

Supplementary notes

Gaussian distributions

1 Introduction

The multivariate normal, or multivariate *Gaussian* probability distribution is central to classical and even to modern statistics. The Gaussian distribution has many mathematical properties that make it a convenient model for statistical inference. Gaussian or approximately distributions arise through the central limit theorem. They also arise as what might be called the *Laplace approximation* (called the *low temperature* approximation in physics and physical chemistry). In the pre-computer classical era of statistics (before, say, 1970), this meant a Gaussian model was used whenever it was remotely plausible. In the computational statistics era, Gaussians are used because they are one of a small class of multivariate distributions for which efficient general samplers are available. In the AI era, Gaussians are used, largely for convenience, in many generative models.

Many of the properties and uses of Gaussians follow from two basic facts. One is that a linear transformation of a multivariate gaussian is also multivariate gaussian. The other is that the distribution is completely determined by the mean and covariance matrix. If X and Y are independent Gaussian, then the pair (X, Y) is jointly Gaussian and still independent. Since the covariance determines the distribution, this implies that uncorrelated components of a multi-variate normal are independent. This makes simple linear estimators optimal in many senses if the randomness is Gaussian.

2 The Gaussian distribution

If $X = (X_1, \dots, X_d)$ is a multi-variate random variable with PDF $p(x)$, we say X is multi-variate normal if there is a positive definite symmetric matrix, H , called the *information* matrix, and a vector $\mu \in \mathbb{R}^d$, called the *mean*, so that

$$p(x) = \frac{1}{Z} e^{-\frac{1}{2}(X-\mu)^T H (X-\mu)} . \quad (1)$$

The PDF is completely determined by the “parameters” H and μ . Much of the time, working with multi-variate normals involves using linear algebra together with linear changes of variable in multi-dimensional integrals.

2.1 The linear algebra of symmetric matrices

Much of this subsection will be review for many readers. Please read it anyway for notation and because there might be small gaps in your background. In addition to the specific content, this subsection illustrates my view that the simple matrix/vector formalism of abstract linear algebra is one of the most useful parts of theoretical applied math.

A $d \times d$ matrix, A , is *symmetric* if $A^T = A$. The *homogeneous quadratic form* associated to A is the function of d variables

$$Q(x) = \frac{1}{2} x^T A x . \quad (2)$$

If you multiply this out, writing a_{jk} for the (j, k) entry of A , you get the sum

$$Q(x) = \frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d a_{jk} x_j x_k . \quad (3)$$

A is symmetric when $a_{jk} = a_{kj}$. In that case, the (j, k) and (k, j) terms in the sum are equal. Then you can write the sum as a sum of *diagonal* terms ($j = k$) and above diagonal terms ($j < k$). The above diagonal terms lose the $\frac{1}{2}$ because each one has a partner below the diagonal. Please check that the index ranges in the above diagonal sum are correct:

$$Q(x) = \frac{1}{2} \sum_{j=1}^d a_{jj} x_j^2 + \sum_{j=1}^{d-1} \sum_{k=j+1}^d a_{jk} x_j x_k .$$

A function like this is called *homogeneous* because every term in the sum has the same degree. It is *quadratic* because the degree is 2. A *form* is a function that is the sum of terms of the same degree, so “homogeneous” is redundant.

The factor $\frac{1}{2}$ in the quadratic form definition is there to make the gradient formula simple. If A is symmetric, then

$$\nabla Q(x) = Ax . \quad (4)$$

This is the multi-variate analogue of the scalar formula

$$\frac{d}{dx} \frac{1}{2} ax^2 = ax .$$

You can verify the gradient formula (one of the ways) component-wise by showing that component i of the gradient is component i of the matrix/vector product:

$$\partial_{x_i} Q(x) = \sum_{k=1}^d a_{ik} x_k .$$

When you different the sum on the right of (3) there are non-zero terms when $i = j \neq k$, when $i = k \neq j$, or when $i = j = k$. These possibilities lead to

$$\begin{aligned} \partial_{x_i} a_{ik} x_i x_k &= a_{ik} x_k & (i = j \neq k) \\ \partial_{x_i} a_{ji} x_j x_i &= a_{ji} x_j = a_{ij} x_j & (i = k \neq j, a_{ij} = a_{ji}) \\ \partial_{x_i} a_{ii} x_i^2 &= 2a_{ii} x_i & (i = j = k) \end{aligned}$$

Combining these leads to

$$\partial_{x_i} Q(x) = \sum_{k \neq i} a_{ik} x_k + \sum_{j \neq i} a_{ij} x_j + 2a_{ii} x_i .$$

The first two terms on the right are the same, so the result is twice the i component of Ax .

