Multivariate Normals

Tom Glinnan

January 2022

Introduction

We all know what a normal distribution is. Its pdf looks like the figure below. The normals are an example of a *family* of distributions - just as the χ^2, the F and the T are also families. To pick out a particular distribution from the family, we have to specify two things: its mean μ and variance σ^2, which is why we write a normal as $N(\mu, \sigma^2)$. μ is also sometimes called the 'location parameter' - it tells us where the middle of the distribution is. Similarly, σ^2 is the 'scale' parameter - it tells us spread out the distribution is

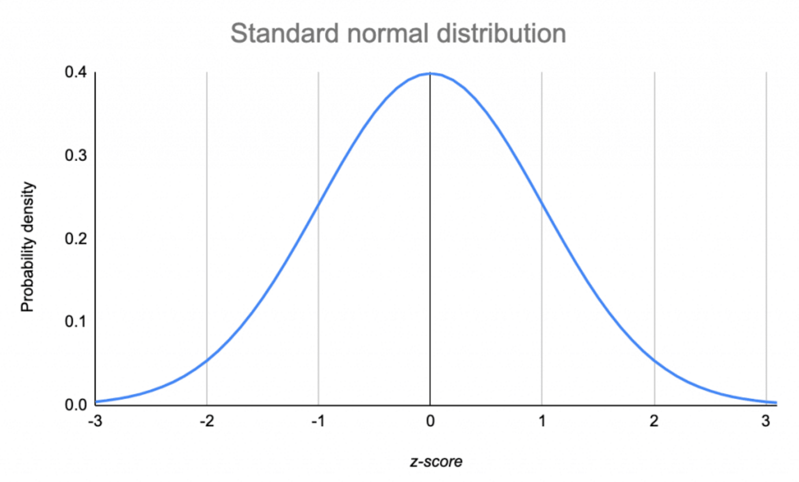

Figure 1: The pdf of a Standard Normal

What does it really mean to say that $x \sim N(\mu, \sigma^2)$? It means that the probability of seeing $x = 5$, for example, is $f(5)$, where f is the pdf of the $N(\mu, \sigma^2)$ distribution. Also, if you ever see the word 'Gaussian', just read 'Normal' instead - they mean exactly the same thing

Now we want to generalise so that x is a vector[1] in $\mathbb{R}^n$ rather than a real number. A typical place where this comes up is when considering the distribution of an estimator, such as the OLS

[1] Appropriately called a random vector instead of a random variable

estimator $\hat{\beta}$. If we know that $\hat{\beta}$ has a multivariate normal distribution, then we can talk about doing tests of the entire vector $\hat{\beta}$ rather than just one component of it (like we did in MT) . This is what's behind the F-test in regression (the test that all coefficients are 0), but we can also use it to test essentially any complicated function of the components of $\hat{\beta}$. For example, we might want to use this to see if the treatment effects are the same for different groups of people, or if one coefficient is bigger than another. In Q3 of PS2, we derive the distribution for the *quadratic forms*: expressions such as $x'Ax$, under the assumption that x is multivariate normal. Many test statistics look like this, so once we've got the results here we can just apply them straight away

What is a multivariate normal? Roughly, it's like a hill. It's easiest to visualise it when $x \in \mathbb{R}^2$, with the probability of seeing (x_1, x_2) on the vertical axis. The pdf looks like the figure shown. The key thing - if we 'slice' the normal in the direction of the x_1 or the x_2 axis, then the cross section looks like a 1-dimensional normal

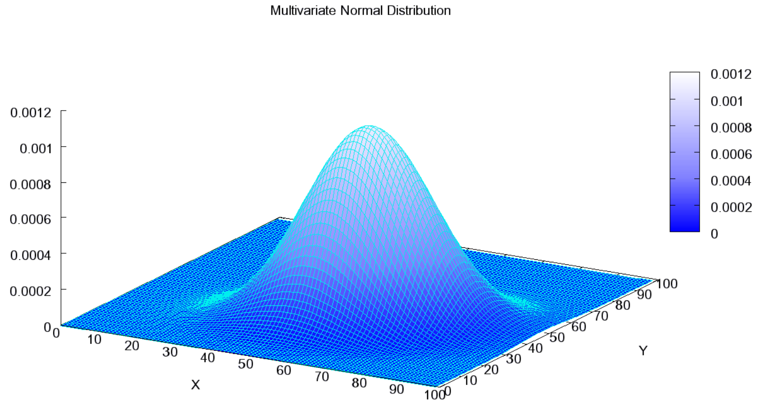

Figure 2: 2-dimensional Multivariate Normal

μ and Σ

How do we specify a particular distribution out of the normal family? In the multivariate case, we need two things, analogous to the 1-d case:

- A 'location parameter' that tells us where the middle of the normal is. If $x \in \mathbb{R}^2$, then this will also be in $\mathbb{R}^2$. Therefore, the mean of the normal is a vector μ
- A 'scale parameter' which tells us about the spread of the normal. Like in the 1-dimensional case, we represent how much the normal is spread out along each axis by the variance terms σ_1^2 and σ_2^2
- So far, all we've done is stack two normals on top of each other. However, there's another way that the normals differ in their spread: whether or not the hill is slanted to one side or

another. The easiest way to see this is by looking at the contours of the pdf: if the pdf is like a hill, then this is the cross-section formed by horizontal slices[2]. If we plot the contours of normals with different values of 'tilt' ρ, we find the following diagram. If $\rho = 0$, then the hill is symmetric on all sides. However, if $\rho \neq 0$ then the hill is tilted to one side - or to put it another way, the contours make an oval, instead of a circle. What $\rho > 0$ means is that x_1 and x_2 are positively correlated - if you see a high value of x_1, then you're likely to also see a high value of x_2. As this is another way in which we can talk about the 'spread' of a distribution, we need to include this in the parametisation of our multivariate normal

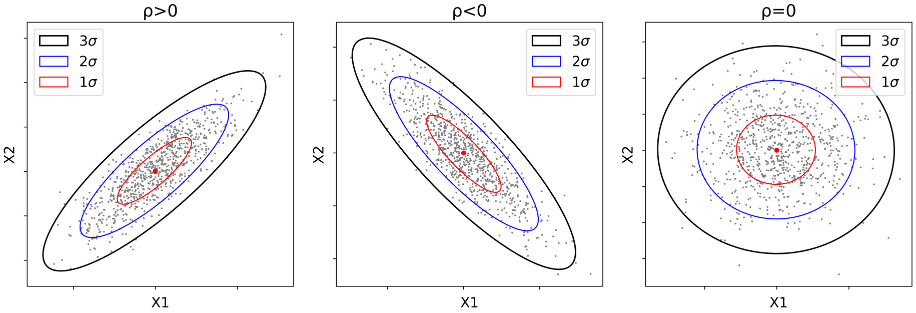

Figure 3: Contours of normals with covariance term ρ

If X is a random vector then from Lecture 1 we can define the *covariance matrix* as:

$$M := \mathbb{E}\big[(X - \mathbb{E}X)(X - \mathbb{E}X)'\big] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

which is a matrix where the diagonal terms are the variances, and the off-diagonal (ie, not on diagonal) terms are the covariances. For example, σ_1^2 is $Var(x_1)$ and $\sigma_{12} = cov(x_1, x_2)$. It turns out the ρ mentioned above is exactly the off-diagonal term σ_{12}. We only need one number to represent this tilting in $\mathbb{R}^2$ because $\sigma_{12} = \sigma_{21}$, which follows since $cov(x_1, x_2) = cov(x_2, x_1)$

Therefore if $x \in \mathbb{R}^n$ is a random vector, then we write x is normally distributed as $x \sim N(\mu, \Sigma)$. Spelling out the components in the 2d case:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right]$$

True for all Σ?

