

Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 14
Lecture 07A - Probability and Statistics Review (Part 2)-4

So, yesterday we celebrated Independence Day, today we will celebrate covariance day. We will learn a lot about covariance today; mostly about correlation, but we will also talk about independence and a lot of things and I should say that these are some of the lectures which are central to your ability to understand the auto covariance functions or cross covariance and even partial auto covariance functions that we will later learn in the context of random signals. So therefore you need to pay attention, apart from that you should also ask any questions that you may have and not wait until the course is over, so feel free, relax and learn.

So, let us begin with the covariance from where we left off essentially in the last class and as I said very often and this is not just in time series analysis, you will see this everywhere right from the animals to the human beings everywhere, in every sphere we are trying to correlate stuff you know I saw this and then this looks similar to that and so on, all of that is based on correlation. Of course, specifically in statistical analysis we are interested in knowing whether the variation in one variable is influenced or is influencing any other random variable or a bunch of random variables, but we will restrict ourselves to a pair of random variables and a statistical measure of such co-variation is a covariance which also happens to be the second moment of the joint pdf.

(Refer Slide Time: 01:50)

Probability & Statistics - Review 2

Covariance

One of the most interesting questions in bivariate analysis and prediction theory is if the outcomes of two random variables influence each other, i.e., whether they co-vary.

The statistic that measures the co-variance between two RVs is given by

$$\sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y)f(x, y) dx dy \quad (9)$$

Covariance, second-order property of the joint p.d.f., can further be shown as

$$\sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - E(X)E(Y) \quad (10)$$

Arun K. Tangirala Applied TSA August 14, 2016 NPTEL 13

(Refer Slide Time: 02:05)

$\hat{\mu}$

$$\sigma_{XY} = E(XY) - E(X)E(Y)$$
$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} \quad \Sigma_{\underline{X}} = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \sigma_{X_1 X_3} & \dots & \sigma_{X_1 X_M} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \dots & \dots & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \sigma_{X_M X_1} & \dots & \dots & \dots & \sigma_{X_M}^2 \end{bmatrix}$$
$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^N V[k] \quad i=1, \dots, R$$

NPTEL

And as I mentioned in the last class, covariance can be also written conveniently as you see on the slide as a difference between expectation of products and the product of expectations.

And once again the expressions that you see here on the slides are useful for theoretical analysis, at a later stage we will learn how to compute covariance's given data those are called sample covariance's and so on. We will not worry about it as of now, hopefully if today, but if not today; the first thing tomorrow morning I will kind of go through an

example in hour for you as to how to compute sample covariance. The purpose of introducing this covariance measure is to see how two variables co vary and we are talking in the context of random variables, covariance can also be thought of between two deterministic signals that say, but we will talk about that later.

Now, very soon we shall learn and as you must have also learned in the NPTEL course on introduction to hypothesis testing that covariance is a measure of linear dependence alright which means that when this quantity covariance is 0, we rule out the absence of any linear relationship between two variables and the lot of other subtle but very important things that we shall learn.

Now, before we move on to understanding the properties of covariance, the significance of covariance we also made a statement that as far as linear random processes are concerned or linear models are concerned, it is sufficient to know only the first and second order moments, we will go through that but before we go through that let us understand how this notion of covariance extends to a vector of random variables, this is how we define for a pair of random variables, but when I have a vector of random variables and I denote this vector X with an underscore here.

Let us say I have here M or n , e , m random variables; it does not matter I can now extend this idea of covariance to the case of vector of random variables by constructing what is known as a covariance matrix that you see on the slide as well and this covariance matrix is made up of individual variances, that is variance of the individual random variables along the diagonal and the co-variances between a pair of variables along the off diagonals. So, if you were to look at the first row; the off diagonal elements you would have $\sigma_{X1, X2}$, $\sigma_{X1, X3}$ up to $\sigma_{X1, XM}$.