I find the elementary but complicated verification reassuring, particularly when I'm not in an abstract mood. But once you have the formula, you might seek a more abstract verification that might be simpler. One such is based on the directional derivative formula

$$s^T \nabla Q(x) = \lim_{h \rightarrow 0} \frac{1}{h} [Q(x + hs) - Q(x)] .$$

The quantity in square brackets is

$$(x + hs)^T A(x + hs) - Q(x) = x^T Ax + hs^T Ax + hx^T As + h^2 s^T As - x^T Ax$$

The “trick” is to recognize that any number (scalar) is its own transpose, so (A being symmetric)

$$(x^T A s)^T = s^T A^T x = s^T A x .$$

Therefore (as in Calculus I)

$$\begin{aligned} s^T \nabla Q(x) &= \lim_{h \rightarrow 0} \frac{1}{h} [Q(x + hs) - Q(x)] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \frac{1}{2} [hs^T A x + hs^T A x + h^2 s^T A s] \\ &= s^T A x + \frac{1}{2} \lim_{h \rightarrow 0} hs^T A s \\ &= s^T A x . \end{aligned}$$

If A is not symmetric, the derivative formula is

$$\nabla x^T A x = A x + A^T x .$$

You can get the non-symmetric gradient formula from the symmetric one using the symmetrization trick

$$x^T A x = (x^T A x) = x^T A^T x .$$

Thus, any quadratic form is equivalent to a quadratic form with a symmetric matrix.

2.1.1 Eigenvalues and eigenvectors, variational principle

Statistical applications make a lot of use of the eigenvalues and eigenvectors of a symmetric matrix. These may be understood using the *Rayleigh quotient* (defined when $x \neq 0$)

$$R(x) = \frac{Q(x)}{\frac{1}{2} \|x\|_2^2} = \frac{x^T A x}{x^T x} . \quad (5)$$

Conversely, when the Rayleigh quotient arises, we understand it by finding the eigenvalues and eigenvectors of A . In this context, they may be called *principle components*. The relationship revolves around the fact that eigenvectors of A are *stationary points* of R and conversely. This is the *variational* characterization of the symmetric eigenvalue problem.

The gradient of a quotient is

$$\nabla \frac{f(x)}{g(x)} = \frac{1}{g} \nabla f - \frac{f}{g^2} \nabla g .$$

A stationary point is where $\nabla \frac{f}{g} = 0$. Stationary points satisfy

$$\nabla f(x) = \frac{f(x)}{g(x)} \nabla g(x) .$$

We write λ for the fraction on the right. In this notation a stationary point of the Rayleigh quotient satisfies

$$A x = \lambda x , \quad \lambda = \frac{x^T A x}{\|x\|_2^2} . \quad (6)$$

Therefore, any stationary point of the Rayleigh quotient is an eigenvector of A and any eigenvector is a stationary point. The value of the Rayleigh quotient is the corresponding eigenvalue.

Instead of taking the Rayleigh quotient as a quotient, you can consider just the numerator – if you *normalize* to set the denominator equal to 1. If x is any non-zero vector, you can normalize (re-scale) x by its length to get a unit vector in the same direction. In formulas, if $x \neq 0$, then

$$y = \frac{1}{\|x\|_2} x$$

is a unit vector because $\|x\|_2 = \sqrt{\|x\|_2^2} = \sqrt{x^T x} = 1$. Furthermore, $R(x) = R(y)$ because the scaling factor $\|x\|_2$ cancels out. Therefore, the range of values of $R(x)$ over all $x \neq 0$ is the same as the range of values of R over normalized vectors $\|x\|_2 = 1$. For that reason, you can find the maximum of the Rayleigh quotient by solving an equality constrained optimization problem

$$\max_{x \neq 0} R(x) = \max_{\|x\|_2=1} R(x) = \max_{\|x\|_2=1} x^T A x .$$

The constrained optimization problem on the right may be solved using the method of Lagrange multipliers. The result is $\nabla x^T A x = \lambda \nabla x^T x$, which is the same as (6). The Lagrange multiplier is the eigenvalue. As (6) says, the value of the Rayleigh quotient is the eigenvalue. That could be checked directly by multiplying $Ax = \lambda x$ by x^T .

This shows that the largest eigenvalue is the max of the Rayleigh quotient. The same reasoning shows that the smallest eigenvalue is the min of the Rayleigh quotient. What about the eigenvalue between? The reasoning above shows that any real eigenvalue/eigenvector pair is a stationary point of R and any stationary point corresponds to a real eigenvalue/eigenvector pair. Can you find all the eigenvalues/eigenvectors this way? What can you learn from this variational point of view? Answers to these questions will lead naturally to the *singular value decomposition*, or *SVD*, which applies to any matrix even if it is not symmetric or even square.

