

CH5230: System Identification

Probability, Random variables and moments: Review 6

Arun K. Tangirala

Department of Chemical Engineering
IIT Madras



Arun K. Tangirala, IIT Madras

System Identification

Arun K. Tangirala: Okay, so this covariance is a measure that you have to be extremely comfortable with. You cannot say, I do not know how to compute covariances, how to interpret covariances, what is the definition, nothing, no such excuses are allowed, when it comes to modeling. It is that ubiquitous quantity that will be with you all the time, and therefore, you should be extremely comfortable with covariances, computing covariances, interpreting them and so on.

Vector case

In the general case, for a vector of random variables,

$$\mathbf{X} = [X_1 \quad X_2 \quad \cdots \quad X_N]^T$$

the matrix is given by,

$$\begin{aligned} \Sigma_{\mathbf{X}} &= E((\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T) \\ &= \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_N} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2 X_N} \\ \vdots & \cdots & \cdots & \vdots \\ \sigma_{X_N X_1} & \sigma_{X_N X_2} & \cdots & \sigma_{X_N}^2 \end{bmatrix} \end{aligned}$$

Now when you have a vector of random variables, we have not yet proved by the way that covariance is a linear measure and we may not prove, but I'll give you some results. You should remember covariance as a probabilistic measure and also prediction interpretation. Probabilistic perspective is that it is a second central moment of the joint p.d.f. That doesn't fetch you much. What the interpretation that fetches you a lot is that it is a measure of linear dependence. That is more useful to us than prediction theory, but you should now slowly understand probabilities and predictions are related.

Okay, so coming back to the vector case. In the vector case, again, you talk of the covariance matrix, the same story, the off diagonals contain the cross-covariances of the pairs of variables in combination, and the diagonals contain variances along the individual variables. Now the other thing at this point I want to broaden your view of random variables. Until now we have been speaking of random variables in the context of random signals. Am I right? That is how we got introduced to random variables, but now you have to graduate from that entry point.

Now you have to learn to look at all other situations or at least a few other situation where you would encounter random variables, which are not appearing in the context of random signals. So what is one such situation? One such situation arises very commonly in parameter estimation, which we are going to do extensively in this course, right. We do keep estimating parameters of a model.

So what is the connection between this covariance and parameter estimation? Of course, the connections are quite deep, but one point that I want to really mention here is these random variables that you see here will take a different interpretation in parameter estimation where the parameter estimates now should be interpreted as random variables. After all, how is the estimate being derived from data and data has a DNA of randomness in it. Therefore, parameter will also has a DNA of randomness in it, right. So every estimate, every parameter estimate is a random variable, because it's a function of random variable. Any function of a random variable is a random variable. There's no doubt about it.

So parameter estimates are also random variables and in estimation theory, we will run into covariance of parameter estimates. So the x_i that you see there, any x_i can be thought of as parameter estimate, estimate of some θ_i , some parameter, right. In which case, a covariance matrix is now the covariance of parameter estimation with diagonals containing variances of the individual parameters, if I have p parameter, I have p elements in the diagonal, and of course, the σ_{θ} is p/p . so here I have $\sigma_{\hat{\theta}_1}$, $\hat{\theta}_2$, and I have $\sigma_{\hat{\theta}_1}$, $\hat{\theta}_p$, and it's a symmetric matrix. So covariance matrix is a symmetric matrix, it's a symmetric positive definite matrix, which means its eigen values are greater than 0 and those are some very useful properties of the variance, covariance matrix. We don't say variance, covariance matrix, we simply say covariance matrix. You have to understand.

The more important rather than just remembering the definition, you should remember the interpretation. Here the interpretation is the same, that is the diagonals contain the variances of the respective parameter estimates, and the off diagonals contain the covariances of the parameter estimates. Both you have to understand carefully, and for that, you have to have a sound interpretation of what is variance. Variance of a random variable is a spread of the outcomes, spread of possibilities for that random variable. Likewise $\sigma^2_{\hat{\theta}_i}$, if I pick i parameter, what is $\sigma^2_{\hat{\theta}_i}$. Do not say $\sigma^2_{\theta_i}$, that is wrong, because for us at least in the classical framework, θ_i , the true parameters are deterministic quantity. It's the estimate that is the random variable.

