does not add up to one. For to add up to 1, then resulting distribution has to be equal to the distribution of $f(X_2)$ which will be satisfied if:

$$f(X_1 = x_1) = \left[\frac{1}{\sqrt{2\pi\sigma_1^2}} exp \left[-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} \right] \right] = 1$$

We can visualize these slices in the next Figure, which shows the function $f(X_1 = x_1, X_2)$ for different values of x_1. As we can see, each of the values of the function has a different scale, this can also be seen in the projection of all of them that is done in the left axis. The scale as we have just seen, is given by the $f(X_1 = x_1)$ value. The smaller $f(X_1 = x_1)$ is, the smaller the slice will be.

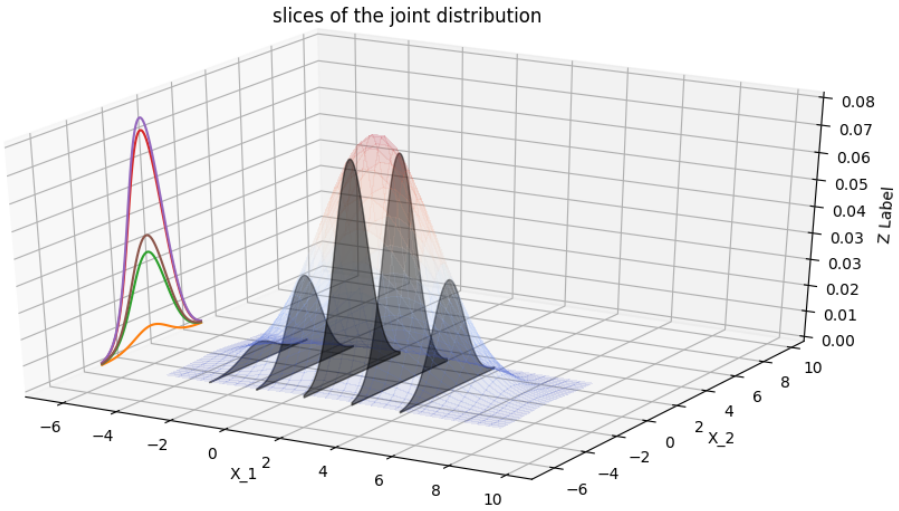


Figure 8: SMA and its window

To transform these slices into a probability distribution over X_2 we just divide the slice by the probability of $f(X_1 = x_1)$ so that they cancel each other out. We thus obtain the conditional distribution:

$$\frac{f(X_1 = x_1, X_2)}{f(X_1 = x_1)} = f(X_2 | X_1 = x_1)$$

In this case, since our variables are independent, the shape of the slices will be the same, $f(X_2)$ multiplied by a constant.

$$f(X_1 = x_1, X_2) = k_i \cdot f(X_2) = f(X_1 = x_1) \cdot f(X_2)$$

By normalizing with $f(X_1 = x_1)$ we will always obtain the same distribution $f(X_2)$. The next Figure shows the result of plotting the normalized slices, that is, the conditional probabilities $f(X_2 | X_1 = x_1)$ which since the variables are independent, it is always equal to $f(X_2)$. We can also appreciate how the peak of the marginal distribution $f(X_2)$ is bigger than the peak of the joint distribution, meaning that normalization constant of $f(X_1)$ is smaller than 1.

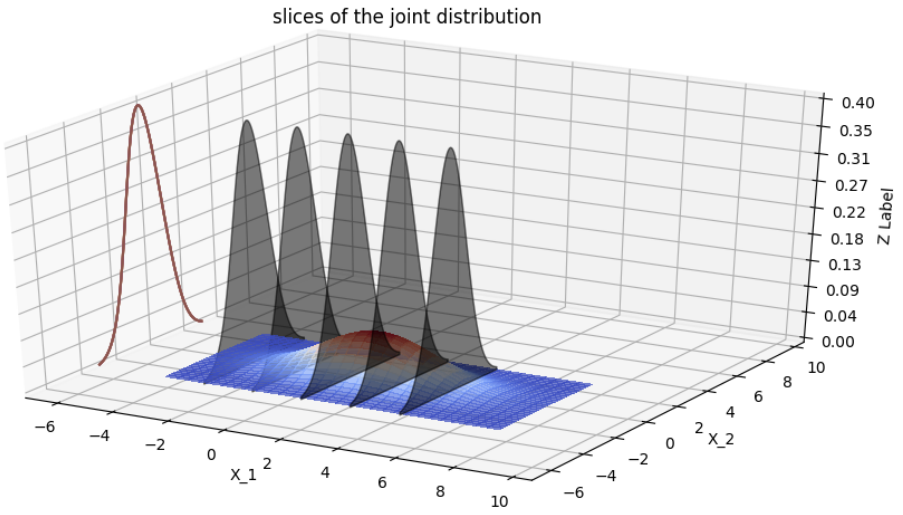


Figure 9: SMA and its window

So basic conclusion is, no matter what slice of $X_1 = x_1$ we take, the resulting shape of the distribution is the Gaussian of X_2 scaled by $f(X_1 = x_1)$. We get analogous results if the condition on X_2 instead, we would obtain horizontal slides in that case.

Back to the original graph where we joinly plotted the 2D distribution and the marginal distributions: When we plot the marginal distributions, we first compute the marginal, $f(X_2)$ for example. As we see, this value will be higher or lower than $f(X_2, X_1 = x_1)$ due to the normalization constant $f(X_1 = x_1)$. To obtain one of the slices $f(X_2, X_1 = x_1)$ from the conditional distribution $f(X_2|X_1 = x_1)$ (equal to the marginal in this case) we just need to multiply by the normalization constant $f(X_1 = x_1)$. In this case, the biggest slice is when $X_1 = \mu_1$, which leads to

$$f(X_1 = \mu_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}}$$

The same analysis holds for any increasing number of variables. If we have 3 random variables, then we can obtain 2D slices of the 2D distribution by seting the value of one of its variables:

$$f(X_1 = x_1, X_2, X_3) = f(X_1 = 1)f(X_2, X_3|X_1 = x_1)$$

In this case, $f(X_2, X_3|X_1 = x_1)$ will correspond to a 2D distribution like the one we saw for $f(X_2, X_1)$

Maybe show 3D

2.4 Multivariate Gaussian

We have seen how to trivially create the Multivariate Independent Gaussian distribution, but... can we create a Multivariate distribution in which its marginal variables are Gaussians and the variables are not independend ?

Using the chain rule we know that the important relationships are in the conditional relationships between the variables. The general equation using the chain rule for this distribution is:

$$f(X_1, X_2, X_3) = f(X_1)f(X_2|X_1)f(X_3|X_2, X_1)$$

In order for $f(X_1)$, $f(X_2)$ and $f(X_3)$ to be Gaussian and dependent of each other in $f(X_1, X_2, X_3)$ not every shape of the conditional distribution is valid, since the marginal distributions when marginalizing still have to be Gaussian. For the case with only 2 R.V, this imposition over the conditional distribution translates to:

$$f(X_1) = \int_{x_2 \in X_2} f(X_1, X_2)dx_2 = \int_{x_2 \in X_2} f(X_1)f(X_2|X_1)dx_2$$

So the shape of the conditional distibution $f(X_2|X_1)$ has to so that the integral will give $f(X_1)$. The analogous condition applies for $f(X_1|X_2)$. As we have seen, one possibility is that the variables are independent in which case the condition will always be fullfilled:

$$f(X_1) = \int_{x_2 \in X_2} f(X_1) f(X_2) dx_2 = f(X_1) \left[\int_{x_2 \in X_2} f(X_2) dx_2 \right] = f(X_1)$$

Can we find another joint distribution which satisfies the condtions ? Well, we can, it is called the Multivariate Gaussian is a distribution where the variables $X = [X_1, X_2, ...X_D]$ are a linear combination of independent normalized Gaussians $U = [U_1, U_2, ...U_D]$.

In the 2D case, let U_1 and U_2 be the independent random variables with 0 mean and variance 1. $U_1, U_2 \sim \mathcal{N}(0, 1)$. Let us define the random variables X_1 and X_2 as a linear combination of the previous variables:

$$X_1 = a_{11} \cdot U_1 + a_{12} \cdot U_2 + \mu_1$$
$$X_2 = a_{21} \cdot U_1 + a_{22} \cdot U_2 + \mu_2$$

We can write this in a matrix form as:

$$X = \mu + AU = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

By the properties of linearity of the Gaussian distribution, we have that the distribution of the new variables is:

$$X_1 \sim \mathcal{N}(\mu_1, a_{11}^2 + a_{12}^2)$$
$$X_2 \sim \mathcal{N}(\mu_2, a_{21}^2 + a_{22}^2)$$

Geometrically, the matrix A has rotated and scaled the initial variables U and the vector μ has later displaced them to somewhere in the space, becoming its new center. To have a better intuition, the next Figure shows the distribution of U and the distribution of X when:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0.9 & 2 \\ 0.8 & 0.7 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} + \begin{bmatrix} -1.5 \\ 2 \end{bmatrix}$$

We can see a set of points drawn from the U in the left figure, along with the controur of the distribution and in the right image we can see the transformation into X. As we can see, the distribution of X is a rotated, scaled and displaced version of the distribution of U. But the most important property is that marginal variables X_1 and X_2 obtained from the linear combination of U_1, U_2 are still Gaussian Distributions. This is the key property that makes the Gaussian Distribution so useful.

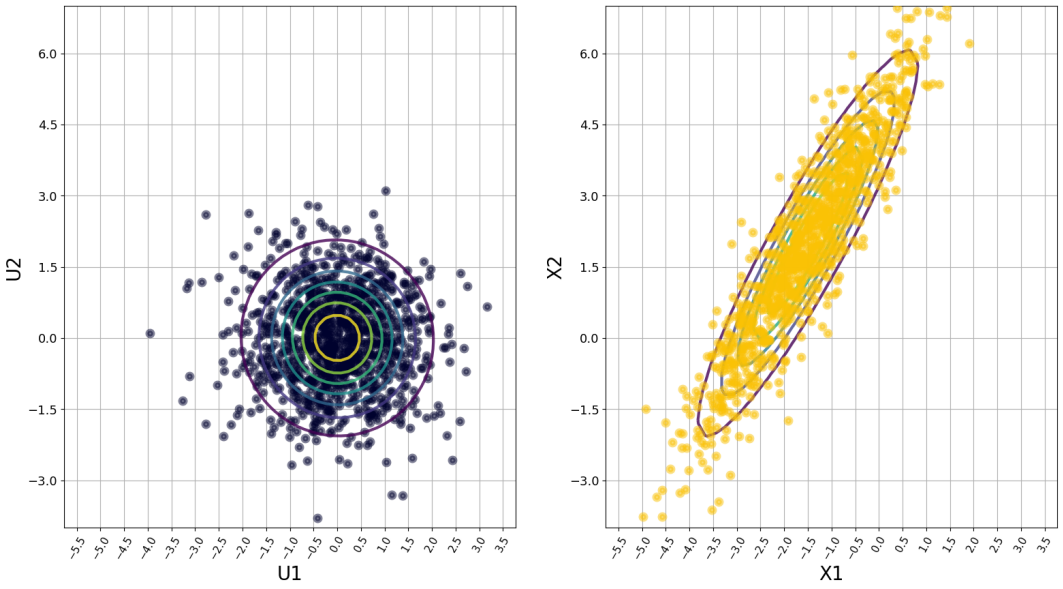


Figure 10: SMA and its window

Make chart like 7 to highlight that the projections are also Gaussian.

We know that the original variables U_1 and U_2 are independent, but now, the transformed variables X_1 and X_2 are dependent of each other, we can see a clear linear relationship in the previous Figure, knowing the value of X_1 gives me information about where X_2 could take place since the bigger X_1, the bigger X_2 usually is.

The multivariate distribution is now characterized by the mean vector μ and the variance-covariance matrix Σ_X.

$$\mu = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} \quad \Sigma_X = AA^T = \begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22}\end{bmatrix}$$

The diagonal elements of the variance-covariance matrix will indicate the variance of the individual gaussian variables, X_1 and X_2 in this case. The non-diagonal elements will indicate the covariance between the random variables, that is, the linear relationship between them, how much variance they share. We will develop on this concept later in this Section.

Notice that the covariance matrix cannot be any matrix, since it comes from the product AA^T, it must have a specific set of properties, we will develop on these properties later but notice for the 2D case we have that the variance-covariance parameters are:

$$\Sigma_X = AA^T = \begin{bmatrix}a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}a_{11} & a_{21} \\ a_{12} & a_{22}\end{bmatrix} = \begin{bmatrix}a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{21}a_{11} + a_{22}a_{12} & a_{21}^2 + a_{22}^2\end{bmatrix}$$

So, what is the joint distribution of $f(X_1, X_2)$?. It can be shown, and we will prove this in a moment, that the distribution of these variables follows the form:

$$f(X_1, X_2) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_X|}} exp\left[-\frac{1}{2}\begin{bmatrix}X_1 - \mu_1 \\ X_2 - \mu_2\end{bmatrix}^T \begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22}\end{bmatrix}^{-1} \begin{bmatrix}X_1 - \mu_1 \\ X_2 - \mu_2\end{bmatrix}\right]$$

Notice that the normalization constant of the distribution is divided by the square root of the determinant of the covariance matrix, $\sqrt{|\Sigma_X|}$. The reason for this is that when performing the linear transformation of the space multiplying by A, it has scaled the space, so we need to renormalize it. Notice in the previous Figure how the area of the ellipses has changed even though it contains the same total probability (and number of samples in the example). The total scaling of the transformation can be computed as the volume of the space, which is the determinant:

$$det(A) = |A| = \frac{Vol(A)}{Vol(I)}$$

Since $\Sigma_X = AA^T$ then we have that $|A| = \sqrt{|\Sigma_X|}$ and that is why we have this term in the normalization constant, so that the integral across the entire hypervolume (area in 2D) of the distribution is equal to 1.

In the general case, we have D-dimensional Multivariate Gaussian distribution X obtained from a vector of independent normalized Gaussian variables U by using a linear transformation:

$$X = \mu + AU = \begin{bmatrix}X_1 \\ X_2 \\ \vdots \\ X_D\end{bmatrix} = \begin{bmatrix}\mu_1 \\ \mu_2 \\ \vdots \\ \mu_D\end{bmatrix} + \begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1D} \\ a_{21} & a_{22} & \cdots & a_{1D} \\ \vdots & \vdots & & \vdots \\ a_{D1} & a_{D2} & \cdots & a_{DD}\end{bmatrix}\begin{bmatrix}U_1 \\ U_2 \\ \vdots \\ U_D\end{bmatrix}$$

And in the same way we have just seen for 2D, now X belongs to a Multivariate Normal Distirbution with mean μ and variance-covariance matrix $\Sigma_X = AA^T$. The equation of the pdf of such distribution can be written as:

$$f(X) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_X|}} exp\left[-\frac{1}{2}(X - \mu)^T \Sigma_X^{-1}(X - \mu)\right]$$

To quickly prove this we just need to have in mind the distribution of the original variables U:

$$f(U) = \frac{1}{\sqrt{2\pi}^D} exp\left[-\frac{1}{2}U^T U\right]$$

Since we can express these independent random variables U from X just by solving the equation:

$$U = A^{-1}(X - \mu)$$

We can just replace U in the argument of the exponent to obtain:

$$U^T U = (A^{-1}(X - \mu))^T A^{-1}(X - \mu) = (X - \mu)^T A^{-1T} A^{-1}(X - \mu) = (X - \mu)^T \Sigma_X^{-1}(X - \mu)$$

Also, if the linear transformation matrix A is not an orthonormal basis of the space then it is expanding or contracting the space, so we need to reescale the projection so that the integral adds up to 1. We have previously seen that in order to do this we just need to divide the pdf of the distribution by $\sqrt{|\Sigma_X|}$.

2.4.1 Dependence between the Random Variables

In the previous case where the variables where independent we saw that knowing one of them $f(X_1 = x_1)$ gave us no extra information on the set of values that the other variable X_2 could take, all the conditional probabilities $f(X_2|X_1 = x_1)$ where equal to $f(X_2)$. In this case the variables depend on each other and now $f(X_2|X_1 = x_1)$ is different for each value of x_1. In this section we will understand the shape of the slices $f(X_2, X_1 = x_1)$ and the conditional probabilities $f(X_2|X_1 = x_1)$, and how these evolve with x_1.

Let's start again seeing the slices of this distribution, shown in the next Figure, as we can see, the mean value of the slices changes as we change the value of x_1. This means that knowing X_1 gives us a better idea of where could X_2 appear. This could have been guessed from looking at the previous Figure where we can clearly see that the y-value of the points depend on X_1. As before, these slices are not probability distributions, we need to normalize them by $f(X_1)$ to get the conditional distribution.

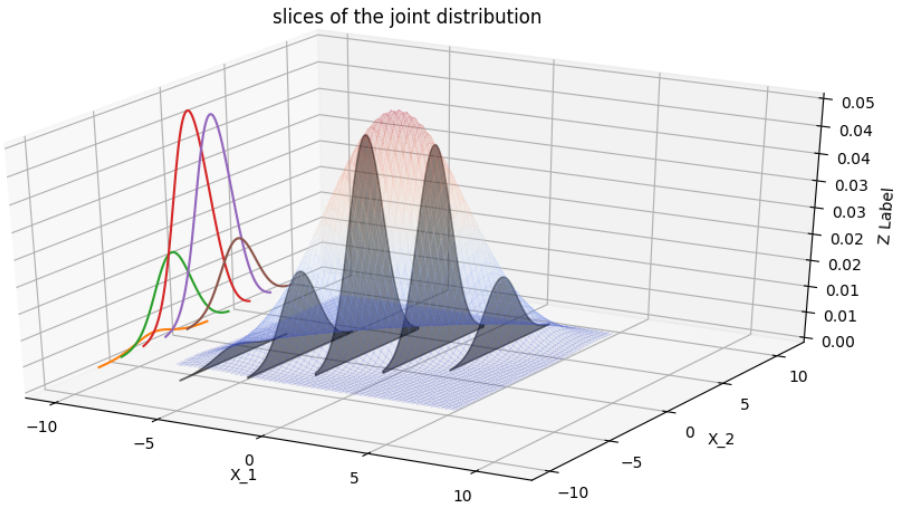


Figure 11: SMA and its window

But what is the equation that governs how the mean of the slice changes with X_2? Does the variance of the gaussian of each slide also changes with $X_1 = x_1$? Graphially it is hard to tell, lets see the actual equations:

$$f(X_1, X_2) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_X|}} exp\left[-\frac{1}{2}\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}\right]$$

Setting $X_1 = x_1$ we can reformulate the exponent as a new Gaussian form where the mean and variance are:

$$E[X_2|X_1 = x_1] = \mu_2 + \sigma_{21}\sigma_{11}^{-1}(x_1 - \mu_1)$$
$$V[X_2|X_1 = x_1] = \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}$$

We can make the following observations from these equations:

- The mean varies linearly with the value that takes the conditioned variable $X_1 = x_1$. It does not depend on the variance of X_2.
- If the covariance between the variables is 0 (independent), then $\sigma_{21} = 0$ and the condition mean is independent of x_1, as in the previous section happened.
- The variance does not depend on x_1
- The variance is always smaller than the marginal variance, we have gained information !! The more correlated the variables are (σ_{21}), the bigger the uncertainty we will remove, and the bigger the variance of the conditioned variable (σ_{11}) the less uncertainty we will remove.

So lets plot the conditional distribution in the next Figure. As we can see, the variance of all the slices is the same and the mean varies linearly with x_1, the prediction of the equations became true !

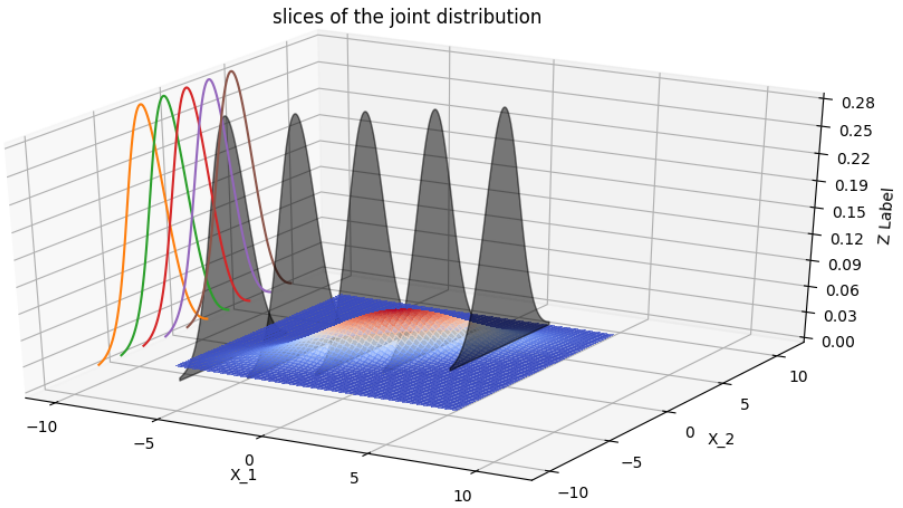


Figure 12: SMA and its window

In the general case where we have a set of variables $X = [X_1, X_2, ...X_{D_x}]$ and $Y = [Y_1, Y_2, ...Y_{D_Y}]$ that are jointly Gaussian, $Z = [Y, X] \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ we can decompose the joint Gaussian distribution as:

$$E[Z] = E\begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} E[Y] \\ E[X] \end{bmatrix} = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}$$

$$VAR[Z] = \Sigma_Z = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}$$

The general equations for the conditioned mean and covariance matrix of Y given X are:

$$E[Y|X] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

$$VAR[Y|X] = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

As we can see from these equations, the conditional mean also varies linearly with X and its covariance matrix does not. The covariance matrix also has reduced uncertainty.

Needless to say that the marginal probabilities of $p(X = X, Y)$ can be computed as just the distribution of those variables, not taking the rest into account. As we have no information about the others, they are marginalized and do not remove any uncertainty from X. The universe does not remove uncetainty from a variable just because it is related to others, if we have no information about the others, it is as if we dont know they exist.

As a graphical example we could use the example with 3 random variables where we condition on X_1 and X_2.

$$\mu = \begin{bmatrix} \mu_Y \\ \mu_{X_1} \\ \mu_{X_2} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{YY} & \sigma_{YX_1} & \sigma_{YX_2} \\ \sigma_{X_1Y} & \sigma_{X_1X_1} & \sigma_{X_1X_2} \\ \sigma_{X_2Y} & \sigma_{X_2X_1} & \sigma_{X_2X_2} \end{bmatrix}$$

From the structure we can identify the terms:

$$\Sigma_{YY} = \sigma_{YY} \qquad \Sigma_{XX} = \begin{bmatrix} \sigma_{X_1X_1} & \sigma_{X_1X_2} \\ \sigma_{X_2X_1} & \sigma_{X_2X_2} \end{bmatrix} \qquad \Sigma_{YX} = \begin{bmatrix} \sigma_{YX_1} & \sigma_{YX_2} \end{bmatrix} \qquad \Sigma_{XY} = \begin{bmatrix} \sigma_{YX_1} \\ \sigma_{YX_2} \end{bmatrix}$$

Where we need to invert the covariance matrix of X resulting in:

$$\Sigma_{XX}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12}} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix}$$

Plugging these values into the equations we have the expectation and variance of the conditional variable Y as:

$$E[Y|X_1 = x_1, X_2 = x_2] = \mu_Y + \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12}} \begin{bmatrix} \sigma_{YX_1} & \sigma_{YX_2} \end{bmatrix} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix} \begin{bmatrix} \mu_{X_1} - x_1 \\ \mu_{X_2} - x_2 \end{bmatrix}$$

$$V[Y|X_1 = x_1, X_2 = x_2] = \sigma_{YY} + \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12}} \begin{bmatrix} \sigma_{YX_1} & \sigma_{YX_2} \end{bmatrix} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix} \begin{bmatrix} \sigma_{YX_1} \\ \sigma_{YX_2} \end{bmatrix}$$

2.4.2 Decomposition of A

As we have just seen, the general Multivariate Distribution comes from a set of independent normalized random variables U to which we applied a linear transformation using A and then we have added a bias μ. Notice that adding this mean is not a linear operation, only the transformation of A, the mean is added to gain expressibility in the statistical distribution. We will first most commonly remove the mean, then perform a linear transformation and then add it back if desired, or add another different mean. In the general case we have:

$$X - \mu = AU$$

We know that μ just produces a displacement of the points, called **translation**, it will not be very important since you can apply this displacement later at any point. The important operation is the linear transformation by A. This transformation will deform the space by both **rotating** and **scaling** the space, as if we were seeing it from another perspective.

The **properties of any Rotation matrix** R is that its projection vectors $r_i = [r_{1i}, r_{2i}, ..., r_{Di}]$ have modulus 1 and are perpendicular to each other $r_i^T r_j = 0$, $||r_i|| = r_i^T r_i = 1$. For the 2D case, the equation of R is thus constrained to:

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$$

As it can be seen, the rotation matrix in 2D only has one degree of freedom, θ due to the constraints of orthogonality: The modulus of the 2 row vectors is 1 and their dot product is 0. In 2D, this is the angle in which we will rotate the space. For the D-dimensional case, we could have any combination of 2 angles in the matrix. Which boils down to the number of combinations out of the D-Dimensions.

The **properties of any scaling matrix** S are that it is a diagonal matrix will tell us the scaling of the dimensions. The i-th element of the diagonal tells us how much we scaled the i-th original dimension of U. In the 2D case we would have:

$$S = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix}$$

We can use different combinations of these rotation and scaling transformations to create any linear projection A. In 2D these transformation will convert squares into romboids and circles into ellipsoids. Is there a way to mathematically know how the space is being transformed by the matrix A ?

The good news are that yes we can. We can **decompose our matrix A into 3 matrices** that each performs a specific task. One that performs an initial rotation V^*, another one that performs the scaling S and a final one that performs the final rotation U. This decomposition can be express in the following formula:

$$A_{nxn} = U_{nxn} S_{nxn} V^*_{nxn}$$

Notice how in this equation the rotation matrix V^* is to the right, meaning that is the first matrix that is multiplying the original space given by U, so in this decomposition we will be first rotating, then scalling and then rotating again. If V is not a complex matrix, then $V^* = V^T$.

The decomposition of A into the previous components is rather complicated, a way of doing so is using the **Singular Value Decomposition (SVD)**, which is a more general approach applicable to matrices that do not need to be squared. We will not be explaining this method in this document, we just need to know that it is possible. In the following Figure we have the 3 transformations of the matrix A, an initial rectangle has been also added to highlight the transformation of the initial space.

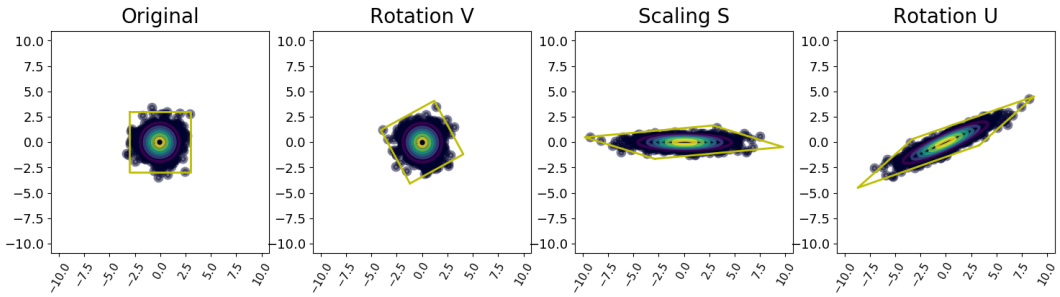


Figure 13: SMA and its window

This decomposition is possible as long as A is a nice transformation of the space, that is, it creates a proper new basis, which is boiled down to the fact that A is positive semidefinite, $det(A) > 0$. The new decomposed matrices have a set of properties that we will explain next.

- **The initial rotation matrix** V^*. This matrix rotates the space so that the later scalling can transform the square into a romboid. Notice that it does not affect the circle, just the square.

- **The scaling matrix** S. It is a diagonal matrix that will tell us the scaling of the dimensions rotated by V^*. The i-th element of the diagonal tells us how much we scaled the i-th original dimension of U.

- **The final rotation matrix** U. It is the final rotation of the space to be able to rotate anywhere. As we can see, the angle of the rotation of U is the angle of the rotation of the ellipse, since the intial rotation does not really change anything regarding the circle, unlike regarding the square. No matter the rotation of V, the ellipse will have the angle given by U. So for all that matters to ellipses, V is not used.

As we can observe, in this 2D example, the 3 decomposed matrices only contain 4 degrees of freedom, the same as the original matrix A. The angle of V^*, denoted as θ, the diagonal of S, denoted by s_1, s_2 and the angle of U, denoted as ϕ. In the previous example these values are:

$$A = USV^* = \begin{bmatrix} -0.91 & -0.415 \\ -0.415 & 0.91 \end{bmatrix} \begin{bmatrix} 2.40 & 0 \\ 0 & 0.40 \end{bmatrix} \begin{bmatrix} -0.48 & -0.88 \\ -0.88 & 0.48 \end{bmatrix}$$

As we can see, the rotation matrices obey the conditions previously described, we can compute the initial angle of rotation as $\theta = cos^{-1}(-0.48) = 0.66\pi[rad]$ radians and the final angle of rotation as $\phi = cos^{-1}(-0.91) = 0.86\pi[rad]$. From the scaling values we can see that this transformation enlarges X_1 since $s_1 > 1$ and compresses X_2 since $s_2 < 1$.

This is not the only way to express the matrix, we could also **decompose** A **into an stretching matrix** P **which will deform the space and a rotation** R. We can directly obtain it from the SVD decomposition as:

$$A = USV^* = (UV^*)(VSV^*) = RP$$

Where it is straight forward to see that:

- **The streaching matrix P**: Is obtained as the product $P = VSV^*$. This transformation rotates the space, scales it and rotates it back to the original dimensions.

- **Rotation R**: This is the final rotation of the space $R = UV^*$. Its angles will tell us the rotation of the romboid or ellipses respect to the original space.

Using the transformation matrix of the previous example, lets see the intermediate results of the the transformation in the next Figure. It can be observed that P both scales and rotates the space and R rotates it back to the correct angle so that $A = RP$.

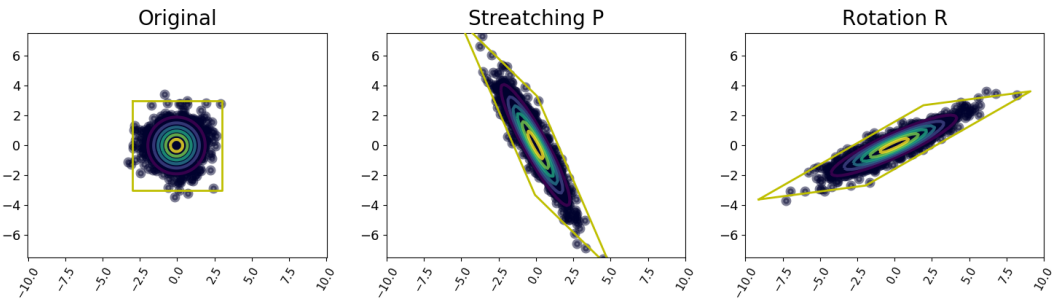


Figure 14: SMA and its window

We can also easily see how this decomposition allows us to express the variance-covariance matrix of the distribution in terms of the rotation and scaling matrices. Using the last decomposition we can reduce the covariance matrix to:

$$\Sigma_X = AA^T = RPP^*R^* = US^2U^T$$

As we can see, the covariance is equal to twice the rotation of the scaling matrix squared. The covariance matrix is a symetric matrix. Later we will draw connections between this fact and other properties of the distribution. Finish Description !! As we can see, the rotation of this matrix is the same as that of the matrix A, what differs is the scalling, which has been squared.

2.4.3 The Variance-Covariance Matrix

So far we know that the variance-covariance matrix indicates the variance of the independent components of the distribution and the covariance (linear relationship) among the variables. In this Section we will develop on more properties of the variance-covariance matrix and how these relate to the geometric properties of the distribution. Let us start with the general covariance matrix Σ_X:

$$\Sigma_X = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{1D} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_{DD} \end{bmatrix}$$

First of all, not every square matrix can be a covariance matrix, for example we know that:

- The covariance matrix has to be equal to its transpose: $\Sigma_X = \Sigma_X^T$.
- The elements of the diagonal, σ_{ii} have to be positive since they are the variance of the marginal variables X.
- We also know that the modulus of the normalized covariances $\frac{\sigma_{ij}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}, i \neq j$ cannot be bigger than 1.

All these properties that Σ_X must have come from the fact that it comes from the transformation matrix A in the form $\Sigma_X = A^T A$. They can be easily seen from the 2D example:

$$\Sigma_X = AA^T = \begin{bmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{21}a_{11} + a_{22}a_{12} & a_{21}^2 + a_{22}^2 \end{bmatrix}$$

As we can see, $\sigma_{12} = \sigma_{21}$ so the matrix is equal to its transpose. The elements of the diagonal are always the sum of squared values so they are positive, and the squared normalized covariance is:

$$\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} = \frac{a_{11}^2 a_{21}^2 + a_{12}^2 a_{22}^2 + 2(a_{11}a_{12}a_{21}a_{22})}{a_{11}^2 a_{21}^2 + a_{12}^2 a_{22}^2 + a_{11}^2 a_{22}^2 + a_{12}^2 a_{21}^2}$$

By saying: $C = a_{11}^2 a_{21}^2 + a_{12}^2 a_{22}^2 + 2(a_{11}a_{12}a_{21}a_{22})$ we have:

$$\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} = \frac{C}{C + (a_{11}a_{22} - a_{12}a_{21})^2}$$

So this value will always be lower than 1, and therefore the normalized correlation has module less than 1. We can also see the what the normalized covariance matrix will be like from this equation.

There properties can be summarized in the fact that Σ_X has to be positive definite, that means that given any vector w, the projection:

$$VAR[Y] = w^T \Sigma_X w = \sum_{i=1}^{D} \sigma_{ii} w_i^2 + 2 \sum_{i=1}^{D} \sum_{j<i} \sigma_{ij} w_i^T w_j > 0$$

Is this a big deal for the Multivariate Normal distribution ? Well, if we perform a linear projection using the vector w of the D-dimensional space of X to create a new univariate variable Y:

$$Y = [y] = w^T X = \begin{bmatrix} w_1 & w_2 & \cdots & w_D \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$$

The new variable Y is a Gaussian univariate random variable, since it is composed of a linear combination of Guassian random variables. Its distribution is easily obtainable by substituting X by $Y = w \cdot X$ in the density functions. Its mean and variance are trivially:

$$\mu_Y = w^T \mu = \begin{bmatrix} w_1 & w_2 & \cdots & w_D \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix}$$

$$VAR[Y] = \sigma_Y^2 = w^T \Sigma_X w = \begin{bmatrix} w_1 & w_2 & \cdots & w_D \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{1D} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_{DD} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

As we can see the fact that the matrix is semipositive difinite implies that the variance of any of the projections is greater than 0 ! Since it is variance cannot be negative, and if it is 0, then it is not a random variable anymore, just a point without uncertainty, this means that corresponding projection matrix A of the distribution does not fill all the space. So, at this point, it is obvious that the covariance matrix should be positive definite respectively positive semidefinite so that the variance of any projection is greater than 0.

This is checked with all the eigenvalues positive !!!

The next image shows the projection of the Multivariate Gaussian for several vectors w. Starting with a 2D Gaussian distribution:

- We plot in black the marginal distributions, from which we can observe the variance. These marginal distributions would be the projection of the original distribution multiplied by unitary vectors $U_1 = [10]U$ and $U_2 = [01]U$.
- Then we plot in green the distribution of multiplying the initial variables by the red and green projection vectors seen. As we can see, the projected variances are also bigger than 0 and while one decreases, the other increases. As we will see the sum of variances is always the same if we multiply the original distribution by a rotation matrix.

as we can see, no matter what w is, the variance of the projections is bigger than 0, unless $w = 0$.The new values will be the distance to they hyperplane they represent.

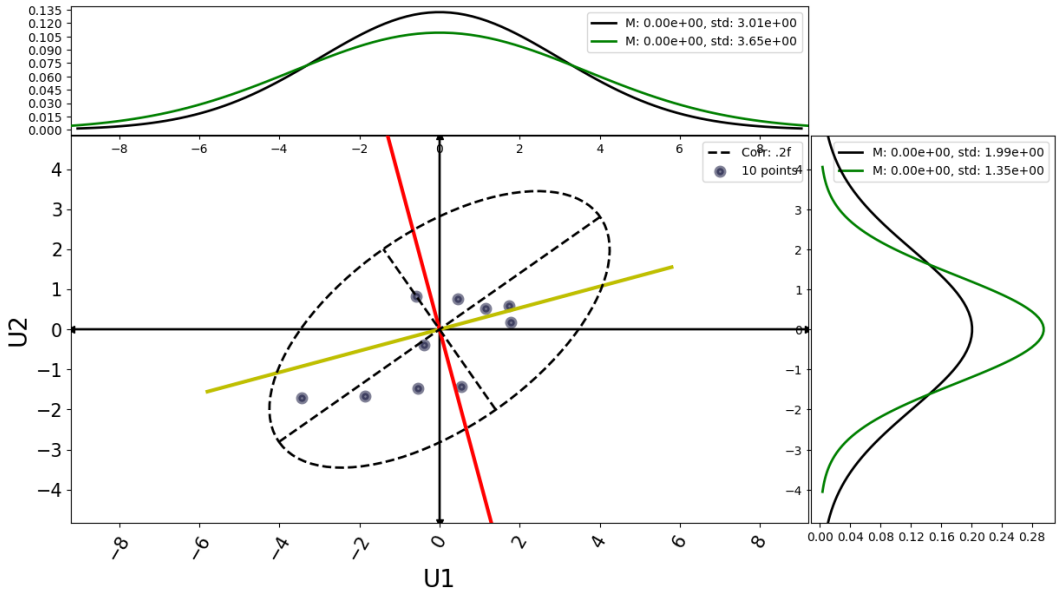


Figure 15: SMA and its window

REWRITE THIS PART In 2D the hyperplane is a line, so the projection is the distance using a vector perpendicular to v to the point. Same as the original axis is the same as the units.