The extra fact is an orthogonality principle related to the eigenvalue problem. Vectors x and y are *orthogonal* if $x^T y = 0$. Geometrically, if x is a vector from the origin and S is the sphere around the origin that contains x , then y is orthogonal to x if y is tangent to S at x . Analytically, this means that the directional derivative of $\|x\|_2^2$ (which is $2x$) is perpendicular to y .

$$\left. \frac{d}{dh} \|x + hy\|_2^2 \right|_{h=0} = 2y^T x = 0 .$$

We saw that the maximizer of the Rayleigh quotient is an eigenvector, which we will call u_d . The corresponding eigenvalue will be λ_d . We will see that there are d real eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$. The orthogonality principle lets us use the variational method to find the next largest eigenvalue λ_{d-1} and corresponding eigenvector u_{d-1} .

What makes this possible is the fact that if y is perpendicular to x then Ay is perpendicular to x also. Perpendicular means $(Ay)^T x = 0$. A simple calculation to verify this uses the fact that $A^T = A$ and that transpose reverses the order of matrix/vector operations¹ :

$$\begin{aligned} (Ay)^T x &= y^T A^T x \\ &= y^T A x \\ &= y^T \lambda x \\ &= \lambda (y^T x) \\ &= 0 . \end{aligned}$$

Let $V_1 \subset \mathbb{R}^d$ be the vector space of x with $X^T u_d = 0$. We just saw that if $x \in V_1$ then also $Ax \in V_1$. This implies that A is a linear transformation from V_1 to V_1 , and is given by a symmetric matrix. Therefore, you can find

$$u_{d-1} = \arg \max_{x \in V_1, x \neq 0} R(x) .$$

As before, there is a Lagrange multiplier, λ_{d-1} , so that $Au_{d-1} = \lambda_{d-1} u_{d-1}$. In other words, λ_{d-1} is not the maximum of $R(x)$, but it is the max of $R(x)$ over the subspace perpendicular to u_d . The

¹There is a more conceptual but possibly more complicated way to see this. If there is a y perpendicular to u_d with Ay not perpendicular to u_d , then (do this first derivative calculation) there is an h small enough (possibly negative) so that $R(u_d + hy) > R(u_d)$. That means, if Ay is not perpendicular to u_d , then u_d was not optimal.

dimension of V_1 is $d - 1$ because it is defined by one linear constraint. Similarly, we could define V_2 to be the perpendicular complement of u_1 and u_2 and see that V_2 is also an invariant subspace for A (meaning $x \in V_2 \implies Ax \in V_2$) and maximize R over V_2 to find u_{d-2} , and so on. In the process, we see that u_{d-1} is perpendicular to u_d (because $u_{d-1} \in V_1$) and so on. The eigenvectors $u_d, u_{d-1}, \dots$, are all orthogonal to each other. There are d linearly independent such eigenvectors and d corresponding real eigenvalues. This is a variational approach to the “symmetric eigenvalue problem”.

2.1.2 Positive & positive definite, Cholesky factorization

The matrix A is *positive* if $x^T A x \geq 0$ for all x . It is positive definite if $x^T A x > 0$ whenever $x \neq 0$. The variance/covariance matrix is an example, always positive, hopefully positive definite. A common way to get a positive matrix is to start with any matrix B , not necessarily square, and take $A = B^T B$ or $A = B B^T$. Such a matrix is positive because matrix/vector multiplication is associative, and

$$\begin{aligned} x^T A x &= x^T (B^T B) x \\ &= (x^T B^T) (B x) \\ &= (B x)^T (B x) \\ x^T A x &= \|B x\|_2^2 \geq 0. \end{aligned}$$

If $A = B B^T$ then $x^T A x = \|B^T x\|_2^2 \geq 0$. If B is a square and non-singular matrix, then $B x = 0$ only if $x = 0$. Therefore, if B is non-singular, then $x^T A x = \|B x\|_2^2 \neq 0$ if $x \neq 0$, so A is positive definite. The vectors x with $B x = 0$ form the *null space*, or *kernel* of B . If B has more columns than rows (a *short fat* matrix) then the null space of B contains a non-zero x and $A = B^T B$ is not positive definite. If B has more rows than columns (a *tall thin* matrix), then it's unlikely (but not impossible) that B has a non-trivial null space (a null space with a non-zero vector in it). Thus, $A = B^T B$ is positive definite if and only if the null space of B is trivial. This is why a covariance matrix is likely to be positive definite.

