Normal distributions on vector spaces

Kristian Wichmann

July 2, 2017

1 Univariate normal distributions

The *standard normal distribution* is given by the density function:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{1.1}$$

This, like all density functions in this section, is understood to be with respect to the Lebesgue integral on $\mathbb{R}$. A random variable Z that follows this distribution is said to be standard normally distributed, and we write $Z \sim N(0,1)$. It can be shown that $E[Z] = 0$ and $\mathrm{Var}(Z) = 1$.

A general, univariate normal distribution with parameters μ and σ, is given by the distribution of $X = \mu + \sigma Z$. If $\sigma \neq 0$, we can invert to find $z = \frac{x-\mu}{\sigma}$. So $\frac{dz}{dx} = \frac{1}{\sigma}$. According to the usual transformation rules, x has the density function:

$$f(x) = \phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{1.2}$$

We write $X \sim N(\mu, \sigma^2)$. From the usual rules of expectation values and variances, it follows that $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

1.1 What if $\sigma = 0$?

The derivation above assumes σ to be non-zero. But even if this is the case, $X = \mu$ still has a distribution - it is simply μ all the time. However, this random variable does not have a density function with respect to the Lebesgue measure. In light of the Radon-Nikodym theorem, this is because the Lebesgue measure does not dominate the probability measure of X: $P_X(\{\mu\}) = 1, m_1(\{\mu\}) = 0$.

In the multivariate case, we will often run into similar problems.

2 Random vectors

Before tackling the multivariate case, we need some basic tools. In this section, we consider vectors of random variables. So a random vector of dimension n is:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{2.1}$$

Here, each X_i is a random variable.

2.1 Variance

The variance of a n-dimensional vector is the $n \times n$ matrix:

$$\mathrm{Var}(X) = E[(X - \mu_X)(X - \mu_X)^t] \tag{2.2}$$

Here $\mu_X = E[X]$, i.e. the vector of expectation values of the X_i's. From the usual definitions of variances and covariances between random variables, we see that the diagonal of $\mathrm{Var}(X)$ contains the variances of each X_i, while the off diagonal elements are the covariances between variables:

$$[\mathrm{Var}(X)]_{ij} = \mathrm{Cov}(X_i, X_j) \tag{2.3}$$

Due to the symmetry of covariance, this means that $\mathrm{Var}(X)$ is a symmetric matrix.

2.1.1 Variance calculation rules

Similarly to ordinary random variables, we might calculate the variance matrix as follows:

$$\begin{aligned} \mathrm{Var}(X) =& E[(X - \mu_X)(X - \mu_X)^t] = & (2.4)\\ & E(XX^t) - E(X)\mu_X^t - \mu_X E(X)^t + \mu_X \mu_X^t = & (2.5)\\ & E(XX^t) - \mu_X \mu_X^t & (2.6) \end{aligned}$$

Here, we've used the linearity of the expectation value and the definition of μ_X.

Adding a constant vector b does not change the variance, since $E[X+b] = \mu_X + b$:

$$\mathrm{Var}(X+b) = E[(X+b-(\mu_X+b))(X+b-(\mu_X+b))^t] = E[(X-\mu_X)(X-\mu_X)^t] \tag{2.7}$$

This is just the variance of X.

If A is a constant $m \times n$ matrix and X is an n-dimensional random vector, then:

$$\begin{aligned} \mathrm{Var}(AX) =& E[(AX - A\mu_X)(AX - A\mu_X)^t] = & (2.8)\\ & E[(A(X - \mu_X))(A(X - \mu_X))^t] = & (2.9)\\ & E[A(X - \mu_X)(X - \mu_X)^t A^t] = & (2.10)\\ & A[(X - \mu_X)(X - \mu_X)^t]A^t & (2.11) \end{aligned}$$

So we have $\mathrm{Var}(AX) = A\ \mathrm{Var}(X)A^t$.

2.2 Covariance

The covariance matrix between two variable vectors X and Y is:

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)^t] \tag{2.12}$$

If X has dimension m and Y dimension n, then $\mathrm{Cov}(X, Y)$ has dimension $m \times n$. Here, the matrix elements reduce to ordinary covariances between X_i's and Y_js:

$$[\mathrm{Cov}(X, Y)]_{ij} = \mathrm{Cov}(X_i, Y_j) \tag{2.13}$$

This also means, that $\mathrm{Cov}(X, Y) = (\mathrm{Cov}(Y, X))^t$

We note, that the variance could have been defined as a special case of covariance, since $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.

2.2.1 Covariance calculation rules

Similarly to the rule for variances, we have:

$$\mathrm{Cov}(X, Y) = E[XY^t] - \mu_X \mu_Y^t \tag{2.14}$$

The proof is essentially the same.

If A and B are constant matrices of appropriate dimesion, we also have:

$$\mathrm{Cov}(AX, BY) = A\ \mathrm{Cov}(X, Y)B^t \tag{2.15}$$

Again, the proof is entirely analogous to the corresponding variance formula.

The covariance is bilinear:

$$\begin{aligned} \mathrm{Cov}(X + Y, Z) =& \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z) & (2.16)\\ \mathrm{Cov}(X, Y + Z) =& \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z) & (2.17) \end{aligned}$$

This follows from the bilinearity of the ordinary covariance.

2.2.2 Addiational variance formulas

Since we noted that $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$, we may use these rules to derive further properties of variances.