So what is σ^2_{θ} I had? It is the variance or the spread of possible values of θ_i , and you have to ask in your mind, why should I have multiple possibilities for $\hat{\theta}_i$. Why do you think that I should have -- that there are multiple possibilities for $\hat{\theta}_i$? Correct, because there are multiple possibilities for data. I am only working with on realization. Had I used another sensor or if I repeat experiment or someone was doing an experiment in parallel, another data record could have been generated another possible value of $\hat{\theta}_i$ could have been obtained. So this thought experiment you should get used to until you are comfortable with variance of the random variable, okay. So there are multiple records possible, therefore, multiple $\hat{\theta}_i$ are possible, and what does $\sigma^2_{\hat{\theta}_i}$ is telling me is what is the spread of all such possibilities.

Now if the estimation algorithm is good, then it should be able to shrink the variability in the data. By the time it gives you $\hat{\theta}_i$, it should have actually shrunk the variability into -- it can't shrink it to 0, because you have finite data, but it should have definitely done a very good job of it, and the more -- the estimator is better at shrinking it, the more efficient phase is estimated. That is the notion of efficiency and estimation theory. The most efficient estimator is the one that gives you the minimum variability in θ , okay.

So, so much about the variance of the -- that is the along the diagonals. When you look at the off diagonals, what is it telling you? It is telling you how the estimate of θ_i is influencing estimate of θ_j . In what ways it influences -- linearly influencing, because it's a covariance measure. See, when I am estimating more than one parameter, one estimate can have an impact on the other one. That means the error in one can affect the error in other. Of course, there are some very, very specific situations where they don't affect each other, in which case $\sigma_{\hat{\theta}}$, the $\sigma_{\hat{\theta}}$ is going to be diagonal, okay, but by and large your $\sigma_{\hat{\theta}}$ is going to be not a diagonal matrix. So the off diagonals tell you how the error in one estimate is influencing the error in other estimate.

Now having discussed so many -- typically when it comes to using this $\sigma_{\hat{\theta}}$, at least for so-called construction of confidence intervals and so on, we generally ignore the off diagonal terms, okay, but that we'll talk about later on. You should note this interpretation very well in your mind.

Properties of the covariance matrix

The covariance (matrix) possesses certain properties which have far-reaching implications in analysis of random processes.

- ▶ Covariance is a second-order property of the joint probability density function
- ▶ It is a **symmetric** measure: $\sigma_{XY} = \sigma_{YX}$, i.e., it is not a directional measure. **Consequently it cannot be used to sense causality** (cause-effect relation).
- ▶ The covariance matrix Σ_X is a **symmetric, positive semi-definite** matrix
 $\implies \lambda_i(\Sigma_X) \geq 0 \quad \forall i$
- ▶ **Number of zero eigenvalues of $\Sigma_Z =$ Number of linear relationships in Z**
 (cornerstone for principal component analysis and multivariate regression)

We've talked about the properties of the covariance matrix. So the point to take home for you is this covariance matrix is ubiquitous, it has a structure, the same structure across all situation. In all situations, Σ_X is going to be the same, but it's interpretation changes with what the random variables stand for. In random -- when they are simply variables, it has a different interpretation; if they are parameter estimates, then they have a different interpretation and so on, okay.


Probability, Random Variables & Moments MATLAB commands

Properties of the covariance matrix

- ▶ Linear transformation of the random variables $\mathbf{Z} = \mathbf{A}\mathbf{X}$ results in

$$\Sigma_{\mathbf{Z}} = E((\mathbf{Z} - \mu_{\mathbf{Z}})(\mathbf{Z} - \mu_{\mathbf{Z}})^T) = \mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}^T \quad (19)$$
- ▶ Most importantly, **covariance is only a measure of linear relationship** between two RVs, i.e.,

When $\sigma_{XY} = \sigma_{YX} = 0$, there is no linear relationship between X and Y



Arun K. Tangirala, IIT Madras System Identification March 1, 2017 54

And there are many other useful properties of covariance matrix that is used in multivariate data analysis. I am not going to talk about it, but the important thing to remember when it comes to covariance is, it is only a measure of linear relationships. You should remember that, which means if the covariance between a pair of random variables is 0, it only rules out linear dependency, it doesn't rule out any other form of dependency. That means it says x and y are not blood relatives or first cousins, that's all. They maybe second cousin, third cousin, some hundred cousin, don't know, all right.