The *Cholesky factorization* theorem is a converse of the above discussion. If A is $d \times d$ symmetric and positive definite, then there is a $d \times d$ matrix L with

$$A = L L^T. \quad (7)$$

If A is positive definite then L must be non-singular, and conversely (L is singular if A is not positive definite). You can think of L as a square root of A . For numbers ($d = 1$), a is non-negative if and only if there is a number ℓ with $a = \ell^2$. For 1×1 matrices, $\ell^2 = \ell \ell^T$. When $d > 1$, $L L^T$ is often more natural and more useful than L^2 . For example, $L L^T$ is symmetric but L^2 probably isn't.

The *Cholesky factor* of A is the upper triangular matrix L that satisfies the factorization equation (7) and has non-negative numbers on the diagonal. You can see directly that L is unique by looking at the equations that the entries of L must satisfy. The factorization equation (7), is (A is symmetric, so $a_{12} = a_{21}$, etc.)

$$\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{d1} \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \\ a_{d1} & \cdots & & a_{dd} \end{pmatrix} = \begin{pmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \ell_{d1} & \cdots & & \ell_{dd} \end{pmatrix} \begin{pmatrix} \ell_{11} & \ell_{21} & \cdots & \ell_{d1} \\ 0 & \ell_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \ell_{dd} \end{pmatrix}$$

The formulas for the matrix product are

$$\begin{aligned} a_{11} &= \ell_{11}^2 \\ a_{21} &= \ell_{21} \ell_{11} \\ a_{22} &= \ell_{21}^2 + \ell_{22}^2 \\ &\text{etc.} \end{aligned}$$

The first equation gives $\ell_{11} = \sqrt{a_{11}}$. The square root exists because $a_{11} = e_1^T A e_1 \geq 0$. Here, e_1 is the “standard” basis vector with 1 in the first entry and 0 in all the rest. If $\ell_{11} \geq 0$, then ℓ_{11} is uniquely determined. If A is positive definite then ℓ_1 is positive. The second equation gives

$$\ell_{21} = \frac{a_{21}}{\ell_{11}} .$$

If A is positive definite and $\ell_{11} > 0$, this uniquely determines ℓ_{21} . The third equation gives

$$\ell_{22} = \sqrt{a_{22} - \ell_{21}^2} .$$

This determines ℓ_{22} uniquely if $\ell_{22} > 0$, but it doesn’t say why the quantity in the square root is positive. In other words, it says the Cholesky factor is unique if it exists but it doesn’t say why it exists. The rest of the entries of L follow this pattern. The *subdiagonal* entries ℓ_{jk} with $j > k$ are found by division. The diagonal entries require a square root,

$$\ell_{kk} = \sqrt{a_{kk} - \ell_{k1}^2 - \cdots - \ell_{k,k-1}^2} ,$$

that needs to be real. It turns out that the quantity in the square root has a statistical interpretation which shows it must be positive if A is positive definite (see below). This is the proof of the Cholesky factorization theorem.

There is good software for computing the Cholesky factor, if it exists. The Python package *numpy* includes `np.cho1`. The algorithm is something like the calculations above, but done in a fancy way using *block* submatrices for efficiency.

The Cholesky factorization theorem has an interesting use. If you have a symmetric matrix A and you want to know whether it’s positive definite, pass it to the Cholesky calculating function. If A is positive definite, you will get an L . If A is not positive definite, you will get an error message (in some form).

The eigenvalues of A also tell you whether A is positive definite. You can see this from the Rayleigh quotient formula for λ_1 . If

$$\lambda_1 = \min_{x \neq 0} \frac{x^T A x}{\|x\|_2^2} > 0 ,$$

then $x^T A x > 0$ whenever $x \neq 0$, which is the definition of positive definite. You can also see it from the factorization formula (8) (below). Define $y = U^T x$. Then

$$x^T A x = y^T \Lambda y = \prod_j \lambda_j y_j^2 .$$

The product is always positive when all the λ_j are positive.

2.1.3 Eigenvalue/eigenvector factorization & functional calculus

The eigenvalues and eigenvectors of a symmetric matrix may be assembled into the eigenvalue matrix, Λ , and the corresponding eigenvector matrix, U , to give a matrix factorization

$$A = U \Lambda U^T . \tag{8}$$

This factorization is equivalent (see below) to the formula

$$U \Lambda = A U . \tag{9}$$

We describe Λ and U in more detail, verify (9), then see that it is equivalent to (8). The factorization (8) gives a quick way to compute functions of A , including powers, the inverse, etc.