The same way that in the original axis you measure the U1 value of a point by its perpendicular distance to the y-axis.
If the initial ellipse had some mean, then the rotation would rotate the same degrees but there would be also a change in mean, specifically, the new mean would be XXXX.
If we multiply by a scaling matrix then as we saw, we are multyplying the diagonal and therefore the angle of the ellipse change, namelly for example the conditional mean slope changes by...
Probably the equation of the combined decompositions !!!
S is increasing the variance of the samples without increasing their covariance ! We are adding or deleting noise (increasing or reducing variance) by multiplying the space by the covariance matrix.

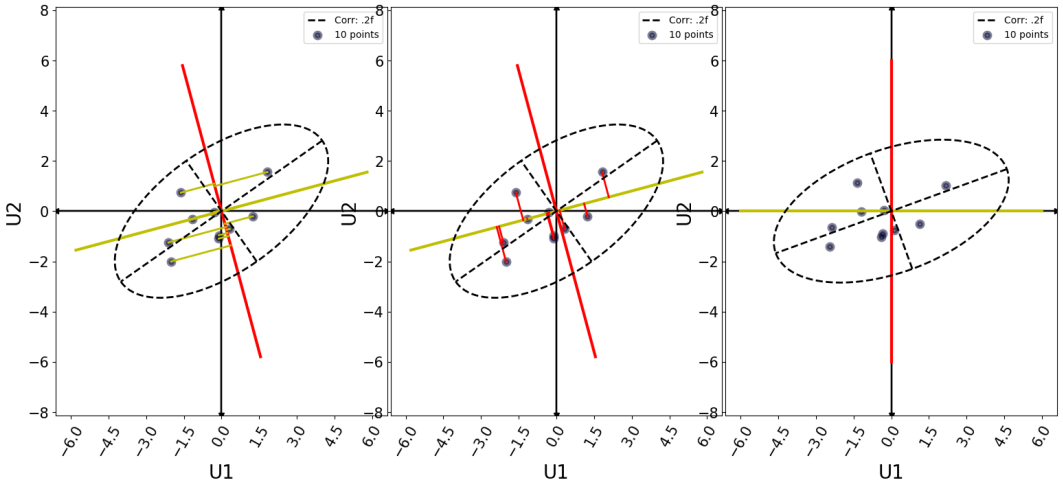


Figure 16: SMA and its window

Show matematically that the sum of diagonal elements does not change upon rotation ? It is equal to the volume of the space.

Perform exercise of Q2 exam 2012 here !

Equation of the associated hyperplane to a projection vector:

2.4.4 Linear Transformation of the Distribution

Now that we gained some knowledge about how we are creating the Multivariate Distribution from scaling, rotation and translations, and we have an understanding of the Covatiance Matrix; let us see some transformations of the Gaussian Distribution. Once we have a Multivariate Gaussian Distribution we could scale and rotate again by multiplying it by more matrices. Let us see how this changes the properties of the distribution.

We could **rotate the space of the Gaussian Distribution** $X \sim N(\mu, \Sigma_X)$ by multiplying the variables with a rotation matrix R. To do so, we first remove the mean μ_X and then we add it back after the linear transformation. In the case of 2D we can write:

$$Y = R(X - \mu_X) + \mu_Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

The next Figure shows the above operation for different angles of rotation θ and same mean, $\mu_Y = \mu_X$. As we can see the result are just rotations of the original space, the scales are not changed, and therefore the total variance of the distribution and normalization constant are kept the same. Looking at this we can argue that we could actually rotate the space in a way that both variables are independent, and we can ! This part will be covered later.

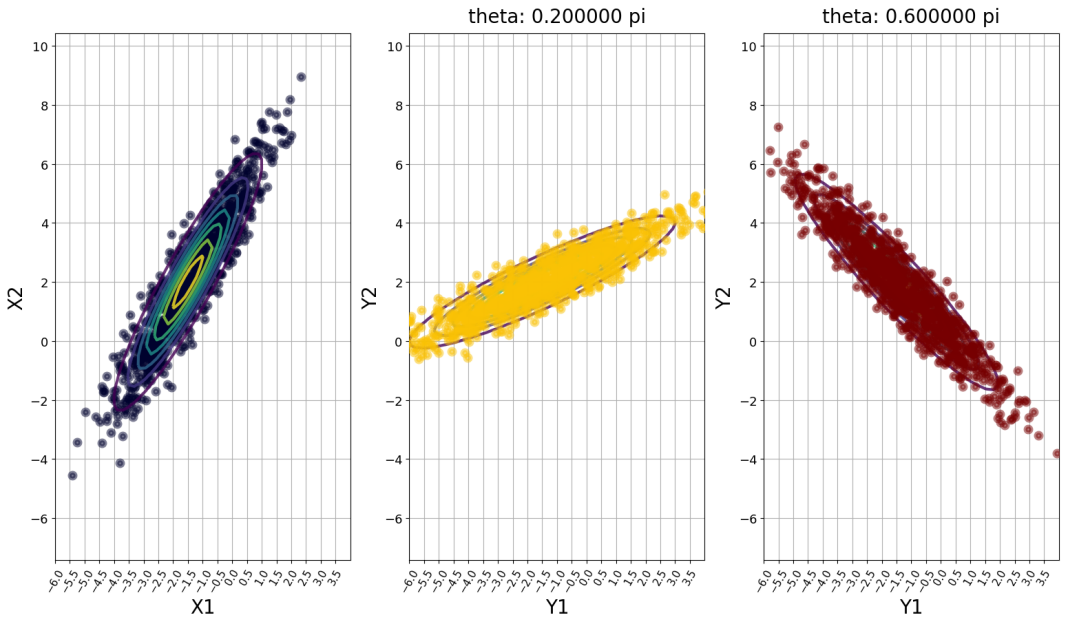


Figure 17: SMA and its window

The covariance matrix of the distribution is being changed by this rotation matrix R. Let us see the effect in the distribution by just plugging the transformed values. Solving the equation we have that:

$$X - \mu = R^{-1}(Y - \mu)$$

Pluging it into the equation, we find:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_X|}} exp\left[-\frac{1}{2}(Y-\mu)^T R^{-1T} \Sigma_X^{-1} R^{-1}(Y-\mu)\right]$$

So the covariance matrix of the new transformed space is:

$$\Sigma_Y = R^T \Sigma_X R$$

And the normalization constant does not change since we have not scaled the space, only rotated it, $|\Sigma_X| = |\Sigma_Y|$. So... in order to find the rotation R that makes the variables independent we have to find the rotation that makes the Σ_Y a diagonal matrix ! We can also observe from this equation that the covariance between 2 projected variables $Y_1 and Y_2$ is:

$$COV(Y_1, Y_2) = \sigma_{12} = P_1^T \Sigma_X P_2^T$$

In the same way, **we could just scale the space of the Gaussian Distribution** $X \sim N(\mu, \Sigma_X)$ by multiplying the variables with a scaling matrix S. To do so, we first remove the mean μ_X and then we add it back after the linear transformation. In the case of 2D we can write:

$$Y = S(X - \mu_X) + \mu_Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} s_1(X_1 - \mu_1) \\ s_2(X_2 - \mu_2) \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

The next Figure shows the above operation for different scalings of the space and same mean, $\mu_Y = \mu_X$. As we can see the result are just scaling. The scaling after a rotation changes the angle of rotation !!!

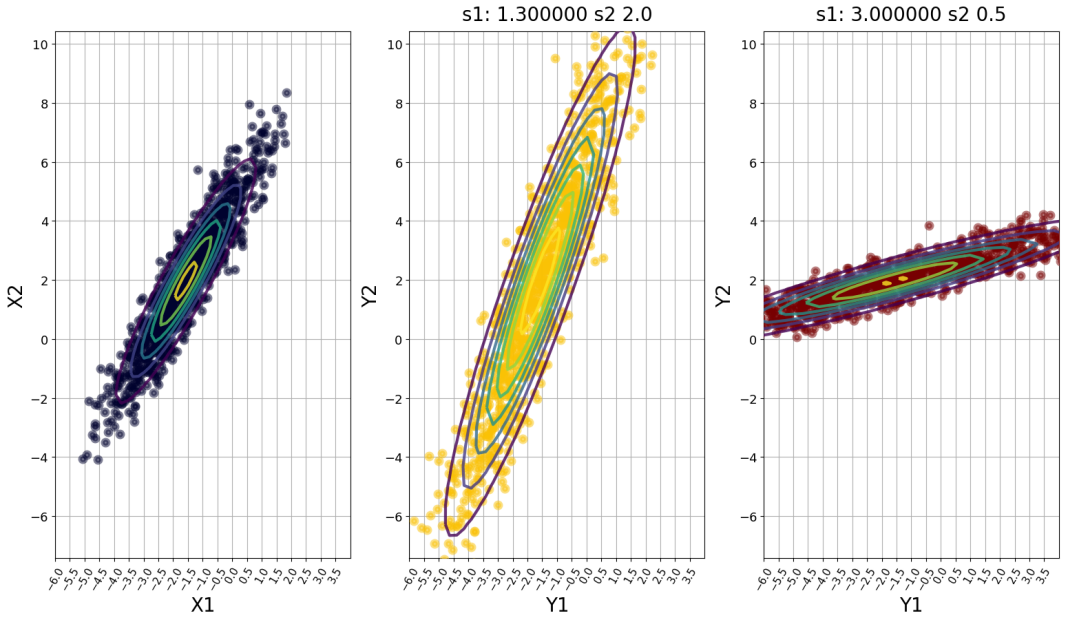


Figure 18: SMA and its window

The covariance matrix of the distribution is being changed by this scaling matrix S. Let us see the effect in the distribution by just plugging the transformed values. Solving the equation we have that:

$$X - \mu = S^{-1}(Y - \mu)$$

Pluging it into the equation, we find:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_X| \prod_{i=1}^{D} s_i^2}} exp\left[-\frac{1}{2}(Y-\mu)^T S^{-1T} \Sigma_X^{-1} S^{-1}(Y-\mu)\right]$$

So the covariance matrix of the new transformed space is:

$$\Sigma_Y = S^T \Sigma_X S = \begin{bmatrix} s_1^2\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & s_2^2\sigma_{22} & \cdots & \sigma_{1D} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & s_D^2\sigma_{DD} \end{bmatrix}$$

As can see, we have multiplied the diagonal of the covariance by the squared coefficients of the scalling. And since we have scaled the space, now the normalization constant has changed by volume of the new transformation which is $det(S) = \prod_{i=1}^{D} s_i$.
As we can see it is not like we have expand the whole space but rather the independent components, we have decreased the angle between the variables, not they are less "correlated" which can be seen from the equation of correlation. It is like dividing all previous correlations by $s_1 \cdot s_2$ so the correlation will increase if this product is less than one and it will decrease if it is more than one.

As we can see, looking at this decomposition we can express the matrices in a useful way... XXXXX

TODO: Say what is the variance of the linear projection !

2.4.5 Ellipsoids

So far we have visually seen that the equiprobably areas of the Gaussian Distribution follow Ellipsoids. Can we find the equation that indicate them ? Yes we can. Lookin at the equation of the pdf we can see that the points with the same argument in the exponential will have the following probability:

$$(X-\mu)^T \Sigma_X^{-1} (X-\mu) = c$$

In the case of 2D, the inverse covariance Matrix has the equation:

$$\Sigma_X^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix}$$

Being $\sigma_{21} = \sigma_{12}$ we obtain the equation:

$$c = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \left[\sigma_{22}(X_1 - \mu_1)^2 + \sigma_{11}(X_2 - \mu_2)^2 - 2\sigma_{21}(X_1 - \mu_1)(X_2 - \mu_2)\right]$$

Resolving the multiplications and ordering terms we get:

$$c = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \left[X_1^2(\sigma_{22}) + X_2^2(\sigma_{11}) + X_1X_2(-2\sigma_{12}) + X_1(-2\mu_1\sigma_{22} - \mu_2) + X_2(-2\mu_2\sigma_{11} - \mu_1) + \sigma_{22}\mu_1 + \sigma$$

Talk about how we relate the angle, the displacement and so on to this equation and plot the generalized 3D version as well, also de 2D.
So pretty much, form the covariance matrix we can find how exactly the variables are linearly related and therefore compute its exact ellipsoids.

2.4.6 The Correlation Matrix

We have already seen that the covariance between 2 variables, σ_{ij} is a meausure of how related are the 2 variables. It gives a meausure of their dependence, in this case, linear dependence. The problem is that its meaning is relative to the variances of the independent variables, σ_{ii}, σ_{jj}. Even if σ_{ij} is very big, the variables could actually be barely related if their variances are very high. For this purpose we have the measure of correlation between two variables.

The correlation is a normalized measure on how linearly related are 2 reandom variables. The correlation between the random variables X_1 and X_2, denoted as ρ_{ij} is computed as the covariance between the variables $COV(X_1, X_2)$ divided by the product of the standard deviation of the independent variables, $STD(X_1)STD(X_2)$.

$$\rho_{ij} = \frac{E[(X_1 - \mu_1)(X_2 - \mu_2)]}{\sqrt{E[(X_1 - \mu_1)^2]E[(X_2 - \mu_2)^2]}} = \frac{Cov(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}}$$

The same way we have the variance-covariance matrix Σ_X, we also have the correlation matrix R_X, whose parameters can be computed directly from the variance-covariance matrix as:

$$R_X = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1D} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2D} \\ \vdots & \vdots & & \vdots \\ \rho_{D1} & \rho_{D2} & \cdots & \rho_{DD} \end{bmatrix} \qquad \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}$$

As we can see, the correlation between a random variable and itself is $\rho_{ii} = 1$ since $\sigma_{ii} = \sigma_i\sigma_i$. For the 2D case, we have the correlation matrix:

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix} = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\sigma_2} \\ \frac{\sigma_{21}}{\sigma_1\sigma_2} & 1 \end{bmatrix}$$

The good news is that now, the correlation ρ_{ij} gives a normalized measure of the linear relationship between X_1, X_2 that takes into account the variance of the marginal variables. The value of the correlation can only go between -1 and 1, that is $\rho_{ij} \in [-1, 1]$.

The angles is given by the ratio ?
But we also miss information with respect to the covariance matrix, the correlation cannot tell us what is slope between the variables, it only tells us how noisy is the linear relationship and if it positive or negative. On the other hand, it only tell us about the linear relationship but there could be many other relationships. The next Figure shows the computed correlation for several sets of data. As we can see:

- In the first row we see how the absolute value of the correlation indicated the noisiness of the linear relation and the sign.
- In the second we see how different slopes give the same correlation, so correlation cannot tell us the angle.
- The last row shows us that linear relationship is only one type of dependence between the variables so correlation cannot be used to model the entire depende etween them.

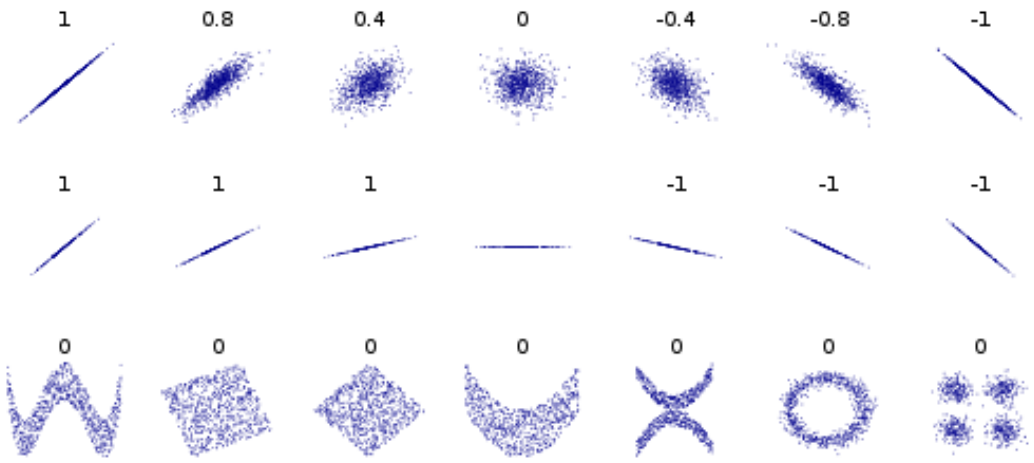


Figure 19: SMA and its window

We can also express the variance-covariance matrix using the correlation coefficients, since the covariance between 2 random variables is $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, therefore we get:

$$\Sigma_X = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1D}\sigma_1\sigma_D \\ \rho_{21}\sigma_2\sigma_1 & \sigma_D^2 & \cdots & \rho_{2D}\sigma_2\sigma_D \\ \vdots & \vdots & & \vdots \\ \rho_{D1}\sigma_D\sigma_1 & \rho_{D2}\sigma_D\sigma_2 & \cdots & \sigma_D^2 \end{bmatrix}$$

Expressing the variance-covariance matrix in this way allow us to make a lot of simplification and gain insight about the distribution and the correlation coefficient. This is shown when computing the conditional distribution $f(X_j|X_i)$, where the correlation coefficient shows its value.

In the 2D case, we saw that the conditional variance $V[X_2|X_1]$ was always smaller than $V[X_2]$, we have gained knowledge about X_2 if we know X_1. We can re-express this conditional variance using the correlation coefficient as:

$$E[X_2|X_1 = x_1] = \mu_2 + \rho_{21}\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$
$$V[X_2|X_1 = x_1] = \sigma_2 - \rho_{21}^2\sigma_2 = \sigma_2(1 - \rho_{21}^2) = \gamma$$

As we can see in the conditional variance equation, the square correlation coefficient ρ_{ij}^2 **is the proportion of variance from variable X_i that can be explained by X_j and viceversa.** This how much uncertainty we can reduce about X_i if we know X_j and viceversa. We can express the squared coefficient as:

$$\rho_{ij}^2 = \frac{V(X_1) - V(X_1|X_2)}{V(X_2)} = \frac{V(X_2) - V(X_2|X_1)}{V(X_2)} = \frac{\sigma_2^2 - \gamma}{\sigma_2^2}$$

So the correlation coefficient can be used as a measure of how much information 2 random variables share. Which also means how much uncertainty we can remove from one, if we know the other.

2.4.7 Partial Correlation

In the D-dimensional case, knowing the value of 1 or several of the variables can reduce the uncertainty of all the rest if they are correlated. As we previously saw, the conditional correlation matrix can be computed as:

$$E[Y|X] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) = \begin{bmatrix} \mu_{Y1} \\ \mu_{Y2} \\ \vdots \\ \mu_{D_Y1} \end{bmatrix}$$

$$V[Y|X] = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1D_Y} \\ a_{21} & a_{22} & \cdots & a_{2D_Y} \\ \vdots & \vdots & & \vdots \\ a_{D_Y1} & a_{D_Y2} & \cdots & a_{D_YD_Y} \end{bmatrix}$$

Now $V[Y|X]$ is the covariance matrix of the variables $Y = [Y_1, Y_2, ...Y_{D_Y}]$ when we know which value took the variables $X = [X_1, X_2, ..., X_{D_X}]$. This covariance matrix has already been reduced the uncertainty due to the correlation between Y and X. The correlation coefficients between two random variables Y_i and Y_j, given the variables X can be obtained from the conditioned covariance matrix $V[Y|X]$ as if it was an independent distribution. We have the conditional correlation coefficients:

$$\rho_{ij|X} = \frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}$$

But this is not very informative for us, nothing we did not know already, the a_{ij} values are too abstract for use to gain knowledge from this as normal humans. More interesting it would be to express $\rho_{ij|X}$ as a function of other correlation coefficients, this is what we will show next.

Lets start with 3 random variables, X_1, X_2, X_3, computing the variance-covariance matrix of $V[X_1, X_2|X_3]$ it can be easily proven that:

$$V[X_1, X_2|X_3] = \begin{bmatrix} \sigma_1^2(1-\rho_{13}^2) & \sigma_1\sigma_2(\rho_{12}-\rho_{13}\rho_{23}) \\ \sigma_1\sigma_2(\rho_{12}-\rho_{13}\rho_{23}) & \sigma_2^2(1-\rho_{23}^2) \end{bmatrix}$$

We can observe that the variance of the variables X_1 and X_2 has decreased according to the square of their correlation coefficient with X_3 as it was previously seen. Regarding the covariance between X_1 and X_2 we see a reduction proportial to the correlation of both variables and the conditioned variable X_3. We can express the conditioned correlation coefficient as:

$$\rho_{12|3} = \frac{a_{12}}{\sqrt{a_{11}a_{22}}} = \frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

As we can see, we can express the conditioned correlation, that is, the correlation between some variables, given the others, using the original correlation coefficients. XXX maybe talk more about the equation.

For any other number of conditioned variables we can iteratively condition using the rule several times... The meaning of $\rho_{ij|X}$ is still unchanged, it is the same concept applied to a conditional probability, it means the correlation between the variables Y_1 and Y_2 given the variables X. And the value $\rho^2_{ij|X}$ means the proportion of variance that can be explained of Y_i if we know Y_j given X.

It should be noted that the correctness of the interpretation of variances and correlations is very dependent on the data actually following a Gaussian distribution, if it doesnt then, things do not really make sense.

Develop the other relation with volumes of the space !
Page 40 of week 2.

2.4.8 Eigenvalues and EigenVectors

Now that we have a good intuition about the Multivariate Gaussian Distribution and its geometrical properties. Let us introduce the eigenVector and eigenValues of a Matrix. This will help up gain new insights of the Gaussian Distribution and how the variances, correlations, dependencies and shape are related.

Lets start first defining what eigenvalues and eigenvectors are. Given a square matrix A, a non-zero vector v is an eigenvector of A if:

$$Av = \lambda v \qquad \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1D} \\ a_{21} & a_{22} & \cdots & a_{2D} \\ \vdots & \vdots & & \vdots \\ a_{D1} & a_{D2} & \cdots & a_{DD} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{bmatrix}$$

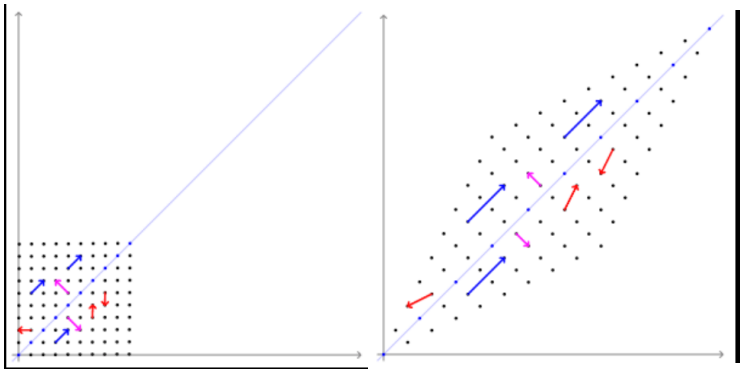
If the condition is satisfied then x is an eigenvector of A and λ is its associated eigenvalue. This condition can also be expressed as:

$$(A - \lambda I)v = 0 \qquad \begin{bmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1D} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2D} \\ \vdots & \vdots & & \vdots \\ a_{D1} & a_{D2} & \cdots & a_{DD}-\lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Geometrically this has a very nice intuition. Being the rows of A the linear transformation of the space into a new space. Then the eigenvectors of A will be the vectors that will not change direction in this transformation. For example the next Figure shows a 2D example where the space has been transformed. As we can see:

- The set of vectors in the direction of dark blue and light blue vectors would be the eigenvectors since their direction is unchanged.
- The rest of vectors, in red, will always change its direction.
- The blue eigenvector are always scaled by the same value. In this case the dark blue eigenvectors are scaled by a big number and the light blue by a small one.

We can think of A as a system, a function that is applied to an input vector v, transforming it in modulus and direction. The eigenvectors of the system are those vectors v that do not change direction, only modulus and thus, the output of the system is equal to the input multiplied by a constant. This concept is similar to the eigenfunction of a linear system.



(a) sizeInches = [2,2]

(b) sizeInches = [10,6]

Figure 20: Effect of the figure size in the saved figures

This is all very nice but...

- How do we compute the eigenvalues and eigenvectors ?
- How are they related to the Multivariate Gaussian distribution and why are they useful ?

Well, lets start with the math regarding computing them. One way of computing them is directly by computing the charactetistic polynomical of the function:

$$det(A - \lambda I) = 0 = f(\lambda) = A + B\lambda + C\lambda^2 + \cdots$$

This will yield the polynomial equation λ whose roots for λ will give us the eigenvalues $\lambda_1, \lambda_2, ...\lambda_D$. We can then compute the eigenvectors $v_1, v_2, ..., v_N$ associated to the eigenvalues by substituting them one by one in the equation $(A - \lambda I)v = 0$, which solution will give us the eigenvectors.

In the case of a 2D transformation we will have:

$$det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda) - a_{21}a_{12} = \lambda^2 + \lambda(-a_{11} - a_{22}) + (a_{11}a_{22} - a_{12}a_{21})$$

The roots of this polynomial and therefore the eigenvalues follow the form:

$$\lambda_i = \frac{(a_{11} + a_{22}) \pm \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}}{2}$$

We can then just plug each of the roots λ_i obtained into the system to obtain the corresponding eigenvectors v_i.

$$(A - \lambda_i I)v_i = \begin{bmatrix} a_{11} - \lambda_i & a_{12} \\ a_{21} & a_{22} - \lambda_i \end{bmatrix} \begin{bmatrix} v_{1i} \\ v_{2i} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

It is work noting that the eigenvalues:

- Can be complex, $\lambda_i = a + bj$. Comment about this.
- Can be repeated. In this case the eigenvector associated is a eigenspace of dimension equal to the one given.

So far nothing too useful, but in fact these **eigenvalues and eigenvectors have some fantastic properties** that we will see next. These properties are even better if we assume that the matrix A is positive semidefinite, as it is the case of the covariance matrix of the Multivariate distribution !

- The set of eigenvectors $v_1, v_2..., v_D$ obtained for all the eigenvalues are **orthogonal,** $v_i^T v_j = 0$ for $i \neq j$ and therefore they form a new basis of the space. To prove it we just need to combine the equations for 2 different eigenvalues $Av_i = \lambda_i v_i$ and $Av_j = \lambda_j v_j$. Multyplying each for the other vector, transposing one and setting equality:

$$x_i^T A^T x_j = \lambda_i x_i^T x_j \qquad \lambda_j x_i^T x_j = \lambda_i x_i^T x_j$$

  The only possibility for the equality to be true is that $v_i^T v_j = 0$, so the vector are orthogonal. If we select the eigenvectors with modulus one, $||v_i|| = v_i^T v_i = 1$ then we have a new **orthogonal basis of the space** ! In 2D.

$$V = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix} & \begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

  This of course trivially implies that $V^T V = I$

- The nice property here is that if we multiply A by this new basis V then we obtain a domain a matrix composed by the eigenvectors multiplied by the eigenvalues:

$$AV = [\lambda_1 v_1, \lambda_2 v_2, \cdots, \lambda_D v_D]$$

  We can express this in a linear way by creating the diagonal matrix Λ which has the eigenvalues as the diagonal values:

$$AV = V\Lambda = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Now things get interesting, if **A is a symetric matrix**, as it is the case of the covariance matrix Σ_X then we can multiply both righ-hand sides of the equation by V^T, and since A is symetric $VV^T = V^T V = I$ and therefore we can express A in terms of V and Λ. In the 2D case we have the equation:

$$\Sigma_X = V\Lambda V^T = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

This is the exact same equation as the SVD decomposition for symmetric Matrices we saw before where $\Sigma_X = US^2U^*$. So computing the eigenvectors and eigenvalues we can decompose the matrix into the rotations of the axis of the ellipses $U = V$ and the scaling of the axis $\Lambda = S^2$. We can also reexpress Σ_X in terms of the sum of the individual basis as:

$$\Sigma_X = \lambda_1 v_1^T v_1 + \cdots + \lambda_D v_D^T v_D = \sum_{i=1}^{D} \lambda_i \begin{bmatrix} v_{1i} \\ v_{2i} \\ \vdots \\ v_{Di} \end{bmatrix} \begin{bmatrix} v_{1i} & v_{2i} & \cdots & v_{Di} \end{bmatrix} = \sum_{i=1}^{D} \lambda_i v_i^T v_j \begin{bmatrix} v_{1i}^2 & v_{1i}v_{2i} & \cdots & v_{1i}v_{Di} \\ v_{2i}v_{1i} & v_{2i}^2 & \cdots & v_{2i}v_{Di} \\ \vdots & \vdots & \vdots & \vdots \\ v_{Di}v_{1i} & v_{Di}v_{iD} & & v_{Di}^2 \end{bmatrix}$$

This partitioning of the symmetrical matrix Σ_X is often called its spectral decomposition, since the eigenvalues $\lambda_1, \lambda_2, ...\lambda_D$ are called the spectrum of the matrix. From this previous equation we can see that the matrix Σ_X is symetric and the diagonal is positive if all the eigenvalues are positive, $\lambda_i > 0$.

We can further trivially say that $S = diag(s_1, s_2, ..., s_D)$ and $S = \Lambda^{1/2} = \sqrt{\Lambda} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ...\sqrt{\lambda_D})$. In this way we can express Σ_X as:

$$\Sigma_X = (V\Lambda^{1/2})(V\Lambda^{1/2})^T = AA^T$$

As we can see, the previous way of expressing the symmetric covariance matrix Σ_X can be. The eigenvectors V are a base of the space in which:

- The projected random variables are independent.
- The marginal variance of the independent components is maximized and equal to the diagonal of Λ.

- The axis of rotation of the hyper-ellipse are the columns of V.

As we can see, this properties are very useful for many applications, for example the PCA where we aim to obtain the independent linear transformations that maximize the projected variance. It is also nice for plotting the ellipse since what we need is to start with the unit hyper-circle, then scale it by $\Lambda^{1/2}$ and then rotate it by V.
CHART OF THE ELLIPSE AND THE ANGLE OF ROTATION.

2.4.9 Correlation, Covariance and Linear Regression

In this Subsection we will see how normal Linear Regression is related to the Correlation of the original X. We will begin by describing the process obtaining the linear regression without any explicit satistical formulation and then we will see how the results relate to the estimation assuming gaussianity in the data.

In simple **Linear regression**, we want to be able to estimate the variable Y as a linear combination of a set of input features $X = [X_1, X_2, ...X_D]$. We want to be able to find the best linear projection with coefficients $\theta = [\theta_0, \theta_1, ..., \theta_D]$ so that:

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_D X_D + \epsilon = \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_D\end{bmatrix} \begin{bmatrix}1 \\ X_1 \\ \vdots \\ X_D\end{bmatrix} + \epsilon$$

Where the term ϵ is the difference between the actual Y and out estimated value $\hat{Y} = \theta X$. In this scenario without explicit statistical assumptions, our goal is to minimize the squared error of the regression, $\epsilon^2 = (Y - \hat{Y})^2$ given the set of N data points that we have $\{Y_i, X_i^D\}^N$. This previous notation means that we have $i = 1, ..., N$ samples of the pair of data points Y_i and X_i^D, the exponent meaning that X_i is D-dimensional.

In order to be able to use complete matrix notation including the bias parameter θ_0, we will add a bias variable X_0 to the D-dimensional X vector, which value will always be 1. So we have a new input variable vector $Z = [X_0 = 1, [X_1, ..., X_N]]$. So, given our dataset, our model in **transposed notation** is:

$$Y^T = \begin{bmatrix}Y_1 \\ Y_1 \\ \vdots \\ Y_N\end{bmatrix} = (\theta Z + \epsilon)^T = Z^T \theta^T + \epsilon^T = \begin{bmatrix}1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1N} \\ \vdots & \vdots & & \vdots \\ X_{D1} & X_{D2} & \cdots & X_{DN}\end{bmatrix}^T \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_D\end{bmatrix}^T + \begin{bmatrix}\epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N\end{bmatrix}^T$$

Aims to find the best θ that minimizes the sum of the squared errors of the regression, noted as S.

$$S = \epsilon\epsilon^T = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - \theta Z_i)^2 = (Y - \theta Z)(Y - \theta Z)^T$$

Mathematically speaking we want:

$$\hat{\theta} = arg\min_{\theta}\{(Y - \theta Z)(Y - \theta Z)^T\}$$

The error term of S is a convex function of θ so to obtain the optimal set of coefficients $\hat{\theta}$ we only need to derivate the term and equal to 0. Performing such operation and solving the equation we have that:

$$\hat{\theta}^T = (ZZ^T)^{-1} Z Y^T$$

These are the optimal values coefficients that minimize the squared error of the data, S. Now we will see its relation to the statistical framework when we assume Gaussianity in the data. We can already can see some relations from this equation to the Gaussian Distribution. The first inversed term is actually the inverse covariance matrix of the data if it data has 0 mean, $\Sigma_{ZZ}^{-1} = (Z^T \cdot Z)^{-1}$ and the later term is the covariance between Z and Y, namely $\Sigma_{ZY} = ZY^T$ if both Z and Y have 0 mean. But lets start the process from the beginning.

In a **Statistical Framework** we are going to assume that the objective variable Y and the input features X all follow a joint Gaussian distribution, $[Y, X_1, X_2, ...X_D] \sim N(\mu, \Sigma)$. For visualization purposes, in a matrix form we will have the parameters:

$$\mu = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix} = \begin{bmatrix} \mu_Y \\ \mu_{X_1} \\ \vdots \\ \mu_{X_D} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} = \begin{bmatrix} \sigma_{YY} & \sigma_{YX_1} & \cdots & \sigma_{YX_D} \\ \sigma_{X_1Y} & \sigma_{X_1X_1} & \cdots & \sigma_{X_1X_D} \\ \vdots & \vdots & & \vdots \\ \sigma_{X_DY} & \sigma_{X_DX_1} & \cdots & \sigma_{X_DX_D} \end{bmatrix}$$

In this scenario, the optimal estimation of Y given the rest of the variables is simply the conditional expectation $E[Y|X]$. We already presented conditional mean and variance in previous sections (XXequation), if we denote the matrix W_{YX} as the matrix resulting from multyplying the covariance between Y and W_{YX} by the inverse of the variance of Y, therefore $W_{YX} = \Sigma_{YX}\Sigma_{XX}^{-1}$ we can express the conditional mean and variance as:

$$E[Y|X] = \mu_Y + W_{YX}(X - \mu_X)$$
$$VAR[Y|X] = \Sigma_{YY} - \Sigma_{YX}W_{YX}^T$$

We can rewrite the expected mean $E[Y|X]$ in a way that it contains a bias and the linear coefficients. By doing this we can identify terms and see that we can express the linear regression coefficients, if we only have one variable Y to estimate, then W_{XY} is a vector and we have:

$$\theta_0 = \mu_Y - W_{YX}\mu_X \qquad \theta_{1:D} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_D \end{bmatrix} = W_{YX}$$

Notice that the linear coefficients θ define a hyperplane, a linear projection, the distance to a hyperplane, such that the square of theses distances is minimized for the given dataset. In 3D, W_{YX} is the equation of a plane, and in 2D, the equation of a line. Notice how the coefficients W_{YX} are different depending on which of the variables we choose as Y and the rest as X. Therefore the slopes of W_{YX} are different for each Y, unlike the main direction of the total distribution of $[Y, X]$, which is always equal to the first eigenvector v_1.

We see that $VAR[Y|X]$ is equal to $S(\hat{\theta} = Z^TY(Z^TZ)^{-1})$ so the linear projection given by the conditional variance minimizes the variance of the projected Y. As we can see, these are the coefficents that minimize the variance of the projected Y.
The central value is the eigenvector, namely:

$$\Sigma_{YX} = US^2U^T = \rho$$

So the slope is $\Sigma_{XY}v_1 = \lambda_1 v_1$
$v_1Y/v1_X$. Probably this equal to something like rho.

In the 2D case where both Y and X are one-dimensional, we have that the XXXX. Also put the . Notice that if the variables are uncorrelated then for one is 0 and for the other infinity.
Do it !!!

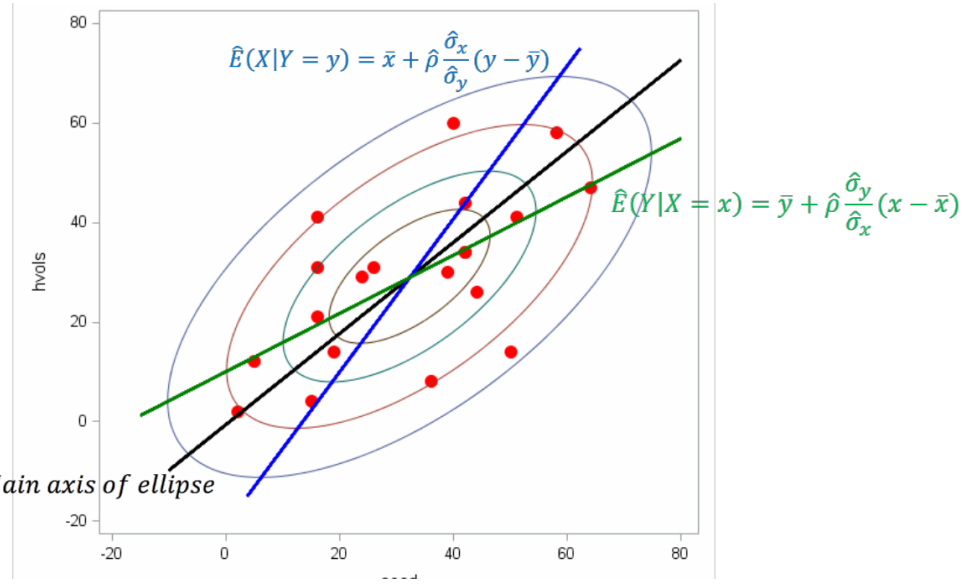


Figure 21: SMA and its window

Final words about exchanging variables. There is no causality in a gaussian distribution. Causal distributions.
In Regression we cannot just rotate the space and make the variables independent, we have to estimate the original Y. If we rotate it, we are note estimatimating Y but a linear combination $Z = aX + bY$.
The statistical framework also offers us optimal estimation in case that we only know a few of the random variables X_1, which can also be computes from the linear regresor by just leaving out the variable we do not know about.
So minimizing the squared error is equal to obtaining the best Expected shit !! Give this theorem !

If the regression line was the center, then the residual would no be gassian with the same variance no matter the x, but it would be dependend on x and the close to the center, the more noise.
PLOT CHART.

2.4.10 Multiple Correlation Coefficient

We have already seen that the correlation between two variables Y and X, denoted ρ_{YX} is a measure of how well we can be predicted one using a linear function of the other. Its squared value ρ^2_{YX} gives a measure of how much of the variance of Y can be explained knowing X and viceversa in the Gaussian Distribution.