For instance, the variance of a sum:

$$\begin{aligned}\mathrm{Var}(X+Y) =& \mathrm{Cov}(X+Y, X+Y) = && (2.18)\\ & \mathrm{Cov}(X,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(Y,Y) = && (2.19)\\ & \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) && (2.20)\end{aligned}$$

This mirrors the formula for the covariance of sums ordinary random variables, but is complicated by the fact that the vector covariance is not symmetric.

2.3 Quadratic forms

If X is an n-dimensional random variable and A an $n \times n$ matrix, then the corresponding quadratic form is $Q = X^t A X$. I.e. a scalar. What is the expectation value of the quadratic form? We can use a trick here. Since Q is a scalar, we can trivially write this as a trace:

$$Q = X^t A X = \mathrm{tr}(X^t A X) = \mathrm{tr}(A X X^t) \tag{2.21}$$

Here, we've used the cyclic property of traces. Now, the expectation value is:

$$E[Q] = E[\mathrm{tr}(AXX^t)] = \mathrm{tr}(E[AXX^t]) = \mathrm{tr}(A\ E[XX^t]) \tag{2.22}$$

But we know, that $\mathrm{Var}(X) = E(XX^t) - \mu_X \mu_X^t$, so $E[XX^t] = \mathrm{Var}(X) + \mu_X \mu_X^t$:

$$E[Q] = \mathrm{tr}(A(\mathrm{Var}(X) + \mu_X \mu_X^t)) = \mathrm{tr}(A\ \mathrm{Var}(X)) + \mathrm{tr}(A \mu_X \mu_X^t) \tag{2.23}$$

The last term may be rewritten:

$$\mathrm{tr}(A \mu_X \mu_X^t) = \mathrm{tr}(\mu_X^t A \mu_X) = \mu_X^t A \mu_X \tag{2.24}$$

In the last step we've used that the contents of the parenthesis is a scalar. So all in all:

$$E[X^t A X] = \mathrm{tr}(A\ \mathrm{Var}(X)) + \mu_X^t A \mu_X \tag{2.25}$$

3 Multivariate normal distributions

A *multivariate normal distribution* in n dimensions is a random vector X that has the property, that any linear combination of its elements is a univariate

normal distribution. In other words $q^t X$ should be a univariate normal for all $q \in \mathbb{R}^n$.

Specifically note, that it is not enough to require each component of the random vector to be normally distributed. This leaves room for some pathologically distributed overall distributions. However, if each of the components are also independent, X is a multivariate normal. To see this, assume that X_i is normally distributed with parameters μ_i and σ_i^2. We can now use the usual properties of normals: First we see that $q_i X_i \sim N(q_i\mu_i, q_i^2\sigma_i^2)$. And because of independence we further have:

$$q^t X = q_1 X_1 + q_2 X_2 + \ldots + q_n X_n \sim N\left(q^t\mu, \sum_{i=1}^{n} q_i^2\sigma_i^2\right) \tag{3.1}$$

3.1 Dependence vs. correlation

Further than that, it turn out that for components of a multivariate normal, independence and uncorrelation is equivalent:

Theorem 3.1. *Let X be a multivariate normally distributed and X_1 and X_2 be components of X. Then X_1 and X_2 are independent if and only if they're uncorrelated.*

Proof. The 'only if' part is true for any distributions. It's the other way that's interesting. Assume therefore, that X_1 and X_2 are uncorrelated. Recall, that the moment-generating function (MGF) of a unidimensional normal distribution Y is given by:

$$M_Y(t) = E\left[e^{tY}\right] = \exp\left(E[tY] + \frac{1}{2}\mathrm{Var}(tY)\right) \tag{3.2}$$

Consider now the bivariate distribution:

$$X' = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{3.3}$$

The MGF for X' is:

$$M_{X'}(t) = E\left[e^{t^t X'}\right] \tag{3.4}$$

Here, t is two-dimensional as well, and tY from before has been replaced by a dot product. But this is equal to:

$$E\left[\exp(\underbrace{t_1 X_1 + t_2 X_2}_{'tY'})\right] \tag{3.5}$$

Now, by the assumption that X is multivariate normal, the underbraced part must be univariate normal. So we can use (the latter half of) equation 3.2 to get:

$$E\left[e^{t^t X'}\right] = \exp\left(E[X']t + \frac{1}{2}\mathrm{Var}X'\right) \tag{3.6}$$

Now, $E[X'] = t_1\mu_1 + t_2\mu_2$, where μ_1 and μ_2 are the means of X_1 and X_2 respectively. And because the two are uncorrelated, $\mathrm{Var}X' = t_1^2\sigma_1^2 + t_2^2\sigma_2^2$. Similarly, here the sigmas are the standard deviations of the X components. So:

$$M_{X'}(t) = \exp\left[t_1\mu_1 + t_2\mu_2 + \frac{1}{2}\left(t_1^2\sigma_1^2 + t_2^2\sigma_2^2\right)\right] \tag{3.7}$$

But this can be rewritten:

$$\exp\left(t_1\mu_1 + \frac{1}{2}t_1^2\sigma_1^2\right)\exp\left(t_2\mu_1 + \frac{1}{2}t_2^2\sigma_2^2\right) = M_{X_1}(t_1)M_{X_2}(t_2) \tag{3.8}$$

Since the joint MGF is equal to the product on the individual ones, we conclude that X_1 and X_2 are independent. □

3.2 The multivariate standard normal

The *multivariate standard normal distribution* is the random vector Z that consists of n independent components, all univariate standard normally distributed. According to the previous section, this is a multivariate standard normal. Because of independence, the density function in $\mathbb{R}^n$ is simply a product of terms like equation 1.1:

$$\phi(z) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-z_i^2/2} = (2\pi)^{-n/2}e^{-||z||^2/2} \tag{3.9}$$

Here $z \in \mathbb{R}^n$, and $||\cdot||$ is the standard Euclidean norm. If a n-dimensional dimensional random vector Z follows this distribution we write $Z \sim N(0, I_n)$, where I_n is the identity matrix in n dimensions. The reason for this is, that the variance matrix of X is equal to I_n.