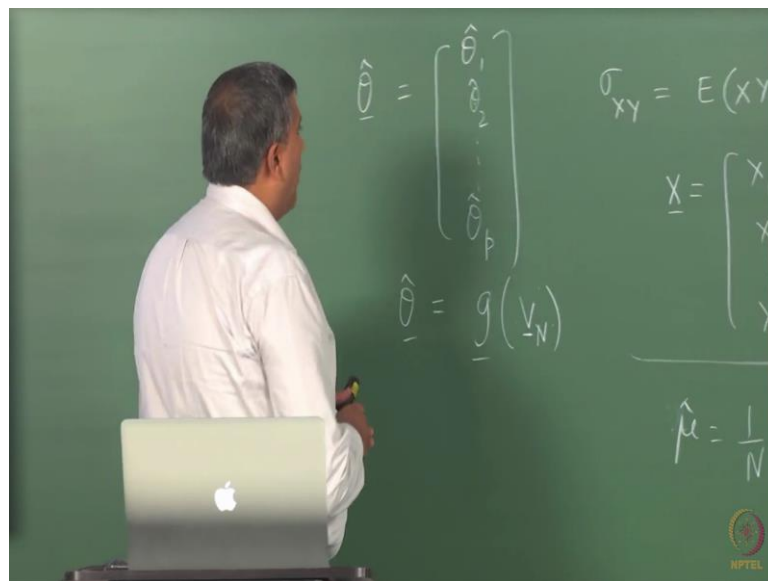
And likewise the elements here as well you can guess what they are and it does not take too much effort to realize that covariance or the variance covariance matrix is a symmetric matrix.

Because covariance is a symmetric measure, it has no notion of directionality embedded in it and will realize that in a different way when we talk of the connections between covariance or correlation and regression, we have not yet spoken about correlation, but we are not too far away from it. So, this variance covariance matrix is one of the central quantities that you will encounter in all forms of statistical data analysis; there is no

escape from it. Therefore you should be comfortable with this quantity, it is a very straightforward quantity; variance along the diagonals and co-variances along the off diagonals.

At this point I also would like you to broaden your treatment or understanding of what is a random variable. Until now you have been thinking of random variables as outcomes of some events but that need not be the case all the time.

(Refer Slide Time: 06:14)



In a lot of situations, particularly in parameter estimation; the estimates parameter estimates that you obtain could also be the random variable vector or you know if you are estimating a couple of parameters then the estimates could also be those random variables and so essentially in any parameter estimation problem. Let us say you are estimating; let us denote the parameter vector by theta and hat denotes the estimate then this vector is made up of the estimates of the individual parameters. Let us say there are p parameters; instead of m you have p that is only difference, but those are just dummy variables these are also random variables.

Because do you know why; why am I claiming that the parameter estimates in any estimation exercise can be thought of as random variables.

Student: (Refer Time: 07:10).

It is sorry.

Student: (Refer Time: 07:15).

It is estimated from random variables; so what kind of random variables participate in the estimation. Can you elaborate a bit more?

Student: (Refer Time: 07:24).

Ok.

Student: The estimates can be different from co different surface.

Does it make it random?

I mean the first answer was pretty close as just seeing if anybody could elaborate on that.

How do you estimate? So, let us take the simplest case where you are estimating mean. We know the theoretical definition of mean and we know how we estimate mean typically which is through the sample mean and how do we calculate the sample mean there is a record of data which is statistically called as a sample. A sample does not mean a single observation; the sample is a collection of observations but one data record. So, if I were to look at the estimation of mean through the sample mean then this is what I would do $\hat{\mu}$ would be $\frac{1}{N} \sum_{i=0}^{N-1} x_i$; let us say I am looking at 0 to N minus does not matter. So V case, are the observations; that I have in that single sample and I just compute the simple average to get the estimate.

Now, on the basis of this equation can we explain why $\hat{\mu}$ is a random variable, μ hat is a parameter estimate. What is the parameter that we are estimating mean? μ hat is an estimator or you can say the value itself is an estimate of the mean. On the basis of this equation, can we come to the conclusion that $\hat{\mu}$ is a random variable right because why because simply each observation is a random variable; by our definition of a random signal.