The **Multiple Correlation coefficient** generalizes this concept. In this case we have a variable Y and a set of variables $X = [X_1, X_2, ..., X_D]$, the Multiple Correlation between them, noted as $\rho_{Y|X_1,X_2,...,X_D}$ is a measure of how well a given variable Y can be predicted using a linear function of a set of other variables X. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables.

As it was proven in the previous section the optimal combination of the X variables to linearly predict Y is given by the linear coefficients θ compued as:

$$\theta_{1:D} = \begin{bmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_D\end{bmatrix}^T = W_{YX} = \Sigma_{YX}\Sigma_{XX}^{-1} = \begin{bmatrix}\sigma_{YX_1} & \sigma_{YX_2} & \cdots & \sigma_{YX_D}\end{bmatrix} \begin{bmatrix}\sigma_{X_1X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_D}\\ \sigma_{X_2X_1} & \sigma_{X_2X_2} & \cdots & \sigma_{X_2X_D}\\ \vdots & \vdots & & \vdots\\ \sigma_{X_DX_1} & \sigma_{X_DX_2} & \cdots & \sigma_{X_DX_D}\end{bmatrix}^{-1}$$

As we can see in the equation, if Y is only one dimensional as it is the case, then the covariance Σ_{YX} is simply a row vector that we name σ_Y^T, therefore we have then the covariance matrix of the combined variables Y and X.

$$\Sigma = \begin{bmatrix}\Sigma_{YY} & \Sigma_{YX}\\ \Sigma_{XY} & \Sigma_{XX}\end{bmatrix} = \begin{bmatrix}\sigma_{YY} & \sigma_Y^T\\ \sigma_Y & \Sigma_{XX}\end{bmatrix}$$

We have that the Multiple Correlation coefficient between the variable Y and the variables X is defined as:

$$\rho_{Y|X_1,X_2,...,X_D} = \frac{\sqrt{\sigma_Y^T\Sigma_{XX}^{-1}\sigma_Y}}{\sqrt{\sigma_{YY}}}$$

Same as the correlation coefficient, this Multivariate Coefficient also tell us the reduction of variance in Y using a linear combination of the variables in X.

$$\rho^2_{Y|X_1,X_2,...,X_D} = \frac{V(Y) - V(Y|X)}{V(Y)} = \frac{\sigma_Y^T\Sigma_{XX}^{-1}\sigma_Y}{\sigma_{YY}}$$

Note as well that **this coefficient is not symmetric**, if we choose another variable to be Y then its multiple correlation coefficient with the rest of the variables will be different.

Notice as well the relation with the determinants of the matrix where we have that:

$$1 - \rho^2_{Y|X_1,X_2,...,X_D} = \frac{|\Sigma|}{\sigma_{YY}|\Sigma_{XX}|} = \frac{V(Y|X)}{V(Y)}$$

So we can compute the correlation coefficient from the determinant of the matrices and it will be valid when these are also valid. Notice this is a great **way to check if 2 sets of random variables Y and X are independent**. Since if the previous fraction is equal to 1 then the multiple correlation is 0 and therefore no linear combination of X can explain any of the variance of Y.

Also we could use the correlation matrix instead of the variation matrix. Since the correlation matrix is computed as $R = \Sigma S^{-1}$ where S is the diagonal matrix of the variances of the individual variances. Given that $|AB| = |A||B|$ we have:

$$1 - \rho^2_{Y|X_1,X_2,...,X_D} = \frac{|\Sigma|}{\sigma_{YY}|\Sigma_{XX}|} = \frac{|S||R|}{\sigma_{YY} \cdot |S_{XX}||R_{XX}|} = \frac{|R|}{|R_{XX}|}$$

It should be noted that in other notations, instead of using Y as the regressed variable, it is denoted as any other variable from the vector $X = [X_1, X_2, ..., X_D]$ where it is selected using the subindex i, and we only use another subset of the X to regress Y. In this scenario we have have that $Y = X_i$ and $X = [X_{m+1}, ..., X_p]$ so the dimensionality of the new X is $D = p - m$.

2.4.11 Correlation and the angle of rotation

In this Section we will see how the variance, covariance and correlation of the variables change as we rotate an Multivariate Gaussian distribution made with independent variables. Let us start with a 2D Multivariate Independent Gaussian distribution $U = [U_1, U_2]$, with 0 mean and variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 1$ respectively. Let us create the random variables $Y = [Y_1, Y_2]$ by means of multiplying U by a rotation matrix $R(\theta)$ where θ is the angle of rotation.

The next Figure shows the initial independent distribution U and the second one, shows how the variance of the projection Y evolves with the rotation $R(\theta)$ with an angle from 0 to 360 degrees. We can observe the following things:

- The evolution of all properties seems sinusoidal. Being the covariance and correlation apparently very similar to a sine functions and the projected variances very similar to stationary waves.
- The sum of individual variances is always equal to 5, which is the sum of the original projections:

$$\sigma^2_{X1} + \sigma^2_{X2} = \sigma^2_{Y1} + \sigma^2_{Y2}$$

- At 45 degrees the variances of the projections are the same, and the correlation is maximixed.
- The variance of Y_1 and Y_2 both oscillate between the initial variances of X_1 and X_1.
- We have 2 cycles of the signals, these properties are symmetric respect 180 degrees! Also the variance of a signal is equal to the other delayed 90 degrees.

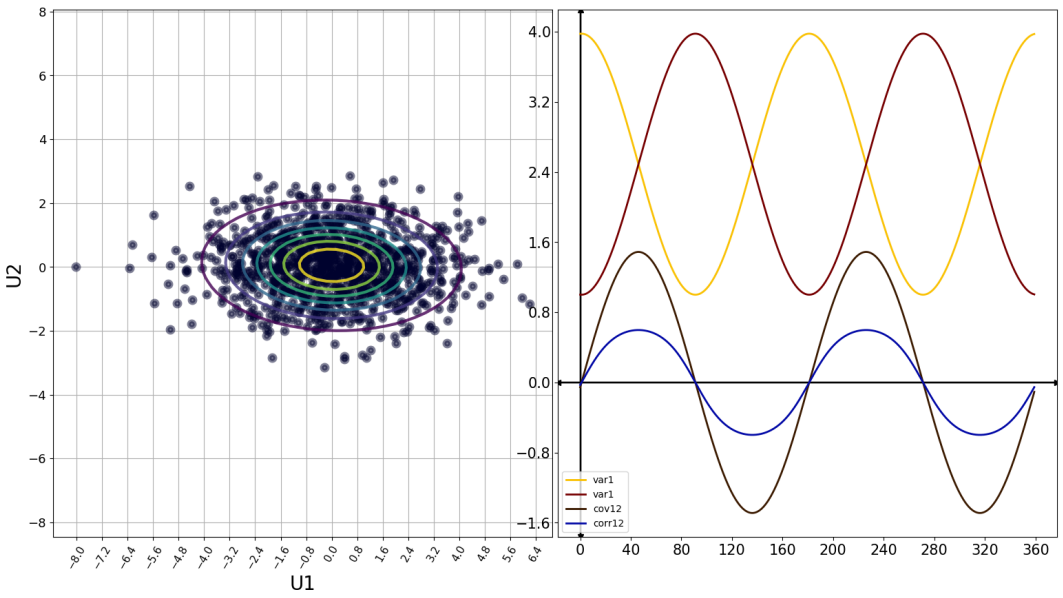


Figure 22: SMA and its window

Let us explain what is happening with the mathematics of the transformation. As we saw before, if we multiply the initial space U by a rotation matrix R, the new covariance matrix will be $\Sigma_Y = R\Sigma_X R^T$ In the 2D case we have the rotation matrix:

$$R(\theta) = \begin{bmatrix} r_{11}(\theta) & r_{12}(\theta) \\ r_{21}(\theta) & r_{22}(\theta) \end{bmatrix} = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$$

So, being the original distribution of X independent variables, its covariance matrix is a diagonal matrix with the variances in the diagonal, the resulting covariance matrix is:

$$\Sigma_Y = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \begin{bmatrix} s_1^2 & 0 \\ 0 & s_2^2 \end{bmatrix} \begin{bmatrix} cos(\theta) & sin(\theta) \\ -sin(\theta) & cos(\theta) \end{bmatrix}$$

$$\Sigma_Y = \begin{bmatrix} s_1^2 cos^2(\theta) + s_2^2 sin^2(\theta) & (s_1^2 - s_2^2) sin(\theta) cos(\theta) \\ (s_1^2 - s_2^2) sin(\theta) cos(\theta) & s_2^2 cos^2(\theta) + s_1^2 sin^2(\theta) \end{bmatrix}$$

As we can see

- The marginal variances are the sum of 2 squared sinusoids, which are 90 degrees separated, explaining the shape. If we add them up together we get will always get $s_1^2 + s_2^2$ since $sin^2(\theta) + cos^2(\theta) = 1$.
- The maximum value of the covariance is limited by $(s_1^2 - s_2^2)$, this means that the more different the variances of the original marginal distributions are, the more covariance there can be between them. In the opposite extreme case, if they are all the same, then the shape is a circle and they can never be correlated. The maximum value of this covariance is given by $\sin(\theta)cos(\theta)$ which has maximum value at $\theta = 45deg$ being 0.5. The next image shows this function.

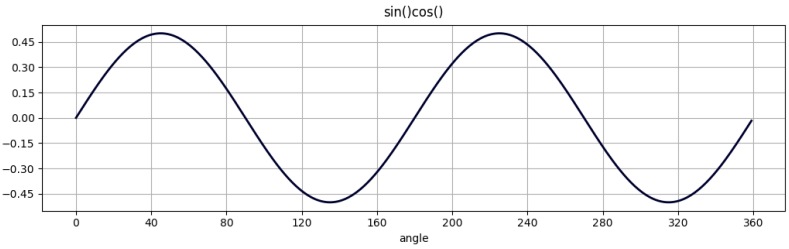


Figure 23: SMA and its window

So the maximum covariance between the 2 projected random variables Y_1 and Y_2 will be

$$\sigma_{12_{max}} = (s_1^2 + s_2^2)/2$$

In our example this is equal 1.5 which is exactly the maximum we can see in the previous graph.

- Respect to the correlation, using trigonometric equalities we can reduce the equation of the correlation to:

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} = \frac{\sigma_{12}}{\sqrt{s_1^2 s_2^2 + \sigma_{12}^2}}$$

So the correlation will me maximum when the covariance is maximum, which we can also see in the graph. depends on the maximum covariance and the product of the independent variances. Being the maximum covariance equal to $(s_1^2 + s_2^2)/2$, the the maximum covariance is:

$$\rho_{max} = \frac{(s_1^2 + s_2^2)/2}{\sqrt{s_1^2 s_2^2 + (s_1^2 + s_2^2)/2}}$$

In our case, the maximum correlation is $1.5/\sqrt{(4 + 1.5^2)} = 0.6$ which can be observed in the graph :).

Throught this, what has remained constant is the determinant of the matrix, that is the size of the space. Elaborate

2.5 Estimators

Having that we can express the covariance matrices as:

$$\Sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)^T] = E[XY^T - X\mu_Y^T - \mu_X Y^T + \mu_X \mu_Y^T]$$

3 Descriptive statistics

In this Section we will learn the basics of how to perform statistical inference from data, that is, we will learn how to obtain useful information from a set of samples, called our dataset $\mathcal{D}$. Most commonly this information will come in the form of a statistic t, a value that we obtain from our dataset by appliying a function to it $t = f_t(\mathcal{D})$, for example the mean or variance of our dataset. We will also make emphasis on how confident are we that the infered information is correct, since our dataset has a variance, it has an uncertainty, and so it does any function applied to it, in this case out statistic t. It is very important to know what is the uncertainty of our statistic and if it relevant at all, in other words, its value is not due to noise but actually significant.

In Science, if you cannot measure something, it does not exist. You cannot prove that something that does not exist, does not exist since you cannot gather any information about it, you cannot measure it, i.e. God. In a standard scenario we start with a dataset $\mathcal{D}$, composed by a set of N samples $x_i, i = 1, ..., N$ that we obtained from measuring a phenomenoun X. These samples in the general case can be D-dimensional.

$$\mathcal{D} = \{x_0, x_1, x_2 ... x_N\} \qquad x_i \in \mathbb{R}^D$$

.

In the most general case, we don't assume anything about the statistical distribution of the samples. Each sample x_i could come from a different distribution $X_i \sim$?? and could be related to the other samples in any allowed way by the laws of probability. These random variables $X_1, X_2, ..., X_N$ have an implicit joint N-dimensional probability distribution $f(X)$ which expresses the marginal distributions of all the samples and all the relations among them.

$$f(X_1, X_2, ..., X_N)$$

If we knew the statistical distribution of $f(X)$, then we could infer anything we want about the variables and we could obtain the best estimations for any event, which is what we want, Science is about control (estimation) and prediction. For example, we can compute the **likelihood of the dataset** $\mathcal{L}(D)$, that is, the probability (or likelihood) of having obtained the set of samples in $\mathcal{D}$.

$$\mathcal{L}(\mathcal{D}) = f(X_1 = x_1, X_2 = x_2, ..., X_N = x_N)$$

Trying to learning such a high-dimensional distribution from data samples is most commonly intractable, since we would need a lot of samples x_i in order to estimate it. The number of samples needed grows exponentially with the number of variables of the distribution we are trying to estimate. If we had an infinite number of samples, we could just compute the histogram of the possible events to compute the distribution but this is not feasible in practice. We need to constraint the possible distributions that $f(D)$ could model. What we usually do is used a parametrized model where we express f(D) as a combination of simpler distributions.

We usually make assumptions that limit the shape of the distributions of the marginal variables X_i and we also constraint the relationships between them. A common simplistic assumption is that:

- **The samples follow a Gaussian distribution**. Due to the nice properties of the Multivariate Gaussian distribution discussed in the previous Section, we most commonly assume that the samples follow a D-dimensional Gaussian distribution.

  $$X_i \sim N(\mu_i, \Sigma_{X_i})$$

  Notice that if we do not assume that the samples are identically distributed, then each X_i will have its own parameters μ_i and Σ_{X_i}.

- **The samples are independent**. That means there is no relation between the samples, knowing the value of a sample x_i does not give us any information about the possible value of another sample x_j.

- **The samples are identically distributed**. This means that all of the samples x_i come from the same distribution X, all the distributions X_i are the same:

  $$x_1, x_2, ..., x_N \sim D$$

In a common scenario we will make one of more of these kind of assumptions. A usual case is to assume that the samples are **independent and identically distributed, (i.i.d)**. This makes it much easier to

learn from data since we can use all the samples to estimate the shape of the distribution in a simple way. Furthermore if we assume Gaussian distribution we have a lot of properties and mathematical foundation to estimate the shape of the distribution by estimating its mean and variance.

Throughout the Section we will make assumptions on the statistical distribution that governs our dataset in order to compute different statistics of its distribution and gain an insight that allows us to estimate and predict events. Of course, **the correctness of our estimates depends on the correctness of the assumptions** that we make. These assumptions must be checked somehow as well. We will describe mechanisms to infer propeties of the distribution under some assumptions and how to check if the assumptions are correct.

DEVELOP: In the general notation we will write X as the general form of the random variable, but it could be $X(t)$ for example if the distribution depends on time, or $X(h)$ if the distribution depends on a set of parameters h.

Regarding the matrix description of our data, we will describe each sample x_i as a D-dimensional column vector, and our dataset $\mathcal{D}$ as the concatenation of these vectors in a DxN matrix, which we will call X (abusing some notation).

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix} \qquad \mathcal{D} = X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1D} \end{bmatrix} & \begin{bmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2D} \end{bmatrix} & \cdots & \begin{bmatrix} x_{N1} \\ x_{N2} \\ \vdots \\ x_{ND} \end{bmatrix} \end{bmatrix}$$

In the same way, to denote the random variables, we will call X_i the distribution

$$X_i = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$$

IGNORE FOR NOW: There are other non-parametric technicques where no assuptions are made about the shape of the distribution of X or the relation among its samples but we will leave that for another section.

3.0.1 Basic Example

Let us start with a basic example where our dataset $\mathcal{D}$ contains the 15 min returns of the CLOSE price of the company AAPL during the timespan of a week. Each 15M return occurring during this week will be a unidimensional sample x_i of our dataset. The global joint distribution X that governs the samples is the 15M return distribution of AAPLE during that week. The dataset used contains the $N = 130$ samples. In the Figure below we can observe the return of all the samples, where we have removed the temporal information, we are indirectly assuming the graph that the samples are independent (well not completely, but temporal relationship is out). The samples could still each have a different distribution but we do not know in the chart, which one is which.

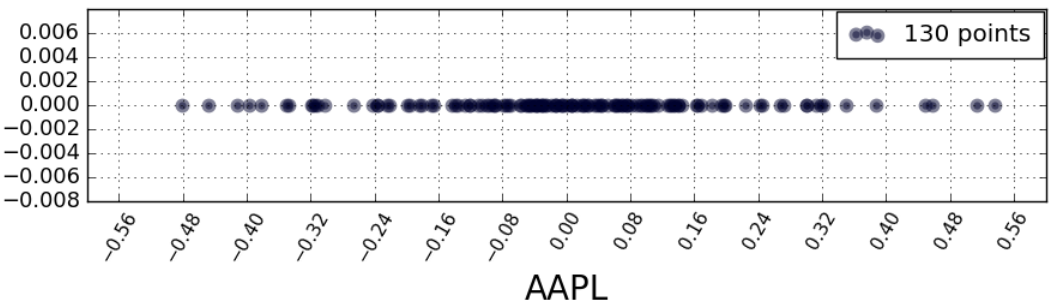


Figure 24: SMA and its window

This **dataset is actually a sequence**, that is, a set of samples in which the order matter, the samples could be related to each other by their time index t_i. ít could be intuitive to assume that the statistical properties of the samples happening around the same time could be more similar than those of samples

happening in very different times. When assuming that the samples are independent from each other, we are destroying this possible temporal relationship that could exist. In this new scenario, any permutation (switching of order) of the samples is equivalent.

This temporal information is visually available in Figure 2, where we plot the temporal sequence of the returns. But since the **temporal relationships can be quite complex and noisy**, it is hard for us as humans to guess a specific relationships from the timeseries datapoints, and so it is for an estimation algorithm as well. Trying to capture robustly the temporal relationships between the samples that could be quite complex and we would need a lot of data.

It easier to visualize all the datapoints and focus on simpler visual information, the scattering of the samples, where we might see some easy patter like for example, all the datapoints clustered around 0. From now on in the example, we will assume that **the datapoints are independent and identically distributed** so that we can use all of them to estimate the distribution.

We could then use a non-parametric way to find a weird shape of the distribution, but to be more robust and find meaningful parameters, we are going to **assume that the samples follow a Gaussian distribution**. This way we know specific estimators to obtain the distribution parameters μ and Σ and we can use all the datapoints to get a trustable estimate.

So, once we have the data and we have a better idea of what we want to obtain from it, we can start the inference process. Ideally we would learn the real statistical distribution of the samples, $f(X)$ but that is actually impossible, **we would need an infinite number samples to be absolutely sure that we learn the true distribution**, so noone will ever know the real distribution of any phenomena by just measuring their values, only GOD knows it.

Once we realized of this, we can be brave and say: "Alright, but I still want to obtain some information from the data, for example the distribution parameters mean and variance, and I would like to know how sure I am of the estimated values". At the end of the day, the 130 samples that we gathered, $\mathcal{D}$, will be a little different to other 130 samples, $\mathcal{D}_{\in}$, that we could obtain from another week.

Since X is a random variable, every time you sample it, you will get a different result (because by definition you do not know what exactly you are goint to get). Therefore these datasets are different $\mathcal{D} \neq \mathcal{D}_2$, there is variance, an uncertainty associated to be datasets $\mathcal{D}_{\rangle}$ obtained. Therefore the computation of any parameter t from the dataset $t = f_t(\mathcal{D}_{\in})$ from the data will also have a variance, an uncertainty. (Unless of course the computed value is independent of the data, or very special conditions happen, which is veeeery unlikely or useless)

To illustrate this point, lets divide the week into its 5 days, each one of them containing $N_{day} = 26 samples$. Even assuming that the samples of each day follow the same statistical distribution, each day will be somewhat different due to the uncertainty in the distribution. Otherwise, it would be so... predictible, and we would be rich. The next Figure illustrates the data points for the those 5 days and also the combination of them.

- We can obviously appreciate that each day is different so every parameter that we compute from them will also be different.
- Even though we can suspect that each of the days actually does not follow the same distribution X since it is likely that it varyes with time, we will assume for now that it does not change. Later we will see ways to determine if it actually varies or not.

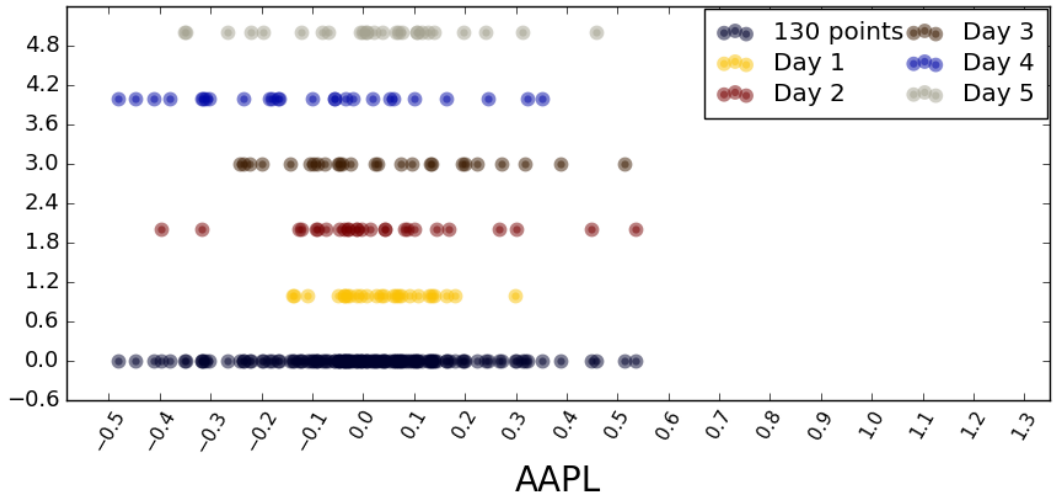


Figure 25: SMA and its window

So now we have 5 datasets, $\mathcal{D}_1$ to $\mathcal{D}_5$, taken from the random variable X_{day}, which represents the random distribution of the 15m return samples of a day for the AAPL company. Let us now compute the sample mean and variance from each of the 5 datasets. Since each dataset is different, each parameter we estimate from the data will be different. Lets assume that the data is Gaussian, so $X \sim \mathcal{N}(\mu, \sigma^2)$. The unbiased estimators for the mean and variance are:

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

Notice how in the estimation of the variance, we are using the previously estimated mean, if we had the actual mean μ instead of the estimator $\hat{\mu} = \bar{x}$ then the unbiased estimator would be:

$$\hat{\sigma}^2 = S^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

In the next Figure we plot the estimated mean and variance obtained from the 5 different days, along with the man and variance of the combined data of the entire week. Since each of the days have a different set of points, each of the estimations is a little different. In this case they could be quite different since the data probably does not follow the assumptions of independence and gaussianity. As we can see in the next Figure both the mean and variance are different for each day.

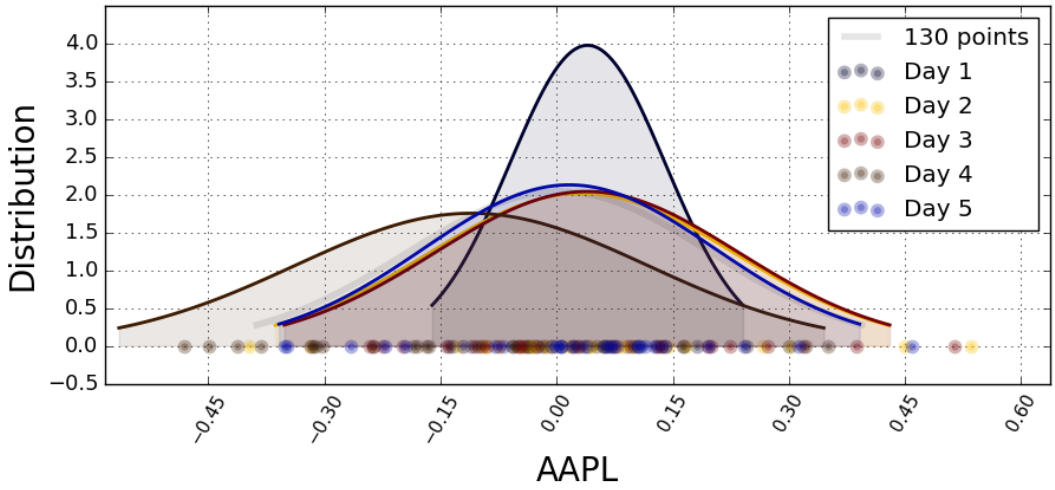


Figure 26: SMA and its window

We can see that the values $\hat{\mu}_1$ to $\hat{\mu}_5$ obtained are 5 different samples from the distribution $\hat{\mu}$ and the same goes for the 5 estimations of the variance $\hat{\sigma^2}$. At this point it is pretty clear that **the estimated parameters mean and variance,$(\hat{\mu}, \hat{\sigma})$ have some uncertainty, they follow some statistical distribution**, but... Can we can we cuantize it ? Can we find the distribution of estimators ? As always, the answer is YES, we can estimate their distribution up to some degree of certainty. The estimation of their distribution can be done mainly in 2 ways:

- Assuming a distribution in the original X. For example we could assume that X is guassian and therefore maybe obtain a closed form estimation of the distribution of the estimators.
- Not assume the distribution and play other triks XXXXX

If we can obtain the distribution of the estimators, in this case $(\hat{\mu}, \hat{\sigma})$, we can do cool things like estimating their confidence interval and testing their statistical significance. The estimator itself obtained from a single dataset D_i will be a sample of this distribution. In the following, we will show the distribution of the sample mean and variance assuming that the samples are Guassian i.i.d. In later sections we will see what we can do when we cannot make such assumptions from the original variable X.

3.1 Distribution of the basic Estimators

3.1.1 Distribution of the sample mean

As we have seen, for each of the different datasets $\mathcal{D}_1$ to $\mathcal{D}_5$, we have obtained a different sample mean $\hat{\mu}_1$ to $\hat{\mu}_5$. So it becomes obvious that these estimated values are realization of a random variable, in this section we will see the close form solution of this distribution assuming that X follows a normal distribution with parameters $X \sim \mathcal{N}(\mu, \sigma^2)$

Under these assumptions, it can be shown that the sample mean $\bar{x} = \hat{\mu}$, from a dataset of size N also follows a normal distribution with the same mean and variance σ^2/N, in math notation $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/N)$. Taking advantage of the linear properties of the Gaussian distribution, the normalized sample mean Z has the standard gaussian distribution with mean 0 and variance 1.

$$Z = \frac{(\bar{X} - \mu)}{\sigma^2/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

The problem here is that we will never know just from sample datapoints the true parameters of the distribution, μ and σ. If we did, this equation will hold, and would represent the distribution of the normalized estimator, but we do not know them. Then we cannot really estimate this distribution, what do we do ? Well, we usually do the following 2 things.

- **We assume that we know the true mean**, for example $\mu = 0$. So we only need to compute σ. This seems like cheating, and it is, but it very useful for testing the statistical significance of values of the parameters (in this case the true mean μ) as we will see later
- **Instead of using the true σ, we will use an estimation of it**, S and put it into the equation. This leads to the normalized equation:

$$T = \frac{(\bar{X} - \mu)}{S^2/\sqrt{N}} \sim t(N - 1)$$

  But since S is not the real σ, it is an estimation, a random variable as we have seen, then what distribution does T follow ? Well it can be proven that it follows a t-student distribution with $N-1$ degrees of freedom $T \sim t(N - 1)$.

How does this t-distribution look like ? It is reasonable that it will be more uncertain than the Guassian distribution, since S is adding randomness. Also, the more samples we compute the sample mean from, the closer is S to σ and the less uncertainty we have ahout S. So the less uncertainty the t-distribution should have. With this said, it is obvious that the distribution has as parameter the number of samples, called the degrees of freedom.

In the next graph, we show the t-distribution for different numbers of degrees of freedom. As we can see, the more degrees of freedom, the more it looks like a Guassian distribution, at 26 samples it is almost identical.

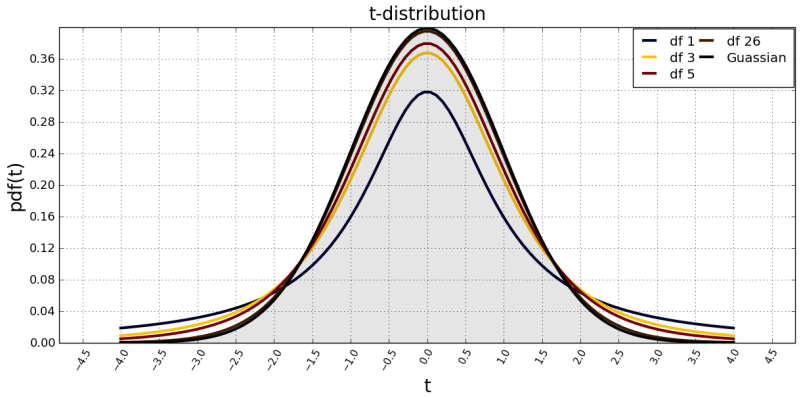


Figure 27: SMA and its window

All in all, this t-distribution is the distribution of the normalized sample mean $\hat{\mu} = \bar{X}$ when we assume we know the true mean μ and we use the estimated variance S^2 in the normalization instead of the true variance σ^2. In practice what you will do is assuming a value for μ and see if your data is likely to be generated from a Gaussian distribution with that mean, this is part of hypothesis testing and will be seen in future sections.

To finish, we would like to point out that our statistic $T = f_t(\mathcal{D})$ is a function of our dataset that we have normalized so that it follows the t-distribution and we can use the standard table so solve, the same way we usually normalize a Gaussian Distribution into $\mathcal{N}(0, 1)$ to be able to compare any Gaussian Distribution with the same table. We could have made another different transformation of the sample mean $\hat{X}$, $T^{'} = f_t^{'}(\mathcal{D})$ that would lead to another distribution that we could also use for analysis. This transformation is the most commonly used in practice due to its nice properties and simplicity.

3.1.2 The CLT for the sample mean

When we cannot assume that the distribution of X is Gaussian, it can be quite hard to have the actual distribution of the sample mean. But we can use the CLT to solve this.

3.1.3 Distribution of the sample variance

We have previously seen how we could find the closed form solution of the distribution of the sample mean $\hat{\mu} = \bar{X}$ when assuming that the dataset comes from the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. And if the distribution is not Gaussian, we simple used the CLT. In this Section we will see the distribution of the sample variance $\hat{\sigma}^2 = S^2$.

We will make the same assumptions about X, but contrary to the estimation of the distribution of $\hat{\mu}$ where the estimator is not very sensitive to deviations from the normal distribution (due to CLT), the method for estimating the sample variance described next depends strongly on the correctness of the normal distribution assumption.

In the same way that we normalized the sample mean $\bar{X}$ into the random variable Z which belongs to an ideal $\mathcal{N}(0, 1)$ if we knew both the true mean and variance (μ, σ^2). **We will normalize the sample variance S^2 into the random variable χ^2.** Since the sample variance S^2 follows a distribution proportional to the χ^2 distribution, we just divide it to normalize it, obtaining the following equation.

$$\chi^2 = \frac{(N-1)}{\sigma^2} S^2 \sim \chi^2(N-1)$$

And in the same way that in the equation of the normalized sample mean Z, we had to assume we know the true mean μ. **In this equation we have the true σ which we also have to assume we know.** As we can see in the equation, the distribution of χ^2 is just the original S^2 multiplied by a constant; unlike what happened with the sample mean, where there is some added randomness due to replacing σ^2 with S^2, which transformed the distribution from Gaussian to t-distribution.

So the distribution of our estimator is the χ^2 distribution with $N-1$ degrees of freedom. This is a non-symmetric distribution, whose domain is the positive axis. We can find this distribution as well as the sum of squared normal random variables X_i^2, where all the variables are independent and follow a normalized Gaussian distribution $X_i \sim \mathcal{N}(0, 1)$, in this case the degrees of freedom is N, unlike our estimate that only has $N-1$.

$$Y = \sum_{i=1}^{N} X_i^2 \sim \chi^2(N)$$

The next image shows the χ^2 distribution for different degrees of freedom in. As we can observe, the more degrees of freedom that we have, the more maximum likelihood point is equal to the

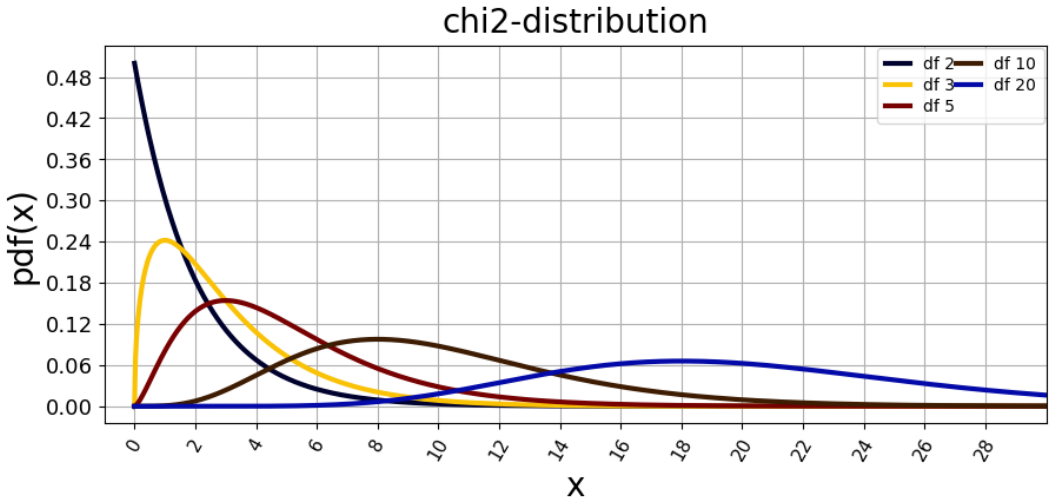


Figure 28: SMA and its window

The sample mean of this distribution will tend to 1 since the mean of X_i is 1. Due to the CLT then the mean will tend to a Gaussian shape as we can see in the previous chart. The next chart shows the distribution of the mean of this χ^2, that is the same distribution divided by the number of degrees of freedom:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} X_i^2 \sim \chi^2(N)$$

As we can observe this distribution, the more degrees of freedom that we have, the more the distribution looks Gaussian.

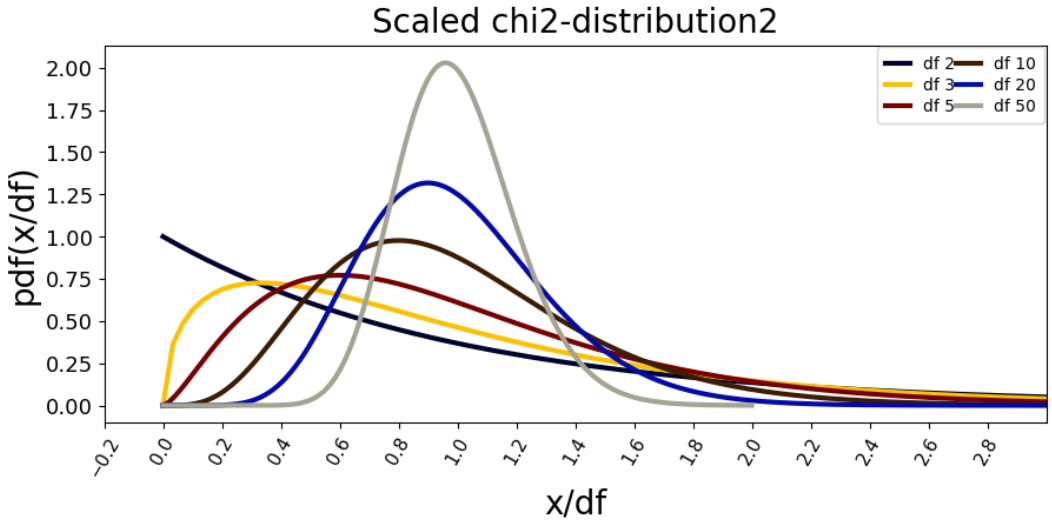


Figure 29: SMA and its window

This comes from CTL theorem

3.2 Statistical Significance

In the previous Section, we have seen that we can find the statistical distribution of the sample mean and sample variance estimators $\hat{\mu}$, $\hat{\sigma^2}$. These estimated will be exact if the data follows a gaussian distribution

and incomplete otherwise. In both estimations of the distributions we actually have a problem:

- In order to know the distribution of the estimator $\hat{\theta}$, we need to know the value of true parameter θ. Paradoxical isn't it? This happened for both μ and σ and holds true for any other estimator.

In practice, this can play to our advantage when we want to test if the true value of a parameter obtained from our dataset follows some specific property, like for example if the mean of our data is equal to a certain value $\mu = c$.

In statistics, **there is usually a hypotesis that we want to reject, this is so called Null Hypotesis** H_0. This Null hypotesis usually implies that nothing interesting is going on and we usually want to prove it wrong, we want to reject it. A common Null hypothesis would be for example that mean value of our data is 0. If the data is the return of a company, then it would mean that the company is not interesting to either buy or sell. Mathematically speaking we express this as:

$$H_0 : \mu = 0$$

The Null Hypothesis could be violated in several ways, the way we are interested on is called the **Alternative Hypothesis** H_A. For example in the previous H_0 we could have 3 alternative hypothesis:

- $H_A : \mu < 0$: The left-tail event
- $H_A : \mu > 0$: The right-tail event
- $H_A : \mu \neq 0$: The double-tailed event.

Ideally we want to prove that the probability (or likelihood) of H_0 being true, when we have observed our dataset $\mathcal{D}$ is very low, in order to reject it. This would provide strong evidence to dischard H_0. Mathematically speaking this would be:

$$P(H = H_0|\mathcal{D}) = \frac{P(H = H_0)P(\mathcal{D}|H = H_0)}{P(\mathcal{D})}$$

But in practice this way of trying to reject H_0 has some **limitations** such as:

- If the hypothesis are continuous, this value is a likelihood which is not very intuitive.
- We need a prior for the hypothesis, $P(H = H_0)$. Which might need to be subjective or computed from insufficient noisy data.
- We need the probability of the data, regardless of the hypothesis $P(\mathcal{D})$. Which might also be intractable, noisy and non-informative for our problem.