3.3 Regular and singular distributions

If a multivariate normal X has a non-singular variance matrix, we call the distribution *regular*. Is this is not the case, the distribution is called *singular*.

4 Affine transformations of euclidean spaces

In order to get to the general, multidimensional normal distribution, we need yet another component:

Let $s : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. This means that the is an $m \times n$ matrix A so $s(x) = Ax$.

An *affine* transformation t is formed by following this linear map by a translation:

$$t : \mathbb{R}^n \to \mathbb{R}^m, t(x) = Ax + v \tag{4.1}$$

Here, $v \in \mathbb{R}^m$. Since translations are always bijective, we note that t is bijective iff A is invertible.

Each component of an affine transformation is composed from measurable function - is is understood that we mean with respect to the Borel algebras of each space) - so the affine transformation itself is measurable as well.

4.1 Transformation properties of the Lebesgue measure

Recall that the Lebesgue measure in n dimensions m_n is invariant under translation: If t is a translation $t : \mathbb{R}^n \to \mathbb{R}^n, t(x) = x + x_0$, where $x_0 \in \mathbb{R}^n$ then:

$$t(m_n) = m_n \tag{4.2}$$

Also, if $s : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto Ax$ is an isomorphism, then:

$$s(m_n) = m_n |\det A^{-1}| \tag{4.3}$$

Combining the two, the formula for affine transformation is the same as for linear ones.

5 Affine transformations of multivariate normal distributions

Let's start by considering the following basic fact:

Theorem 5.1. *Let X be a multivariate normal distribution in n dimensions. Consider an affine transformation $Y = AX+b$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then Y is a multivariate normal distribution in m dimensions.*

Proof. Consider a linear combination of components of Y:

$$q^tY = q^t(AX+b) = \sum_{i,j=1}^{m} q_i\left(A_{ij}X_j + b_i\right) = \underbrace{\sum_{i,j=1}^{m} q_iA_{ij}X_j}\ + q^tb \tag{5.1}$$

The underbraced part is a linear combination of components of X and hence univariate normal. Adding the constant term q^tb doesn't change this fact. Hence, Y is a multivariate normal. □

Using our knowledge of random vectors we can see how such an affine transformation changes the mean and variance:

Theorem 5.2. *If X is a random vector with mean μ and variance Σ, then the mean and variance of $Y = AX + b$ are $A\mu + b$ and $A\Sigma A^t$ respectively.*

Proof. We need to prove both:

- Mean: $E[Y] = E[AX + b] = E[AX] + b = AE[X] + b = A\mu + b$.
- Variance: $\mathrm{Var}(Y) = \mathrm{Var}(AX + b) = \mathrm{Var}(AX) = A\mathrm{Var}(X)A^t = A\Sigma A^t$.

□

How does such a transformation affect the regularity of the distribution?

Theorem 5.3. *Assume X is an n-dimensional regular normal distribution. If $A \in \mathbb{R}^{m\times n}$ has rank m, then $Y = AX + b$ is also regular. Conversely, if Y is regular, then A has rank m.*

Proof. Since b has no bearing on the variance, we need to show that the variance matrix of AX is invertible. The variance of AX is equal to $A\Sigma A^t$, where Σ is the variance of X. Since both A and Σ have full rank, so does $A\Sigma$. So the total rank is the rank of A^t, which is m.

Conversely, assume Y to be regular, i.e. $A\Sigma A^t$ has rank m. But, the rank of a product is greater than or equal to the minimum rank of the factors. So $\min\{\mathrm{rank}A, \mathrm{rank}\Sigma\} \geq m$. But then $\mathrm{rank}A \geq m$, which is only possible when the equality holds. □

5.1 Affine transformation of standard univariate normal

Given an n-dimension random vector $Z \sim N(0, I_n)$, we consider the random variable $X = \mu + AZ$, where $A \in \mathbb{R}^{m\times n}$. From the previous section, we now know that X is a multivariate normal with:

$$E(X) = \mu, \quad \mathrm{Var}(X) = AA^t = \Sigma \tag{5.2}$$

Here, as usual Σ is the variance matrix of X[1].

A plays the roles that σ played for the univariate case, and hence we might expect that special care needs to be taken when it is "zero", which will turn out to mean "non-surjective" in the multivariate case in accordance with theorem 5.3.