A small note is that we have conditions on what Σ can be like. We have two conditions, both of which generalise from the 1d case:

- The (i, j)-component of Σ, the entry in the i-th row and j-th column, is defined as $cov(x_i, x_j)$. This even holds on the diagonal, as $cov(x_i, x_i) = Var(x_i)$. Since $cov(x_i, x_j) = cov(x_j, x_i)$ for

[2] Contours come up in other areas of Economics too. For example, indifference curves are the contours of a utility function

any i and j, we must have that the (i, j)-th component is the same as the (j, i)-th component, and so Σ must be symmetric. This holds trivially in the 1d case, as σ^2 is just a number, so it must be the transpose of itself

- In the $1d$ case we require that $\sigma^2 \geq 0$, as negative variances don't really make sense. It's therefore not surprising that we also have that Σ must be positive semi-definite. The formal proof is only a little harder. Remember for psd we must show that for any vector $a \neq 0$, $a'\Sigma a \geq 0$. This holds because we can write:

  $$a'\mathbb{E}\big[(X - \mathbb{E}X)(X - \mathbb{E}X)'\big]a = \mathbb{E}\Big(\big[(X - \mathbb{E}X)'a\big]'\big[(X - \mathbb{E}X)'a\big]\Big) = \mathbb{E}(z'z) \geq 0$$

  since we can write $z'z = \sum z_i^2$, where $z = (X - \mathbb{E}X)'a$ is a vector of real numbers

We can use these two properties to show the existence of a matrix square root $\Sigma^{1/2}$, as we did in question 3b. Namely: the fact that V is symmetric means that it can be (orthogonally) diagonalised, and the fact that it's psd means that it has non-negative eigenvalues, so we can define the matrix:

$$D^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_n} \end{bmatrix}$$

and don't have to deal with complex numbers or anything like that when we take the square roots. Therefore it's perfectly valid to define $\Sigma^{1/2} = PD^{1/2}P^{-1}$, where (P, D) are the matrices we get from diagonalising Σ

Why do we care?

After the initial leap of moving to matrix and vector notation, it becomes a lot easier to talk about doing complicated things than if we just kept on using scalars, such as more advanced hypothesis tests. Ultimately EC221 isn't about testing your ability to do a lot of the 'pure' linear algebra exercises in PS1 and PS2. The aim is just to get you (hopefully!) a bit more fluent with the tools so that we can start doing things that are a lot more interesting for metrics itself

The other (implicit) takeaway from PS2 is that when we are trying to do hypothesis tests of vectors our test statistics usually look like quadratic forms. As a look ahead, suppose we want to test the hypothesis:

$$H_0 : \hat{\beta} = b \qquad\qquad H_1 : \hat{\beta} \neq b$$

where $\hat{\beta}$ is the OLS estimator and b is a vector. Both lie in $\mathbb{R}^k$. Suppose that we know that:

$$\hat{\beta} \sim N(\beta, V)$$

for some V, were β is the 'true' β in the DGP[3]. If we were doing a hypothesis test in one dimension, we would standardise to make a test statistic which has a distribution that we know. We do a similar

[3] We get this either by assuming that we have normally distributed errors, or we say that it should hold roughly when we have a lot of data by the central limit theorem

thing here. A first guess might be to use the 'test statistic':

$$z = V^{-1/2}(\hat{\beta} - b) \sim N(0, I)$$

The left-hand side does have this distribution under H_0, so it might be hard to see what the problem is. The problem comes in when we try to do the next step. In a one-dimensional test, we would calculate the value of the test statistic, then see if that value is greater than the critical value of the distribution on the right hand side. However, the idea of a vector z being 'greater than' another vector isn't very cleanly defined. We could do it[4], but this doesn't give us a 'nice'[5] notion of 'greater than'. Therefore we want to work with a test statistic that's a number instead of a vector, so we consider the quadratic form:

$$W = z'z = [V^{-1/2}(\hat{\beta} - b)]'V^{-1/2}(\hat{\beta} - b) = (\hat{\beta} - b)'V^{-1}(\hat{\beta} - b)$$

which is known as the **Wald statistic** (after the inventor). From the work done in PS2, we know that this has a χ^2 distribution, so we can use it for tests. In fact, the argument I just gave is basically the same as Q2 of PS2 in reverse!

[4] Say, by saying a vector is greater than another if all its components are greater than the corresponding ones of the other vector

[5] For those of you into technicalities: the word is a *complete ordering*. This is the same 'complete' we talk about in micro when we say that you have a utility function iff you have complete and transitive preferences - it means here that for any two vectors I have either $a \geq b$ or $b \geq a$. If you think hard about the definition in the previous footnote, this gives an incomplete ordering. If you don't understand this don't worry at all - it's not at all important