In order to reject H_0 what we usually do is to compute the probability (or likelihood) that the observed data was generated under the null hypothesis $P(\mathcal{D}|H = H_0)$. But this data alone is also not very informative since the importance of the probability of our dataset depends greatly on many factors like the assumed distribution or the dimensionality of the data. It also does not account for the Alternative Hypothesis H_A. We need a more descriptive value.

What we do is, we compute an statistic t from the distribution $t = f_t(\mathcal{D})$ and we compute the probability of obtaining a less likely dataset, a more extreme dataset that the one we have $\mathcal{D}$ in the case that Null Hypothesis is true.

$$P(t > t_0|H_0)$$

So in order to compute this probability asscіated to H_0 we need to compute:

- The distribution of the statistic $f(t)$
- The value of the statistic given our dataset $t_{\mathcal{D}} = f_t(\mathcal{D})$.
- The way we define a "more extreme value" than t. Thay is the way we are breaking H_0, in short, the Alternative Hypothesis H_A

The p-value is the probability of obtaining a dataset less likely that the one have, given that the null hypothesis is true. Depending on how we define a "more extreme than the one observed", given by H_A we can compute 3 different p-values:

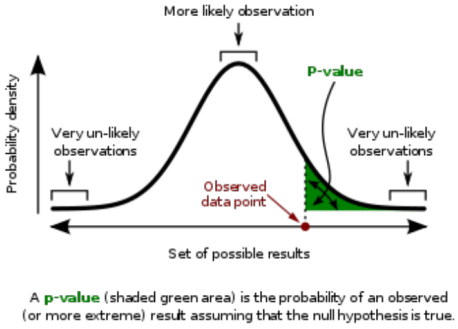
- $P(t > t_{\mathcal{D}}|H_0)$: The left-tail event

- $P(t < t_{\mathcal{D}}|H_0)$: The right-tail event
- $2 \cdot min\{P(t > t_{\mathcal{D}}|H_0), P(t < t_{\mathcal{D}}|H_0)\}$: The double-tailed event.

Being the p-value the probability of obtaining a statistic t more extreme that the one computed from our dataset $t_{\mathcal{D}}$ given that H_0 is true. If we resampled our dataset a big number of times then "p-value" proportion of the times we would get a t statistic more extreme than the one we got. We can use the p-value to reject H_0, the p-value reflects the strength of evidence against the null hypothesis.

- If the p-value is too small, then we have week evidence about the Null Hypothesis and therefore H_0 is rejected.
- If the p-value big enough, we do not have strong evidence against H_0 and we cannot reject it. This does not mean that H_0 is true and H_A is false, it just means that we cannot reject H_0 with the data that we have.

There must be a subjective threshold for which we can stablish that the null hypothesis can be rejected or accepted, in practice there is a set of thresholds that tell us how much evidence we have againt the H_0. The next figure summarizes graphically the obtaining of the p-value of a right-tail event and provides a table with common valued of the this explanation.



$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

https://02402.compute.dtu.dk/enotes/chapter3-StatisticsNormalAssumption

Section 3.1.7 onwards

Figure 30: SMA and its window

Of course, no matter how low the p-value is, there is always the probability that the dataset we obtained is an outlier, not statistically significant, a conspiration of the universe, and therefore we wrongly accept or reject the H_0. DEVELOP MATHEMATICALLY with false alarm and so on.

In the following we will perform statistical significance tests to the mean and variance estimates of the daily prices return obtained in the previous Sections.

3.2.1 Statistical Significance for the sample mean

In the case of the sample mean, we know that the statistical distribution of the normalized variable $T = f_T(\mathcal{D})$ follows the t-distribution with $N-1$ degrees of freedom, given that we know the true value of the mean μ. We can write this mathematically as:

$$f_T(\mathcal{D}|\mu = \mu_0) = T = \frac{(\bar{X} - \mu_0)}{S^2/\sqrt{N}} \sim t(N-1)$$

This fits perfectly within the p-value foundation since we can set the Null Hypothesis H_0 as:

$$H_0 : \mu = \mu_0$$

And therefore we already have the distribution of the statistic $f(t)$ and a way to obtain the statistic from the dataset given by T. We just need to stablish what our alternative hypothesis H_A is to compute the p-value. Which could be:

- $P(t > t_{\mathcal{D}}|\mu = \mu_0) = 1 - cdf(t_{\mathcal{D}})$: The left-tail event
- $P(t < t_{\mathcal{D}}|\mu = \mu_0) = cdf(t_{\mathcal{D}})$: The right-tail event
- $2 \cdot min\{P(t > t_{\mathcal{D}}|\mu = \mu_0), P(t < t_{\mathcal{D}}|\mu = \mu_0)\}$: The double-tailed event.

Since the t-student distribution is symmetric, the p-value of the double-tailed event event can be written as:

$$p - value = 2 \cdot P(t > |t_{\mathcal{D}}||\mu = \mu_0) = 2(1 - cdf(|t_{\mathcal{D}}|))$$

In the following example we will test the Null hypothesis that the mean is 0, $H_0 : \mu = 0$, for all of the days that we have. Each of the 5 days will give a different statistic T_i. The Alternative Hypothesis is that the mean is not equal to 0, $H_A : \mu \neq 0$, meaning that we can use the company to either buy or sell shares and therefore we have the double-tailed case. The following Figure shows the T statistic and p-values associated to all the days.

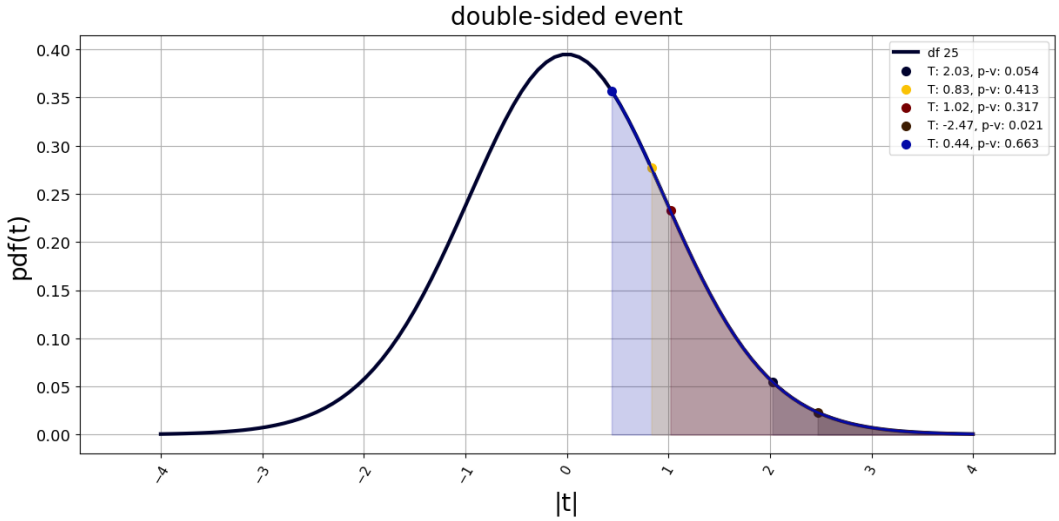


Figure 31: SMA and its window

We could also be interested in any of the other 2 possible H_A, for example, if we can only buy assets of APPL, then we would only be interested if the mean is bigger than 0, so we are interested in the right-sided event. In the opposite case, if we can only sell share, we are interested in the left-sided event. The next Figure shows the p-values for these alternative cases.

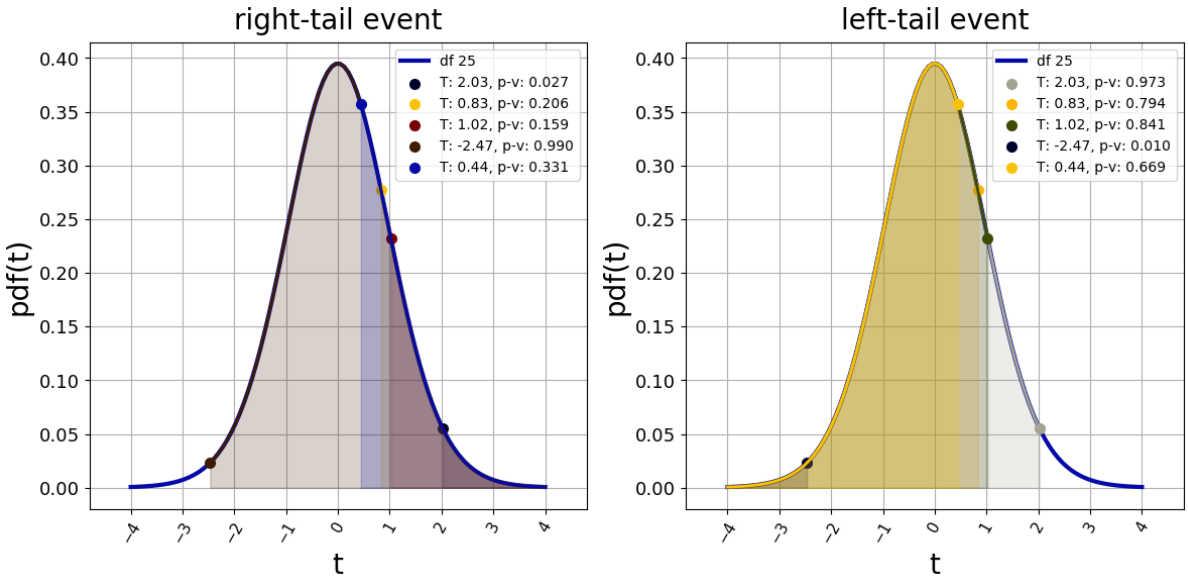


Figure 32: SMA and its window

Final comments regarding strength of evidence and complementary p-values.

3.2.2 Statistical Significance for the sample variance

In the case of the sample variance, we know that the statistical distribution of the normalized variable $\chi^2 = f_{\chi^2}(\mathcal{D})$ follows the χ^2-distribution with $N-1$ degrees of freedom, given that we know the true value of the variance σ^2. We can write this mathematically as:

$$f_{\chi^2}(\mathcal{D}|\sigma^2 = \sigma_0^2) = \chi^2 = \frac{(N-1)S^2}{\sigma_0^2} \sim \chi^2(N-1)$$

This fits perfectly within the p-value foundation since we can set the Null Hypothesis H_0 as:

$$H_0 : \sigma^2 = \sigma_0^2$$

And therefore we already have the distribution of the statistic $f(t)$ and a way to obtain the statistic from the dataset given by χ^2. We just need to stablish what our alternative hypothesis H_A is to compute the p-value. Which could be:

- $P(\chi^2 > \chi^2_{\mathcal{D}}|\sigma^2 = \sigma_0^2) = 1 - cdf(\chi^2_{\mathcal{D}})$: The right-tail event
- $P(\chi^2 < \chi^2_{\mathcal{D}}|\sigma^2 = \sigma_0^2) = cdf(\chi^2_{\mathcal{D}})$: The left-tail event
- $2 \cdot min\{P(\chi^2 > \chi^2_{\mathcal{D}}|\sigma^2 = \sigma_0^2), P(\chi^2 < \chi^2_{\mathcal{D}}|\sigma^2 = \sigma_0^2\}$: The double-tailed event.

In this case, the distribution is not symetric so we cannot apply the trick from before.

In the following example we will test the Null hypothesis that variance of the each day is equal to the variance of the week, that is, we are gonna take the estimate of the variance of the week as the true variance and we will see if the variance of the independent days vary significantly with respect to it.

$$H_0 = \sigma^2 = \sigma_w^2 = [xxxx]$$

Each of the 5 days will give a different statistic χ_i^2. The Alternative Hypothesis is that the daily variance is not equal to the weekly variance, $H_A : \sigma^2 \neq \sigma_w^2$, meaning that days do not behave in the same way as weeks. The following Figure shows the χ^2 statistic and p-values associated to all the days.

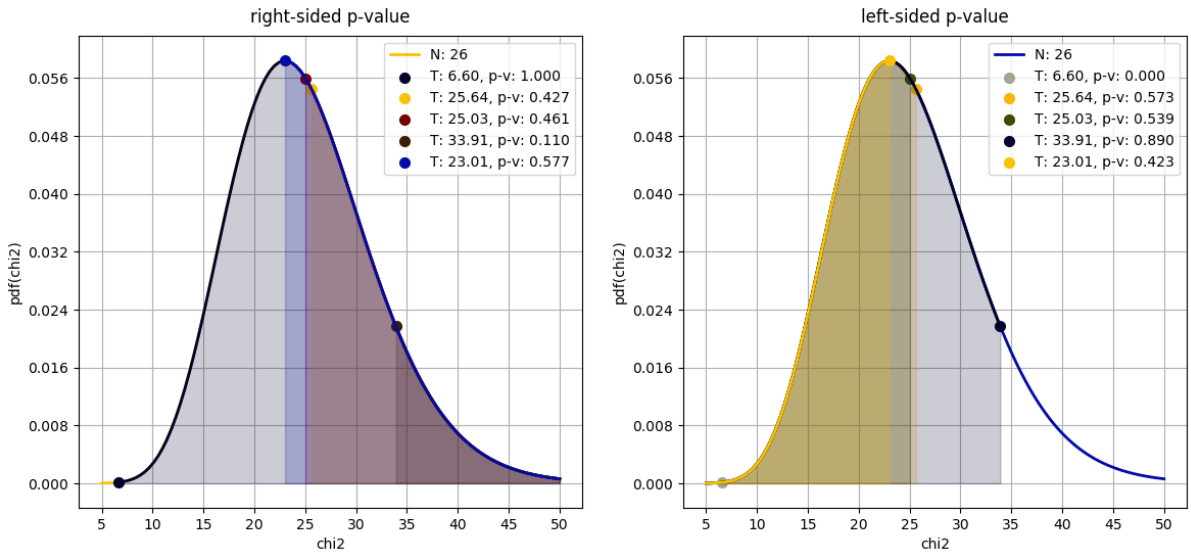


Figure 33: SMA and its window

In the both sided case, we choose the chi2 quared value min probability . Plot this shit

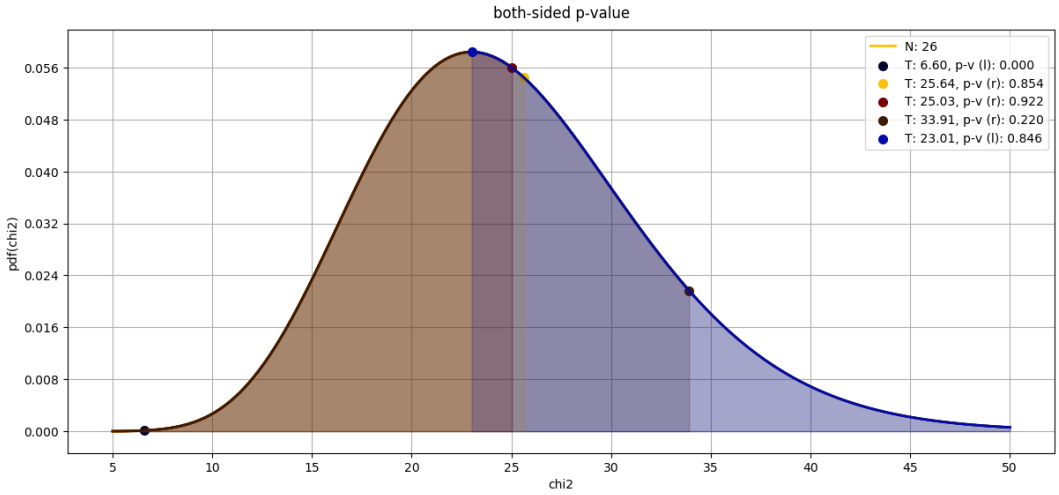


Figure 34: SMA and its window

Final comments regarding strength of evidence and complementary p-values.

3.3 Confidence Interval

So far we have seen that the datasets $\mathcal{D}$ have an uncertainty and how, under gaussianity assumptions we can estimate their true mean and variance, (μ, σ), using the sample mean and variance $(\bar{X}, S^2)$. Furthermore we saw that due to the uncertainty of $\mathcal{D}$, the estimators also follow a statistical distribution, $Z \sim t(N-1)$, $\chi^2 \sim \chi^2(N-1)$. But these distributions made the assumption that we know their corresponding true parameter. We can also ask ourselves, how close are these estimates to their true parameter ?

Using knowledge about probability distributions, we are able to quantify the uncertainty in our estimate even without knowing the true values. Statistical practice is to quantify precision (or, equivalently, uncertainty) with a confidence interval (CI). We compute the interval of this distribution in which the true parameter could be.

The CI of a value at a $(1-\alpha)$ level of confidence is the range of values of the parameter θ which contain the $(1-\alpha)$ probability of the distribution of the parameter. For example, if the 95% CI of a parameter is [-2.5, 4], it means that the true parameter will be inside that range 95% of the times.

$$CI = P\left[\theta_{(\alpha/2)} < \theta < \theta_{(1-\alpha/2)}\right] = (1-\alpha)$$

The boundaries are the points in the distribution that cover $(1-\alpha)\%$ of the probability around the mean:

- The left boundary $\theta_{(\alpha/2)}$ is the value of θ that covers $(\alpha/2)\%$ of probability to the left of the mean of the parameter.
$$P(\theta < \theta_{(\alpha/2)}) = \alpha/2$$
- The right boundary $\theta_{(1-\alpha/2)}$ is the value of θ that covers $(\alpha/2)\%$ of probability to the right of the mean of the parameter.
$$P(\theta > \theta_{(1-\alpha/2)}) = \alpha/2$$

In a common scenario we will normalize the distribution of θ and use precomputed tables to know the boundaries in the normalized domain. Later we will denormalize the boundaries to obtain the real CI of the parameter.

In order to be able to compute this CI, we need to know the distribution of the true parameter θ. This can be computed from the distribution our estimator $\hat{\theta}$, in fact the distribution of the real parameter is relative to the distribution of our estimator. The following examples that we will see are mean and variance of course.

3.3.1 Confidence Interval of the Sample mean

For the sample mean estimator $\bar{X}$, we know that its normalized version T follows a t-distribution. From the equation, we can switch around terms to obtain the distribution of μ. Solving for μ in the equation of T we obtain that:

$$\mu = -T \cdot \frac{S}{\sqrt{N}} + \bar{X}$$

Where:

- T follows a t-student distribution with $N-1$ degrees of freedom.
- S is the estimated value of the sample standard deviation. A deterministic function of the dataset.
- $\bar{X}$ is the sample mean, a deterministic function of the dataset.
- N is the number of samples

So we have that μ is a deterministic function of the random variable T. We can see that the distribution of μ is a mirrored and scaled t-distribution with mean in $\bar{X}$ and since the t-distribution is symmetric, the - sign does not affect its shape.

Now, we want to compute the range of values of μ that covers $(1-\alpha) = 95\%$ of its probability around its mean. That is, we want to find the boundaries of the range, $\mu_{\alpha/2}$ and $mu_{(1-\alpha/2)}$, so that the proability that the true mean μ is in that range is 95%.

$$P(\mu_{\alpha/2} < \mu < \mu_{(1-\alpha/2)}) = 95\%$$

We could compute this directly over distribution of the true μ but it is easier to do it over the distribution of T and then convert the boundaries to the domain of μ. So in the domain of T, we would find the boundary values $t_{\alpha/2}$ and $t_{(1-\alpha/2)}$ around the mean that contain 95% of the probability, in other words we have to find those values such that:

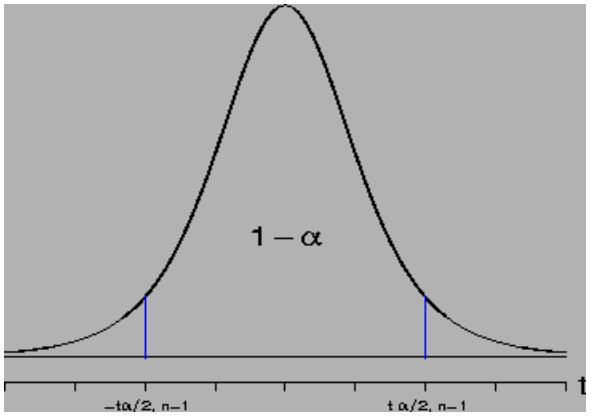
$$P(t_{\alpha/2} < t < t_{(1-\alpha/2)}) = 95\%$$

And then we use equation relating μ and T to transform the point values of the distribution in T to the point values in the distribution in μ, that is, we transform $t_{\alpha/2}$ to $\mu_{\alpha/2}$ and $t_{(1-\alpha/2)}$ to $\mu_{(1-\alpha/2)}$. It is easy as the following equations.

$$\mu_{\alpha/2} = -t_{\alpha/2} \cdot \frac{S}{\sqrt{N}} + \bar{X}$$

$$\mu_{(1-\alpha/2)} = -t_{(1-\alpha/2)} \cdot \frac{S}{\sqrt{N}} + \bar{X}$$

This way we can just have the $t_{\alpha/2}$ and $t_{(1-\alpha/2)}$ in the normalized domain which does not depend on the actual distribution parameters of X. Since the t-distribution is symetric, it is clear that in this case $t_{\alpha/2} = -t_{(1-\alpha/2)}$ being usually called $t_{1-\alpha/2}$. The value value of $t_{1-\alpha/2}$ will vary with the precision of the confidence interval given by α and the degrees of freedom. We will use normalized tables like the one below



(a) sizeInches = [2,2]

$1-\alpha$	.80	.90	.95	.99
t : n-1=5	1.48	2.02	2.57	4.03
t : n-1=15	1.34	1.75	2.13	2.95
t : n-1=25	1.32	1.71	2.06	2.79
t : n-1=35	1.31	1.69	2.03	2.72
t : n-1=50	1.30	1.68	2.01	2.68
t : n-1=100	1.29	1.66	1.98	2.63
t : n-1=500	1.28	1.65	1.96	2.58

(b) sizeInches = [10,6]

Figure 35: Effect of the figure size in the saved figures

As we can see in the table:

- The more degrees of freedom that we have, the smaller is the range for the same α, so we have less uncertainty. Also the values tend to the gaussian ones.
- The smaller α the bigger the interval since we have to contain more probability.

Let use finally use some of our data to compute the confidence intervals of the mean. Let us choose day1 and day2. The next image shows their computed distribution of μ and their range. Since we have $df = N - 1 = 25$, the value for 95% is $t_{0.95} = 2.06$. The shaded area contains the CI. Both distributions of μ follow the t-distribution. Day one has less estimated variance S and therefore, its CI is lower.

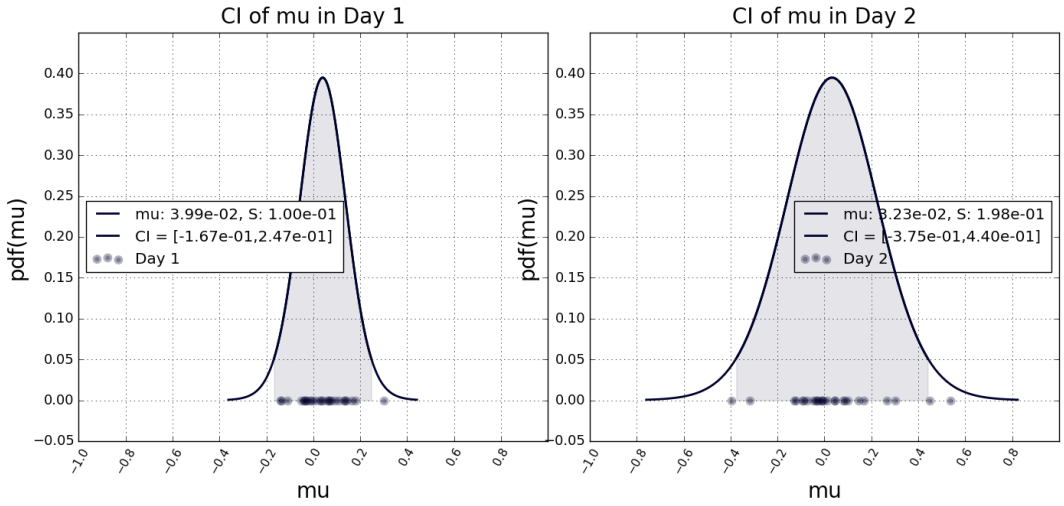


Figure 36: SMA and its window

We have also plotted the samples of the days. This is not completely correct since in this domain we should only plot the means obtained from sets of $N = 26$ samples. But it visually helps with the notion of the variance of the estimator. The central point of the distribution is the obtained sample mean $\bar{X}$.

TODO: I think i am not normalizing the PDF.

3.3.2 Confidence Interval of the Sample variance

In pretty the same way as before, we can obtain the distribution of the true parameter σ^2 from the distribution of its normalized estimator χ^2 using the deterministic function that relates them. Solving for σ^2 we have that:

$$\sigma^2 = \frac{(N-1)S^2}{\chi^2}$$

Where:

- χ^2 follows a χ^2 distribution with $N - 1$ degrees of freedom.
- S^2 is the estimated value of the variance. A deterministic function of the dataset.
- N is the number of samples

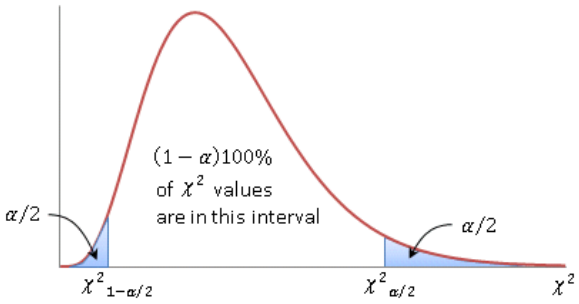
So the distribution of σ^2 is the inverse of χ^2 multiplied by a constant. The procedure to follow is pretty much the same. We want to know $\sigma^2_{(1-\alpha/2)}$ and $\sigma^2_{(\alpha/2)}$ such that they contain the quantiles around the mean of σ^2 that contain $(1 - \alpha)$ of the probability.

$$P(\sigma^2_{(1-\alpha/2)} < \sigma^2 < \sigma^2_{\alpha/2}) = 95\%$$

Likewise with the sample mean, we could compute this directly over distribution of the true σ but it is easier to do it over the distribution of χ^2 and then convert the boundaries to the domain of σ^2. So in the domain of χ^2, we would find the boundary values $\chi^2_{(1-\alpha/2)}$ and $\chi^2_{\alpha/2}$ around the mean that contain 95% of the probability, in other words we have to find those values such that:

$$P(\chi^2_{(1-\alpha/2)} < \chi^2 < \chi^2_{\alpha/2}) = 95\%$$

And then we use equation relating χ^2 and σ^2 to transform the boundaries as we did before. This way we can just have the $\chi^2_{(1-\alpha/2)}$ and $\chi^2_{\alpha/2}$ in the normalized domain which does not depend on the actual distribution parameters of X. As before, we can obtain the values from a table.



(a) sizeInches = [2,2]

Chi-Square Table

Table 5-2
Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

(b) sizeInches = [10,6]

Figure 37: Effect of the figure size in the saved figures

As we can see in the table, the more degrees of freedom that we have, the WIDER? is the range for the same α, so we have less uncertainty. Also the values tend to the gaussian ones. Also, the smaller α the bigger the interval since we have to contain more probability. We can see the XXX

Finally lets compute the confidence intervals of the variance for day1 and day2 from our dataset. The next image shows their computed distribution of σ^2 and their range. Since we have $df = N - 1 = 25$, the value for 95% is $\chi^2_{0.025} = XX$ and $\chi^2_{0.975} = XX$. The shaded area contains the CI. Both distributions of μ follow the t-distribution. Day one has less estimated variance S and therefore, its CI is lower.

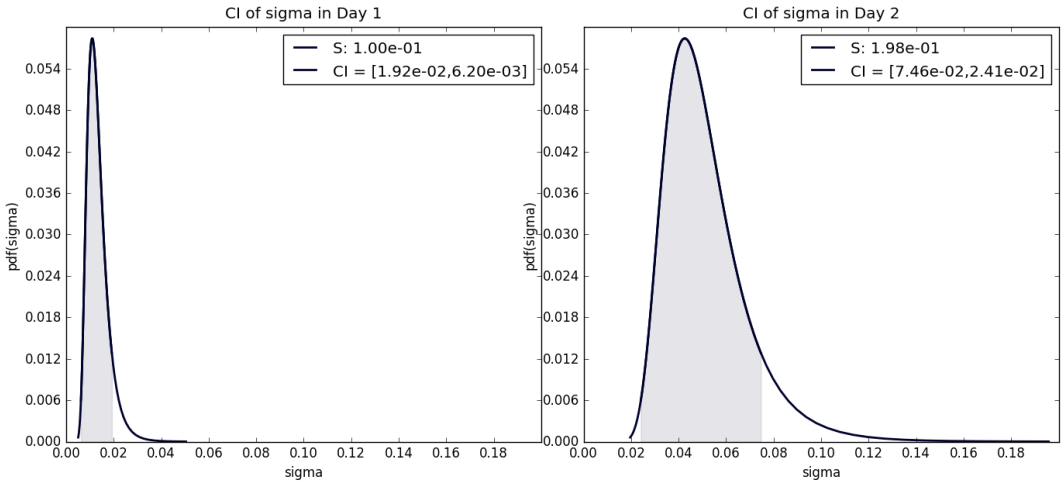


Figure 38: SMA and its window

As you can see, the lower sigma, the lower the uncertainty about it as well.
TODO: Equations for obtaining the number of samples needed for obtaining the desired level of certainty. Page 139 E-noter.
Talk about Type I and Type II error.

3.4 2 Sample descriptive statistics

So far we have seen statistical tests that arise from having a single population X and we were performing statistical tests on its properties when we assume Gaussianity. We have been focused on the Univariate case and then shown the Multivariate Gaussian parts.

Now we will see the 2 sample tests, in this case we have not only one population Y_1 but also another one Y_2. Both can be any distribution, but we will usually assume they are both Gaussian Multivariate distribution. In the genr

3.4.1 F-Distribution

The F-distribution comes from comparing the variances of 2 samples of distributions. It also comes from many other phenomena but this is probable the most straight forward and intuitive.
If we have two populations X and Y then the

$$caca$$

Test statistic for correlation

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

t distribution with N-2 degrees of freedom.

Test for partial correlation:
similar to the one we used for an ordinary correlation. This test statistic is shown below:

$$H_0 : \rho_{ij|X_1,X_2,...,X_k} = 0 \qquad H_A : \rho_{ij|X_1,X_2,...,X_k} \neq 0$$

The test statistic is:

$$t = r_{ij|X_1,X_2,...,X_k}\sqrt{\frac{n-2-k}{1-r^2_{ij|X_1,X_2,...,X_k}}} \dot{\sim} t_{n-2-k}$$

Statistical test for Multiple correlation:

$$F = \frac{r^2_{Y|X_1,X_2,...,X_D}}{1-r^2_{Y|X_1,X_2,...,X_D}} \frac{N-D-1}{D} \sim F(D, N-D-1)$$

In order to assess the significance of a given R. Under the usual assumptions of normality of the error and of independence of the error and the scores, this F ratio is distributed under the null hypothesis as a Fisher distribution.

Notice the relation with the future test that if the coefficients of our linear regression are 0. In the future this test cannot be done because we do not assume the underlying X variables to be Gaussian ? Remember

$$r^2_{Y|X_1,X_2,...,X_D} = \frac{SS_{res}}{SS_{Total}} = \frac{V[Y] - V[Y|X]}{V[Y]}$$

3.5 Tests for the Multivariate Gaussian

3.6 Tests for Gaussianity

Instead of using histograms, one can construct a quantile to quantial plot, usually called **q-q plot**. The idea of the q-q plot is a comparison of the empirical cumulative distribution function with the best possible normal distribution, usually the use estimated from the data.
In many scenarios we have the assumtion that the data samples $D = \{x_0, x_1, x_2...x_N\}$ come from the gaussian distribution X. Meaning that the samples:

- Are independent. Definitions of independence XXX.
- Follow the Gaussian distribution. X dim in

are gaussian and independent. We usually say that the sample are i.i.d. Independent and Identically Distributed. In this Seciton we will see forms to check such a thing.

The assumption about independent observations can be difficult to check, since the relation between the samples could come in many flavours, being one of the most common, the correlation. But there are many other forms of relation that the correlation cannot capture. For example if X and $X^2 + Z$
GRPAH The distribution could also depend on time or hidden variables.

The gaussian distribution assumption can be checked graphically, for example with a histogram or using statistical tests over parameters that characterize the Gaussian shape. For example the historgram is not as good since its shape depends on the size of the boxes.

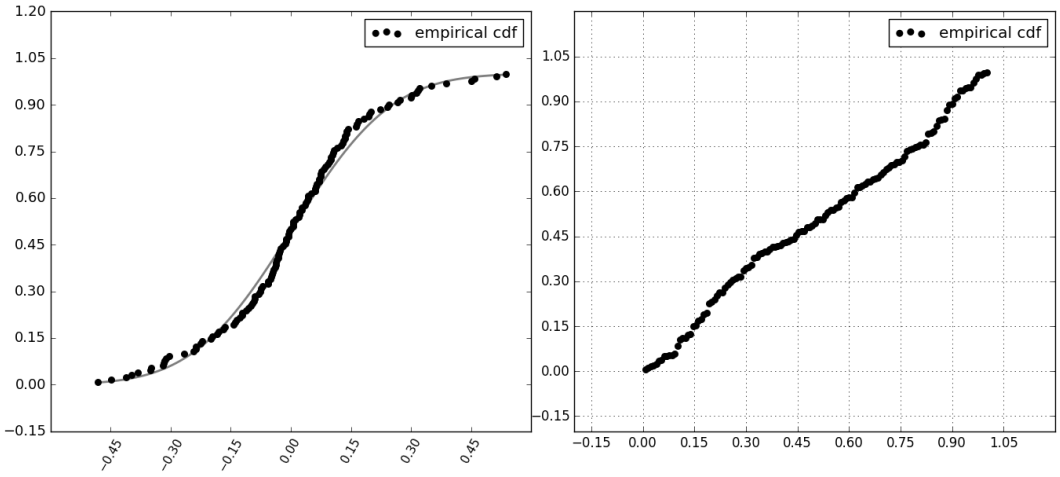


Figure 39: SMA and its window

As we can see, it looks like pretty gaussian, does that mean there is nothing to predict in the market ? Nooo ! We still have dependence of time and other latent variables ! Also the tails are larger. By CLT, the sum of a lot of random variables ends to be gaussian, but we may with other informaiton like time, in its simple form, separate the independent components that form the big gaussian mass.

3.7 Tests for independence

4 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised feature extraction method that, given a set of random variables $X = [X_1, X_2, ..., X_D]^T$ with 0 mean, finds the linear projections $P = [P_1, P_2, ..., P_D]$ that maximize the variance of the projected variables $Y = [Y_1, Y_2, ...Y_D]^T$. PCA obtains a new basis of the space, P, with component row vectors $P_i, i = 1, ..., D$, which projections have maximum variance. The new basis P has as rows the projection vectors P_i.

$$P = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_D \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1D} \end{bmatrix} \\ \begin{bmatrix} p_{21} & p_{22} & \cdots & p_{2D} \end{bmatrix} \\ \begin{bmatrix} \vdots & \vdots & & \vdots \end{bmatrix} \\ \begin{bmatrix} p_{D1} & p_{D2} & \cdots & p_{DD} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1D} \\ p_{21} & p_{22} & \cdots & p_{2D} \\ \vdots & \vdots & & \vdots \\ p_{D1} & p_{D2} & \cdots & p_{DD} \end{bmatrix}$$

The new variables Y are the linear transformation expressed with the new base are computed simply applying the linear projection. The projection for one sample X can be computed as:

$$Y = PX = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1D} \\ p_{21} & p_{22} & \cdots & p_{2D} \\ \vdots & \vdots & & \vdots \\ p_{D1} & p_{D2} & \cdots & p_{DD} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$$

Of course, since the set of projection P is a new basis of the initial dimensions, all of the components $p_i, i = 1, 2, ...D$ are orthogonal and have module 1.

$$p_i \perp p_j \forall i \neq j | p_i | = 1$$

As we previously saw in the Multivariate Gaussian, the variance of the linear projection Y is defined as:

$$VAR(Y_d) = P_d^T \Sigma_X P_d = \sum_{i=1}^{D} \sum_{j=1}^{D} \sigma_{ij} p_{di} p_{dj} = \sum_{i=1}^{D} \sigma_{ii} w_i^2 + 2 \sum_{i=1}^{D} \sum_{j<i} \sigma_{ij} w_i^T w_j$$

And the covariance between two projections Y_i and Y_j is:

$$COV(Y_i, Y_j) = P_i^T \Sigma_X P_j$$

The **first principal component** P_1 is the linear projection that maximizes the variance of Y_1 while having modulus equal to 1. So this projection accounts for the maximum variation in the data possible. This formal definition can be expressed using the following equation:

$$P_1 = \underset{P}{\operatorname{argmax}} \, VAR(Y) = \underset{P}{\operatorname{argmax}} \, VAR(PX)$$

subject to the constraint that $P^T P = 1$.

The **second principal component** P_2 is the linear combination of x-variables that accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0.

$$P_2 = \underset{P}{\operatorname{argmax}} \, VAR(Y_2) = \underset{P}{\operatorname{argmax}} \, VAR(P_2 X)$$

subject to the constraint that $P^T P = 1$ and $COV(Y_1, Y_2)$. The covariance constraint implies that P_1 and P_2 are perpendicular. A way to ensure this is to remove the first projection from the original data.

$$X_{remaining} = X - P_1 X$$

This way we fullfill the uncorrelation condition and we just need to compute the same optimization problem as before.

The **i-th principal component** P_i is the linear combination of x-variables that accounts for as much of the remaining variation as possible, that has not been explained yet by the previous $i-1$ components. The correlation with the previous components must be 0.

$$P_i = \underset{P}{\operatorname{argmax}} \, VAR(Y_i) = \underset{P}{\operatorname{argmax}} \, VAR(P_i X)$$

subject to the constraint that $P^T P = 1$ and $COV(Y_i, Y_j) = 0, j = 1, ..., i-1$. As for the second component we just iteratively substract the previous components.