5.1.1 Invertible A

So we need for A to be surjective in order to get a density function with respect to the Lebesgue measure. For now, let's assume that $m = n$, and that A is also injective and therefore invertible. Then we can solve for Z:

$$Z = A^{-1}(X - \mu) \tag{5.3}$$

We may now write

$$||z||^2 = z^t z = \left(A^{-1}(x-\mu)\right)^t \left(A^{-1}(x-\mu)\right) = (x-\mu)^t \left(A^{-1}\right)^t A^{-1}(x-\mu) \tag{5.4}$$

According to equation 4.3 the density function for X is:

$$f(x) = (2\pi)^{-n/2}\det(A^{-1}) \exp\left[-\frac{1}{2}(x-\mu)^t \left(A^{-1}\right)^t A^{-1}(x-\mu)\right] \tag{5.5}$$

But the transpose of an inverse is the inverse of a transpose:

$$\left(A^{-1}\right)^t A^{-1} = (A^t)^{-1}A^{-1} = (AA^t)^{-1} = \Sigma^{-1} \tag{5.6}$$

So the density function is:

$$f(x) = (2\pi)^{-n/2}\det(A^{-1}) \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right] \tag{5.7}$$

We may express this solely in terms of Σ, since $\det \Sigma = \det AA^t = (\det A)^2$. So $\det(A^{-1}) = (\det \Sigma)^{-\frac{1}{2}}$. One handy way to express this is:

$$f(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right] \tag{5.8}$$

This closely mirrors the univariate formula.

[1]Since Σ corresponds to σ^2 in the univariate case, Σ^2 would in some sense have been a more logical naming choice, but such is tradition.

5.1.2 Non-invertible A

If A is not surjective, X does not have a density with respect to the Lebesgue measure in $\mathbb{R}^m$. This is because the rank of A is less than m, and so X only takes on values in an affine, proper subspace of $\mathbb{R}^m$. The m-dimensional Lebesgue measure of such a space is zero.

5.1.3 Single value decomposition

However, A may still be surjective but not invertible. How to calculate the density function in this case? To do so, it is useful to rewrite A using *single value decomposition*. I.e. if A is a $m \times n$ matrix, then there exists orthogonal matrices $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ such that:

$$A = UDV^t \tag{5.9}$$

Here D is a diagonal (but generally not square) matrix with non-negative diagonal entries, the eigenvalues of AA^t. These are known as the *singular values* of A. The number of non-zero singular values is equal to the rank of A.

5.2 A from variance Σ

Often, we want to specify a univariate normal from a variance Σ instead of a transformation matrix A. As we saw above, we only need Σ to write down the density (if it exists), but what if we want A?

To do this, we note that Σ is symmetric and positive semi-definite because it is a covariance matrix. So there is an orthogonal matrix O such that $\Sigma = ODO^t$, where D is a diagonal matrix with non-negative diagonal entries. Hence, we can construct another diagonal matrix $D^{\frac{1}{2}}$ where the entries are the square roots of the ones in D. Now we obviously have:

$$\Sigma = ODO^t = \underbrace{OD^{\frac{1}{2}}}_{A}\underbrace{D^{\frac{1}{2}}O^t}_{A^t} \tag{5.10}$$

By setting $A = OD^{\frac{1}{2}}$ we have achieved a decomposition of Σ that will bring about all the results above (though this is not necessarily the only one). A is sometimes written as the "square root of Σ": $A = \Sigma^{\frac{1}{2}}$.

5.2.1 Sphering

One use of the decomposition $A = OD^{\frac{1}{2}}$ is to gain geometrical insight into multivariate Gaussians by what is called *sphering*.

How can we describe the linear map A? It is a composite of the map $D^{\frac{1}{2}}$ followed by O. $D^{\frac{1}{2}}$ is a pure scaling of each axis. If none of the diagonal entries are zero - i.e. if Σ is positive-definite - this turns an origin-centered sphere (which is a contour curve of a multivariate standard normal) into an ellipsoid. This is followed by an orthogonal operation, which means that distance is preserved. The ellipsoid is now rotated and/or reflected around the origin. These are the contour curves of a regular multidimensional Gaussian. If the distribution is singular, one or more dimensions are "squashed out" - we get an elliptic "pancake" instead of an ellipsoid.

6 Multivariate normals in block form

6.1 Definition

Consider a multivariate normal on $\mathbb{R}^n$. As shown above, we can characterize such a distribution by a vector of means $\xi \in \mathbb{R}^n$ and a symmetric variance matrix $\Sigma \in \mathbb{R}^{n \times n}$. We now split ξ and Σ into *block form* as follows:

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{6.1}$$

Here $\xi_1 \in \mathbb{R}^{n_1+}, \xi_2 \in \mathbb{R}^{n_2+}$ and $\Sigma_{11} \in \mathbb{R}^{n_1 \times n_1}, \Sigma_{12} \in \mathbb{R}^{n_1 \times n_2}, \Sigma_{21} \in \mathbb{R}^{n_2 \times n_1}$, and $\Sigma_{22} \in \mathbb{R}^{n_2 \times n_2}$. The transpose of Σ is:

$$\Sigma^t = \begin{pmatrix} \Sigma_{11}^t & \Sigma_{21}^t \\ \Sigma_{12}^t & \Sigma_{22}^t \end{pmatrix} \tag{6.2}$$

From this, we immediately see, that since Σ is symmetric, we have:

$$\Sigma_{11} = \Sigma_{11}^t, \Sigma_{12} = \Sigma_{21}^t, \Sigma_{22} = \Sigma_2^t \tag{6.3}$$

6.2 Results about multivariate normal distributions in block form

Theorem 6.1. *Let X_1 and X_2 be random vectors in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively. Now assume:*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \tag{6.4}$$

Then $X_1 \sim N(\xi_1, \Sigma_{11})$ and $X_2 \sim N(\xi_2, \Sigma_{22})$.

Proof. It is trivial, that both X_1 and X_2 are multivariate normal, since all linear combinations of components from each must be univariate normal because of the assumption of the block vector being multivariate normal.