So when it comes to using covariance in practice, there are two shortcomings. One is that covariance, as you must recall from the definition, is sensitive to the choice of units for x and y , and two, it's an unbounded measure. That means that covariance can take on from $-\infty$ to ∞ . So the utility of such a measure is not so great, always bounded measures are easier to use and work with, and also we need a measure that is invariant of the choice of units of x and y .

Correlation

In order to overcome the unbounded and scaling-sensitive nature of σ_{XY} , we work with a normalized version of covariance known as **correlation**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (20)$$

Boundedness

For all bivariate distributions with finite second order moments,

$$|\rho_{XY}| \leq 1 \quad (21)$$

with equality if, with probability 1, there is a linear relationship between X and Y .

If x and y are temperature and pressure. Changing the unit should not change the value, because it's a measure of linear dependence; therefore, we introduce correlation, which is normalized covariance, and this correlation is bounded in magnitude by 1, which means the maximum value correlation can take on is unity in magnitude, it can have [Indiscernible 00:10:37].

That's beautiful, so it addresses two issues in one shot. You can see from the definition that it is invariant to the choice of units of x and y , and two, [Indiscernible 00:10:49] with what proof of I am going you. You can look at the proof anywhere in the literature. It's standard proof that's available that correlation is bounded. So what is the interpretation for correlation that is extremely important. You know the definition of correlation, but what is the interpretation. When correlation is 0, then x and y have no linear relationship, interpretation number one.

That means non-linear relationships cannot be ruled out. It cannot be -- we do not know, it cannot detect necessarily. When correlation hits the maximum, that is 1, then x and y are perfectly related in a linear relationship including an affine form, that is $\alpha x + \beta$ also, $\alpha x + \beta$ is not truly a linear relationship, but it's okay, in functional analysis $\alpha x + \beta$ is called a linear also, I mean in polynomial function analysis. But anyway, remember that y is a linear function of x with the admissibility of an intercept. Perfect.

And when you -- the third is the most practical case that correlation takes on values less than 1? Typically when you estimate correlation, you may see values of 0.8, 0.5, 0.6 or some value less than 1, it will never be equal to 1.

So what is the interpretation there? Well, when correlation is less than 1 in magnitude, then there are at least two different possibilities, and it's not exclusive. It could be a combination of these two. One that x and y are non-linearly related. That's one possibility, which can lower the correlation. That means you have a situation where $y = \alpha x(\epsilon)$. It could be some function of x or some other variable, but it's definitely not a linear function of x , right.

It's definitely not that and this epsilon could contain non-linearities, non-linear functions of x of pure noise, okay. It could be pure noise also. That means the true relation between y and x is linear, but epsilon could stand for measurement noise or it could be combination of both, okay. So that is a situation that you encounter in practice. And normally, again, you would conduct hypothesis test, typical hypothesis test that are connected in the context of correlation, that the true correlation is 0 against the alternative that it isn't. This is the standard -- I mean hypothesis tests are standard in statistical inferencing. Any parameter that you estimate, you would have a hypothesis test of this form, that the true value of parameter 0 against the alternative that it isn't. Hypothesis tests of these forms are called significance tests, [Indiscernible 00:13:58] estimate is significant or not. So remember that. You can look up the videos that I have on introduction to statistical hypothesis testing.

So the bottom-line here, coming back to the discussion, is when correlation is less than 1 in magnitude or if it is extremely low, let us say, 0.1 or 0.05 and so on, it could be that there's heavy amount of noise or it could be that there is non-linearity, and it's hard to figure out what is happening. The only thing that is for sure is when correlation values are extremely high, then a linear model will do a very good job of prediction. That is guaranteed, all right.