So the question is how do we solve the independent maximization problems ? We could... or we could see that this is the same as the eigenvalu.

4.1 PCA by Eigendecomposition

The important thing here is that we are maximizing the variance of the orthogonal projections, this equations reduce to the same problem of finding the eigenvectors of the covariance matrix of the data. Being the mean of X = 0, if we decompose $Y = P^{'} X$ we arrive to the expression.

$$p_1 = \underset{p}{\text{argmax}} \, ||P'X||^2 s.t. ||p|| = 1 = \underset{p}{\text{argmax}} \, ||P'XXP||^2$$

Since P has modulus 1, this is equal to:

$$p_1 = \underset{p}{\text{argmax}} \, \frac{||P'XXP||^2}{P'P}$$

Which solution is equivalent to the first eigenvector of the covariance (dispersion) matrix. So this method is basically is equivalent to assuming that the D-dimensional data X follows a D-Dimensional Gaussian Distribution, and we want to find the axis of the components. We are basically just changing the base (rotation) to another one where the axis components are the ones that maximizes the variance.
The sucessive eigenvectors are computed by first substracting the new component. WRITE MATHEMATICALLY THAT THESE ARE THE EIGENVECTORS AND EXPLAIN THE EIGENVALUES RELATION TO VARIANCE.
TALK ABOUT THE SPECTRAL THINGY

To build up a more intuitive idea, **let us see a 2D example**, in this case we have as data the 15M returns of the companies "AAPL" and "Google" during a week timespan, therefore our data matrix X has dimensions $(D, Nsam) = (2, 130)$. As we saw before, the joint distribution of these 2 stocks is similar to a Multivariate Gaussian distribution. We are going to compute the 2 PCA components of this data, P_1, P_2, and compute the transformed variances Y_1, Y_2. The next Figure shows the returns of both assets, first in the original basis X and then in the projected space Y.

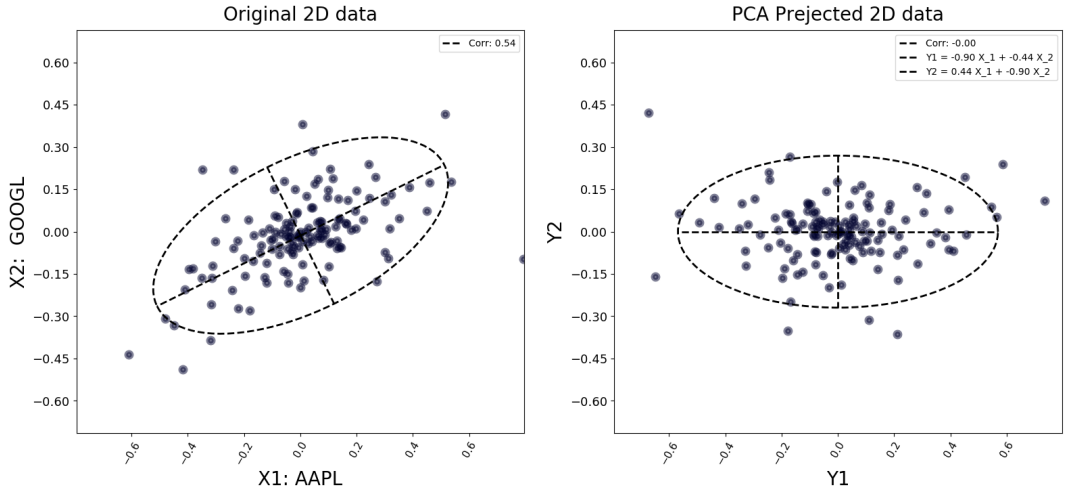


Figure 40: SMA and its window

As we can observe:

- In the originial domain, X_1, X_2, the direction of maximum variance is given by the first eigenvector V_1 and the second direction is given by V_2.
- In the projected data Y_1, Y_2, we can see how the data forms a Multivariate Gaussian with independent components. There components correspond to the original eigenvectors V_1, V_2 which have rotated the space to stablish themselves as the axis of the new domain.

- Notice that the values of the eigenvectors are plotted in the legend. We can check that they are perpendicular and have modulus equal to 1. Also notice that the projection has been mirrored respect to Y_1, this is because, since the distribution is symetric, the $V_i = -V_i$ in the decomposition and therefore sometimes the vectors computed are the inverted.

Notice once again that the projection Y_i over any of the vectors P_i geometrically means computing the perpendicular distance between any point x_0 and the hyperplane defined by p_i.

$$Y_i = \frac{|P_i \cdot x_0 + b|}{||P_i||}$$

In the case of 2D, this hyperplane is a line, in the case of 3D it is a plane and for bigger dimensions it is called a hyperplane. In this case the projection over p_i would be the distance of the points to the line drawn by p_2. We could visualize pretty much the same in 3D dimensions, where now the projections will be the distance to each point to the plane defined by the projection vectors p_i, there are the surfaces perpendicular to the vectors.

Lets see now a **bigger example** where have the returns of $D = 25$ companies as our dataset, in this case we have a data matrix X of dimensions $(D, Nsam) = (25, 130)$. We cannot see in the 25-dimensional space, but we can summarize it using the 2 projections Y_i, Y_j, the same way we could naively summarize it with 2 of the original variables. Since many markets are correlated, it reaches a point where some of the projections have very little variance and most likely are non-informative, they contain residual noise.

The next Figure shows the projection of several components, the first graph shows the projection between the first and second component, and the second graphs shows it for the 3rd and 9th. These type of plots are called **Component Scores plot** where we show the values of the principal components Y_i for each of the original observations $X_i = x_i$.

- As described in the theory, the gaussian distributions fitted to the projections P_i are always independent among them. Here we can see the independence between Y_1, Y_2 and Y_3, Y_9.
- The variance of the projections decreases with the number of the projection. Y_1 and Y_2 have similar variance, but we can see a big difference in variance between Y_3 and Y_9.
- There are some outliers in the data could cause a very distorted covariance matrix. As we can see the variables Y_1 and Y_2 seem quite correlated but we have 2 extreme points that the overall estimation to be independent. This points probably do not belong to the distribution and should be removed from the analysis. It is always good to visualize the data and test the assumptions of the distribution. There are other techniques to reduce the effect this outliers and also more robust estimation techniques agains outliers that we will see later in the document.

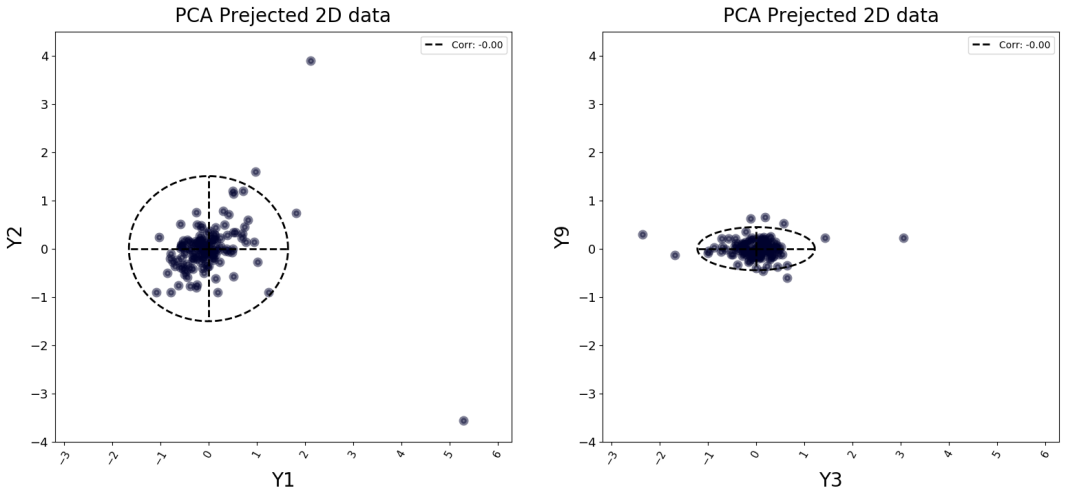


Figure 41: SMA and its window

4.1.1 PCA as dimensionality Reduction

As discussed in previous sections, the eigenvalues λ_i associated to the eigenvectors P_i indicate the variance σ_{ii} of its projection Y_i. The PCA projections for a Multivariate Gaussian Distribution with covariance matrix:

$$\Sigma_Y = P^T \Sigma_X P = P^T \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{1D} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_{DD} \end{bmatrix} P = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix}$$

The trace of the variance-covariance matrix, made of the sum of the variances of the individual variables is always constant given that P is a rotation matrix:

$$Total\_var = tr(\Sigma_X) = tr(\Sigma_Y) = \sum_{i=1}^{D} \sigma_{ii} = \sum_{i=1}^{D} \lambda_i$$

Therefore the eigenvalues are the explained variance of each of the Y_i. The proportion of variation explained by the ith principal component is the eigenvalue for that component divided by the sum of the eigenvalues. The ith principal component explains the following proportion of the total variation:

$$\lambda_i^{rel} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_D} = \frac{\lambda_i}{\sum_{i=1}^{D} \lambda_i}$$

The next Figure shows in the left chart the explained variance λ_i of all the components, and in the right chart the cumulative relative explained variance. As it was expected, the explained variance decreases with the i-th component, this decreasing seems to have a decreasing exponential shape. The cummulative chart shows that we actually only need the first 10 components to explain 90% of the total variance.

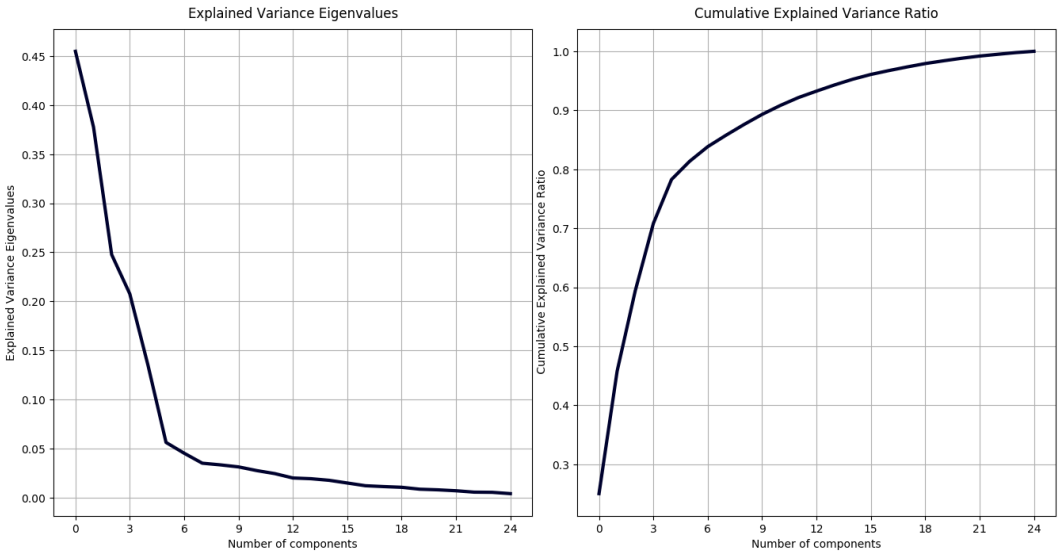


Figure 42: SMA and its window

It is common to use **PCA as a dimensionality reduction technique** where we will delete the projection Y_i with the lowest variance. Notice that this approach is not necessarily the best since we do not know if those low-variance projections actually have the discriminatory information of our particular problem, but it is usually a good practice.

The **number of components to select** will depend on the problem, as a rule of thumb we could pick the components that explain 90% of the variance. In this case we would obtain for example the first 10 components. We can also perform statistical tests where the hypothesis is that the first m eigenvalue are different and the last $m - D$ eigenvalues are the same, and therefore equal to a small value that could be considered insignificant. The mathematical hypotesis can be writen as:

$$H_0 = \lambda_1 \geq \lambda_1 \geq \cdots \lambda_m \geq \lambda_{m+1} = \lambda_{m+2} = \cdots = \lambda_D$$

Notice that the maximum value of m is equal to $D - 2$, since we at least need to select the last 2 eigenvalues to be equal, so m is equal to the total dimensions minus the number of eigenvalues that we want to test equality for. Without getting into much detail, given the dispersion matrix $\hat{\Sigma}_X$. We have that the statistic Z_1

$$Z_1 = -n^{'} log\left(\frac{|\hat{\Sigma_X}|}{\hat{\lambda}_1 \cdot \cdots \cdot \hat{\lambda}_m \cdot \hat{\lambda}^{D-m}} \right)$$

Where we have:

- that the determinant of $\hat{\Sigma_X}$ will be the same as the $\hat{\Sigma_Y}$, equal to the volume to the space, which is the product of the estimated lambdas and we have:

$$Z_1 = -n^{'} log\left(\frac{\hat{\lambda}_{m+1} \cdot \cdots \hat{\lambda}_{D} \cdot}{\hat{\lambda}^{D-m}}\right)$$

- The variable $n^{'}$ is equal to:

$$n^{'} = n - m - \frac{1}{6}\left(2(D-m) + 1 + \frac{2}{D-m}\right)$$

- The estimated lambda for the insiginificant labdas is the trace of X, which is equal to the trace of Y minus the significant lamndas normalized

$$\hat{\lambda} = \frac{tr(\hat{\Sigma_X}) - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m}{D-m} = \frac{\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_D}{D-m}$$

We have that Z_1 follows a χ^2 distribution with a number of degrees of fredom equal to $n_1 = \frac{1}{2}(D-m+2)(D-m-1)$. And we use the right statistic so we have a p-value:

$$p - value = P(\chi^2 > Z_1 | H_0) = 1 - cdf(Z_1)$$

If we have instead the estimation of the correlation matrix, the result is similar, now we use the estimator Z_2 that follows the same distribution as Z_1 and they are difference in the constant, where now, the n is just the number of samples we have.

$$Z_2 = -nlog\left(\frac{\hat{\lambda}_{m+1} \cdot \cdots \hat{\lambda}_{D} \cdot}{\hat{\lambda}^{D-m}}\right)$$

The number of degrees of freedom is the same. Is there a statistical way to know when to stop ? There is. Since the lambdas are the variance of the projection. We can perform the statistical test that XXXXx.

Once we have selected the m biggest principal components $Y_1, Y_2, ...Y_m$, we can **reconstruct our original D-dimensional distribution** to which we have removed the unwanted Y_m to Y_D components. This recovered X^{rec} distribution can be computes as:

$$X^{rec} = P^{mT} Y^m = \begin{bmatrix} X_1^{rec} \\ X_2^{rec} \\ \vdots \\ X_D^{rec} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & & \vdots \\ p_{D1} & p_{D2} & \cdots & p_{Dm} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} = Y_1 P_1 + \cdots + Y_D P_D$$

The reconstructed covariance matrix has XXXX Page 11 of week 4.
The next Figure shows the reconstruction of AAPL and Google when we are using a different number m of variables to reconstruct it. As we can see, the less componets we use:

- The less variance the initial componets have
- The more correlated the variables are

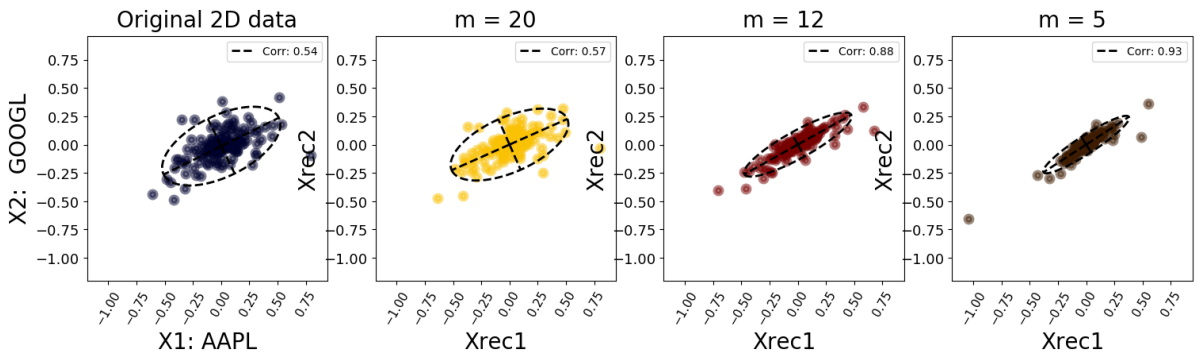


Figure 43: SMA and its window

4.1.2 Covariance and Correlations between X and Y

Since the projected variables Y are a linear combination of the original variables X, there will be a correlation between Y and X. In the general case, the correlation between 2 linear combinations of X given by the vectors V_i and V_j is:

$$\sigma_{ij} = V_1^T \Sigma_X V_2 = \begin{bmatrix} v_{i1} & v_{i2} & \cdots & v_{iD} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{1D} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_{DD} \end{bmatrix} \begin{bmatrix} v_{j1} \\ v_{j2} \\ \vdots \\ v_{jD} \end{bmatrix}$$

In our case where we express $\Sigma_X = P \Lambda P^T$. Also notice that we can easily express X in terms of Y as well. Since the projection matrix has the property $P^T = P^{-1}$ then we have that:

$$X = P^{-1}Y = P^T Y = \begin{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1D} \end{bmatrix} \begin{bmatrix} p_{21} \\ p_{22} \\ \vdots \\ p_{2D} \end{bmatrix} \cdots \begin{bmatrix} p_{D1} \\ p_{D2} \\ \vdots \\ p_{DD} \end{bmatrix} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_D \end{bmatrix}$$

Which means we can express an original variable X_j as a linear combination of the projected variables $Y = [Y_1, ..., Y_D]$. We therefore have:

$$X_j = \sum_{i=1}^{D} p_{ij} Y_i = p_{1j} Y_1 + \cdots + p_{Dj} Y_D$$

And since all the variables of Y are uncorrelated, we can express the variance X from the components of Y as:

$$VAR(X_j) = \sum_{i=1}^{D} p_{ij}^2 \lambda_i = p_{1j}^2 \lambda_1 + \cdots + p_{Dj}^2 \lambda_D$$

The covariance matrix between the selected variables Y^m and the original variables X is given by:

$$Cov(X, Y^m) = \Sigma_X P^m = P^T \Lambda P P^{mT}$$

We can separate this into 2 components: The transposed projection matrix P^T that has as columns the projections vectors P_i and the eigenvalues matrix Λ resulting in a matrix made of the projection vectors scaled by their corresponding eigenvalues:

$$P^T \Lambda = \begin{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1D} \end{bmatrix} \begin{bmatrix} p_{21} \\ p_{22} \\ \vdots \\ p_{2D} \end{bmatrix} \cdots \begin{bmatrix} p_{D1} \\ p_{D2} \\ \vdots \\ p_{DD} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} = \begin{bmatrix} \lambda_1 P_1 & \lambda_2 P_2 & \cdots & \lambda_D P_D \end{bmatrix}$$

And the product between the projection matrix P and the transpose of the projection matrix made with the selected principal components P^{mT}. Notice that if we selected all the components, then we would have $PP^T = I$. In this case, since some of the components are not selected, then the result is a Dxm matrix with the initial I mxm identity and the rest zeros since different eigenvectors are perpendicular.

$$PP^{mT} = \begin{bmatrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1D} \end{bmatrix} \\ \begin{bmatrix} p_{21} & p_{22} & \cdots & p_{2D} \end{bmatrix} \\ \begin{bmatrix} \vdots & \vdots & & \vdots \end{bmatrix} \\ \begin{bmatrix} p_{D1} & p_{D2} & \cdots & p_{DD} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1D} \end{bmatrix} \begin{bmatrix} p_{21} \\ p_{22} \\ \vdots \\ p_{2D} \end{bmatrix} \cdots \begin{bmatrix} p_{m1} \\ p_{m2} \\ \vdots \\ p_{mD} \end{bmatrix} \end{bmatrix}$$

So the covariance matrix between the selected principal components and the original variables is made of the selected eigenvectors P_i multiplied by its selected eigenvalues λ_i.

$$Cov(X, Y^m) = \begin{bmatrix} \lambda_1 P_1 & \lambda_2 P_2 & \cdots & \lambda_m P_m \end{bmatrix}$$

Which means that the covariance between the orginal variable $X_j, j = 1, ..., D$ and the i-th component $Y_i, i = 1, ..., m$ is the variance of Y_i, being $VAR(Y_i) = \sigma_i^2 = \lambda_i$ weighted by the linear constant between Y_i and X_j, that is $X_j = p_{ij} Y_i + \cdots$. So we basically have:

$$COV(X_j, Y_i) = \sigma_{ij} = COV\left(\sum_{i=1}^{D} p_{ij} Y_j, Y_j\right) = p_{ij} VAR[Y_j] = \lambda_i p_{ij}$$

In the same way, the correlation between the selected components and the the originial is:

$$Corr(X, Y^m) = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_D} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} P_1 & \sqrt{\lambda_2} P_2 & \cdots & \sqrt{\lambda_m} P_m \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \frac{\sqrt{\lambda_1}}{\sigma_1} p_{11} \\ \frac{\sqrt{\lambda_1}}{\sigma_2} p_{12} \\ \vdots \\ \frac{\sqrt{\lambda_1}}{\sigma_D} p_{1D} \end{bmatrix} & \begin{bmatrix} \frac{\sqrt{\lambda_2}}{\sigma_1} p_{21} \\ \frac{\sqrt{\lambda_2}}{\sigma_2} p_{22} \\ \vdots \\ \frac{\sqrt{\lambda_2}}{\sigma_D} p_{2D} \end{bmatrix} & \cdots & \begin{bmatrix} \frac{\sqrt{\lambda_m}}{\sigma_1} p_{m1} \\ \frac{\sqrt{\lambda_m}}{\sigma_2} p_{m2} \\ \vdots \\ \frac{\sqrt{\lambda_m}}{\sigma_D} p_{mD} \end{bmatrix} \end{bmatrix}$$

Which means that the covariance between the orginal variable $X_j, j = 1, ..., D$ and the i-th component $Y_i, i = 1, ..., m$ is the variance of Y_i,

$$\rho_{ij} = \frac{\sqrt{\lambda_i}}{\sigma_j} p_{ij} = \frac{\sigma_i}{\sigma_j} p_{ij}$$

We can represent this correlations in a **Component Pattern plot**. In this plot, we will compute the correlation between each original dimension $X_d, d = 1, ..., D$ and 2 chosen projections variables Y_{i_1}, Y_{i_2}. This will show us patterns of correlations between the original variables and the projections that might give us discriminatory information. In the next Figure we can see the Component Pattern plot between the first 10 companies $X_j, j = 1, ..., 10$ different pairs of principal components.

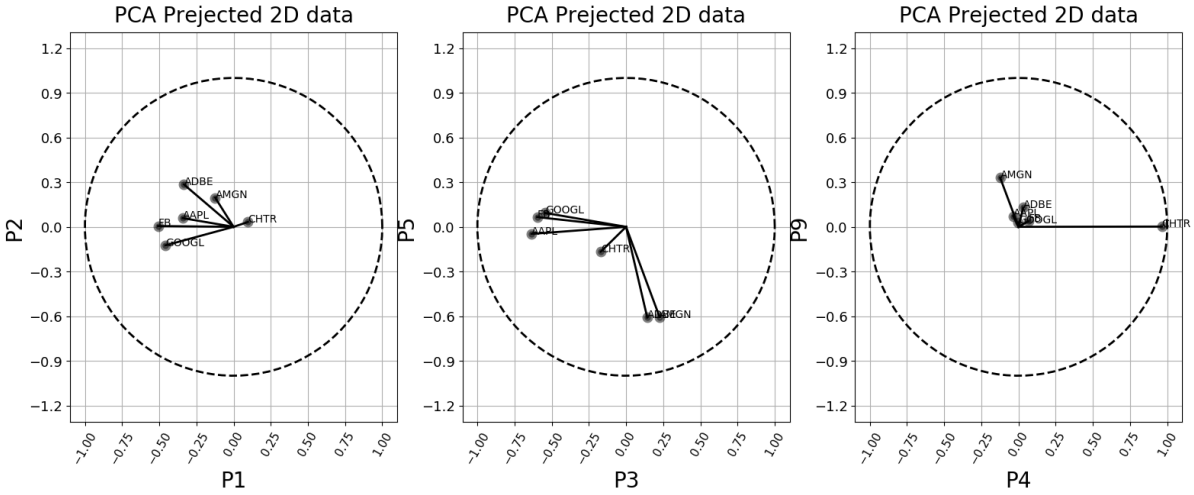


Figure 44: SMA and its window

It up to us to try to give some meaning to this projections, maybe they will give an insight on how the companies are clustered together. For example we can observe in the image that:

- In the first chart all the companies are positively correlated with Y_1 except CHTR. The correlations are not very strong.
- In the second chart we see a clear differentiation between FB,GOOGL and AAPL, which are technological companies and AMGN and ADBE.
- In the last chart we see how CHTR has the most of its component in Y_4.

At the end of the day, each of the original variables X_j will have some degree of correlation with the different projected independent variables Y_i, these correlations depend on the noisiness of both variables σ_i and σ_j and the linear relation between them p_{ij}.

Everything is interconnected, variances, correlations and linear coefficients, and they must obey some set of specific rules. For example the variances, the correlations are the explained variances so some shit. The projection vectors sum 1, and the variances relationship

In a **Component Pattern Profiles plot** maps the correlations between the original variables and the principal components. We can see for each original variable X_j, its correlation with the projected variables Y_i. We can try to spot some other patterns in this plot as well.

For example the next Figure shows this kind of plot for 7 companies and 10 principal components. It can be hard to see patterns since the variables in the X axis do not have to be related to a line chart might not be the most appropriate. We can see things that we saw before like that GOOGLE, APPLE and FB have similar principal component of that most of the correlation of CHTR comes from Y_4.

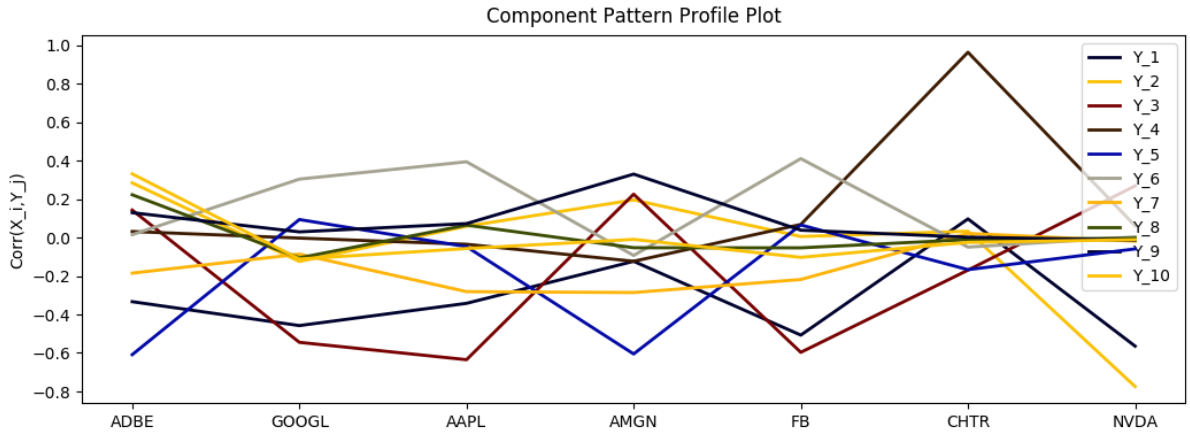


Figure 45: SMA and its window

Instead we could plot the information in another way, like a batchart, even though the charts might overlap, maybe separate small bars for all of them.

Chart

We could also plot other information like the squared correlation coefficients ρ_{ij}^2 which gives us an idea of the explained variance between the variables. Or we could just the weights p_{ij}, or weights squared since their constraints might show some pattern. The next Figure shows the squared correlation in which we can more clearly see the explained variance by the projections.

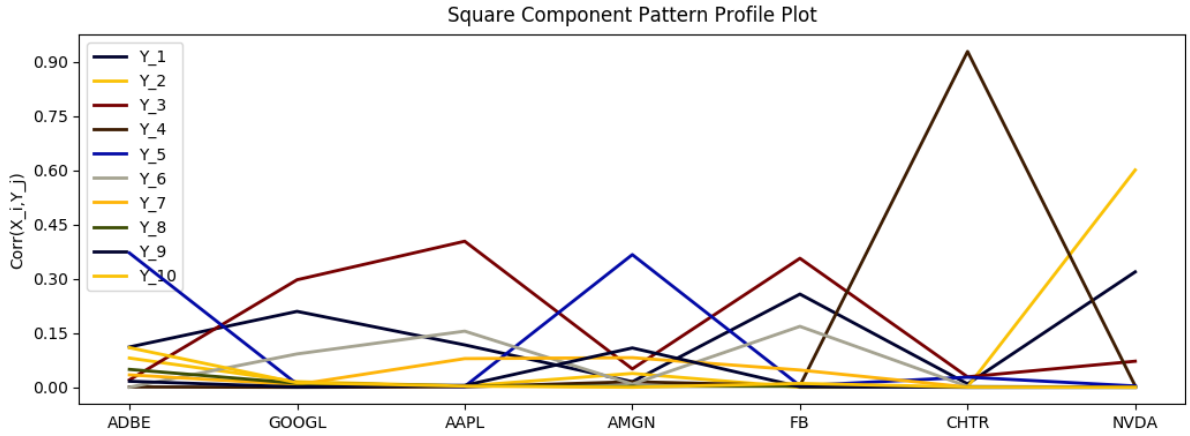


Figure 46: SMA and its window

4.1.3 Too many dimensions and too little samples

We can just invert it. A training matrix with N samples and D dimensions will form a matrix being , the matrix will have rank N. The N D-divisional vectors form a basis of the subspace where the images are.

4.1.4 The effect of normalizing variables

5 Factor Analysis

Factor Analysis is a method for modeling the observed variables $X = [X_1, X_2, ..., X_D]$ in terms of a linear combination of some unobservable latent variables called factors $F = [F_1, F_2, ..., F_D]$ plus some noise term $G = [G_1, G_2, ..., G_D]$ that models the error of the regression. The linear combination of the latent variables is given by the matrix A. The resulting equation is:

$$X = AF + G = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{1m} \\ \vdots & \vdots & & \vdots \\ a_{D1} & a_{D2} & \cdots & a_{Dm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} G_1 \\ G_2 \\ \vdots \\ G_D \end{bmatrix}$$

This **structure is equivalent to the previously seen in the PCA** where we selected a subset of m principal components $Y_1, ..., Y_m$ and we could reconstruct the original data as $X = X^{rec} + G = P^{mT}Y + G$ where G accounts for the variance that we lost due to not selecting the remaining principal components.

These factors F_i are typically viewed as broad concepts or ideas that may describe an observed phenomenon, the same way that we tried to give a meaning to the principal components Y_i of PCA. For example we could try to find a factor F_i that modeled the Country or the Sector of the companies. Factor analysis is generally an exploratory/descriptive method that might require subjective choices to craft the factors and their interpretation.

In PCA, the interpretation of the principal components is often not very clean, a given original variable X_j can contribute contribute significantly to more than one of the components Y_i. And there is nothing we can do since PCA is deterministic. Ideally we like each X_j to contribute significantly to only one Y_i. As we will see, in FA we can rotate the projections F using a technique called factor rotation in order to obtain variables that satisfy our subjective views.

EQUATION WITH DATA

In order to solve the FA structure, we need to find the matrix A, the latent variables F and the noise variables G, and of course they are all interrelated. There is a set of **constraints** in the parameters of the model:

- The factors have the identity matrix as covariance, that means that they are uncorrelated and have variance 1. This differs a little from PCA where the components were uncorrelated but had variance equal to their eigenvalue.

$$\Sigma_F = I_m = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- The noisy variables G have a diagonal covariance matrix, that is, they are uncorrelated but each dimension can have a different variance. This is different to PCA, where the covariance matrix of the removed components is symmetric covariance matrix.

$$\Sigma_G = \triangle = \begin{bmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \delta_D \end{bmatrix}$$

- The variables of F and G are uncorrelated.

- The observations X_j are standardised in such a way that they have variance 1, $V(X_i) = 1$. This can be achieved simply by normalizing the original variables. This implied that the variance-covariance matrix for X is equal to its correlation matrix which is denoted as:

$$R_X = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1D} \\ \rho_{21} & 1 & \cdots & \rho_{2D} \\ \vdots & \vdots & & \vdots \\ \rho_{D1} & \rho_{D2} & \cdots & 1 \end{bmatrix} \qquad \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- The matrix A is not constrained at it was the case of PCA where its columns must be independent projection vectors. So this model has more expressivity, even though we had to standarize at first the original variables X.

Now that we have seen the properties and constraints of the model, let us see its **consequences**:

- The variance-covariance matrix of X can be express in terms of A and variance of the noise $\triangle$.

$$\Sigma_X = R = AA^T + \triangle$$

Since R is the correlation matrix, its diagonal is equal to 1 and therefore:

$$Var(X_j) = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + \delta_j = h_j^2 + \delta_j = 1$$

Where h_j^2 is the j-th **communality**, the fraction of the variation in X_j that is explained by the common factors F. Correspondingly δ_j **gives the uniqueness in X_j's variance**, the proportion of X_js variance which is not due to the m common factors.

- The Covariance between an original variable X_j and a factor F_j is equal to the associated factor weight a_{ji}. This is trivial since the latent variables are uncorrelated and the noise is also uncorrelated

$$Cov(X_j, F_i) = Cov(\sum_{k=1}^{K} a_{ki}F_i + G_j, F_i) = a_{ji}$$

In the Matrix form, the covariance matrix between X and F is:

$$\Sigma_{XF} = Cov(X, F) = A$$

The value a_{ji} is called the factor loading, it is the covariance (and also correlation since the variance of both X_j and F_i is set so 1) between our variable X_j and our factor F_i. And therefore a_{ji}^2 is how much variance of X_j is explained by factor F_j.

- **The conditional covariance of X given the Factors F** is equal to the covariance of the residual. This is obsvious from the previously seen, if we know the Factors, then the remaining uncertainty is the one providaded by the error variables.

$$VAR(X|F) = R - AA^T = \triangle$$

Trivially for a given observed variable X_j we have as we saw before:

$$V(X_j) = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + \delta_j = h_j^2 + \delta_j = 1$$

Notice the relation between these equations and the general conditional variance equation where:

$$V(X|F) = \Sigma_{XX} - \Sigma_{FX}\Sigma_{FF}^{-1}\Sigma_{FX}$$

Notice that both expressions are equivalent since $\Sigma_{XF} = A$ and $\Sigma_{FF} = 1$. Since the factors are independent, for a single known factor F_i, the conditional variance is:

$$V(X_j|F_i) = 1 - a_{ji}^2$$

That is, that a_{ji}^2 is how much variance of X_j is known by knowing factor F_i, and since all factors are independent:

$$V(X_j|F) = V(X_1|F) + V(X_2|F) + \cdots + V(X_m|F) = 1 - a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 = \delta_j$$

- The expected value of the Factors given the observations is:

$$E[F|X] = \mu_F + A^T R^{-1}(X - \mu_X) = A^T R^{-1} X$$

Assuming that the mean of the X can be 0, otherwise we just temporarely remove it. Notice the relation with the general conditional expectation. Also notice that the other expectation is trivial:

$$E[X|F] = \mu_X + A(F - \mu_F)$$

- EXTRA: jth communality is always larger than or equal to the square of the multiple correlation coefficient between Xj and the rest of the variables

There are several algorithms to try to come up with the matrix A from which we can later trivially obtain the statistical properties of F and G.

5.1 Principal Factor solution

We can easily find a solution to the FA model by means of the PCA of the data. If we first normalize the individual random variables, then their covariance matrix is equal to their correlation matrix.

The solution is the projection vectors, eigenvectors, scaled by the eigenvectors so that the variance of the Factors is equal to 1.

$$A = P^{(m)T}\Lambda^{(m)\frac{1}{2}} = \begin{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1D} \end{bmatrix} \begin{bmatrix} p_{21} \\ p_{22} \\ \vdots \\ p_{2D} \end{bmatrix} \cdots \begin{bmatrix} p_{m1} \\ p_{m2} \\ \vdots \\ p_{mD} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_m} \end{bmatrix}$$

Where we have that the covariance matrix of the recovered data is:

$$\Sigma_{XX} = AA^T + \triangle = (P^{(m)T}\Lambda^{(m)\frac{1}{2}}(P^{(m)T}\Lambda^{(m)\frac{1}{2}})^T + \triangle = P^{(m)T}\Lambda P^{(m)} + \triangle$$

As we can see from PCA, AA^T is the recovered variance and $\triangle$ the variance we lost in the components we left. Notice in this case that the residual has a diagonal covariance matrix. This happened since we initially normalized the data.

The Principle factor solution for A is not unique, given a mxm rotation matrix Q which has a columns the rotation vectors:

$$Q = \begin{bmatrix} \begin{bmatrix} q_{11} \\ q_{12} \\ \vdots \\ q_{1m} \end{bmatrix} \begin{bmatrix} q_{21} \\ q_{22} \\ \vdots \\ q_{2m} \end{bmatrix} \cdots \begin{bmatrix} q_{m1} \\ q_{m2} \\ \vdots \\ q_{mm} \end{bmatrix} \end{bmatrix}$$

we have that the solution $A_2 = AQ$ is also a solution of the constrains of the factor model since:

$$A_2A_2^T = (AQ)(AQ)^T = AQQ^TA^T = AA^T$$

Since the factors have a covariance matrix equal to the identity matrix, they have equal variance and are uncorrelated, so the distribution is symmetric respect to any rotation. This is related to the rotation of the svd and shit.

The new factor loadings a'_{ji} can be computed from the previous ones a_{ji} and the rotation matrix Q.

$$a'_{ji} = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jm} \end{bmatrix} \begin{bmatrix} q_{1i} \\ q_{2i} \\ \vdots \\ q_{mi} \end{bmatrix}$$

What rotation matrix Q should we choose ? Well, the one that allows us to say that we found some factors that have a meaning with what we can publish a paper of course. This is the wonderful world of **factor rotation**.

A common technique is to use the rotation Q that maximizes One of the most often used criterions is the one introduced by Kaiser, the Varimax
criterion. It says that we must choose Q in such a way that the quantity. Page 287 from the Book.