Now consider the following $n_1 \times n$ block matrix:

$$\begin{pmatrix} I_1 & 0 \end{pmatrix} \tag{6.5}$$

Here I_1 is the $n_1 \times n_1$ unit matrix. Now we may express X_1 as follows:

$$X_1 = \begin{pmatrix} I_1 & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{6.6}$$

The mean must then be:

$$\begin{pmatrix} I_1 & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \xi_1 \tag{6.7}$$

And the variance:

$$\begin{pmatrix} I_1 & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_1 \\ 0 \end{pmatrix} = \Sigma_{11} \tag{6.8}$$

Similarly for X_2. □

Theorem 6.2. *As above, let:*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \tag{6.9}$$

Now assume Σ_{22} to be invertible. Then the conditional distribution of X_1 given X_2 is:

$$X_1|_{X_2=x_2} \sim N(\xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{6.10}$$

Proof. Consider the following random vector:

$$\begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} \tag{6.11}$$

Call the upper block random vector Z:

$$Z = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \tag{6.12}$$

Now, we know that the random vector $\begin{pmatrix} Z \\ X_2 \end{pmatrix}$ must be multivariate normal, as it's a linear transform of the multivariate normal $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Let us calculate the mean:

$$E\left[\begin{pmatrix} Z \\ X_2 \end{pmatrix} \right] = \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \xi_1 - \Sigma_{12}\Sigma_{22}^{-1}\xi_2 \\ \xi_2 \end{pmatrix} \tag{6.13}$$

And the variance:

$$\begin{aligned}
\mathrm{Var}\left(\begin{pmatrix} Z \\ X_2 \end{pmatrix}\right) &= \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix}^t = && (6.14) \\
& \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_1 & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_2 \end{pmatrix} = && (6.15) \\
& \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ \Sigma_{21} - \Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{22} \end{pmatrix} = && (6.16) \\
& \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ 0 & \Sigma_{22} \end{pmatrix} = && (6.17) \\
& \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} \\ 0 & \Sigma_{22} \end{pmatrix} = && (6.18) \\
& \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} && (6.19)
\end{aligned}$$

Since the off-diagonal elements are all zero, we see that all components of Z are uncorrelated with all components of X_2. Since the composite block vector is multinormally distributed, it follows that they are also independent. Hence, the conditional distribution of Z given X_2 is simply the corresponding marginal distribution:

$$Z|_{X_2=x_2} \sim N(\xi_1 - \Sigma_{12}\Sigma_{22}^{-1}\xi_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{6.20}$$

Here, we've utilized theorem 6.1. But remember, that $Z = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$, or equivalently that $X_1 = Z + \Sigma_{12}\Sigma_{22}^{-1}X_2$. So:

$$X_1|_{X_2=x_2} = Z|_{X_2=X_2} + (\Sigma_{12}\Sigma_{22}^{-1}X_2)|_{X=X_2} = Z|_{X_2=x_2} + \Sigma_{12}\Sigma_{22}^{-1}x_2 \tag{6.21}$$

This simply means a shift of the mean, so we end up with:

$$X_1|_{X_2=x_2} \sim N(\xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{6.22}$$

□

7 Formulation in terms of precision

7.1 Inner products on $\mathbb{R}^n$

An inner product can be seen as a map $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. By the inner product axioms, this map must be linear in both variables, and so it can be expressed as a quadratic form:

$$\langle x, y \rangle = x^t B y \tag{7.1}$$

Here B is a $n \times n$ matrix. By symmetry of the inner product, $\langle e_i, e_j \rangle = \langle e_j, e_i \rangle$ implying $B_{ij} = B_{ji}$, so B is symmetric. Since $\langle x, x \rangle = x^t B x \geq 0$, B is semi-positive definite. Since the equality sign only holds when $x = 0$, B is also positive definite.

We will refer to such an inner product as a *precision*.

7.2 Density function

Now, we may rewrite equation 5.8 as follows:

$$f(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left[-\frac{1}{2}||x-\mu||_\Sigma^2\right] \tag{7.2}$$

Here, $||\cdot||_\Sigma$ is the norm induced by using the inner product as described above, using Σ as B. So for regular multivariate normals, specifying a variance matrix or a precision are two different ways to express the same thing.

8 Orthogonal complement

We now wish to develop a theory of normal distributions on arbitrary inner product spaces, not just Euclidean ones. We will do this in a number of steps.

Let V be a finite-dimensional vector space with an inner product $\langle \cdot, \cdot \rangle$. Let U be a subspace of V. Then we define the *orthogonal complement* of U as:

$$U^\perp = \{v \in V | \forall u \in U : \langle u, v \rangle = 0\} \tag{8.1}$$

Theorem 8.1. *$U^\perp$ is a subspace of V.*

Proof. According to the subspace theorem, we need to show three things:

- $U^\perp$ is not empty: Clearly $0 \in U^\perp$.
- Closed under addition: If $v_1, v_2 \in U^\perp$, then for all $u \in U^\perp$:

$$\langle v_1 + v_2, u \rangle = \langle v_1, u \rangle + \langle v_2, u \rangle = 0 \tag{8.2}$$

- Closed under scalar multiplication: If $v \in U^\perp$ and $c \in \mathbb{R}$ then for all $u \in U^\perp$:

$$\langle cv, u \rangle = c\langle v, u \rangle = 0 \tag{8.3}$$

□

Since the only vector perpendicular to itself is 0, we further conclude that $U \cap U^\perp = \{0\}$.