Uncorrelated variables

Uncorrelated variables

Two RVs are said to be **uncorrelated** if $\sigma_{XY} = 0 \implies \rho_{XY} = 0$. Alternatively,

$$E(XY) = E(X)E(Y) \quad (22)$$

- ▶ Uncorrelatedness \iff NO **linear** relationship between X and Y .
 - Independence \implies Uncorrelated condition but NOT vice versa.**
- ▶ Thus independence is a stronger condition.
- ▶ Determining the absence of non-linear dependencies requires the **test of independence**.



So let's conclude the discussion here. There is this notion of uncorrelatedness and there's this notion of independence. When two variables are uncorrelated, it only means that there's no linear relationship between them that is from a prediction viewpoint. The probabilistic viewpoint is expectation of the product of a product of expectations, whereas when it comes to independence, the f -- so look at this nice similarity in the form of result. This is uncorrelatedness condition and this is independent condition. Of course, we are not looking at $f(x, y)$. Nevertheless, there's some similarity, in fact, I should note marginals.

And what this slide is telling you is that uncorrelatedness does not necessarily imply independence. Just because moments can factorizable, the p.d.f.s cannot be factorizable. On the other hand, if the p.d.f.s are factorizable, then the moments are factorizable, joint moments are factorizable. So that means independence naturally rules out all forms of relations including linear ones, whereas uncorrelatedness only rules out linear this thing.

Remarks, limitations, . . .

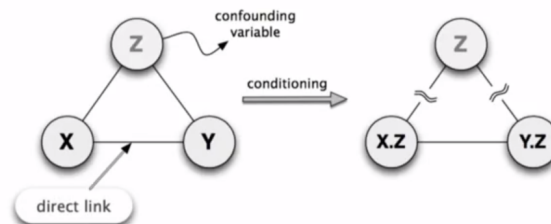
- ▶ **Correlation is only a mathematical / statistical measure.** It does not take into account any physics of the process that relates X to Y .
- ▶ High values of correlation only means that a linear model can be fit between X and Y . It does not mean that in reality there exists a linear process that relates X to Y
- ▶ Correlation is symmetric, i.e., $\rho_{XY} = \rho_{YX}$. Therefore, it is not a cause-effect measure, meaning it cannot detect direction of relationships
- ▶ Correlation is primarily used to determine if a linear model can explain the relationship. **Absence of correlation only implies that no linear model can be fit.**

So just to conclude, correlation is -- you know correlation is a very ubiquitous measure that's used in data analysis. It has its own limitations, despite its mighty use, it has its limitations. One, it doesn't tell you which causes the other. It doesn't tell you whether x causes y or y causes x , because it's a symmetric measure, that's point number one. Secondly, it has no physics involved in it. I can take any two variables and who that there's very high correlation. It doesn't tell you anything about physics. You have to choose, therefore, x and y carefully in a manner that is significant. And thirdly, absence or correlation only implies, again, I keep repeating, absence of correlation only implies no linear model can be built. You can possibly build a non-linear model and do the job, okay.

Confounding

When two variables X and Y are correlated, a question that begs attention is:

Q: Are X and Y connected to each other **directly** or **indirectly**?



Correlation measures **total** (linear) connectivity, whereas **conditional** or **partial** version measures **"direct"** association.

So the other point that we will discuss, in the interest of time I am going to stop, but I just want to leave you with this thought, this is important as well. When I have correlation between x and y , it could be confounded by a third variable. So which means, I can have a non-zero value of correlation between x and y , because there is a common variable influencing them. Truly x and y may not be related directly, that is what we call as confounding, and the way we resolve confounding is what is known as conditioning, that is instead of evaluating plain correlation between x and y , you say now that z is given and possibly influencing both x and y , are x and y correlated. This is called condition correlation or partial correlation, and I'll just give you the expression straight away, we'll not go into the proof of derivations and so on.