Talk abour red, blue and dark, regarding the thinking fast and slow. Modelling everything mathematically. Hypothetical situation thinking how I would feel or react. Difference to reality.
HERE

6 Canonical Correlation Analysis

As we previously saw, the Multiple Correlation coefficient between the variable Y and the set of variables X, noted as $\rho_{Y|X_1,...,X_D}$ gives a measure of the linear relationship between Y and X, what it the maximum correlation there can be between Y and a linear combination of the variables in X. And therefore, how much variance about Y we can express with the optimal linear combination of X, since the squared correlation gives us this quantity.

In this case we will relate a set of variables $Y = [Y_1, ..., Y_p]$ with another set $X = [X_1, ..., X_q]$ where $p \leq q$ for numerical reasons (if otherwise we can always rename the variables, switch Y for X ?).

In this scenario we will have the joint Multivariate Gaussian Distribution $Z \sim N_{p+q}(\mu, \Sigma)$ which is the joint distribution between the two sets of random variables Y and X. We can express this distribution as:

$$Z = \begin{bmatrix} Y \\ X \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}$$

Similarly to PCA where we want find linear combinations of the variables that maximize the variance of the projections, here we want to find the linear projections of X and Y that maximize the correlation between the projected variables. In this scenario we define the **pairs of cannonical variables** (V_r, W_r) which are obtained linearly from the original variables X and Y using the set of coefficient (a_r, b_r). We can write this combination of variables as:

$$V_r = a_r Y = \begin{bmatrix} a_{r1} & a_{r2} & \cdots & a_{rp} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \sum_{i=1}^{p} a_{ri} Y_i$$

$$W_r = b_r X = \begin{bmatrix} b_{r1} & b_{r2} & \cdots & b_{rq} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} = \sum_{i=1}^{q} b_{ri} X_i$$

Similar for the PCA, the $r-th$ cannonical pairs (V_r, W_r), given by the $r-th$ cannonical coefficients (a_r, b_r) seek to **maximize the correlation between** V_r **and** W_r, named $\rho_{V_r W_r}$ under the following constraints:

- The variances of the projections V_r and W_r is equal to 1.

$$V(V_r) = 1 \qquad V(W_r) = 1$$

- The pairs (V_r, W_r) are uncorrelated to any of the previous $r-1$ pairs. This means:

$$\begin{aligned} Cov(V_i, V_j) &= 0 \qquad i \neq j \\ Cov(W_i, W_j) &= 0 \qquad i \neq j \\ Cov(V_i, W_j) &= 0 \qquad i \neq j \end{aligned}$$

The maximum correlation between V_r and W_r given the previous constraints is called the **canonical correlation** ϱ_r

So we are left in a similar scenario as PCA where in PCA we found subsequent linear projection p_i that maximized the variance of the projection and were independent to the previous projections. Now we want to find subsequent pairs of linear projections, a_r and b_r which maximize the correlation between the projections under the constraints that the projecitons are independent and their variance is 1.

I guess the coefficients a_1 and a_2 do not need to be perpendicular so that the correlations are 0. If they are, then we are sure of it, it is sufficient but not necesary condition. Playing properly with the covariances of the variables we can achieve it ?. Also a_r and b_r must project independent spaces but over different variables X and Y so they dont need to be perpendicular.

6.1 Computation of solutions

About how to solve this problem, we will not show any deep proof, it is a maximisation problem with restrictions and can be solved by using Lagrange multipliers. Here we will state the solutions, it can be proven that:

- The r-th canonical correlation ϱ_r is equal to the r-th largest root λ_r of:

$$f(\lambda) = det\left(\begin{bmatrix} -\lambda \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & -\lambda \Sigma_{XX} \end{bmatrix} \right) = 0$$

  So we could use eigendescomposition for example to obtain all of the cannonical correlations $\varrho_1, ..., \varrho_p$ but we still need to find the projections a_r and b_r. Notice we only compute this for $p \leq q$ roots of the funtion f(λ).

- The coefficients in the r-th pair of canonical variables satisfies the constraints so that the pairs a_r and b_r are uncorrelated.

$$\begin{bmatrix} -\varrho_r \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & -\varrho_r \Sigma_{XX} \end{bmatrix} \begin{bmatrix} a_r \\ b_r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

  And since the variance of the projections must be 1 then it is constrained by:

$$a_r^T \Sigma_{YY} a_r = 1$$
$$b_r^T \Sigma_{XX} b_r = 1$$

Under this type of solution, we can solve the linear system independently for a_r and b_r and we have that:

- The previous is the same as solving for a_r:

$$\Big(\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \varrho_r^2 \sigma_{YY}\Big) a_r = 0$$

  Subject to:

$$det\Big(\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \varrho_r^2 \sigma_{YY}\Big) = 0$$

- Analogous for b_r we have that:

$$\Big(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \varrho_r^2 \sigma_{XX}\Big) b_r = 0$$

  Subject to:

$$det\Big(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \varrho_r^2 \sigma_{XX}\Big) = 0$$

So in order to find the solutions we first find the cannonical correlation ϱ_r, by means of either the first determinant of the 2nd, and once we have them we have an homogenoues linear system, which we solve reducing the space to a triangular matrix, then it becomes trivial to solve. After that we standarize a_r and b_r so that the porjected variances are 1.

The value of the coefficients of a and b still measures how much they are representing the original variables X and Y.

Also keep in mind that the sets of variables X and Y will be independent if the following is equal to 1

$$1 - \rho^2_{Y|X_1,X_2,\ldots,X_D} = \frac{|\Sigma|}{|\Sigma_{YY}||\Sigma_{XX}|} = \frac{V(Y|X)}{V(X)}$$

7 Linear Discriminant Analisys

In this section we will see the classification systems that arise when we assume that the datapoints X associated to different classes Y_j, $j = 1, ..., J$ belong to different Multivariate Gaussian distributions. In the literature these classes are called populations. Our final goal is to build a decisor function that can tell us for each sample X_i to what class belongs to, and if the statistical framework allows it, with what probability it belongs to that and the other classes.

In a normal classification problem XXX

In the case of Linear Distriminant Analysis, we will assume that the samples of each class C_j follow a Multivariate Gaussian distribution with its own mean and variance. This is express in the following formula.

$$X_j \in C_j \sim N(\mu_j, \Sigma_j)$$

And therefore the probability likelihood of any sample X_i belonging to any of the possible classes, it is the likelihood of those samples being generated by the Gaussian distribution of the class with parameters μ_j, Σ_j. That being:

$$f_j(X_i) = f(X_i|\mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_j|}} exp\left[-\frac{1}{2}(X_i - \mu_j)^T\Sigma_j^{-1}(X_i - \mu_j)\right]$$

In general we will have a discriminator function $d(X)$ that will tell us for each sample X to which class it belongs to. The Region of a class R_j is the set of points, the part of the space in R^D which we classify as class j.

$$R_j = x \in R^D|d(X) = C_j$$

The classifier function is the one that determines these regions, and we usually want to find the classifier that maximizes some criteria. In this document we will see in increasing order of complexity the classifiers that:

- Maximizes the likelihood the data X_i given the class $f_j(X_j)$
- Maximizes the probability of the data X_j belonging to the class $P(C_j|X_j)$
- Maximizes a Loss function over the probabilities of missclasification $\mathcal{L}(X)$.

Since the probability distribution of the different classes overlap, sometimes we will missclasify a sample, even though overall that is the best decision to maximize the probability of classification. The probability of misclassification for the class C_j will be the probability of the areas where C_j has probability density but other class is chosen by the discriminator function since other class hay a higher probability in that area. The probability of missclasification of the class C_j, the probability that the discriminator will wrongly say that the sample belong to other class when it belongs to class j is:

$$Pe_j = \int_{X \notin R_j} P(C = C_j|X)dx$$

Given 2 classes C_1 and C_2 we will define the discriminator function Z as a function that will help us discriminate between 2 different classes. For a given sample X, this function will generate a number:

$$Z = f(X, mu1, simga1)$$

If $Z > 0$ then we classify the sample as belonging to Class 1, if $Z < 0$ then we classify as class 2. The discriminant function is therefore:

$$d(X) = \begin{cases} C_1 & if Z \geq 0 \\ C_2 & if Z < 0 \end{cases}$$

Different decisors will have different discrination funciton, even when the distribution of the classes is the same. Therefore this function also takes into account in general any other information that we want to use in the classfication.

7.1 The ML decisor

The resulting ML classification problem is therefore:

$$C_j = argmaxfj$$

Given two classes C_1 and C_2 then we can compute the MLE discrination function by obtaining the equation of $f_1(X_i) > f_2(X_i)$ which we usually solve by simplicity by passing the second term to the left and taking logarithm, since its a monotonically increasing function, the relation still holds:

$$log\left(\frac{f_1(X_i)}{f_2(X_i)}\right) > log(1) = 0$$

TODO: Needed for QUIZ 5.2 !! Do over the weekend.
Just stablish proper definition for:
Then the Quadratic Discriminant Function (QDF) is (for equal loss
and the Linear Discriminant Function (LDF) is (for equal losses)

Here we will differentiate 2 cases:

- Where we assume that both classes have the same covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma$ in which case the discrimination is called LDA and the disrimination funciton is a linear function of the dimensions of X. In this case we can simplify the fraction since normalization constants cancel out and we can combine the tensors of the exponent in the Gaussian distribution:

  $$log\left(\frac{f_1(X)}{f_2(X)}\right) = -\frac{1}{2}\Big[(X-\mu_1)^T\Sigma^{-1}(X-\mu_1) - (X-\mu_2)^T\Sigma^{-1}(X-\mu_2)\Big]$$

  Rearranging terms we end up with the equivalent expression:

  $$log\left(\frac{f_1(X)}{f_2(X)}\right) = -\frac{1}{2}\Big[X^T\Sigma^{-1}(\mu_1-\mu_2) - \mu_1^T\Sigma^{-1}\mu_1 + \mu_2^T\Sigma^{-1}\mu_2\Big]$$

  As we can see, the discrimination function is a linear function of X, specfically the Malbolis dintance between X and the vector obtained by the substraction of the means, that is the vector going from μ_1 to μ_2. **The discrimination boundary will be a hyperplane with mean in the middle point between the means and direction equal to the first eigenvector** !

- Where we assume that both classes have the different covariance matrix $\Sigma_1 \neq \Sigma_2$ in which case the discrimination is called QDA (Quadratic Analysis) and the disrimination funciton is a quadratic function of X. In this case we obtain the following discrimination function:

  $$log\left(\frac{f_1(X)}{f_2(X)}\right) = log\left(\sqrt{\frac{|\Sigma_1|}{|\Sigma_2|}}\right) - \frac{1}{2}\Big[(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1) + (X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)\Big]$$

  In this case we cannot simplify the terms and therefore this is a quadratic function of X. We can also see how determinants will bias the discrimination. If the space of C_1 is more expanded that the one of C_2, given by the determinant of their covariance matrices, then the decision is biased to unfavour C_1 with respect to C_2 with respect to the exponent, since this is negative.

7.2 The MAP and general decisor

In the Maximum a Posterior Discrimination case, we maximice the probability of the class, given the sample instead. The classes and the samples form a joint mixed distribution where we have the possible class as a discrete random variable that can take any of the possible values $C \in [C_1, C_2, ..., C_j]$. Using bayes we can compute That is:

$$P(C = C_j|X) = \frac{f(X|C = C_j)P(C_j)}{f(X)}$$

Where $P(X)$ is the probability of having obtained the sample overall. It can be computed by just maginalizing over the classes:

$$f(X) = \sum_{j=1}^{L} f(X, C = C_j) = \sum_{j=1}^{L} f(X|C = C_j)P(C_j)$$

In this case $f(X)$ is a mixture of Gaussians. Since its value does not depend on the class C_j, it does not affect the maximization problem and therefore the maximization problem is:

$$armax$$

The values of the prior probabilities can be chosen as desired, a common practice is to select them as the proportion of samples of the classes that we have for example.

$$P(C_j) = \frac{N_j}{N}$$

In this cirunstances then the Discrimination problem between 2 classes is $f_1(X_i)P(C_1) > f_2(X_i)P(C_2)$ resulting in the solution

$$log\left(\frac{f_1(X_i)}{f_2(X_i)}\right) > log\left(\frac{P(C_2)}{P(C_1)}\right)$$

Where the left function is the same as viewed before. As we can see, if the prior probability of the first class is bigger than the second, then the decision is biased towards C_1 since now the ratio of likehoods need to be lower than 1 to choose the first class.

Futhermore, we can establish costs to classification and missclasification events. In the MAP estimation of the classfication, a missclassficaiton from C_1 to C_2 has the same weight as the opposite. In some situations we do not want this since the cost of having, for example the C_1 is "There is a XX....

We can add this asymetry in a probabilistic way, where our discrimanator maximizes the expected cost, that is the sum of each cost multplied for the proabability that that will cost will happen, which happen in the missclassification situations. We usually assing no weight to the right classification. For the 2 class case, we have the costs:

- L_{12}: Cost of classifiying C_2 when it was C_1
- L_{21} Cost of classifiying C_2 when it was C_3

$$E[\mathbf{L}(X)] = L_{12}Pe_1 + L_{21}Pe_2$$

Where in the 2 Class problem, the regions are complementary. If we use the probability of missclassification using the MAP then we have:

$$E[\mathbf{L}(X)] = L_{12}\int_{X\in R_2} P(C = C_1|X)dx + L_{21}\int_{X\in R_1} P(C = C_2|X)dx$$

So the new classification using bayes becomes:

$$log\left(\frac{f_1(X_i)}{f_2(X_i)}\right) > log\left(\frac{P(C_2)L_{21}}{P(C_1)L_{12}}\right)$$

As we can see, if the cost missclassifying the first class L_{12} is bigger than the second, then the decision is biased towards C_1, so that we make less errors in C_1, we do not care that much if we missclassify more samples from class C_2 as class C_1.
And using LDA, we have that Minimizing this cost implies in the binary case, that

7.3 Linear Discrination function

Finally, our discrimination function will be at the end of the day, when we assume the same covariance matrix is given by the most general equation:

$$Z = -\frac{1}{2}\left[X^T\Sigma^{-1}(\mu_1 - \mu_2) - \mu_1^T\Sigma^{-1}\mu_1 + \mu_2^T\Sigma^{-1}\mu_2\right] - log\left(\frac{P(C_2)L_{21}}{P(C_1)L_{12}}\right)$$

If $Z > 0$ then we classify the sample as belonging to Class 1, if $Z < 0$ then we classify as class 2. The discriminant function is therefore:

$$d(X) = \begin{cases} C_1 & if Z \geq 0 \\ C_2 & if Z < 0 \end{cases}$$

This function will determine the regions R_1 and R_2 from which we can compute later the different probabilities of correct classification and misclassification for each of the classes. In the 2 classes case we have the probability of missclassification for class 1 as:

$$Pe_1 = \int_{X \notin R_1} P(C = C_1|X)dx = \int_{X \in R_2} P(C = C_1|X)dx = \int_{Z<0} P(C = C_1|X)dx$$

Every sample X_i will yield a different Z_i from which it will be classified into either C_1 or C_2 by the discriminant function $d(X_j)$. Since Z_i is a mere deterministic function of X_i which is a random variable belonging to a mixture of Gaussians, then Z is also a random variable. Actually Z in the LDA case, is just a linear projection of the input space X with coefficients w and bias b given by:

$$w = -\frac{1}{2}\Sigma^{-1}(\mu_1 - \mu_2) \qquad b = \frac{1}{2}\left[\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right] - log\left(\frac{P(C_2)L_{21}}{P(C_1)L_{12}}\right)$$

Since X belongs to a mixture model, the linear projection Z will also be a mixture model of the projections of the individual Multivariate Gaussians of each class. But the distribution of Z given that we only projects samples from a single class C_j will be a gaussian, the Gaussian obtained by this linear projection of the Gaussian of the class $f_j(X)$. Of course we could project any of the Classes into this hyperplane, but it only makes sense for the 2 classes C_1 and C_2 from which it was computed. The Resulting distributions are

$$Z_1 \sim N\left(\frac{1}{2}||\mu_1 - \mu_2||^2_{\Sigma^{-1}} - log(c), ||\mu_1 - \mu_2||^2_{\Sigma^{-1}}\right) \qquad X \sim N(\mu_1, \Sigma)$$

$$Z_2 \sim N\left(-\frac{1}{2}||\mu_1 - \mu_2||^2_{\Sigma^{-1}} - log(c), ||\mu_1 - \mu_2||^2_{\Sigma^{-1}}\right) \qquad X \sim N(\mu_2, \Sigma)$$

Where we call c the biases added by the priors and costs

$$c = \frac{P(C_2)L_{21}}{P(C_1)L_{12}}$$

We can see that the variance of the projection is very close to the distance between the means. The more distance there is between the means, the more variance we have in the projection. **Elaborate** !

We can rewrite the probability of missclassification using Z_1 and Z_2 now, since the probability of missclassfiying C_1 is the probability that decision value Z_1 is negative, $Z_1 < 0$ when the sample was generated by the distribution of C_1. That is:

$$Pe_1 = \int_{Z<0} f(C = C_1|X)dx = \int_{z_1<0} f(Z_1)dz_1$$

Since Z_1 follows the Gaussian distribution seen previously, we can normalize it substracting the mean and dividing by the standard deviation to express the probability of missclassification using the normalized gaussian distribution $N(0, 1)$ as:

$$Z_1^{'} = \frac{Z_1 - \frac{1}{2}||\mu_1 - \mu_2||^2_{\Sigma^{-1}} + log(c)}{\sqrt{||\mu_1 - \mu_2||^2_{\Sigma^{-1}}}} \sim \mathcal{N}(0, 1)$$

$$Pe_1 = P[N(0, 1) < Z_1^{'}] = \Phi(Z_1^{'})$$

Equivalently we have that the proability of missclassification for the second class is:

$$Pe_2 = P[N(0, 1) > Z_2^{'}] = 1 - \Phi(Z_2^{'})$$

Where $Z_2^{'}$ is the corresponding normalized Z_2.

$$Z_2^{'} = \frac{Z_2 + \frac{1}{2}||\mu_1 - \mu_2||^2_{\Sigma^{-1}} + log(c)}{\sqrt{||\mu_1 - \mu_2||^2_{\Sigma^{-1}}}}$$

7.4 Best linear discriminator

Way of seeing it as finding the best hyperplane given by... Say that we could always select a subset of X, adding X will always add discrinimatory information ? Could it make the distance more noisy and therefore less probability of good discrimination ? Or the lower bound is that it is just random and there is no discrimination. Graph with 1,2,3 dimensions.

7.5 Estimating the Parameters

If we do not have the distribution parameters of the populations, then we need to estimate them. Focusing in the 2 classes case but extendable to all classes, if our samples are labeled from which class they come from, we have the pairs (X_i, C_i) then we can compute the parameters using normal estimation. If we assume Gaussianity on the classes, then we have:

$$\mu_i = ESTIMATOR \qquad Estimator$$

In the case of LDA, we can estimate the common Dispersion Matrix from the individuals, called the **pooled estimate** of the dispersion matrix

$$\hat{\Sigma} = \frac{1}{N-2}\left[(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2\right]$$

This will be the weighted mean of the individual Dispersion matrices of the classes. For the multiclass case we have the general case:

$$\hat{\Sigma} = \frac{1}{N-C}\sum_{j=1}^{C}(n_j - 1)\hat{\Sigma}_j$$

Now that we have the estimated parameters but not the real ones, lets see how this affects the discrimination function Z between two classes C_1 and C_2, since now X_j belongs to a estimated distribution $N(m\hat{u}_j, \hat{\Sigma}_j)$ where the parameters are also uncertain, since mu_j follows a multivariate T student distribution, and $Sigma_j$ follows a multivariate χ^2 distribution.

The actual distribution of Z is very complicated but for large sample sizes it is asymptotically equal to the distribution seen before.

$$Z$$

reasonable sample sizes we can use the theory we have derived.

So for computing things like the false alarm probabilities and so on, we will assume that the parameters are the true ones. But we can also make other simpler statistical tests where we dont make this assumption.

- Is there an actual distance between the means of two classes ? $\mu_1 = \mu_2$
- Do the last q dimensions of X contribute to the classification ? We could actually select a subset of functions

7.6 Test of significant distance between classes

For the first one, we want to know whether the difference between the means of both Multivariate Distributions is significant $H_0 : \mu_1 = \mu_2$, here we do not assume that estimated parameters are the real ones, otherwise there would be no uncertainty. To test this, we compute the empirical Mahalanobis' distance between the means, the distance using the inverse covariance matrix.

TODO: There is a notation problem here with the values of D.

$$d^2 = ||\hat{\mu_1} - \hat{\mu_2}||_{\hat{\Sigma}^{-1}} = (\hat{\mu}_1 - \hat{\mu}_2)^T\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

Notice this is the empirical distance, not the ideal one since we are using the estimated parameters that add uncertainty to this distance. We see that D^2 is closely related to Hotelling's T^2 statistic for the two sample situation. In such case, we have that:

$$d^2 = \frac{n_1 + n_2}{n_1 n_2}T^2$$

Therefore we can test whether $mu_1 = mu_2$ by means of d^2 by computing its corresponding T^2 statistic. The final normalized statistic is:

$$z = \left[\frac{n_1 + n_2 - D - 1}{D(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2}\right] d^2$$

Where d^2 is the sample of D^2 that we have obtained. The statistic belongs to the F distribution, the same way as the normalized T^2 statistic with degrees of freedom:

$$z \sim F(D, n_1 + n_2 - D - 1)$$

The alternative hypothesis is that the distance is bigger since a distance cannot be negative.

7.7 Test of better estimator between classes

We can also test which one is the best discriminator between the 2 classes, that is finding the best hyperplane δ that reduced the Expected loss. In our case we have that:

$$\delta = \Sigma^1(\mu_1 - \mu_2)$$

Which it is the best discriminator because it is the linear projection d that maximizes the expected distance between the classed divided by the variance $g(d)$

$$g(d) = \frac{||E_1[X^T d] - E_2[X^T d]||^2}{V(X^t d)} = \frac{[(\mu_1 - \mu_2)^T d]^2}{d\Sigma d}$$

Mahalanobis' D^2 is the maximum value of g(d)

Since we readily get that a hyperplane d is the same as the same hyperplane multiplied by a constant k, we have that $g(kd) = g(d)$. We can determine the maximum for the numerator under the constraint $d\Sigma d = 1$, by using Lagrange multipliers λ we arrive to the solution that the d that maximizes the numerator and therefore the function is:

$$d = \left[\frac{1}{\lambda}(\mu_1 - \mu_2)^T d\right] \Sigma^1(\mu_1 - \mu_2) = k\delta$$

This implies that the discrination function that we obtained in LDA, and which hyperplane is defines as:

$$X^T \delta = \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_D X_D$$

is the linear function that is the projection that "moves" C_1 furthest possible away from C_2 measured in units of the standard deviation of the projected distributions or - in analysis of variance terms - the projection which maximizes the variance between populations divided by the total variance.

REPLACE WITH REAL CHART

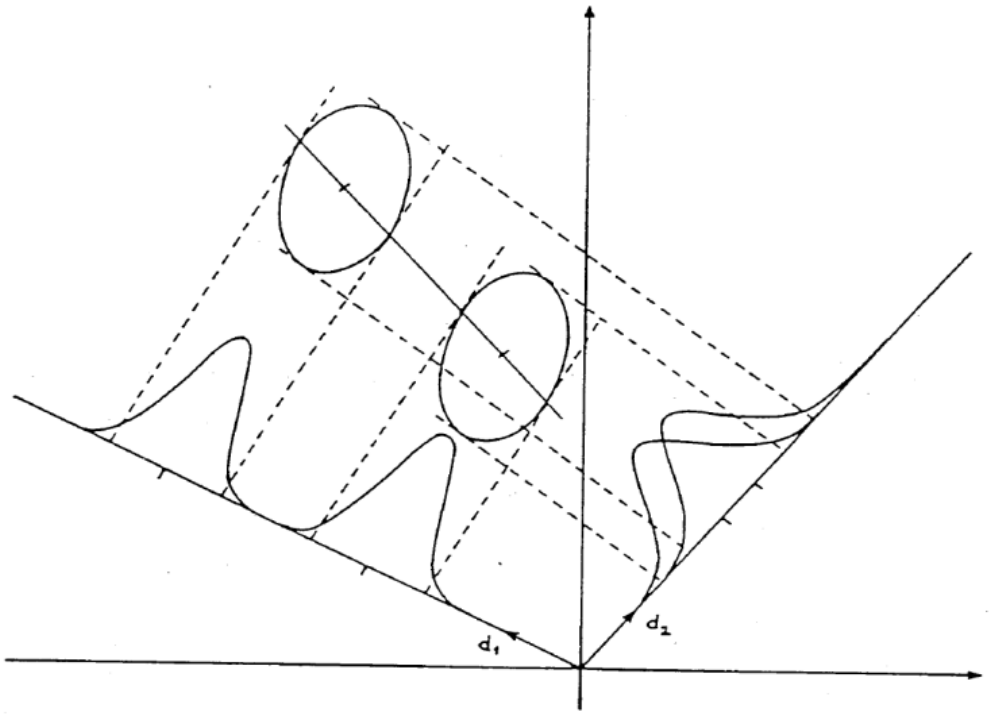


Figure 47: SMA and its window

Since now we have uncertainty in our estimated, the best discriminator function $g(d) = D^2$ has an associated uncertainty. We can use that uncertainty to make an statistical test if the difference between the goodness of a discriminant hyperplane d is statistically significant respect to that of another hyperplane d. We have

$$D^2 = \hat{\sigma}(d) = \frac{[(\hat{\mu_1} - \hat{\mu_2})^T d]^2}{d\hat{\Sigma}d}$$

which is noisy version of the Malahobis distance. So we can check if this value signigcantly different for two linear projects d_1 and d_2 with the same number of dimensions D and with squared distance between the means D_1^2 and D_2 respectively, we have the test statistic:

$$z = \frac{n_1 + n_2 - D - 1}{D - 1} \frac{n_1 n_2 (D_2^2 - D_1^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_1^2}$$

This statistic is distributed according to:

$$z \sim F(D - 1, n_1 + n_2 - D - 1)$$

z gives a measure of how much the "distance" between the two populations is reduced by using d_1 instead of d. We want this dintance to be as maximum of possible so that the distributions are far apart. If this reduction is too big i.e. if z is large, we are bringing the two classes too close together and therefore d_1 is not a good alternative to d since it impoverishes the solution too much.

7.8 Test of reduction of dimensionality

In the same like of reasoning, we could test if a subset of the variables of X is enough to have a good discriminator. If we reduce the number of variables, we will bring the means closer together since the distance between them can only decrease. The less dimensions we have, the less distance between two points. But maybe that decrement of distance between the means is insignificant and we can decide not to use the variables as they are just noise that impoverishes the real classification.

$$\sqrt{x_1^2 + x_2^2 + \cdots + x_D^2} \leq \sqrt{x_1^2 + x_2^2 + \cdots + x_D^2 + x_{D+1}^2}$$

We can use the same statistic to check if the last m variables are insignificant, by means of the following statistic. Now the two projections will have different number of dimensions, one is D and the other

$D - m$. The distance of D_D is suposed to be bigger since the more dimensions you have, the distance is always increased !!!

$$z = \frac{n_1 + n_2 - D - 1}{m} \frac{n_1 n_2 (D_D^2 - D_{D-m}^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_{D-m}^2}$$

This statistic is distributed according to:

$$z \sim F(D - m, n_1 + n_2 - D - 1)$$

under the hypothesis that both discriminators are as good. High values of Z are critical meaning that there is a significant difference between their means and therefore their discrimination capabilities.

8 Canonical Discriminant Analisys

Not much about this, LDA already does it, generalization of LDA. I dont have to get deeply.

9 Linear Regression

In this Section we will describe different ways solving the Linear Regression problem, going from the non-statiscal ways of computing the parameters, to the purely statistical framework where we assume the joint distribution of the input data X and the output data Y. As we will see, the solutions of each approach are similar, but adding a statistical framework allows us to express the uncertainty about our estimations, and it enables natura extensions within Bayesian statistics.

In a general Linear Regression scenario, we want to estimate the value of a variable Y, by using a linear combination of a set of input features $X = [X_1, X_2, ...X_D]$. In other words, we want to find the best linear projection $\hat{Y} = \theta X$ with coefficients $\theta = [\theta_1, \theta_2, ..., \theta_D]$ applied to the observed features X. The estimation $\hat{Y}$ should be the "best" one possible according to some criterion that we establish, different criteria will lead to different optimal parameters θ^*.

$$\hat{Y} = f(X) = \theta_0 + \theta_1 X_1 + \cdots + \theta_D X_D = \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_D\end{bmatrix} \begin{bmatrix}1 \\ X_1 \\ \vdots \\ X_D\end{bmatrix} = \theta Z$$

It is also common to add a bias term in order to give more expressivity to the model, capturing the bias term... This bias can be modeled like a new input variable X that takes always the value 1

Notice how we have created new input variable vector $Z = [X_0 = 1, [X_1, ..., X_D]]$ by adding a 1 to each of the samples in order to be able to use complete matrix notation including the bias parameter θ_0. We could also substract the mean of both X and Y, therefore removing the need to have bias term θ_0.

In this scenario we will start with a dataset of N samples, each sample being composed by the desired output Y_i and its associated input X_i, we denote our dataset as $\mathcal{D} = \{X_i, y_i\}_{i=1}^N$, where X_i is a D-dimensional vector, $X_i \in \mathcal{R}^D$ and Y_i is one-dimensional in this simple case. Our ultimate goal is to find these linear coefficients θ to express Y as a linear function of X. The estimation for each of the samples in matrix form is:

$$\hat{Y}^T = (\theta Z)^T = Z^T \theta^T \qquad \begin{bmatrix}\hat{Y}_1 \\ \hat{Y}_1 \\ \vdots \\ \hat{Y}_N\end{bmatrix} = \begin{bmatrix}1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1N} \\ \vdots & \vdots & & \vdots \\ x_{D1} & x_{D2} & \cdots & x_{DN}\end{bmatrix}^T \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_D\end{bmatrix}^T$$

Notice in this notation how:

- The size of the rows of the matrix Z is the dimension of the input variables D plus one to account for the bias. A sample vector is a column vector.
- The columns of the matrix correspond to each of the N samples.

This regression is not going to be perfect, our estimates $\hat{Y}_i$ are not gonna exactly match the dataset values Y_i that we are trying to estimate, therefore there is going to be a residual, an error associated to each estimated sample Y_i. We denote as ϵ_i the residual of the regression of the $i-th$ sample:

$$\epsilon_i = Y_i - \hat{Y}_i$$

Thefore we can express the residual in matricial form as:

$$\epsilon = Y - \hat{Y} = Y - \theta Z \qquad \epsilon^T = \begin{bmatrix}\epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N\end{bmatrix} = \begin{bmatrix}Y_1 - \hat{Y}_1 \\ Y_2 - \hat{Y}_2 \\ \vdots \\ Y_N - \hat{Y}_N\end{bmatrix} = \begin{bmatrix}Y_1 \\ Y_2 \\ \vdots \\ Y_N\end{bmatrix} - \begin{bmatrix}1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \cdots & x_{ND}\end{bmatrix} \begin{bmatrix}\theta_0 \\ \theta_1 \\ \vdots \\ \theta_D\end{bmatrix}$$

In this scenario $\hat{Y}$ is our estimation of Y and ϵ is the residual, the error we have in our estimation. Notice that **so far we have not made any statistical assumption**, and we do not explicitly need them in order to solve the problem.

In the following we will see 3 different ways of modeling the data, our assumptions and desires. We will the 3 following models:

1 When we do not make any statistical assumption and our goal is just to minimize a function over the data.

2 When we assume the joint distribution of the data (Y, X). This will be related to the conditional distribution in the Multivariate Gaussian case that we saw in earlier Sections.

2 When we assume the distribution of the residual $\epsilon = Y - \theta X$, but not necesarily the distribution of X and Y. The case of the GLM model.

9.1 Non statistical framework

In this model, we do not consider that our data points follow any given distribution, they are just pair of values (X_i, Y_i) from which we can compute a determinist function that we want to minimize, called the Loss function $\mathcal{L}(\theta, \mathcal{D})$.

In the Linear Regression problem, given a Loss function to minimize, we could just build an algorithm that tries to find the best linear projection $\theta = \theta^*$, which associated Loss function value $\mathcal{L}(\theta^*, \mathcal{D})$ is the minimum for all possible θ values using the dataset $\mathcal{D}$.

$$\theta^* = arg\min_{\theta}\{\mathcal{L}(\theta, \mathcal{D})\}$$

Notice that this $theta^*$ will be optimal for the dataset D that we have, we know that for a different dataset we would get a different value, but since this has no statistical framework we are not able to obtain out certainty about the parameters. In this case $theta^*$ is our best estimated parameters $\theta^* = \hat{\theta}$

Different Loss functions will lead to different optimal values of $theta^*$. A typical example of Loss function $S(\theta)$ is the Mean Squared Error (MSE) where we minimize the average of the sum of the individual squared errors.

$$S(\theta) = \epsilon\epsilon^T = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - \theta Z_i)^2 = (Y - \theta Z)(Y - \theta Z)^T$$

This is a convex measure of error over the parameters θ and therefore it has a closed form solution as long as the dispersion matrix among the samples is positive definite.

$$\hat{\theta}^T = (ZZ^T)^{-1} ZY^T$$

Notice that for any other Loss for which we might not have a closed form solution, we could always use a greedy algorithm in which we draw parameter vectors θ at random and just keep the one that minimizes the Loss. Or we could use Local Search methods in which we start from an initial solution θ^{t_0} and we move towards the direction that minimizes the loss, reaching the solution θ^{t_T} after T iterations.

Also notice that we could extend this non-statistical framework further. For example we could give a weight $w_i, i = 1, .., N$ to each sample which will weight its contribution to the Loss function. This way we consider the error of some samples more important than others the error of others. An example of application of this scheme is Time Series Analysis, where you want to penalize more the error of nearby samples so that the model learns local parameters, specialized for a given part of the series.

In order to include these weights to the Loss function in matrix notation we can place them as the diagonal of a matrix Λ where:

$$\Lambda = diag(W) = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_N \end{bmatrix}$$

Our loss function then changes to:

$$S(\theta, W) = \sum_{i=1}^{N} w_i \epsilon_i^2 = \epsilon\Lambda\epsilon^T = (Y - \theta Z)\Lambda(Y - \theta Z)^T$$

And our optimal estimator now changes to:

$$\hat{\theta}^T = (Z\Lambda Z^T)^{-1} \Lambda ZY^T$$

Notice again that no statistical assumptions have been made explicitly. In the following we will stablish statistical assumptions on the data and then derivate their optimal estimators θ, as we will see, the result is very similar to the ones we have just seen, but thanks to the mathematical framework we can compute things like the uncertainty of our esrtimation and we have straigh forward extensions of the regression with statistical foundation.

9.2 Full statistical framework

In the first statistical model, we will assume that our data points (X_i, Y_i) are samples from a joint probability distribution $f(X, Y)$. As usual, we will assume this joint distribution to be a Multivariate Gaussian distribution $(Y, X) \sim \mathcal{N}(\mu, \Sigma)$. Furthermore we will initially assume that all the samples are independent and identically distributed. The parameters of the Joint Distribution are:

$$\mu = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix} = \begin{bmatrix} \mu_Y \\ \mu_{X_1} \\ \vdots \\ \mu_{X_D} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} = \begin{bmatrix} \sigma_{YY} & \sigma_{YX_1} & \cdots & \sigma_{YX_D} \\ \sigma_{X_1Y} & \sigma_{X_1X_1} & \cdots & \sigma_{X_1X_D} \\ \vdots & \vdots & & \vdots \\ \sigma_{X_DY} & \sigma_{X_DX_1} & \cdots & \sigma_{X_DX_D} \end{bmatrix}$$

Since X follow a Multivariate Gaussian distribution, our estimation $\hat{Y}$ will have a Gaussian Distribution as well since it is a linear projection X given by the parameters θ. And since the error ϵ is a difference of Univariate Gaussians, then its distribution is also Gaussian.

$$\epsilon = Y - \theta X \sim \mathcal{N}(\mu_\epsilon, \Sigma_\epsilon)$$

Where using the linear properties of the Multivariate Gaussian distribution we have that the statistical parameters of the noise are:

$$\mu_\epsilon = \mu_Y - \theta\mu_X \qquad \sigma_\epsilon = \theta\Sigma_{XX}\theta^T + \sigma_Y^2$$

Given that we know the true distribution parameters μ, Σ, we can compute the likelihood of a given error sample ϵ_i from the values X_i and Y_i, for a given projection θ. The probability density function of the residual can be expressed as:

$$\begin{aligned} f(\epsilon|\theta) = g(X, Y|\theta) &= \frac{1}{\sqrt{(2\pi)^D\sigma_\epsilon}} exp\left[-\frac{1}{2}(\epsilon - \mu_\epsilon)^T\sigma_\epsilon^{-1}(\epsilon - \mu_\epsilon)\right] \\ &= \frac{1}{\sqrt{(2\pi)^D\sigma_\epsilon}} exp\left[-\frac{1}{2}\Big((Y - \mu_Y) - \theta(X - \mu_X)\Big)^T\sigma_\epsilon^{-1}\Big((Y - \mu_Y) - \theta(X - \mu_X)\Big)\right] \\ &= \frac{1}{\sqrt{(2\pi)^D\sigma_\epsilon}} exp\left[-\frac{1}{2\sigma_\epsilon}\Big((Y - \mu_Y) - \theta(X - \mu_X)\Big)^2\right] \end{aligned} \tag{2}$$

Notice we use $g(X, Y|\theta)$ notation and not $f(X, Y, |\theta)$ since the later one could be confused with the joint pdf of X, Y. Without loss of generality we could remove the mean to both the X and Y samples so that now we have X:

$$f(\epsilon|\theta) = g(X, Y|\theta) = \frac{1}{\sqrt{(2\pi)^D\sigma_\epsilon}} exp\left[-\frac{1}{2\sigma_\epsilon}(Y - \theta X)^2\right] \tag{3}$$

In this set up, we would like to find the linear projection θ that maximizes the likelihood of our error data samples ϵ_i, where we assume that they are $i.i.d$ and follow the previous distribution. We can also apply the $log()$ function to the element that we are maximizing since it is a monotonically increasing function, so anything that maximizes $f(z)$ will also maximize $log(f(z))$ and viceversa. This will turn the multiplications into sums. Furthermore we can negate the function to convert it into a minimization problem. This process can be seen in the following equations:

$$
\begin{aligned}
\hat{\theta} &= arg\max_{\theta}\{f(\epsilon_1, \epsilon_2, ..., \epsilon_N|\theta)\} \\
&= arg\max_{\theta}\{\prod_{i=1}^{N} g(X_i, Y_i|\theta)\} \\
&= arg\min_{\theta}\{-\sum_{i=1}^{N} log(g(X_i, Y_i|\theta))\} \\
&= arg\min_{\theta}\{\sum_{i=1}^{N}\left[\frac{D}{2}log(2\pi) + \frac{1}{2}log(\sigma_\epsilon) + \frac{1}{2\sigma_\epsilon}(Y_i - \theta X_i)^2\right]\}
\end{aligned}
\tag{4}
$$

Finish this but not now. How we optimize θ and the estimated parameters of the distribution at the same time ? Same as normal loglikelihood I guess, in the derivation they should either be independent or cancel each other. I can just say that the solution holds for the estimated values as well I guess.