Theorem 8.2. *If $e_1, e_2, \ldots, e_m$ is an orthonormal basis for U, then for any $v \in V$:*

$$v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i \in U^\perp \tag{8.4}$$

Proof. Let $u \in U$. Then we can write $u = \sum_{j=1}^{m} \lambda_j e_j$ for some coefficients λ_j. Now calculate the inner product with the vector above:

$$\langle v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i \ , \ \sum_{j=1}^{m} \lambda_j e_j \rangle = \sum_{i=j}^{m} \lambda_j \langle v, e_j \rangle - \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_j \langle v, e_i \rangle \langle e_i, e_j \rangle \tag{8.5}$$

Since $\langle e_i, e_j \rangle = \delta_{ij}$ this vanishes. □

This means that we may write any $v \in V$ as a sum of vectors from U and $U^\perp$ respectively:

$$v = \underbrace{\sum_{i=1}^{m} \langle v, e_i \rangle e_i}_{\in U} \ + \ \underbrace{v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i}_{\in U^\perp} \tag{8.6}$$

Theorem 8.3. *The decomposition into elements from U and $U^\perp$ from equation 8.6 is unique.*

Proof. Let $v = u_1 + u_1^\perp$ and $v = u_2 + u_2^\perp$ be two such decompositions. Then $u_1 + u_1^\perp = u_2 + u_2^\perp$ and hence $u_1 - u_2 = u_2^\perp - u_1^\perp$. But this means that this vector is a member of both U and $U^\perp$, and hence it must be 0. This means $u_1 = u_2$ and $u_1^\perp = u_2^\perp$. □

8.1 The orthogonal projection

The previous section motivates the following:

Definition 8.1. *Let V be a finite-dimensional inner product vector space and U a subspace of V. The orthogonal projection from V onto U is the map $p : V \to V$ which satisfies:*

$$\forall v \in V : \quad p(v) \in U, \quad v - p(v) \in U^\perp \tag{8.7}$$

As we see, one could also define the co-domain of p to be U. Usually, the distinction will not matter much.

Theorem 8.4. *The orthogonal projection operator is linear.*

Proof. We need to show additivity and homogeneity:

- Additivity: Let $v_1, v_2 \in V$. Then $p(v_1) + p(v_2) \in U$ and:

$$v_1 - p(v_1) + v_2 - p(v_2) = v_1 + v_2 - (p(v_1) + p(v_2)) \in U^\perp \tag{8.8}$$

  Adding the two we get $v_1 + v_2$. So $p(v_1 + v_2) = p(v_1) + p(v_2)$.

- Homogeneity. Let $v \in V$ and $c \in \mathbb{R}$. Then $cp(v) \in U$ and $c(v - p(v)) = cv - cp(v) \in U^\perp$. Adding the two we get cv, so $p(cv) = cp(v)$.

□

Theorem 8.5. *The orthogonal projection operator $p : V \to V$ is idempotent. I.e. $p \circ p = p$.*

Proof. Let $v \in V$. Then $p(v) \in U$. But this means that the decomposition of $p(v)$ is $p(v) + 0$. So $p \circ p(v) = p(v)$. □

8.2 The decomposition theorem

The following theorem is very important in the theory of the general linear model. It (or variations/corollaries of it) is sometimes known as Cochran's theorem.

Theorem 8.6. *Let $V = \mathbb{R}^n$ be a vector space with the usual inner product. Let U be an m-dimensional subspace of V and p_U the associated orthogonal projection. Let X be a regularly normally distributed random vector on V with mean ξ and variance Σ. Then the following is true:*

- *The random vectors $p_U(X)$ and $X - p_U(X)$ are independent.*
- *$p_U(X)$ is normally distributed with mean $p_U\xi$ and a variance matrix of rank m.*
- *Similarly, $X - p_U(X)$ is normally distributed with mean $(1 - p_U)\xi$ and a variance matrix with rank $n - m$.*

Proof. Let $e_1, \cdots, e_m$ be a basis for U. Expand to a basis for V: $e_1, \ldots, e_n$. Now, the projection operator applied to such a basis vector is:

$$p_U e_i = \begin{cases} e_i \text{ for } i \leq m \\ 0 \text{ otherwise} \end{cases} \tag{8.9}$$

This means that:

$$(1-p_U)e_i = \begin{cases} 0 \text{ for } i \leq m \\ e_i \text{ otherwise} \end{cases} \tag{8.10}$$

The random vectors $p_U(X)$ and $X - p_U(X) = (1-p_U)X$ are clearly both normally distributed, being linear transformations of the normally distributed X. That the means are $p_U\xi$ and $(1-p_U)\xi$ is equally trivial from previous results. But to check that they are independent, we need to consider the covariance matrix:

$$C = \text{Cov}(p_U X, (1-p_U)X) = p_U \text{Cov}(X,X)(1-p_U)^t = p_U \text{Var}(X)(1-p_U)^t \tag{8.11}$$

But p_U is symmetric, and the variance of X is Σ by definition. So we end up with $p_U\Sigma(1-p_U)$. Since the variance matrix is symmetric, we could also write this as its transpose:

$$C = p_U\Sigma(1-p_U) = (1-p_U)\Sigma p_U \tag{8.12}$$

Now consider this covariance matrix acting on a basis vector e_i:

$$Ce_i = p_U\Sigma(1-p_U)e_i = (1-p_U)\Sigma p_U e_i \tag{8.13}$$

If $i \leq m$, equation 8.9 shows that this is zero. On the other hand, if $i > m$, equation 8.10 shows that it is zero. Since the e_i forms a basis, the only possibility is, that $C = 0$. So the two random vectors are uncorrelated, and as normals this means they are also independent.