And this randomness as you had said actually propagates to $\hat{\mu}$. So, the inherent DNA is randomness and that DNA propagates to $\hat{\mu}$. But in a different form how it manifests, we will learn in the estimation theory, how the uncertainty in V K propagates, how we mean by meaning how we are asking a quantification of uncertainty; that means, if I know for example, the mean and variance of V case, can I determine the mean and

variance of mu hat that is what we mean by how, we will learn to do that later on but for now it is pretty clear that mu hat is a random variable.

Likewise any parameter estimate that normally one constructs is some function of your data set v ; let us say $V \ N$ is your collection of N observations and you have a vector of functions giving you vector of parameter estimates. So, the randomness in these observations propagates through this transformation to theta hat. Therefore, theta hat in its own right is also a random variable vector or you know in the simplest case of scalar. So, whatever we are learning here, whether it is mean, covariance, distribution any other moment; all those concepts equally apply to a parameter estimates as well and it is very important to get this clear in our minds. Which means for example, when I move to parameter estimation; I will run into quantity known as the variance covariance matrix of the parameter estimates, so which would look like this.

(Refer Slide Time: 11:03)

The image shows two handwritten equations on a green chalkboard. The first equation, labeled with a circled 'y', shows a variance-covariance matrix for variables x_1, x_2, \dots, x_m . The matrix is symmetric, with diagonal elements $\sigma_{x_1}^2, \sigma_{x_2}^2, \dots, \sigma_{x_m}^2$ and off-diagonal elements representing covariances like $\sigma_{x_1 x_2}, \sigma_{x_1 x_3}, \dots, \sigma_{x_m x_1}$. The second equation shows a similar matrix for parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$. The diagonal elements are $\sigma_{\hat{\theta}_1}^2, \sigma_{\hat{\theta}_2}^2, \dots, \sigma_{\hat{\theta}_p}^2$, and the off-diagonal elements represent covariances between estimates, such as $\sigma_{\hat{\theta}_1 \hat{\theta}_2}, \dots, \sigma_{\hat{\theta}_1 \hat{\theta}_p}$. A small logo is visible in the bottom right corner of the chalkboard image.

So sigma theta hat would like the variance covariance matrix of x ; would consists variances of the individual parameters along the diagonals and the off diagonals would contain the covariance between a pair of estimates and of course, you know it is a symmetric matrix; I am not going to fill the elements here and one has to be comfortable in interpreting that sigma theta hat.

The diagonals contain the variance of the individual parameter estimates. So, sigma square theta 1 hat for example, is a variance or the variability that you see in the estimate

number 1. Remember we are talking of estimates not the parameters, so should not be confused with between the parameters and their estimates. Parameters for now are being treated as deterministic quantities when we move to the Bayesian world; that is when we talk of Bayesian estimation. At that point we will treat even these parameters as random variables but let us not worry about that. At the moment, the parameters are random are fixed quantities which we do not know and parameter estimates are random variables and we have already explained why they are. So, it is important to understand this sigma square theta 1 hat or any of these elements in sigma theta hat may be upfront at this point itself.

So, let me actually ask you how would you interpret sigma square theta 2 hat, what is a correct interpretation of sigma square theta 2 hat.

Student: As we varied the data sets we can, 2 hat.

And then what is sigma square theta 2 hat measured.

Student: Spread of them.

Spread of that, so precision correct, so the simple answer is remember we are computing for example, here I am computing mu or mu hat or estimating mu from one data record. So, suppose I denote the data record by I or the realization by I , I can repeat this for all the R realizations that I may have. Now when you talk of sigma square theta 2 hat, what you are looking at is the spread or the variance of theta 2 hat, across all possible realizations that you can ever generate; that means, entire ensemble of your theta 2 hat.

It is not possible to do that experimental because in general you may need to have infinite number of realizations; that means, entire ensemble which you may not have but theoretically if you had it as a thought experiment then for each realization you compute your theta 2 hat and then you have as many theta 2 hats. As the number of realizations and then it has its own spread, it has its own distribution and you can do this even in R for example, Monte Carlo simulations are kind of a simulation way of replicating this theoretical experiment.

(Refer Slide Time: 14:30)


Probability & Statistics - Review 2

Covariance for vector quantities

The covariance and variance of a pair of random variables X_1 and X_2 are collected in a single matrix, known as the **variance-covariance** matrix

$$\Sigma_X = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 \end{bmatrix}$$

Arun K. Tangirala Applied TSA August 14, 2016



So, sigma square theta I hat in general is the variance of the theta I hat across whatever you see across all the records.