This is the condition to minimize the variance of the projection. Is it the same as the previous problem ? Can it be proven that it Reexplain relationship with the Conditional Gaussian We previously saw that the optimal estimation of Y given X is:

$$E[Y|X] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

$$VAR[Y|X] = \Sigma_{YY} - \Sigma_{YX}\Sigma_{YX}\Sigma_{XX}^{-1T}$$

Ans we can relate it to the linear coefficients as:

$$\theta_0 = \mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\mu_X \qquad \theta_{1:D} = \begin{bmatrix}\theta_1 & \theta_2 & \cdots & \theta_D\end{bmatrix} = \Sigma_{YX}\Sigma_{XX}^{-1}$$

Of course we do not have the real values of the covariances matrix, but using the ML estimators of θ and plugging them into the formula we obtain:

$$\hat{\theta_0} = \mu_Y - YX^T(X^TX)^-1\mu_X \qquad \hat{\theta}_{1:D} = \begin{bmatrix}\theta_1 & \theta_2 & \cdots & \theta_D\end{bmatrix} = YX^T(X^TX)^-1$$

If we identify terms with the solution we obtained in the previous non-statistical scenario we can see that they are the same.
in this case, our previous estimate would be the mean of the distribution and our squaed residual would be a sample from its variance. Estimate the uncertainty about the estimation.

$$a$$

As we saw earlier, there is a close relationship between this and the Conditional Distribution of Y. It is easy to show that in this case, the solution is:

This model is good, but also incomplete, here we are assuming that all of the samples X_i come from the same distribution and that all the samples are independent. Each of the samples in X_i and Y_i could come from a different distribution with parameters μ_i, Σ_i and be correlated with the other samples (a part from the usual correlation between the dimensions of X). Now it is when things get complicated

In the most generalistic (and unrealistic) case, each sample X_i comes from a different Multivariariate Gaussian distribution with parameters Σ_i, μ_i. Under this assumption we just multiplied the number of parameters of the model by N. This would make it impossible to estimate the distributions already since we only have 1 sample of each of $D+1$-dimensional Multivariate distributions $(X_i, Y_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$.

Furthermore, we can assume there is a correlation between the samples, which implies that there is a correlation between each of the $D-1$ dimensions of all the N samples. In our dataset it consists of NxD samples of univariate gaussians. In the most general case they could be different and correlated in any way so the number of parameters of the model would be:

$$(NxD)(NxD) + NxD$$

TODO: Finish. We dont have really (D+1)*(D+1) covariance between 2 samples right ? It is true for the covariance between 2 samples (X,Y) but not for the samples of the noise, since those do not depend directly on the correlation between Y and X right ? Other thing is that the optimal value of theta depends on the covariance between X and Y.
(Of course there is the redundanci in the Cov matrix)

So we have way more parameters than samples N, this makes the inference process very hard so either we must know the true governing distribution or we can assume some structure.
Say that we can just impose how the samples are correlated, that way we dont need to estimate them. Maybe there is an actual formal way in which the parameters change with the samples but in priciple they could be completely independent. This would be rather unsual and the statisitcal properties of samples would change rather smoothly for neighbouring samples $j = \cdots, i-2, i-1, i+1, i+2, \cdots$. Question to myself, could there be a way of deterministically evolcing the parameters in a way that the samples will still be incorrelated ? Is it possible to be incorrelated given that you dont know the model and correlated (or event the same) given the model ?

Instead of taking this approach of directly assuming that our variables X_i and Y_i have a specific joint distribution (and that it can change with i), other approach is just to assume that the residuals $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_N]$ follow a joint distribution, where the marginal distribution of every sample and the way it relates to the others could take any allowable form by the laws of probability. Depending on what assumptions we make on the distribution of ϵ we can get different solutions of the optimal linear coefficients θ.

9.3 General Linear Model

In the GLM, we will assume that the residuals of the samples $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_N]$ follow a joint Gaussian Multivariate distribution with mean 0 and covariance matrix Σ_ϵ.

$$\epsilon^T = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \sim \mathcal{N}(0, \Sigma_\epsilon) \qquad \mu_\epsilon^T = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \Sigma_\epsilon = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2N} \\ \vdots & \vdots & & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{NN} \end{bmatrix}$$

Where the covariance matrix of the noise Σ_ϵ could in principle take any form, meaning that the residuals of the samples could have a different variance σ_i for each sample and that they could be correlated, given by the covariance value σ_{ij}. The pdf of the joint statistical distribution of all noise values ϵ_i is:

$$\begin{aligned} f(\epsilon|\theta) &= \frac{1}{\sqrt{(2\pi)^D |\Sigma_\epsilon|}} exp\left[-\frac{1}{2} \epsilon \Sigma_\epsilon^{-1} \epsilon^T \right] \\ &= \frac{1}{\sqrt{(2\pi)^D |\Sigma_\epsilon|}} exp\left[-\frac{1}{2} \Big(Y - \theta X \Big)^T \Sigma_\epsilon^{-1} \Big(Y - \theta X \Big) \right] \end{aligned} \tag{5}$$

Very important: Notice also that this model does not assume any distribution on X and Y directly. We simply consider them as values in a deterministic function to generate our random variable $\epsilon = Y - \theta X$. Now the values of the covariance matrix Σ_ϵ do not depend explicitly in X, Y or θ as it happened in the previous full statistical framework. The covariance matrix will be same for all projection vectors θ.

Since ϵ is random and the values of X are considered deterministic, one can easily see that the regressed variable $Y = \theta X + \epsilon$ follows a gaussian univariate distribution $Y \sim \mathcal{N}(\theta X, \Sigma_\epsilon)$ given a set of coefficients θ.

$$Y = \hat{Y} + \epsilon = \theta X + \epsilon \sim \mathcal{N}(\theta X, \Sigma_\epsilon)$$

So our measurements Y do come from a Gaussian distribution given the projection coefficients θ. If X is non Gaussian then for many projections of θX, then Y will not follow a Gaussian distribution. The point is that our measurements X are just point estimates ?.

Of course the variables of X do actually come from a given distribution, they are random, but in this model we just consider them as point values ? For our assumptions to be true, the real joint distribution

of X does not need to be Gaussian, and also Y alone does not need to be Gaussian, the only thing that needs to be gaussian is the residual given by the projection over θ. The actual distribution of X can to be non-Gaussian as long as the linear combination of its dimensions given by θ is Gaussian. In a basic example, imagine we have 2 random variables A and B, where A is Gaussian and B belongs to any distribution. If our input vector is bidimensional $X = [X_1.X_2]$ ans we have:

$$X_1 = \frac{1}{2}A + B \qquad X_2 = \frac{1}{2}A - B$$

Then the sum of these two non-gaussian variables can be Gaussian:

$$Y = X_1 + X_2 = A \sim \mathcal{N}(\mu_A, \Sigma_A)$$

Where in our mode, μ_A would be our value $\hat{Y} = \theta X$ and the variance Σ_A would be our noise term $\Sigma_A = \Sigma_\epsilon$. It can also be proven that if for example the dimensions of X are independent, then a linear combination $\hat{Y} = \theta X$ can only be Gaussian if all the variables of X are Gaussian.

This differs from our previous approach in which (X, Y) formed a joinly gaussian distribution and we try to find the best linear projection to reduce the uncertainty of Y given X. Now our X and Y are allowed to have any distribution that they want, as long as the difference $Y - \theta X$ can be considered Gaussian, which must be checked after the regression.

In general it is us who impose the form of the Σ_ϵ, although some other methods could initially try to estimate its structure from the data as well. This will be, the structure of the noise that we are imposing. Can this noise strcture Σ_{XX} be considered a prior ? Maybe a model prior, but not a prior on the parameters to estimate as it is normally called in the literature. We also dont really need to specify the actual value of the variance ? is this because when you solve it then it comes out of the box ? Or we actually obtaining the solution also assuming that the value of sigma is the one we later obtain ?

Once we have stablished the distribution of the noise, then **it is time to estimate θ.** We want the projection which associated residual values $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_N]$ have the maximum likelihood given the noise covariance matrix Σ_ϵ that we impose. In other words we will be using the maximum likelihood estimator of θ:

$$\begin{aligned}
\hat{\theta} &= arg\max_\theta \{f(\epsilon_1, \epsilon_2, ..., \epsilon_N|\theta)\} \\
&= arg\max_\theta \{\frac{1}{\sqrt{(2\pi)^D|\Sigma_\epsilon|}} exp\left[-\frac{1}{2}\epsilon\Sigma_\epsilon^{-1}\epsilon^T\right]\} \\
&= arg\min_\theta \{\frac{D}{2}log(2\pi) + \frac{1}{2}log(|\Sigma_\epsilon|) + \frac{1}{2}(Y - \theta X)\Sigma_\epsilon^{-1}(Y - \theta X)^T\} \\
&= arg\min_\theta \{(Y - \theta X)\Sigma_\epsilon^{-1}(Y - \theta X)^T\}
\end{aligned} \tag{6}$$

The ML estimator of θ in the general case for any Σ_ϵ can be obtained as the solution of the so-called normal equations:

$$(X\Sigma_\epsilon X^T)\hat{\theta}^T = X\Sigma_\epsilon Y^T$$

If our data matrix X has full rank then we can invert the matrix to the left and obtain the estimated coefficients as:

$$\hat{\theta}^T = (X\Sigma_\epsilon X^T)^{-1}X\Sigma_\epsilon Y^T$$

Notice from the equation how a scaling of the covariance matrix does not affect the solution. The imposed covariance matrices Σ_ϵ and $k \cdot \Sigma_\epsilon$ will yield the same estimated $\hat{\theta}$ so when setting our prior, we should not care about the maginitude of the variance and covariance elements but only in their relative magnitude among them.

Notice this is the exact same solution we got in the non-statistical framework, where we stablished a weight w_i for each sample and created the diagonal matrix Λ. This is a generalization of that case where the constribution of a given error ϵ_i also depends on the error of the other residuals that it is correlated to.

If we assume all the noise samples are independent, then we have diagonal matrix Λ where each weight will be the inverse of the assumed variance for each sample ϵ_i. In an analogy with the non-statistical framework:

$$w_i = \frac{1}{\sigma_i^2}$$

So the more variance we assume for each sample ϵ_i, the less weight it will have in the loss function. This means that if we are more uncertain about specific samples, they contribute less to the computing of θ. Furthermore if we assume all the samples to have the same variance, in which case our noise samples are i.i.d, then our noise covariance matrix becomes $\Sigma_\epsilon = \sigma_\epsilon \mathcal{I}$ and we have the basic case seen in the previous models.

$$\hat{\theta}^T = (XX^T)^{-1}XY^T$$

Where the initial assumed σ_ϵ does not influence the estimator.

9.4 Analysis of the optimal estimator and its residual

As we can read from the previous equation, the ML estimator of the parameters $\hat{\theta}$ is a function of $Y = \theta X + \epsilon$ which is a random variable. Therefore our estimator is also a random variable that depends on the statistical properties of ϵ, that is, it depends on the Σ_ϵ that we imposed.

We can see that the distribution of $\hat{\theta}^T$ will also be Gaussian since it is equal to a linear combination of the individual noised of the samples.

$$\hat{\theta}^T = \Big[(X\Sigma_\epsilon X^T)^{-1}X\Sigma_\epsilon\Big]Y^T = \Big[(X\Sigma_\epsilon X^T)^{-1}X\Sigma_\epsilon\Big](\theta X + \epsilon)^T$$

The the maximum likelihood estimator $\hat{\theta}$ is central:

$$E[\hat{\theta}^T] = E[(X\Sigma_\epsilon X^T)^{-1}X\Sigma_\epsilon(X^T\theta^T + \epsilon^T)] = \theta^T$$

$$V[\hat{\theta}^T] = V[(X\Sigma_\epsilon X^T)^{-1}X\Sigma_\epsilon\epsilon^T] = \sigma^2(X\Sigma_\epsilon^{-1}X^T)^{-1}$$

The term σ^2 is the squared length of the error term, under the matrix Σ_ϵ^{-1} normalized by the number of degrees of freedom. It is the square distance between the observed values Y and our estimates $\hat{Y}$ according to the matrix Σ_ϵ^{-1}. Since we do not know σ^2, we need to estimate it, we will see two estimators, the Maximum Likelihood Estimator and the Unbiased Estimator.

- ML estimator of σ^2 is given by:

$$\tilde{\sigma}^2 = \frac{1}{N}||Y - \hat{Y}||^2_{\Sigma_\epsilon^{-1}} = \frac{1}{N}(Y - \hat{Y})\Sigma_\epsilon^{-1}(Y - \hat{Y})^T = \frac{1}{N}(Y - \hat{\theta}X)\Sigma_\epsilon^{-1}(Y - \hat{\theta}X)^T$$

- The unbiased estimator is:

$$\hat{\sigma}^2 = \frac{1}{N - rkX}||Y - \hat{\theta}X||^2_{\Sigma_\epsilon^{-1}} = \frac{1}{N - rkX}(Y - \hat{\theta}X)\Sigma_\epsilon^{-1}(Y - \hat{\theta}X)^T$$

The following Figure offers a geometrical vision of the situation. In it we can appreciate how the different elements interact. Represented in the $N - dimensional$ space of the samples we have that:

- Y is the set of values that we measured.
- $p_M(Y) = \hat{Y} = \hat{\theta}X$ is our estimated values of Y using a linear regresion of X with our estimated coefficients $\hat{\theta}$
- $E[Y] = \theta X$ is the ideal mean values of Y if we could compute the true θ. We cannot compute the true parameters since our estimation is based on noisy observations, there will always be an uncertainty.

We define the squared length of the regression of our model $\mathcal{M}$ as the squared distance between the measured Y and the estimated one $\hat{Y}$ over the assumed inverse covariance matrix of the noise:

$$SS_{res}(\mathcal{M}) = ||Y - \hat{Y}||^2_{\Sigma_\epsilon^{-1}} = (Y - \hat{Y})\Sigma_\epsilon^{-1}(Y - \hat{Y})^T = (Y - \hat{\theta}X)\Sigma_\epsilon^{-1}(Y - \hat{\theta}X)^T$$

We can see how it is pretty straight forward to compute the ML estimator and unibiased estimator of σ from this quantity as:

$$\tilde{\sigma}^2 = \frac{1}{N} SS_{res}(\mathcal{M}) \qquad \hat{\sigma}^2 = \frac{1}{N - rkX} SS_{res}(\mathcal{M})$$

Important. The sum of residuals is equal to 0, also why the estimation $\hat{Y}$ and the noise ϵ are uncorrelated ?

$$Cov(\hat{Y}, \epsilon) = 0$$

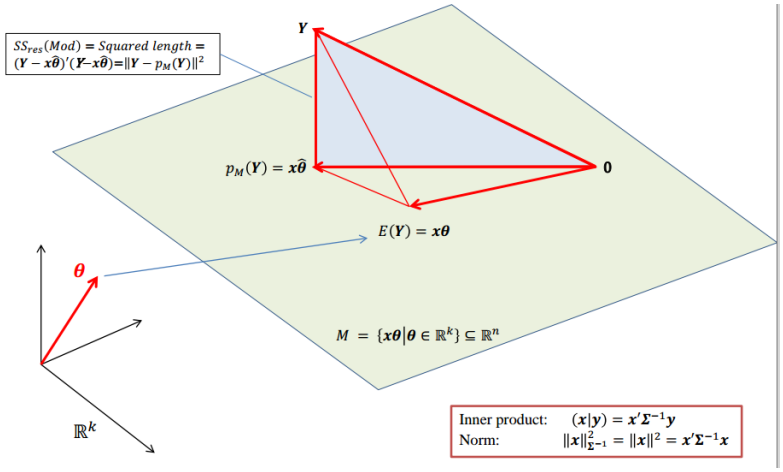


Figure 48: SMA and its window

It should be noted as well that the estimated vector $\hat{Y} = \hat{\theta}X$ and the residual vector $\epsilon = Y - \hat{\theta}X$ are perpendicular so we have that $||\hat{\theta}X||^2 = ||Y\hat{\theta}X||$.

$$||\epsilon||^2 = SS_{res}(\mathcal{M}) = ||Y - \hat{\theta}X||^2 = ||Y||^2 - ||\hat{\theta}X||^2$$

In general we make assumptions on Σ_ϵ being the simplest one that all the samples of the noise are i.i.d. In this case, the covariance matrix has the form $\Sigma_\epsilon = \sigma^2 \mathcal{I}$ and the maximum likelihood estimator of the linear coefficients is:

$$\hat{\theta}^T = (XX^T)^{-1} XY^T$$

Notice how in this case, the assumed σ_ϵ of the noise is not present since it is canceled out by the terms. Therefore, if we assume that the samples are i.i.d, the assumed noise of the samples does not change the estimator. What changes is the variance of the estimator, the more noise σ_ϵ we assume, the more variance

9.5 Test for the individual parameters obtained

Once we found our estimated $\hat{\theta}$, we can finally use it for regression, but we could ask ourselves if all of the values in $\hat{theta}$ are significant. After all, the estimated values $\hat{theta}_i$ have some variance and maybe some of them could have been obtained due to noise. It would be desirable to make the model as simpler as possible, so if we could delete variables, we always should.

We know that our estimated parameters $\hat{\theta}$ follow a Multivariate Gaussian distribution given by the equation:

$$\hat{\theta} \sim N(\theta, \sigma^2 (X \Sigma_\epsilon X^T)^{-1})$$

The estimated covariance matrix of the ML estimator $\hat{\theta}$ is computed as:

$$\hat{D}(\hat{\theta}) = \hat{\sigma}^2 (X \Sigma_\epsilon X^T)^{-1} = \frac{SS_{res}(\mathcal{M})}{N - rkX} (X \Sigma_\epsilon X^T)^{-1} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1D} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2D} \\ \vdots & \vdots & & \vdots \\ \hat{\sigma}_{D1} & \hat{\sigma}_{D2} & \cdots & \hat{\sigma}_{DD} \end{bmatrix}$$

Where $\hat{\sigma}_{ij}$ is the estimated covariance between the i-th and j-th parameter. Notice we use the notation $\hat{D}(\hat{\theta})$ to indicate that it is the estimated covariance matrix. If we had the true value σ^2 instead of

the unbiased estimator $\hat{\sigma}^2$ then we could have the true covariance matrix. Needless to say that the variance-covariance matrix $\hat{D}(\hat{\theta})$ will have dimensions DxD when we do not include the bias term and $(D+1)(D+1)$ when we do.

So we can just perform **statistical tests over the individual coefficients** θ_i to test weather or not the estimated value we obtained $\hat{\theta}_i$ is significantly different to 0, given the estimated variance of the parameter, σ_{ii}. The statistical test is the double sided test of the hypothesis that the true parameter value is 0.

$$H_0 : \theta_i = 0 \qquad H_A : \theta_i \neq 0$$

where the estimated variance is the one obtained from the $\hat{D}(\hat{\theta})$ matrix. For the $i-th$ component we use $\hat{\sigma}_{ii}$ as the estimated variance. We therefore can compute the statistic as:

$$T = f_T(\mathcal{D}|\theta_i = 0) = \frac{(\hat{\theta}_i - 0)}{\hat{\sigma}_{ii}} \sim t(N-1)$$

And finally the p-value for the statistical test can be computed as:

$$p - value = 2 \cdot P(t > |t_{\mathcal{D}}||\theta_i = 0) = 2(1 - cdf(|t_{\mathcal{D}}|))$$

EXAMPLE

We could make for complicated tests about a subset of θ, we just obtain the covariance matrix of the subset and apply the Multivariate Gaussian statistical test equations.

9.6 Test for lower dimensions of the parameter space

It could happen that a subset of the parameters are considered not significant and we would like to get rid of them, since they offer little predictive power and their value is actually noisy, not reliable.

By removing these parameters, we obtain a different model θ_H that is a subset of the previous one. If the previous model consisted of the estimated parameters $\hat{\theta} \in \mathcal{R}^D$, and now we remove a subset of dimensions, the new estimated parameters $\hat{\theta}_H \in \mathcal{R}^r$, being $r < D$. Each set of estimated parameters has its corresponding projected values $\hat{Y}$ which are out estimation of the regressed variable Y. We will call this vectors $p_M(Y)$ and $p_H(Y)$ for the projections $\hat{\theta}$ and $\hat{\theta}_H$ respectively

$$p_M(Y) = \hat{\theta}X \qquad p_H(H) = \hat{\theta}_H X$$

Notice that it is not the same to:

- Just erase the the parameters obtained
- Reobtain the parameters in the lowe dimensional case

In both cases, the new set of parameters obtained $\hat{\theta}_H$ will have less expressivity than the original one $hat\theta$, it this regard, the new $\hat{\theta}_H$ is a subset of $\hat{\theta}$, all the possible values that $\hat{\theta}_H$ could take are contained in all the possible values that $\hat{\theta}$ can take, and the later one has more possible values.

In mathematical notation we have that the original vector is D-dimensional, so $\theta \in \mathcal{R}^D$, and the new reduced vector is r-dimensional, so $\theta_H \in \mathcal{R}^r$, with $r < D$. Therefore in the space $\theta_H \in \theta$. And therefore all the possible values $Y = \theta X$ that we can explain is bigger when using θ than when usnig θ_H.

If we call M all the possible values that $\hat{Y}$ can take when we use θ; and H all the possible values that Y can take when using θ_H:

$$M = \{x\theta|\theta \in \mathcal{R}^D\} \qquad H = \{x\theta|\theta \in \mathcal{R}^r\}$$

We have that H is a subset of the possible values that M can take, that is $H \in M$. Therefore the estimator for θ_H will have less possibilities to get close to the set of values we want to estimate Y.

The new projection vector θ_H will not explain the data Y as good as the original θ, and therefore its residual $SS_{res}(H)$ will be bigger (or at least equal) to the residual of the previous model $SS_{res}(M)$.

$$SS_{res}(\mathcal{H}) = (Y - \hat{\theta}_H X)\Sigma_{\epsilon}^{-1}(Y - \hat{\theta}_H X)^T > SS_{res}(\mathcal{M})$$

The following Figure illustrated this fact geometrically where we can see how $p_H(Y)$ will always be further away from the obtained values Y than $p_M(Y)$. We can also see how the projection XXXXX everything is perpendicular and we can apply pytagoras. Why ?

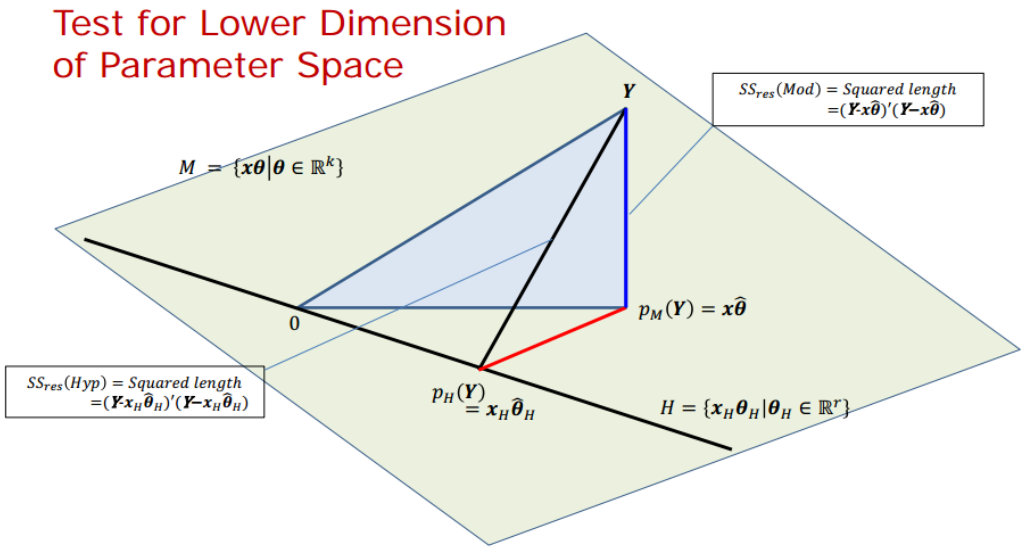


Figure 49: SMA and its window

For the statistical test we want to answer the question: Is the increase in error of the new parameters $\hat{theta}_H$ significant compared to the previous error level ? Maybe the increase in error is small compared to the one we already had and since we prefer a simpler model, we are willing to remove the variables.

In hypothesis notation, our null hypothesis will be that the real mean of Y, that is $E(Y) = \theta Y$ lies in the set of points that can be reached with H, that is,the actual parameters θ lie in the r-dimensional space. The alternative hypothesis is that they lie in a region of the D-dimensional space that is not contained by H.

$$H_0 : \theta \in \mathcal{R}^r \qquad H_A : \theta \in \mathcal{R}^D \ R^r$$

The statistic obtained for this test is the F test statistic and it is the variance increase in the residual from choosing H_0 over H_A divided by the variance of XXX, both normalized by their degrees of freedom.

$$F = \frac{(SS_{res}(\mathcal{H}) - SS_{res}(\mathcal{M}))/(D-r)}{SS_{res}(\mathcal{M})/(N-D)} \sim F(D-r, N-r)$$

Open question to solve. Is it that it actually comes from the equation of relating the variance of the remaining residual. Calling $\epsilon_H = Y - p_H(Y)$ the residuals of using $\hat{\theta}_H$ and ϵ_M the residuals of using θ then we have that the distance between the 2 is:

$$||\epsilon_H - \epsilon_M||^2 = ||(Y - p_H(Y)) - (Y - p_M(Y))||^2 = ||p_M(Y) - p_H(Y)||^2$$

If the residuals of the models are perpendicular then we have that:

$$||\epsilon_H - \epsilon_M||^2 = ||\epsilon_H||^2 - ||\epsilon_M||^2 = ||p_M(Y)||^2 - ||p_H(Y)||^2 = SS_{res}(\mathcal{H}) - SS_{res}(\mathcal{M})$$

And we arrive to the previous equation which is equal to:

$$F = \frac{||\epsilon_H - \epsilon_M||^2/(D-r)}{||\epsilon_M||^2/(N-D)} = \frac{(SS_{res}(\mathcal{H}) - SS_{res}(\mathcal{M}))/(D-r)}{SS_{res}(\mathcal{M})/(N-D)}$$

when we consider the residuals to be perpendicular, and therefore independent ? Is this something that we are assuming, imposing, or it is like that always. We do not consider the estimations from the model $p_H(Y)$ and $p_M(Y)$ independent but we consider the residuals $p_H(Y)-Y$ and $p_M(Y)-Y$ to be independent ? No I think we consider all independent. Is this because the vector $\hat{\theta}_H$ is the same as $\hat{\theta}$ but will less dimensions ? Will this independence hold for any other vector θ_H that is not obtained as a subset of values of θ. Each of the forms is computed different in the computer, and if we already had compted the residual sum of squares then using $(SS_{res}(\mathcal{H}) - SS_{res}(\mathcal{M})$ is preffered.

In this case we have right-sided statistical test where we can reject the null hypothesis H_0, that a lower number of dimensions still explains the data good enough, if the value of F is too big, that is, if the increase in prediction error by using less dimensions is too big.

$$p - value = P(F > F_{\mathcal{D}}) = 1 - cdf(F_{\mathcal{D}})$$

In order to accept of reject the hypothesis we will stablish a limit on F and if F is bigger than than that then, we reject that we can use smaller number of parameters.

9.7 Confidence interval of a prediction

Given a new sample $X^* = [X_1^*, X_2^*, ..., X_D^*]$, we can use our model M to compute its estimated regressed value $\hat{Y}^*$. Given a trained model M with estimated coefficients $\hat{\theta}$, then the estimated value of the regressed variable is $\hat{Y}^* = \hat{\theta}Z$. Since our estimated parameters $\hat{\theta}$ follow a Multivariate Gaussian distribution:

$$\hat{\theta} \sim N(\theta, \sigma^2(X\Sigma_\epsilon X^T)^{-1})$$

Then so does our estimate $\hat{Y}^*$ which is just a linear projection of the multivariate Gaussian, given by the vector X^*.

$$\hat{Y}^* = \hat{\theta}X^* \sim \mathcal{N}(\theta X^*, \sigma^2 X^{*T}(X\Sigma_\epsilon^{-1}X^T)^{-1}X^*)$$

So we have that the mean and variance of the prediction are:

$$E[\hat{Y}^*] = \theta X^* \qquad V[\hat{Y}^*] = X^{*T}V[\hat{\theta}]X^* = \sigma^2 X^{*T}(X\Sigma_\epsilon^{-1}X^T)^{-1}X^*$$

If we do not have the true variance σ^2, but we use the unbiased estimator instead, $\hat{\sigma}^2$, then the normalized estimation follows **the t-student distribution with** $(N - D)$ **degrees of freedom**. We can therefore find the confidence interval of the estimation $C = [\hat{Y}^*_{\alpha/2}, \hat{Y}^*_{(1-\alpha/2)}]$ given by:

$$\hat{Y}^*_{\alpha/2} = E[\hat{Y}^*] - t_{\alpha/2} \cdot \sqrt{\hat{V}[\hat{Y}^*]}$$

$$\hat{Y}^*_{(1-\alpha/2)} = E[\hat{Y}^*] + t_{\alpha/2} \cdot \sqrt{\hat{V}[\hat{Y}^*]}$$

Where we express the estimated standard deviation $\sqrt{\hat{V}[\hat{Y}^*]}$ as:

$$\sqrt{\hat{V}[\hat{Y}^*]} = \hat{\sigma}\sqrt{c} \qquad c = X^{*T}(X\Sigma_\epsilon^{-1}X^T)^{-1}X^*$$

This is the confidence interval of the estimation $\hat{Y}^*$. If we want to compute the confidence interval of the observation, we can do it from the formula of the observation Y^* which is equal to the estimation $\hat{Y}^*$ plus the noise of the residual of that sample ϵ_{X^*}:

$$Y^* = \hat{Y^*} + \epsilon_{X^*} \qquad V[Y^*] = V[\hat{Y^*}] + \sigma^2_{X^*}$$

Therefore the variance of the observation is bigger then the varince of the estimatio since we have the observation noise ϵ_{X^*}. It is up to our model, or our personal belief what is the value of the variance of the sample $\sigma^2_{X^*}$. In the case that we assume that all the samples have the same variance. This is equal to σ so we have:

$$V[Y^*] = V[\hat{Y}] + V[\epsilon_{X^*}] = \sigma^2[1 + X^{*T}(X\Sigma_\epsilon^{-1}X^T)^{-1}X^*]$$

FOR A SET OF SAMPLES... if they are correlated. We will have the sample covariance matrix XX. And we have that the estimation belongs to a multivariate Gaussian.

9.8 The Hat Matrix

The hat matrix H a NxN matri xobtained from a set of samples aggreegated in the matrix $X_{(DxN)}$ when we perform the following equation:

$$H = X^T(XX^T)^{-1}X$$

This matrix as we will see appears naturally in many equations of the GLM. It has a set very good qualities.

- It is independent on the choice of the generalized inverse $(XX^T)^-1$

- It is idempotent $H = HH$
- It is symmetric $H = H^T$
- Its rank is equal to its trace, and equal to the rank of X. $rk(H) = tr(H) = rk(X)$

We also define the matrix M as the identity matrix of N dimensions minus H.

$$M = I - H = I - X^T(XX^T)^{-1}X$$

M is also idempotent and it can be easily seen that M projects on the orthogonal complement:

$$MH = IH - HH = H - H = 0$$

Next we will see the appearance of the H matrix in several equations when we consider that the sample noises are i.i.d then we have that $\Sigma_\epsilon = \sigma_\epsilon \mathcal{I}$. Under these assumptions the estimated parameters and the distribution of the real parameters are:

$$\hat{\theta}^T = \Big[(XX^T)^{-1}X\Big]Y^T$$

$$\hat{\theta} \sim N(\theta, \sigma^2(XX^T)^{-1})$$

We can **express the different error terms from the H and M matrices**. Given a model M with its associated parameters $\hat{\theta}$, there are 3 Squared Sum terms to consider:

- The total variation:

$$SS_{Tot}(Y) = ||Y||^2 = YY^T = \sum_{i=1}^{N} Y_i^2$$

  This is the initial variation of our variable to regress Y. If we can consider it Gaussian, it would be the uncertainty about Y, the more variance, the more uncertainty. We will try to use our model to explain this variability using the variables X. This value has N degrees of freedom.

  Since Y might have a mean that is non-zero, then the mean of Y is contributing to this variation and it shouldnt, we can correct it by just substracting the sample mean $\bar{Y}$.

$$SS_{Tot}(Y) = ||Y||^2 - N\bar{Y}^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

  This value has $N - 1$ degrees of freedom since now we substract the mean.

- The variation of the model

$$SS(M) = ||p_M(Y)||^2 = (\hat{\theta}X)(\hat{\theta}X)^T = YHY^T$$

  It is how much variation of the original Y our model is explaining (removing). This quantity has $rk(X)$ degrees of freedom. If we removed the mean from Y then it has $rk(X)-1$ degrees of freedom.

- The variation of the residual

$$SS_{res}(\mathcal{M}) = ||Y - p_M Y||^{=}(Y - \hat{\theta}X)(Y - \hat{\theta}X)^T = YMY^T$$

  This quantity has $N - rk(X)$ degrees of freedom.

As we know, the 3 variations are related being:

$$SS_{Tot}(Y) = SS_{res}(\mathcal{M}) + SS(M)$$

In short, our model explains certain variance of the orgininal source Y, being this variance independent of the residual variance.

Also, it should be noted that the estimated values of the training samples $\hat{Y} = \hat{\theta}X$ are computed from the estimated parameters $\hat{\theta}$ so they also follow a Multivariate distribution, as we saw with the predicted samples. The multivariate distribution has variance:

$$D(\hat{Y}) = D(\hat{\theta}X) = \sigma^2 X^T (XX^T)^{-1} X = \sigma^2 H$$

And in the same way, the distribution of the error terms is:

$$D(\epsilon) = D(Y - \hat{\theta}X) = \sigma^2 [1 - X^T (XX^T)^{-1} X] = \sigma^2 M$$

Notice from these equations that the uncertainty of each individual estimation $\hat{Y}_i = \hat{\theta}X_i$ is given by the diagonal elements of the covariance matrix of the estimation $D(\hat{Y})$ being:

$$D(\hat{Y}) = \sigma^2 \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1N} \\ h_{21} & h_{22} & \cdots & h_{2N} \\ \vdots & \vdots & & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{NN} \end{bmatrix}$$

The uncertainty about the regressed training data points X_i is thus given by $V[\hat{Y}_i] = \sigma_i^2 = \sigma^2 h_{ii}$ where σ^2 is the total error of the regression and it is the same for all samples. The value h_{ii} **is called the leverage** and it gives us an idea of the relative uncertainty in the prediction of the sample X_i.

If a sample X_i lies in an area of the space close to $\hat{\theta}$, then its uncertainty will be lower and it will have a lower value of h_{ii}. If the sample lies in a zone of the space where we barely have samples, an area where we didnt have data to learn from, and the vector $\hat{\theta}$ was not influenced much by that area, then the prediction will have high uncertainty and a higher value of h_{ii}. The next Figure shows this situation.

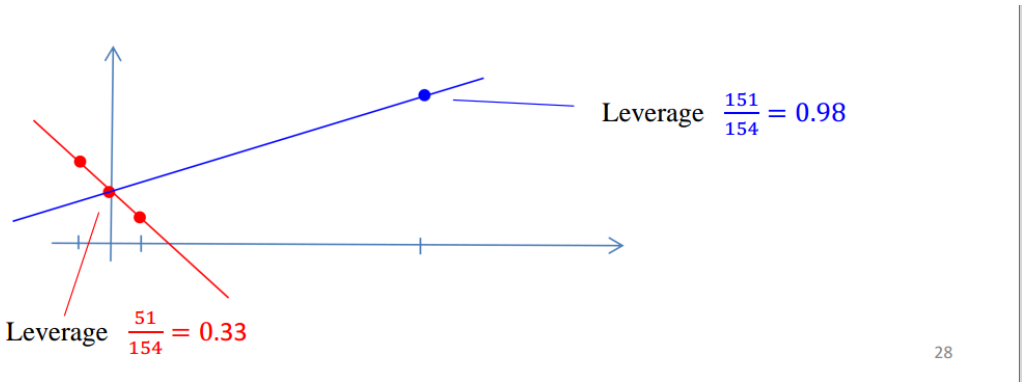


Figure 50: SMA and its window

Samples with a high leverage h_{ii}:

- Are samples that are far away from the center of the big cluster of samples
- They have a big influence in the computation of the parameters $\hat{\theta}$ since these are made in a way to minimize the squared distance of each sample the hyperplane defined by the parameters.
- They are usually outliers, samples that do not really belong to the distribution we are estimating, they are just adding uncertainty and deviating our estimated parameters. Data points with $h_{ii} > 2D/N$ should be investigated.

Notice as well that the variance of the residual of the i-th sample is computed as $V[\hat{\epsilon}_i] = \sigma^2(1 - h_{ii})$. From this we can see that h_{ii} is always less than 1. The variance of each prediction and the variance of its noise are complementary. They are equal to the assumed variance of each independent sample.

$$V[\hat{Y}_i] + V[\epsilon_i] = \sigma^2$$

WILD GUESS:
If our prior has a different covariance matrix for the error, we can create the modifified data $Z = \Sigma_\epsilon^{-1} X$ and then everything applies ?