Now, consider the variance matrix of $p_U(X)$:

$$\Sigma_U = p_U \Sigma p_U^t = p_U \Sigma p_U \tag{8.14}$$

From equations 8.9 and 8.10 we know that p_U has rank m. Now we need to show that the same is true for Σ_U. We will do this by showing them to have the same kernel. If $p_U x = 0$ it immediately follows that $\Sigma_U x = 0$. In other words $\ker(p_U) \subseteq \ker(\Sigma_U)$. Now, conversely assume $\Sigma_U x = 0$. This means:

$$p_U \Sigma (p_U x) = 0 \tag{8.15}$$

Now multiply both sides by x^t:

$$x^t p_U \Sigma (p_U x) = 0 \Leftrightarrow (p_U x)^t \Sigma (p_U x) = 0 \tag{8.16}$$

Since Σ is full rank and positive definite, this can be seen as a norm $||p_U x||^2 = 0$. But a norm is only zero when the vectors is, so $p_U x = 0$. Hence by the rank-nullity theorem, Σ_U has rank m.

The result for $X - p(X)$ follows by noting that it is the orthogonal complement of $p_(X)$, and hence nullity and rank are reversed. □

9 Lebesgue measures on vector spaces

9.1 Coordinate maps

Let V be a finite-dimensional vector space of dimension n. Our question is, if we can turn V into a measure space in a natural way. Since we know that V is isomorphic to $\mathbb{R}^n$, it makes sense to tweak the usual Lebesgue measure in N dimensions:

Let $e_1, e_2, \ldots, e_n$ be a basis for V. Then we can define the *coordinate map* as follows:

$$\phi : \mathbb{R}^n \to V, \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \sum_{i=1}^{n} x_i e_i \tag{9.1}$$

This is obviously an isomorphism. Specifically, it is invertible with inverse $\phi^{-1} : V \to \mathbb{R}^n$.

The coordinate map depends on the chosen basis. If we had chosen another basis $e_1^*, e_2^*, \ldots, e_n^*$ we would get another isomorphism ϕ^*.

9.2 Borel algebra on V

We can now use ϕ^{-1} to induce a σ-algebra on V. Set $\mathbb{B}_V$ to the smallest σ-algebra that makes ϕ^{-1} measurable when $\mathbb{R}^n$ is equipped with the Borel algebra $\mathbb{B}_n$. We call $\mathbb{B}_V$ the *Borel algebra on* V.

At first this object seems to depend of the choice of basis for V. But it turns out that the use of definite article in the definition is justified:

Theorem 9.1. *If $e_1, e_2, \ldots, e_n$ and $e_1^*, e_2^*, \ldots, e_n^*$ are bases for V, then the induced σ-algebra $\mathbb{B}_V$ and $\mathbb{B}_V^*$ is the same thing.*

Proof. We know that ϕ^{-1} is $\mathbb{B}_V - \mathbb{B}_n$ measurable by definition. We have:

$$(\phi^*)^{-1} = (\phi^*)^{-1} \circ \mathrm{id}_V = (\phi^*)^{-1} \circ (\phi \circ \phi^{-1}) = \big((\phi^*)^{-1} \circ \phi\big) \circ \phi^{-1} \tag{9.2}$$

$((\phi^*)^{-1} \circ \phi)$ is a linear operator on $\mathbb{R}^n$ and so according to section 4.1 is measurable. So $(\phi^*)^{-1}$ must be $\mathbb{B}_V - \mathbb{B}_n$-measurable. Since $\mathbb{B}_V^*$ is the smallest σ-algebra to make $(\phi^*)^{-1}$ $\mathbb{B}_V - \mathbb{B}_n$-measurable, we must have $\mathbb{B}_V^* \subseteq \mathbb{B}_V$.

But by a totally symmetric argument, we must also have $\mathbb{B}_V \subseteq \mathbb{B}_V^*$. Hence $\mathbb{B}_V = \mathbb{B}_V^*$. □

It turns out, that ϕ must be measurable too. This is a direct consequence of the pipeline lemma.

Theorem 9.2. *Given two finite-dimensional vector spaces V and W, then:*

$$\mathbb{B}_{V\times W} = \mathbb{B}_V \otimes \mathbb{B}_W \tag{9.3}$$

Proof. Let $e_1, e_2, \ldots, e_n$ be a basis for V with corresponding coordinate map ϕ. And $f_1, f_2, \ldots, f_m$ a basis for W with corresponding coordinate map ψ. Then $(e_1, 0), (e_2, 0), \ldots, (e_n, 0), (0, f_1), (0, f_2), \ldots, (0, f_m)$ is a basis for $V \times W$. The corresponding coordinate map is:

$$\phi \times \psi : (x_1, x_2, \ldots, x_{n+m}) \mapsto \left(\sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{m} x_{n+j} f_j \right) \tag{9.4}$$

The inverse is $(\phi \times \psi)^{-1} = \phi^{-1} \times \psi^{-1}$. Since ϕ^{-1} is $\mathbb{B}_V - \mathbb{B}_n$-measurable and ψ^{-1} is $\mathbb{B}_W - \mathbb{B}_m$-measurable, $\phi^{-1} \times \psi^{-1}$ must be $\mathbb{B}_V \otimes \mathbb{B}_W - \mathbb{B}_n \otimes \mathbb{B}_n$-measurable. But $\mathbb{B}_n \otimes \mathbb{B}_m = \mathbb{B}_{n+m}$. Since $\mathbb{B}_{V\times W}$ is the smallest σ-algebra to make $\phi^{-1} \times \psi^{-1}$ measurable, we must have $\mathbb{B}_{V\times W} \subseteq \mathbb{B}_V \otimes \mathbb{B}_W$.