Partial covariance

The conditional or partial covariance is defined as

$$\sigma_{XY.Z} = \text{cov}(\epsilon_{X.Z}, \epsilon_{Y.Z}) \quad \text{where } \epsilon_{X.Z} = X - \hat{X}^*(Z), \epsilon_{Y.Z} = Y - \hat{Y}^*(Z)$$

where $\hat{X}^*(Z)$ and $\hat{Y}^*(Z)$ are the **optimal predictions** of X and Y using Z .

Partial correlation (PC) for scalar Z

$$\begin{aligned} \rho_{XY.Z} &= \frac{\sigma_{XY.Z}}{\sigma_{\epsilon_{X.Z}} \sigma_{\epsilon_{Y.Z}}} \\ &= \frac{\rho_{XY} - \rho_{XZ} \rho_{ZY}}{\sqrt{(1 - \rho_{XZ}^2)} \sqrt{(1 - \rho_{ZY}^2)}} \end{aligned} \quad (23)$$

The way you construct conditional correlation is you remove the effects of z from both x and y , and then you compute the correlation between the residual. So if you look at $\epsilon_{x,y}$, you have $X - \hat{X}^*(Z)$. What is $\hat{X}^*(Z)$? Optimal prediction of X using Z . Therefore, $\epsilon_{x,y}$ contains all the effects -- that is that part of X which is devoid of the influence of Z . Likewise, $\epsilon_{y,z}$ contains that component of Y that is influenced by Z .

Now I am going to look at correlation between or covariance, if it is partial covariance, between $\epsilon_{x,z}$ and $\epsilon_{y,z}$. That is idea, and you can derive the -- with this definition of partial covariance, you can derive the expression for partial correlation here as $\sigma_{xy,z}/\sigma_{\epsilon_{x,z}}$ and $\sigma_{\epsilon_{y,z}}$, and this is expression that you can derive. It's not difficult at all. The only thing that you have to do is replace \hat{X}^* with a linear of Z and \hat{Y}^* with a linear function of Z , and you can show the proof. You can refer to my videos on time series analysis, the result is derived.

Partial correlation: Example

Consider two random variables $X = 2Z + 3W$ and $Y = Z + V$ where V , W and Z are zero-mean RVs. Further, it is known that *i.e.*, $\sigma_{VW} = 0 = \sigma_{VZ} = \sigma_{ZW}$.

Evaluating the covariance between X and Y yields

$$\sigma_{YX} = E((2Z + 3W)(Z + V)) = 2E(Z^2) = 2\sigma_Z^2 \neq 0$$

although X and Y are not “directly” correlated.

However, applying (23), it is easy to see that they are not **directly** correlated, *i.e.*,

$$\rho_{YX.Z} = 0$$

I'll just conclude with an example on this. So here I have X and Y , $2Z + 3W$ and $Z + V$. Now obviously, by the statement of the problem, V , W and Z are all uncorrelated, which means the only reason why X and Y should be correlated in this example is through Z . If I take out the effects of Z , X and Y should be uncorrelated. So when I compute the unconditional covariance, it turns out to be some $2\sigma_Z^2$, which means it's a non-zero quantity. But if I discount for the effects of Z , and to do that, in practice, I have to be given a measurement of Z , remember that. That means I have to be given the measurement of the confounding variable. When I do that, then I get the conditional correlation to be 0, straight away telling me that Z was the confounding variable.

Still if this conditional correlation turns out to be non-zero, it only means that still they may be confounded by other variables, we don't know, but definitely what it says is even after you remove effects of Z , X and Y are related.

So that brings up to an important point that will end this thing. Whenever two variables are correlated, you can never ever resolve in life whether there is confounding or not, unless you have taken into account all the confounding variables in the universe, okay. So that is one way of looking at limitations of statistics, where you may not have science necessarily involved, but if you have made sure through the knowledge or the physics of the process that all the confounding variables have been taken into account, then you can trust that correlation to be a measure of direct dependence, okay.

When we next meet, we'll move into the random signal world, where we'll apply all of these concepts to the random signal. We'll look at auto correlation, cross-correlation, models for random signals and spectral densities, okay. Thank you.