9.9 Outliers tests

9.10 R square value and Multicolinearity

R-square - the coefficient of determination in a regression model - measures the proportion of variability in the response that is explained by the regressor variables. In a linear regression model with intercept, it is defined as

$$R^2 = 1 - \frac{SS_{res}(M)}{SS_{Tot}}$$

The adjusted R-square has been adjusted for degrees of freedom. It is calculated as

$$\bar{R}^2 = 1 - \frac{(N-i)(1-R^2)}{N-D}$$

where i is equal to 1 if there is an intercept and 0 otherwise, n is the number of observations used to fit the model, and p is the number of parameters in the model including a possible intercept.
Explain relation with correlation and multicolinearity

10 Using Discrete Variables

So far, almost all the variables we have seen were continuous, and they could take any value in the real domain $\mathcal{R}$, from -infinity to infinity, of course only with the degree of precision and range of numbers that our computer would allow us to store, but lets assume that the precision is accurate enough to not having to take this effect into account. In this Section we will different types of variables that exist and they properties.

Different types of variables will require different types of analysis and the way we obtain information from them and combine them with other variables can be different. We can initially divide the variables in two types: The continuous variables and the discrete variables.

A variable X is continuous if it can take values in the real domain $\mathcal{R}$ or a subset of it, like for example if X is a waiting time, it will always by bigger or equal than 0. So the values that it can take form an interval:

$$X \in \mathcal{R}$$

Also, the values of the domain of X are **ordered**, that means that the value 1.5 is more similar to 2.0 than to 3.0. There is a define sequence for the countable or finite, uncountable. There are many examples of continuous variables in nature, as it could be temperature, mass, time, velocity... In finance we also find continuous variables such as returns, price, interest...

The next figure shows an example of price and return as continuous variables. Notice that we represent them different and that time has a contnuity and smoothness that we do not assume for the return. But both variables are continous since they can take values in the real domain and the values have an order.
TODO: Change graph.

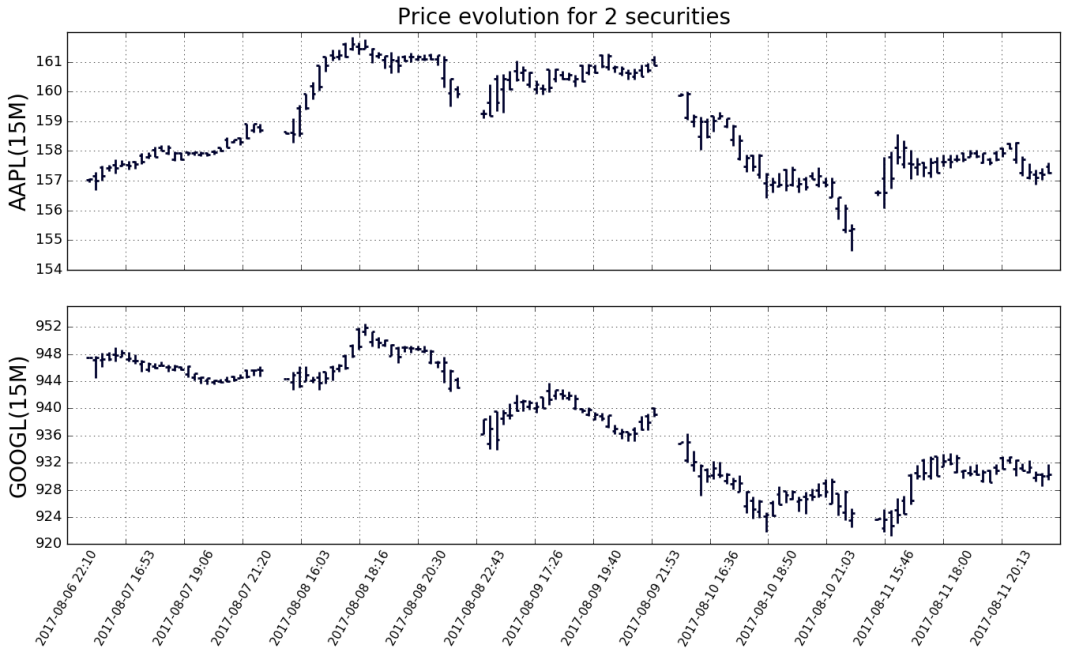


Figure 51: SMA and its window

A categorical variable F, is a variable that can only take a specific value among a set of C discrete values. These are the the domain of the variable F.

$$F \in \mathcal{C} = [C_1, C_2, .., C_C]$$

Categorial variables can be divided into 2 important types:

- **Nominal**: The discrete values that the variable can take do **not have an intrinsic order**. If we have class C_1 is not more similar to C_2 than to C_{10}. We could permutate the number we give to each variable and it should be alright. Examples of these variables could be the country of a person, the gender, a specific medical treatment... In finance we also have Nomical categorical variables, such as the Sector of a Company or the type of an instrument (Bond, Asset, Fund)...
- **Ordinal**: The discrete values have an **not have an intrinsic order or rank**. A common example is any kind of ranking or poll where you can grade you level of satisfaction. In this case the variables themselves have a possible order. In finance we also have these kind of variables such as the Rating of a Bond (AA, AA+, BBB, B) or even just the ranking of the instruments within an Index.

Categorial variables F can be useful on their own, as for example we can compute the number of people in a country by knowing the value of the "country" variable of each person. Also, if a process if outputing a discrete variable over time, like for example the state of XXXX. Then we can also find patterns in that: Examples
This is useful information but it is pretty limited, just like when continuous variables we gained information such as "correlation" when combining them, **categorical variables are most informative when combined with other continuous and discrete variables**.

For example, if our discrete variable is the Country of a person, we would like to maybe see if there is a difference in the XXX

Of course the different classes could be related to each other in some way, according to a similarity measure given by another variables, like for example probably X1 and X2 are more close in XXX than. But this similarity measure, this order, this clustering, depends on other variables. The discrete variables are only that discrete. And we will not associate any property to their "tag".

Now lets get into more shady areas, so, now we know that categorial values are just that and if we have a set of samples $\mathcal{D}$ with just this property then we can obtain some information. XXX. Lets relate it with another one.

Ambiguities in classifying a type of variable In some cases, the measurement scale for data is ordinal, but the variable is treated as continuous. For example, a Likert scale that contains five values - strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree - is ordinal. However,

where a Likert scale contains seven or more value - strongly agree, moderately agree, agree, neither agree nor disagree, disagree, moderately disagree, and strongly disagree - the underlying scale is sometimes treated as continuous (although where you should do this is a cause of great dispute).
It is worth noting that how we categorise variables is somewhat of a choice. Whilst we categorised gender as a dichotomous variable (you are either male or female), social scientists may disagree with this, arguing that gender is a more complex variable involving more than two distinctions, but also including measurement levels like genderqueer, intersex and transgender. At the same time, some researchers would argue that a Likert scale, even with seven values, should never be treated as a continuous variable.

Use the example of a dice, if the question is that XXX.

10.1 Examples of modelling variables in Finance

In finance, categorical variables also ocurr naturally as it is the case of:

- The instrument: Each security is itself an element of the security class.
- Security tyepe: It could for example be Bond, Asset, Fund, Emerging Market...
- Sector: Healthcare, Financials.

When the elements should be treated as Factors or just as independent random variables ? well it depends on the assumtions that we make, which of course can be measure if the assumtions are true. For example, if we have the returns of several C companies, we could model it either:

- Forming and D-dimensional vector of continuous samples $\mathcal{D} = [X_1, X_2, ..., X_D]$. Notice that in this case we will need each time to have the aligned values of each variable. This is how we have seen the data so far in the Linear Regression model. If we have N samples then we will have NxD total values.
- Just creating the variable pair composed by the return and the company $\mathcal{D} = [X, F]$. In this case we just have a 2-dimensional vector of NxD samples. The advantage of this model is that now we do not need to have all the securities all the time. We might lose information if the data were related or a timeseries existed.

Both approaches have advantages and disadvantages and the one to choose depends on our assumptions on the data, the data available (missing data, little data) and so on.
A given variable Y will have a different distribution for the different categories. For example in our case we will have:

- Y: The return of all the 15 Min samples of the week.
- C: The day of return.

We can see that for each day, Y_j has a different set of points, this could be either because of noise or because of.
If we are in an upper trend in the market, there would be a correlation between the days, but at 15 min level there shouldnt ?,
Example of violation:
What if we considered as well the company as a Factor model ? In this case XX. The problem is that we are assuming that the samples are independent, and in this case we can see that they are not.

10.2 Using Continuous variables to learn about Y

Given a set of measurement of a continuous random variable Y, $\mathcal{D} = [Y_1, Y_2, ..., Y_N]$, we would like to characterize its statistical distribution, since knowing its distribution will allow us to predict the most likely future events to happen regarding the variable. In this case for example Y is the 15-Min returns of a given Asset, GOOGLE for example.

Since we do not have anymore any more information than Y we will use its values to estimate the distribution. As we previously saw in this section, initially we can consider that the distribution changes at each sample Y_i and the relation between each pair of samples could take any form that is allowable by the laws of statistics. But since we do not have enough data to estimate reliably these distributions we are forced to make assumptions. Which could be:

- **We constraint the shapes of the marginal distributions**. We assume that they belong to a family of distributions. This family could be anything, such as Gaussian, MoG, exponential... But this is already constraining the shape that the marginal variables can take, thus we need less data to characterize it.
- **We constraint the ways in which the marginal distribution changes over time**. Maybe we only allow the marginal distributions to change in a specific way, like for example the mean of the distribution decreases in a linear way according to a slope that we need to estimate.
- **We constraint the way in which the variables the variables depend on each other**. This is also related to the previous one.

In a common scenario we will assume that Y follows a Gaussian distribution and that the samples are identically distributed and independent. The problem is that many scenarios will not fullfill these requirements.

When we have other sources of information asociated to Y, such as continuous variables X, then the dataset that we have is $\mathcal{D} = [(Y_1, X_1), (Y_2, X_2), ..., (Y_N, X_N)]$. Ideally we would like to compute the joint distribution (Y_i, X_i), which also could change over time and XXX. This estimation problem is even harder that the previous one since now we need to estimate join probabilities.

But usually we are just interested in modeling the distribution of our variable Y given the rest of the variables, that is, we want to estimate the marginal distribution $f_i(Y_i|X_i)$ for all the samples, where $f_i()$ means that is the statistical distribution of the i-th sample and it could in principle change for all the samples. Also, if the samples are dependent of each other, we would like to estimate the joint conditional distribution:

$$f(Y_1, Y_2, ..., Y_N|X_1, X_2, ..., X_N)$$

Also, since the variables Y_i can also be related, we would like to know what is the distribution of a given Y_i given all the data available:

$$f_i(Y_i|X_i, (Y_1, X_1), (Y_2, X_2), ..., (Y_N, X_N))$$

These are the type of questions that we are usually interested in solving. And to do so, we will need to make assumption on the variables how these relate to each other. **The assumptions on the variables and its relations is called in a broad term, our model $\mathcal{M}$.**

When the model contains certain parameters that we need to estimate, such as the linear coefficients θ in Linear Regression. We can also make assumptions on the distribution of these parameters to estimate, in this case, our assumption is called **our prior over the parameters** $f(\theta)$ **.** These are the so-called Bayesian models.

In this case that we have extra information, we usually do not need to assume that Y is Gaussian anymore, which is a big constraint on the shape that Y can take, but **we assume that the distribution of** Y **given the other information** X **is Gaussian**.

$$f_i(Y_i|X_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

Where this statistical distribution $f_i(Y_i|X_i)$ will be estimated by our model $\mathcal{M}$, in notation sometimes we say:

$$f_i(Y_i|X_i, \mathcal{M})$$

Meaning the statistical distribution of Y_i given the rest of the information associated to the sample X_i, and the way in which they are related, which is given by out model $\mathcal{M}$.

For example in the case of GLM, when we the have continuous variables X, we have that the conditional distribution of the sample Y_i is given by:

$$f_i(Y_i|X_i) \sim \mathcal{N}(\mu_i, \sigma_i^2) \qquad \mu_i = \hat{\theta} X_i \quad \sigma_i^2 = \sigma^2[X^{*T}(X\Sigma_\epsilon^{-1}X^T)^{-1}X^*] + \sigma_{\epsilon_i}^2$$

And if the distribution of $f_i(Y_i|X_i)$ have less uncertainty than the distribution of Y_i then we can say that X_i explains part of the uncertainty of Y_i and therefore it can be used to improve our estimated of Y_i for that X_i.

In the continuous case, each vector X_i has its associated distribution $f_i(Y_i|X_i)$. In the case of GLM we performed a series of assumptions that yielded a closed form solution for the conditional distribution

$f_i(Y_i|X_i)$. In general, if we dont have this closed form solution, we could also generate several value pairs Y_i XXXX

10.3 Using Discrete variables to learn about Y

When instead of continuous variables X, we have discrete variables F, then our database consists of the samples $\mathcal{D} = [(Y_1, F_1), (Y_2, F_2), ..., (Y_N, F_N)]$. Just as before, in the most general case (Y, F) could have a joint distribution that changes over time in any form. In this case for the distribution of the discrete variable F, the probability of each of its elements could change with time. And in this case the relation between the continuous Y and the discrete F is:

$$f(Y_i, F_i) = f(Y_i|F_i)P(F_i)$$

Where now:

- $P(F_i)$ is the probability of having observed that class, given the model.
- $f(Y_i|F_i)$ is the probability of having observed the value Y_i when the value of the variable F is F_i.

Remember that we are always implicitly conditioning on the model $\mathcal{M}$ since out estimation of the distributions is done with the model and its parameters.

Our goal is the same as before, we would like to be able to reduce the uncertainty about the possible values of Y_i when we know the other information, in this case, the other information is the class from which the variable came from. If the distribution of Y_i is the same for all classes F_i, we are not reducind the uncertainty of Y_i by knowing F_i. This is true even if the classes themselves have a pattern that we can learn, like a sequential pattern or the probability of the classes is very different.

What we want to investigate is if maybe, the distribution of Y is dependent on the class of C_j from which it was drawn. The statistical distribution of Y can be seen as a marginalization of the distributions of Y_j.

$$f(Y) = \sum_{j=1}^{nC} f(Y/c_j)P(c_j)$$

If we know from which conditional distribution was drawn each sample Y_i, by knowing its associated variable C_i then we can gain information about Y_i, that is, the uncertainty (variance for just gaussian variables) is reduced. Also notice that any momentum over Y can be obtained as the expected momentum over the marginalization of its variables, that is:

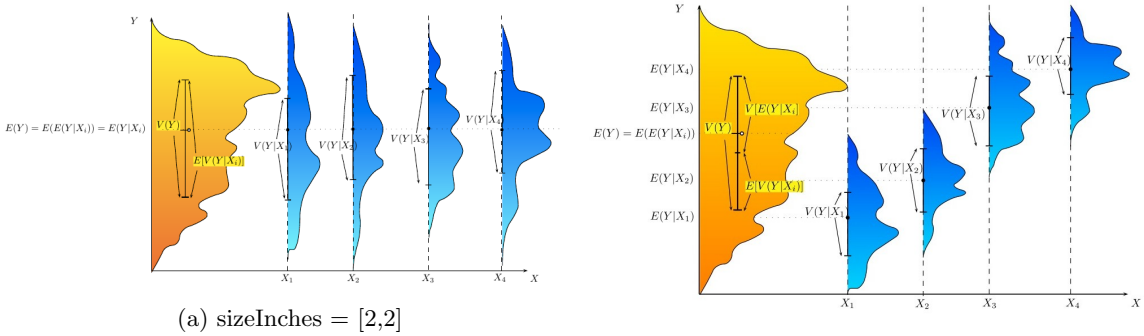
$$\begin{aligned} E_Y[Y] &= \int_{-\infty}^{\infty} y \cdot f(y)dy = \int_{-\infty}^{\infty} \left[\sum_{j=1}^{nC} y \cdot f(y/c_j)P(c_j) \right] dy \\ &= \sum_{j=1}^{nC} \left[\int_{-\infty}^{\infty} y \cdot f(y/c_j)dy \right] P(c_j) = \sum_{j=1}^{nC} E[Y|c_j]P(c_j) \\ &= E_F[E_Y[Y|X]] \end{aligned} \tag{7}$$

We can compute any momentum of Y by computing the momentums over the conditional distributions and add them together with the probabilities as weights. As we will see, different statistical tests will look for if any of the marginal momentums $E[Y|c_j]$ is different from the total momentum $E[Y]$.

The next Figure shows a generalistic case, where we have the variable Y with a given cumbersome distribution, and then we have that each sample Y_i has an associated class F_i. As we can see from both charts:

- The distribution of Y is a marginalization of the individual distributions.
- The momentums of Y, in this case the mean is represented, are weighted momentums of the individual conditional distributions.
- In the left chart, all the conditional distributions $f(Y|c_j)$ have same mean, therefore knowing from which distribution a sample came from does not give us information about the mean of the asociated Y_i.

- In the right chart, each class has a different mean. Therefore the variables have discriminative power with respect to Y. In a normal scenario we also will have to take into account if the difference of mean is statistically signifinicant with respect to the variance.



(a) sizeInches = [2,2]

(b) sizeInches = [10,6]

Figure 52: Effect of the figure size in the saved figures

This is making already two assumtions:

- Samples are independent
- The distributions are independent. Meaning that the distributions are not related ? Y_i, Y_j are not related. This is similar with the LDA. Put in a statistical way.

IF If we assume that the samples depend on each other, then
Counter Examples: - HMM for non-independnet samples
If we assumed for example HMM then equation of it !!

In the general case, we will not have a so defined model and we have an algorithm that in general just will tell us, for each variable. Without has having to marginalize. If our estimations have less uncertainty than not using the models then we can use the variables to gain some knowledge about Y.

$$f(Y|X_i, Datos, model) =$$

10.3.1 Coupled Variables

The previous general case, covers all possibilities but there are problems in which if we have extra information, it will make it easier. This is the case of couple variables. CASE OF STOCKS.

In the previous model it can be seen as just, we stack all the variables and then there is this clear lagged correlations. Then is it not better directly to model them as XXXX ?

The samples will be coupled if we draw two sets of variables at the same time and then there is a statistical relation between them.

- Correlated variables in the clases. Then the distribution is more like a correlated gaussian over the pairs Y_i, Y_j, better solvable with a GLM. But since many correlations are unlikely and we might not have enough samples or missing couples (only one value) then we use the assumptions of ANOVA to try to learn something. If the statistical distribution of Y does not change.

TREAT THE PAIRED VARIABLES THING !!

10.4 ANOVA test

ANOVA is a statistical technique that stands for Analysis of Variance. It is an statistical procedure in which we have an independent variable Y, which is continuous and one-dimensional $Y \in \mathcal{R}$ and we have a set of dependent discrete variables F. Its goal is to detect possible differences in the mean of Y for the different classes.

These kind of analysis arises naturally in many studies:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

Lets start with the simplest case, which is called the **one sided ANOVA.** In this case, we have the variable Y, which has some uncertainty, a variance; and we have a categorical variable F. The database is formed by a set of samples $\mathcal{D} = [(Y_1, F_1), (Y_2, F_2), ..., (Y_N, F_N)]$. Usually these are represented in the form of a table:

$$TableANOVa$$

Where the columns are the different experiments
Like in other statistical test, this model will make assumptions. The following is the list of assumptions from the less restrictive to the most restrictive:

- First of all it assumes that the samples are not coupled, so samples from the different classes are not related.

  $$COV(Y_i, Y_j) = 0$$

  This implies that we can have different number of samples for each class. Each sample of a given class does not have its homologous sample from the other classes.

- Second of all, it assumes that all the samples are independent:

  $$COV(Y_i, Y_i) = 0$$

  This implies that the likelihood of a set of samples is just the product of their individual likelihood. Develop this equation, reasoning of the incomplete and the complete likelihood if we know the samples.

  $$f(Y|D) = \prod_{i=1}^{N} f(y_i|c_j)$$

- It assumes that the conditional distribution for each class of the factor is a gaussian variable.

  $$Y|C \sim \mathcal{N}(\mu_j, \sigma_j^2) \qquad f(y|c_j) = \frac{1}{\sqrt{(2\pi)\sigma_j^2}} exp\left[-\frac{1}{2\sigma_j^2}(y-\mu_j)^2\right]$$

- Furthermore, it assumes that the variance of all the classes is the same:

  $$V[Y|C_1] = V[Y|C_2] = \cdots = V[Y|C_{nC}]$$

One could see already the similarity between this set up and the Linear Discriminant Analysis. Here, we are also interested if there is a significant difference between the means of the XXX.

The next Figure shows an example with 3 days of trading ! Our goal is to see if different days of the week have different mean.
CHART

In order to compute the individual means $\mu_1, .., \mu_{nC}$ and variance of all samples σ^2 we can use all the samples. For computing the mean of each class, we use the samples belonging to the class, and the variance is computing using all samples ?

$$\hat{\mu}_j = \hat{E}[Y|c_j] = \frac{1}{Nc_j} \sum_{i=1}^{Nc_j} Y_{ij}$$

$$\hat{\sigma_j^2} = \frac{1}{N}$$

10.5 GLM formulation of ANOVA

We can then formulate the ANOVA model by means of a Linear Regression using the discrete labels. In this model, the estimated value for a sample $Y_{i,j}$ that comes from the class c_j is simply the mean of the class $\hat{Y}_{i,j} = \mu_j$. And the sample obtained is $Y_{i,j}$ modeled as the estimation plus some noise of the sample ϵ_i

$$Y_{i,j} = \mu_j + \epsilon_i$$

In this way we can model each discrete sample as a C dimensional vector of zeros which has a one in the position of the class it belongs to. This will be our representation of the samples. And then the parameters θ of the model would be the vector of computed means μ.

$$Y = \theta X + \epsilon = \mu F + \epsilon$$

Our estimator $\hat{Y}$ for a given sample that belongs to the second class would be:

$$\hat{Y} = \mu F = \begin{bmatrix} \mu_0 & \mu_0 & \cdots & \mu_C \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Notice that in this reformulation, the number of classes of the factor C is like the number of dimensions of the variables that we had in the purely continuous GLM model, noted previously as D.

I guess there is an easy extension of this in the case that we have noisy samples of F meaning that we do not know certainly form which distribution it came from but we have a probability distribution over the possible classes it was generated from, then we can just estimate the value by putting that distribution as our sample, in that case:

$$\hat{Y} = \mu F = \begin{bmatrix} \mu_0 & \mu_0 & \cdots & \mu_C \end{bmatrix} \begin{bmatrix} P(c_i = 1) \\ P(c_i = 2) \\ \vdots \\ P(c_i = C) \end{bmatrix}$$

This is pretty related to the EM for Gaussian mixtures probably since Y is actually a Gaussian mixture

10.6 Decomposition of error

As in the GLM, we can decompose the total error of not having a model and just predicting the mean of Y all the time.

$$SS_{Tot}(Y) = (N-1) \cdot \hat{V}[Y] = \sum_{i=1}^{Nc} \sum_{j=1}^{Nc_j} (Y_{i,j} - \bar{Y})^2$$

Having the error using the model being:

$$SS_{res}(Y) = \sum_{i=1}^{Nc} \sum_{j=1}^{Nc_j} (Y_{i,j} - \bar{Y}_j)^2$$

In terms of the sum of squares error. Being:

$$\hat{Y} = \frac{1}{Nc} \sum_{i=1}^{Nc} \frac{1}{Nc_j} \sum_{j=1}^{Nc_j} Y_{i,j} = \frac{1}{Nc} \sum_{j=1}^{Nc} \bar{Y}_j$$

10.7 Two-way analysis of variance

10.8 MANOVA formulation

In the multivariate GLM we have more than one variable Y to estimate basically, the main equation is still the same. Lets transpose the parameters and data dimensions and then we have, for M output variables and D input variables. The notation is:

$$\begin{bmatrix} Y_1 & \cdots & Y_M \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & \cdots & X_D \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1M} \\ \theta_{21} & \cdots & \theta_{2M} \\ \vdots & \vdots & \vdots \\ \theta_{D1} & \cdots & \theta_{DM} \end{bmatrix} + \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_M \end{bmatrix}$$

For a given dataset we will have the notation of the estimation:

$$\begin{bmatrix} Y_{11} & \cdots & Y_{M1} \\ Y_{12} & \cdots & Y_{M2} \\ \vdots & \vdots & \vdots \\ Y_{1N} & \cdots & Y_{MN} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1M} \\ \theta_{21} & \cdots & \theta_{2M} \\ \vdots & \vdots & \vdots \\ \theta_{D1} & \cdots & \theta_{DM} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} & \cdots & \epsilon_{1M} \\ \epsilon_{21} & \cdots & \epsilon_{2M} \\ \vdots & \vdots & \vdots \\ \epsilon_{N1} & \cdots & \epsilon_{NM} \end{bmatrix}$$

Where we can express the MANOVA and continuous case with this notation:

- When the input variables X are continuous: Their values will be vectors of real values. And the parameters are obtained by means of computing the ML estimator of the continous regression. We have one D-dimensional vector of parameters θ_m per output variable, since we need to regress all of them. The statistical properties of the regressed variables can be correlated but each just has its parameters. Their parameters will be correlated by a correlation matrix.
- When the input variables X is a discrete variable F with C classes, then we have as parameters a vector of means μ_m or each output continuous variable Y_i. In this case the parameters are the mean of the $m-th$ output variable for each of the C classes, to it is a C dimensional vector f.

(IN GLM THE OUTPUT TO REGRESS IS ALWAYS CONTINUOUS OTHERWISE IT IS CLASSIFICATION, LDA)

The computation of the parameters is similar to the previously seen GLM and ANOVA. For the GLM we can compute them independently for each of the classes. Their estimation will be correlated of course (since they depend on the same X data).

Now the computed noises are correlated and we express them as a covariance matrix where the elements of the diagonal are the variances of the individual noises we would get from solving the Linear Regression for the different outputs separately and the other elements are the covariance between the noises of the different output dimensions. The more correlated they are, the more correrlated will be Y_{j1} and Y_{j2}.

The equation to compute them is the same, it is just that now the error are matrices and not vectors as before so the result is a covariance matrix of the noise. We will assume that the noise samples are i.i.d so the covariance matrix is the identity. The unbiased of this covariance matrix is:

$$\hat{\Sigma} = \frac{1}{N - rk(X)} ||Y - \hat{\theta}X||^2_{\Sigma_\epsilon^{-1}} = \frac{1}{N - rk(X)} (Y - \hat{\theta}X)(Y - \hat{\theta}X)^T$$

$$\hat{\Sigma} = \frac{1}{N - rk(X)} \begin{bmatrix} \epsilon_{11} & \cdots & \epsilon_{1M} \\ \epsilon_{21} & \cdots & \epsilon_{2M} \\ \vdots & \vdots & \vdots \\ \epsilon_{N1} & \cdots & \epsilon_{NM} \end{bmatrix}^T \begin{bmatrix} \epsilon_{11} & \cdots & \epsilon_{1M} \\ \epsilon_{21} & \cdots & \epsilon_{2M} \\ \vdots & \vdots & \vdots \\ \epsilon_{N1} & \cdots & \epsilon_{NM} \end{bmatrix} = \frac{1}{N - rk(X)} \begin{bmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1M} \\ \vdots & & \vdots \\ \hat{\sigma}_{M1} & \cdots & \hat{\sigma}_{MM} \end{bmatrix}$$

Regarding the covariance between the parameters, now we have $D \cdot M$ parameters, since for each output dimension M we have a D-dimensional vector. The covariance between vectors from output variables i and j is:

$$COV(\hat{\theta}_i, \hat{\theta}_j) = \sigma_{ij}(XX^T)^{-1} = \sigma_{ij} \left[\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}^T \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix} \right]^{-1}$$

We obtain our estimation of the covariance matrix by using our estimated σ_{ij} computed above.

10.9 MANOVA decomposition of error

Then we have:

$$\begin{aligned}\mathbf{T} &= \sum_{\mathbf{i=1}}^{\mathbf{g}}\sum_{\mathbf{j=1}}^{\mathbf{n_i}}(\mathbf{Y_{ij}}-\bar{\mathbf{y}}_{\cdot\cdot})(\mathbf{Y_{ij}}-\bar{\mathbf{y}}_{\cdot\cdot})' \\ &= \sum_{\mathbf{i=1}}^{\mathbf{g}}\sum_{\mathbf{j=1}}^{\mathbf{h_i}}\{(\mathbf{Y_{ij}}-\bar{\mathbf{y}}_{\mathbf{i}})+(\bar{\mathbf{y}}_{\mathbf{i}}-\bar{\mathbf{y}}_{\cdot\cdot})\}\{(\mathbf{Y_{ij}}-\bar{\mathbf{y}}_{\mathbf{i}})+(\bar{\mathbf{y}}_{\mathbf{i}}-\bar{\mathbf{y}}_{\cdot\cdot})\}' \\ &= \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{y}_{i\cdot})(Y_{ij}-\bar{y}_{i\cdot})'}_{\mathbf{E}} + \underbrace{\sum_{i=1}^{g}n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})'}_{\mathbf{H}}\end{aligned}$$

Where:

- E: Is the Error Sum of Squares and Cross Products. The (k, l)th element of the error sum of squares and cross products matrix E is:

$$E(k,l)=\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ijk}-\bar{y}_{i.k})(Y_{ijl}-\bar{y}_{i.l})$$

  For $k=l$, this is the error sum of squares for variable k, $SS(M)$**,** and measures the within treatment variation for the kth variable. It is how much variance we can explain from the k output if we use the mode. For $k\neq l$, this measures the dependence between variables k and l after taking into account the treatment.

- H: H is the Hypothesis Sum of Squares and Cross Products. The (k, l)th element of the hypothesis sum of squares and cross products matrix H is

$$H(k,l)=\sum_{i=1}^{g}n_i(\bar{y}_{i.k}-\bar{y}_{..k})(\bar{y}_{i.l}-\bar{y}_{..l})$$

  For $k=l$, this is the treatment sum of squares for variable k, and measures the between treatment variation for the kth variable, that is the total sum of errors $SS_{res}(M)$,. For $k\neq l$, this measures dependence of variables k and l across treatments.

So before when we had a model, we checked if the model was any worth it by computing the F statistic between using the model and not using anything, that is, all the parameters are 0 ?.

$$F=\frac{(SS_{res}(\mathcal{H})-SS_{res}(\mathcal{M}))/(D-r)}{SS_{res}(\mathcal{M})/(N-D)}\sim F(D-r,N-r)$$

When we test the entire model then the hypothesis H has $r=0$ parameters and its error is therefore $SS_{res}(H)=SS(Y)$ And therefore in the numerator we end up with:

$$F=\frac{(SS(\mathcal{H})-SS_{res}(\mathcal{M}))/D}{SS_{res}(\mathcal{M})/(N-D)}=\frac{SS(\mathcal{M})/D}{SS_{res}(\mathcal{M})/(N-D)}$$

Where as we saw, $SS(M)$ is the variation of the model. How much the predictions of the model vary.

$$SS(M)=||p_M(Y)||^2=(\hat{\theta}X)(\hat{\theta}X)^T=YHY^T$$

It is how much variation of the original Y our model is explaining (removing). This quantity has $rk(X)$ degrees of freedom. If we removed the mean from Y then it has $rk(X)-1$ degrees of freedom.

In the following image we can see a MANOVA example where we have 3 output variables $aracell, areanucl, avectypo$ and one unique input variable *donor* which is discrete and have 2 values. In this case $M=3$ and $C=2$ so we would have 6 parameters. We have 40 samples and we can see in the tables that:

- **Table of the Dependent Variable aracell**: This is the normal univariate GLM model with only one input variable donor. In this case we have 2 parameters, the mean for donor = 0 and the mean for donor = 1. But since we are also using the intercept, the degrees of freedom of the model is reduced by 1. In the table we can observe:
  - The degrees of freedom of the different sum of squares.
  - The SS of the Model plus the SS of the error is equal to the SS of the Total (Without using any model).

- The normalized SS, dividing by their degrees of freedom.
- The F statistic computed by dividing the Normalized Model Error (variance we remove) by the Normalized Error of our mode.

- In the table of E we can see that its value is the same as the SS of the Error.
- In the table of H we can see that its valis is the same as the SS of the Model.

This is the global test that SAS does of our model !! And we can obtain these values from the E and H matrix for a given output variable since:

SAS-ANOVA Table – without intercept

Source of variation	Sum of Squares	Degrees of Freedom
Model	SS(Model)	$\text{rk}(\mathbf{x})$
Error	SSRes(Model)	$n - \text{rk}(\mathbf{x})$
Uncorrected Total	SSTot(Uncorrected)	n

$$\text{SS(Model)} = \|p_M(\boldsymbol{Y})\|^2 = (\mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{x}\hat{\boldsymbol{\theta}}) = \boldsymbol{Y}'\mathbf{H}\boldsymbol{Y}$$

$$\text{SSRes(Model)} = \|\boldsymbol{Y} - p_M(\boldsymbol{Y})\|^2 = (\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = \boldsymbol{Y}'\mathbf{M}\boldsymbol{Y}$$

$$\text{SSTot(Uncorrected)} = \|\boldsymbol{Y}\|^2 = \boldsymbol{Y}'\boldsymbol{Y} = \sum_{i=1}^{n} Y_i^2$$

(a) sizeInches = [2,2]

SAS-ANOVA Table - with intercept

Source of variation	Sum of Squares	Degrees of Freedom
Model	SS(Model)	$\text{rk}(\mathbf{x}) - 1$
Error	SSRes(Model)	$n - \text{rk}(\mathbf{x})$
Corrected Total	SSTot(Corrected)	$n - 1$

$$\text{SS(Model)} = (\mathbf{x}\hat{\boldsymbol{\theta}} - \bar{Y}\mathbf{1})'(\mathbf{x}\hat{\boldsymbol{\theta}} - \bar{Y}\mathbf{1}) = \boldsymbol{Y}'\mathbf{H}\boldsymbol{Y} - n\bar{Y}^2$$

$$\text{SSRes(Model)} = (\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = \boldsymbol{Y}'\mathbf{M}\boldsymbol{Y}$$

$$\text{SSTot(Corrected)} = \boldsymbol{Y}'\boldsymbol{Y} - n\bar{Y}^2 = \sum_{i=1}^{h} (Y_i - \bar{Y})^2$$

(b) sizeInches = [10,6]

Figure 53: Effect of the figure size in the saved figures

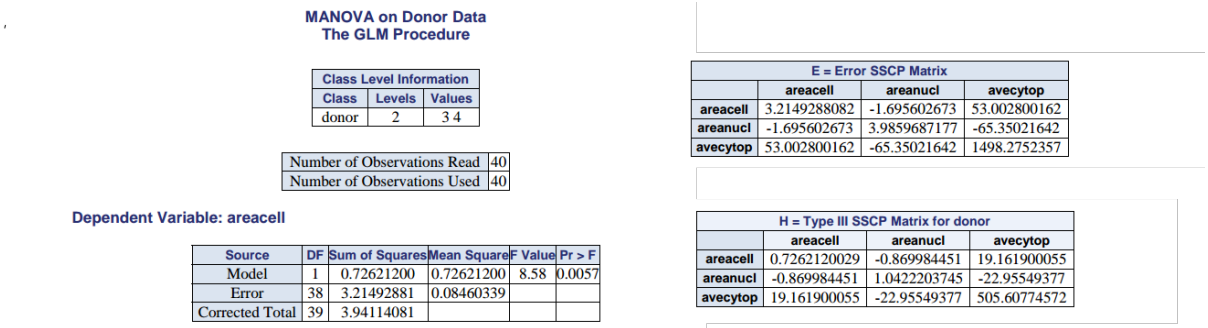


Figure 54: SMA and its window

$$SS(M)$$

10.10 Wilks Lambda test

The Wilk's Lambda test is an statistical test to check if a subset of our parameters θ is equal to some value. We select the subset of parameters by means of the matrices A and B and we indicate the value of the equality by matrix C. So the hypothesis we can test are:

$$H_0 : A\theta B = C \qquad H_A : A\theta B \neq C$$

The matrices A, B and C have the following properties:

- A: The input dimensions selector. This is a $r \times D$ matrix where each row has a 1 in the dimension of θ that we want to include in the test and the rest of 0. Here r is the number of dimensions of θ we are including in the test. As an example the next matrix would include the 2nd and 4th rows of θ.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \end{bmatrix}$$

- B: The output dimensions selector. This is a $M \times s$ matrix where each column has a 1 in the dimension of θ that we want to include in the test and the rest of 0. Here p is the number of output variables we are including in the test. As an example the next matrix would include the 1nd and 2th columns of θ.

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

- C: The values of the selected parameters we want to test. This is a $r \times s$ matrix where each element is the test value for one of the parameters selected. Since we selected 2 output variables and 2 parameters, then it has dimensions $2x2$

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

Lets rewrite it all together to have a better picture:

$$A\theta B = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1M} \\ \theta_{21} & \cdots & \theta_{2M} \\ \vdots & \vdots & \vdots \\ \theta_{D1} & \cdots & \theta_{DM} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \theta_{21} & \theta_{23} \\ \theta_{41} & \theta_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

The number of degrees of freedom is:

$$U(s, r, N - D)$$

Notice we can for example do:

- Global test if all of the parameters are 0, in which we select all the parameters and equal to 0 so $r = D$ and $s = M$.
- In the MANOVA, one could test if the parameters are the same for all the classes, that is, all the $D = C$ inputs. In this case we would select all of the parameters except the ones from one of the classes equal the rest to those parameters so we have $r = D - 1$ and $s = M$.

several elements of the regression values theta . Given by selection matrix A,B and value matrix C. It is the preferred method to carry statistical test since it has nice properties. It can check any combination of parameters almost to see if they are 0.
Here, the determinant of the error sums of squares and cross products matrix E is divided by the determinant of the total sum of squares and cross products matrix T = H + E. If H is large relative to E, then —H + E— will be large relative to —E—. Thus, we will reject the null hypothesis if Wilk's lambda is small (close to zero).

$$\Lambda^* = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

It first gives an U distribution that we trasnform into F. If we were only computing for one we would only need t-student distribution.

In the multivariate GLM we will have an estimated covariance matrix that is composed by the Kronecker product of the H and the covariance matrix of the regressed variables.
One sided ANOVA. B = How much the group vary and the other one is how much within group vaty.
When having a Manova problem, then use the MANOVA ways of testing, not the AB C thing, it can be done but it rather complicated.

IDEA: On the building more complex systems that learn more specific patterns. As noise increases those patterns are too hidden. The fact that the amount of noise influences how much we can learn, and the 4th anf 5th order terms we usually cannot since the noise is too big. Same with hidden MC ?