On the other hand, consider the projection operators:

$$\pi_V : V \times W \to V, (v, w) \mapsto v \tag{9.5}$$
$$\pi_n : \mathbb{R}^{n+m} \to \mathbb{R}^n, (x_1, \ldots, x_{n+m}) \mapsto (x_1, \ldots, x_n) \tag{9.6}$$

Now consider $\pi_V \circ (\phi \times \psi)$. Applied to an $x \in \mathbb{R}^{n+m}$ we have:

$$\pi_V \circ (\phi \times \psi)(x) = \pi_v \left(\left(\sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{m} x_{n+j} f_j \right) \right) = \sum_{i=1}^{n} x_i e_i \tag{9.7}$$

But this is the same as:

$$\phi \circ \pi_1(x) = \phi\left((x_1, \ldots x_n)\right) = \sum_{i=1}^{n} x_i e_i \tag{9.8}$$

So $\pi_V \circ (\phi \times \psi) = \phi \circ \pi_1$. Now apply $\phi^{-1} \times \psi^{-1}$ from the right to get:

$$\pi_V = \phi \circ \pi_1 \circ (\phi^{-1} \times \psi^{-1}) \tag{9.9}$$

Since all the three functions on the right side are measurable, π_V must be $\mathbb{B}_{V\times W} - \mathbb{B}_V$-measurable. By a similar argument the corresponding projection operator $\pi_W : V \times W \to W$ is $\mathbb{B}_{V\times W} - \mathbb{B}_W$-measurable. Since $\mathbb{B}_V \otimes \mathbb{B}_W$ is the smallest σ-algebra to make both π_V and π_W measurable, we must have: $\mathbb{B}_V \otimes \mathbb{B}_W \subseteq \mathbb{B}_{V\times W}$. □

9.3 Lebesgue measures on V

We now want to define a measure on the measurable space $(V, \mathbb{B}_V)$. If $e_1, e_2, \ldots, e_n$ is a basis for V, we will use the associated coordinate map ϕ to define a measure:

$$\lambda_V = \phi(m_n) \tag{9.10}$$

Here, m_n is the usual Lebesgue measure in n dimensions. The problem is, that this measure depends on the chosen basis! Consider another basis $e_1^*, e_2^*, \ldots, e_n^*$ and associated coordinate map ϕ^*. Then the measure is:

$$\lambda_V^* = \phi^*(m_n) = (\phi \circ \phi^{-1}) \circ \phi^*(m_n) = \phi \circ (\phi^{-1} \circ \phi^*(m_n)) \tag{9.11}$$

Now $\phi^{-1} \circ \phi^*$ is an isomorphism $\mathbb{R}^n \to \mathbb{R}^n$, so according to section 4.1, there is a constant c such that $(\phi^{-1} \circ \phi^*(m_n)) = cm_n$. So:

$$\lambda_V^* = c\phi(m_n) = c\lambda_V \tag{9.12}$$

So while there are many Lebesgue measures on V they only differ from each other by a constant factor. This means that they all agree on what constitutes a null set, and on which functions are integrable. They disagree on the integral, but agree on whether it is finite or not. The also agree on whether a measure μ has a density with respect to λ_V or not.

10 Regular normal distributions on V

10.1 Gaussian integrals on V

Let V be an inner product space of finite dimension n. Let $a_1, a_2, \ldots, a_n$ be an orthogonal basis for V, $\phi : \mathbb{R}^n \to V$ be the corresponding coordinate transformation, and λ be the corresponding Lebesgue measure on V. We're now interested in evaluating the following integral:

$$\int_V \exp\left[-\frac{1}{2}||v||^2\right] d\lambda(v) \tag{10.1}$$

Let's start by writing a general vector v as $v = \sum_{i=1}^n x_i a_i$. Then:

$$||v||^2 = \langle v, v\rangle = \langle \sum_{i=1}^n x_i a_i, \sum_{j=1}^n x_j a_j\rangle = \sum_{i=1}^n \sum_{j=1}^n x_i x_j \langle a_i, a_j\rangle \tag{10.2}$$

Since all the a_i's are mutually orthogonal this means:

$$||v||^2 = \sum_{i=1}^n x_i^2 ||a_i||^2 \tag{10.3}$$

So we may write:

$$\int_V \exp\left[-\frac{1}{2}||v||^2\right] d\lambda(v) = \int_{\mathbb{R}^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2||a_i||^2\right] dm_n(x) \tag{10.4}$$

Using Tonelli's theorem this is equal to:

$$\prod_{i=1}^{n} \int_{\mathbb{R}} \exp\left[-\frac{1}{2}x_i^2||a_i||^2\right] dm_1(x_i) = \prod_{i=1}^{n} \frac{\sqrt{2\pi}}{||a_i||} = \frac{(2\pi)^{n/2}}{\prod_{i=1}^{n}||a_i||} \tag{10.5}$$