Now what about the off diagonal elements, how do you interpret the off diagonal elements. So, I pick sigma theta 1 hat, theta 2 hat; it was easier to explain sigma square theta I hat. Of course you know you understand why we are looking at the spread because the spread gives us a measure of the precision and we will talk about more that technically later on. So, let us ask now; how do you interpret sigma theta 1 hat theta 2 I hat or in general sigma theta I hat, theta j hat.

The same way as you interpret this; how do you interpret sigma x 1, x 2; it is a covariance. So, now we have to be careful right it is covariance; what is covariance measuring how x 1 and x 2 vary together and to jump the gun a bit linearly very together. Likewise when I am estimating more than one parameter, it is highly likely that the estimate of one parameter that is the error in the; in fact, sigma square theta I hat is a measure of the error in your theta I hat; the theta I is or you can say the error in the estimate theta I hat. When you are estimating more than one parameter each of this parameter estimates will have an error. Sigma theta I hat theta, j hat tells us how the error in one parameter estimate is influencing the error in another parameter estimate.

Because you are estimating them jointly right which means that there is a degrees of freedom issue, you have n observations with you and you are trying to estimate p

unknowns; it is likely that when you try to estimate one, the other one also is influenced by your ability to estimate this correctly. So, it is likely that in any situation the error in one parameter estimate can affect the error in another parameter estimate but it is also possible that your sigma theta hat is a diagonal.

What do you make of such situations?

Student: (Refer Time: 16:56).

That is but with respect to parameter estimation what do you mean?

Student: (Refer Time: 17:03).

Correct. So, you can estimate them separately, you do not have to estimate them jointly. We learn all of this in technical details and so on and we will come across Cramer of bound and Fisher information matrix and so on when we talk of estimation theory. But I am just showing the seeds now, so in summary whatever you are learning for the case of random variables, whether it is in mean, variance or any other moment or your joint moments like covariance or the notion of independence and so on applies to any random variable. It need not be specific to the outcome of some event and so on. So, that is something to keep in mind any questions on this variance covariance matrix fine.

(Refer Slide Time: 18:00)

Probability & Statistics - Review 2

Vector case


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

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

the matrix is given by,

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

Arun K. Tangirala Applied TSA August 14, 2016



So, let us move on and now also of course in passing let me also show you the actual definition of your variance covariance matrix is nothing but expectation of X minus μ where X is your vector times X minus μ transpose but that is kind of a straightforward extension of the univariate definition of variance for the univariate case or covariance for the bivariate case. So, we will of course keep making use of this expression time and again in our theoretical analysis.

(Refer Slide Time: 18:24)

Probability & Statistics - Review 2

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
 $\Rightarrow \lambda_i(\Sigma_X) \geq 0 \quad \forall i$
- ▶ **Number of zero eigenvalues of $\Sigma_Z =$ Number of linear relationships in Z** (cornerstone for principal component analysis and multivariate regression)

Arun K. Tangirala Applied TSA August 14, 2016 NPTEL 16

Any question yes.

Student: Why measuring the parameter so.

Estimating the parameters.

Student: Estimating the parameter when we get the cross steps.

Yes.

Student: So is there possibility that symmetric of the matrix can be broken from the possible measurement estimating the parameters.

Even in process of estimation.

Student: In process of estimation.

Well it is a good question; I will answer the core part of your question and then later on technically will answer that question later on. It is possible but you can avoid that by making sure that whatever expression or so called formula; recruit term for it that you will use to estimate this, in practice again this is a theoretical quantity, from data you will also be able to estimate this. So, from data you will construct an estimate of σ θ hat. Whatever expression that you use for obtaining that σ hat, θ hat should it should ensure that you have a symmetric matrix.

