Normal distributions on vector spaces

Kristian Wichmann

November 29, 2016

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 $Z \sim N(0,1)$.

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 $X \sim N(\mu, \sigma^2)$.

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 distribution

The standard normal distribution in n dimensions is the distribution of n independent standard normals. Hence, the density function in $\mathbb{R}^n$ is simply

a product of terms like equation 1.1:

$$\phi(z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = (2\pi)^{-n/2} e^{||z||^2/2} \tag{3.1}$$

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 will soon be clear.

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 General multivariate normal distribution

Given an n-dimension random vector $Z \sim N(0, I_n)$, we now define a general standard normal as the distribution of the random variable $X = \mu + \Sigma^{\frac{1}{2}} Z$.

Some notes are in order, before we procees. For now "$\Sigma^{\frac{1}{2}}$" should simply be considered a matrix of dimensions $n \times p$, not necessarily a square root of anything (we will get back to that later). It plays the roles that σ played for the univariate case[1], and hence we might expect that special care needs to be taken when it is "zero", whatever that might signify here.

6 Orthogonal complement

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:

$$U^{\perp} = \{v \in V | \forall u \in U : \langle u, v \rangle = 0\} \tag{6.1}$$

Theorem 6.1. *$U^{\perp}$ 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 $u \in U^{\perp}$:

$$\langle v_1 + v_2, u \rangle = \langle v_1, u \rangle + \langle v_2, u \rangle = 0 \tag{6.2}$$

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

$$\langle cv, u \rangle = c\langle v, u \rangle = 0 \tag{6.3}$$

□

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

Theorem 6.2. *If $e_1, e_2, \ldots, e_m$ is an orthonormal basis for U, then for any $v \in V$:*

$$v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i \in U^{\perp} \tag{6.4}$$

[1] Hence, simply calling it Σ would be more be clearer, but such is the tradition notation.

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:

$$\langle v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i \ , \ \sum_{j=1}^{m} \lambda_j e_j \rangle = \sum_{i=j}^{m} \lambda_j \langle v, e_j \rangle - \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_j \langle v, e_i \rangle \langle e_i, e_j \rangle \quad (6.5)$$

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:

$$v = \underbrace{\sum_{i=1}^{m} \langle v, e_i \rangle e_i}_{\in U} \ + \ \underbrace{v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i}_{\in U^\perp} \quad (6.6)$$

Theorem 6.3. *The decomposition into elements from U and $U^\perp$ from equation 6.6 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 $u_1^\perp = u_2^\perp$. □

6.1 The orthogonal projection

The previous section motivates the following:

Definition 6.1. *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:*

$$\forall v \in V : \quad p(v) \in U, \quad v - p(v) \in U^\perp \quad (6.7)$$

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

Theorem 6.4. *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:

$$v_1 - p(v_1) + v_2 - p(v_2) = v_1 + v_2 - (p(v_1) + p(v_2)) \in U^\perp \quad (6.8)$$

  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 6.5. *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 $p \circ p(v) = p(v)$. □

7 Lebesgue measures on vector spaces

7.1 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:

$$\phi : \mathbb{R}^n \to V, \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \sum_{i=1}^{n} x_i e_i \tag{7.1}$$

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^*.

7.2 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 7.1. *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:

$$(\phi^*)^{-1} = (\phi^*)^{-1} \circ \mathrm{id}_V = (\phi^*)^{-1} \circ (\phi \circ \phi^{-1}) = \left((\phi^*)^{-1} \circ \phi\right) \circ \phi^{-1} \tag{7.2}$$

$((\phi^*)^{-1} \circ \phi)$ 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 7.2. *Given two finite-dimensional vector spaces V and W, then:*

$$\mathbb{B}_{V\times W} = \mathbb{B}_V \otimes \mathbb{B}_W \tag{7.3}$$

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:

$$\phi \times \psi : (x_1, x_2, \ldots, x_{n+m}) \mapsto \left(\sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{m} x_{n+j} f_j\right) \tag{7.4}$$

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:

$$\pi_V : V \times W \to V, (v, w) \mapsto v \tag{7.5}$$

$$\pi_n : \mathbb{R}^{n+m} \to \mathbb{R}^n, (x_1, \ldots, x_{n+m}) \mapsto (x_1, \ldots, x_n) \tag{7.6}$$

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

$$\pi_V \circ (\phi \times \psi)(x) = \pi_v \left(\left(\sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{m} x_{n+j} f_j\right)\right) = \sum_{i=1}^{n} x_i e_i \tag{7.7}$$

But this is the same as:

$$\phi \circ \pi_1(x) = \phi\left((x_1, \ldots x_n)\right) = \sum_{i=1}^{n} x_i e_i \tag{7.8}$$

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

$$\pi_V = \phi \circ \pi_1 \circ (\phi^{-1} \times \psi^{-1}) \tag{7.9}$$

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: $\mathbb{B}_V \otimes \mathbb{B}_W \subseteq \mathbb{B}_{V\times W}$. □

7.3 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:

$$\lambda_V = \phi(m_n) \tag{7.10}$$

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:

$$\lambda_V^* = \phi^*(m_n) = (\phi \circ \phi^{-1}) \circ \phi^*(m_n) = \phi \circ (\phi^{-1} \circ \phi^*(m_n)) \tag{7.11}$$

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:

$$\lambda_V^* = c\phi(m_n) = c\lambda_V \tag{7.12}$$

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.