And if you do not then; that means, it does not qualify to be an estimate,, but is a good question yes; when estimating certain quantities you have to respect and preserve certain properties of the theoretical ones, but yes we should make sure that that is the properties preserve. So, here are some quick properties of the covariance matrix and in general covariance as I said it is a symmetric measure which means it does not know which cause the other that is very important, I will not be able to infer both the directionality of causation and the physical causation. I will not be able to say that if x is correlated with y , then x has physically cause y ; you need not be true, I can take any two random variables in the world and see that it is likely that I will get a very strong correlation that does not mean that they are strongly correlated.

So, covariance is always a statistical measure of the interdependence but not a physical measure necessarily. So as scientists, as engineers and as people with common sense we should not be correlating any two things just because I have a computer and I can compute correlation but of course I mean if you want to do for the fun of it, you can do it and also the covariance matrix itself is a positive definite matrix and this covariance matrix as I said is ubiquitous in data analysis and if you take for example, principal component analysis, some of you may be familiar with it. An eigenvalue analysis of the covariance matrix tells us how many linear relationships exist between the random variables, so we do not go into that there is a course on multivariate data analysis or maybe another course in machine learning where you will learn all of this stuff.

But as you can see the quantity variance covariance matrix has far reaching presence in data analysis. So, let us move on and discuss the more interesting part which is that the covariance is a measure of linear relationship.

(Refer Slide Time: 21:55)

Probability & Statistics - Review 2

Properties of the covariance matrix

- ▶ Linear transformation of the random variables $\mathbf{Z} = \mathbf{A}\mathbf{X}$ results in
$$\Sigma_{\mathbf{Z}} = E((\mathbf{Z} - \mu_{\mathbf{Z}})(\mathbf{Z} - \mu_{\mathbf{Z}})^T) = \mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}^T \quad (11)$$
- ▶ 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 Applied TSA August 14, 2016 NPTEL 17

I am not going to prove that covariance is, I mean it in a very rigorous way but we will at least go through a part of the proof that covariance is a measure of linear dependence which means that what we will do is; we will show that when two variables are linearly related covariance kind of standardized covariance reaches a maximum and the other part the reverse part is not something that we prove and we will buy just way of inference, we will say that from the relation that we see that when the x and y are linearly related, covariance reaches a maximum in a standardized sense. From there we will infer that when two variables have no linear relationship then x and y are uncorrelated.

So, whatever see you see on the slide when two variables are uncorrelated, there is no linear relationship and the vice versa is also true; we may not prove the entire part of that statement but you should just remember that, but we will do a bit more analysis now.

(Refer Slide Time: 23:04)

Probability & Statistics - Review 2

Correlation


Two issues are encountered with the use of covariance in practice:

- Covariance is sensitive to the choice of units for the random variables under investigation. Stated otherwise, **it is sensitive to scaling**.
- It is not a bounded measure**, meaning it is not possible to infer the degree of the strength of the linear relationship from the value of σ_{XY}

To overcome these issues, a normalized version of covariance known as **correlation** is introduced:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (12)$$

Arun K. Tangirala Applied TSA August 14, 2016



So, in order to show that covariance is a measure of linear dependence and also to be able to use covariance practically, it is important to introduce this notion of correlation which addresses two deficiencies or shortcomings of covariance when it comes to using it in practice; one is that as you can see in the expression here covariance is sensitive to the choice of units of x and y . For example, if I am looking at temperature and pressure; we know from thermodynamics that temperature and pressure of a gas are highly correlated might assume fixed volume or whatever.

Now, think of x as a temperature and previous pressure does not matter and you can compute covariance; obviously, the value of the covariance will depend on what units I choose for temperature and pressure, I could choose Fahrenheit and bar for pressure, Fahrenheit for temperature or I could use Kelvin and some other unit maybe Pascal for pressure and so on. So, the value of the covariance is going to be sensitive to the choice of units that I choose for the random variables; that is one issue which means that by a clever choice of units I can make look covariance very small or very large.

Secondly again as a corollary of that, covariance is not a bounded measure and it is hard to work with bounded measures in data analysis, I would like to know have some ceiling and also you know some flooring for this whatever measure that I use so that I know whether dealing with a low dependence or a high linear dependence or no dependence and maximum or linear dependence and so on. For these two reasons, we introduced this

notion of correlation which is nothing but a standardized covariance as you see and you must have seen this expression many a times in various say different situations.