The eigenvector matrix is the matrix whose columns are eigenvectors

$$U = \begin{pmatrix} | & \cdots & | \\ u_1 & \cdots & u_d \\ | & \cdots & | \end{pmatrix} \quad (10)$$

The eigenvalue/eigenvector relation $Au_k = \lambda_k u_k$ may be written in matrix form as

$$\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{d1} \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \\ a_{d1} & \cdots & & a_{dd} \end{pmatrix} \begin{pmatrix} | & \cdots & | \\ u_1 & \cdots & u_d \\ | & \cdots & | \end{pmatrix} = \begin{pmatrix} | & \cdots & | \\ \lambda_1 u_1 & \cdots & \lambda_d u_d \\ | & \cdots & | \end{pmatrix}$$

The eigenvalue matrix is all zeros except for the eigenvalues on the diagonal

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \lambda_d \end{pmatrix}$$

You can check by multiplying out the right side that the last matrix is

$$\begin{pmatrix} | & \cdots & | \\ \lambda_1 u_1 & \cdots & \lambda_d u_d \\ | & \cdots & | \end{pmatrix} = \begin{pmatrix} | & \cdots & | \\ u_1 & \cdots & u_d \\ | & \cdots & | \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \lambda_d \end{pmatrix}$$

This verifies (9).

Going from (9) to (8) requires that we normalize the eigenvectors so that $\|u_j\|_2^2 = 1$. This may be done by scaling: multiplying u_j by a scalar factor to change its length without changing its direction. The variational construction automatically makes the eigenvectors orthogonal: $u_j^T u_k = 0$ if $j \neq k$. One detail here: some simple algebra shows that eigenvectors with distinct eigenvalues are orthogonal (look in any linear algebra book), but the strategy in subsection (2.1.1) makes u_j orthogonal to u_k for all $k > j$ even if the eigenvalues turn out to be the same. A basis of vectors normalized and orthogonal is an *ortho-normal* basis. The matrix U constructed from them is an *orthogonal matrix*. The definition of orthogonal matrix is

$$U^T U = I. \quad (11)$$

This is equivalent to the identities $u_j^T u_j = 1$ (normalization) and $u_j^T u_k = 0$ if $j \neq k$ (orthogonality, the zero elements of the identity matrix). To see this, write it out more explicitly:

$$U^T U = \begin{pmatrix} -- & u_1^T & -- \\ \vdots & \vdots & \vdots \\ -- & u_d^T & -- \end{pmatrix} \begin{pmatrix} | & \cdots & | \\ u_1 & \cdots & u_d \\ | & \cdots & | \end{pmatrix} = \begin{pmatrix} u_1^T u_1 & u_1^T u_2 & \cdots & u_1^T u_d \\ u_2^T u_1 & u_2^T u_2 & \cdots & \vdots \\ \vdots & & & \\ u_d^T u_1 & u_d^T u_2 & \cdots & u_d^T u_d \end{pmatrix}.$$

It is a fact of linear algebra for square matrices (and only square ones) that if $BC = I$ then $CB = I$. Thus, if $U^T U = I$ then $U U^T = I$ also. Now, starting with (9), we multiply both sides from the right (matrix multiplication is not commutative so multiplying from the right is not the same as multiplying from the left) and get (8):

$$\begin{aligned} U \Lambda &= A U \\ U \Lambda U^T &= A U U^T = A. \end{aligned}$$

Orthogonal matrices are special partly because the transpose is the inverse, (11). They also have the property of not changing the 2-norm. If U is an orthogonal matrix, then

$$\|Ux\|_2 = \|x\|_2 . \quad (12)$$

This is a special property of the 2-norm that is not shared by other norms. For example, consider the matrix/vector pair

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} , \quad x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} .$$

You can multiply out the matrix product to check that $UU^T = I$. Also $\|x\|_1 = |x_1| + |x_2| = 1$. However,

$$y = Ux = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

has (check this) $\|y\|_1 = \sqrt{2} > \|x\|_1 = 1$. You can verify the *isometry* formula (12) by computing the square: $\|Ux\|_2^2 = \|x\|_2^2$:

$$\begin{aligned} \|Ux\|_2^2 &= (Ux)^T (Ux) \\ &= x^T U^T U x \\ &= x^T (U^T U) x \\ &= x^T x \\ &= \|x\|_2^2 . \end{aligned}$$

The *functional calculus*² is about applying functions (square, inverse, sine, etc.) to matrices. The basic idea is that $f(A) = B$ defines B by using the same eigenvectors and applying f to the eigenvalues

$$f(A) = U f(\Lambda) U^T = U \begin{pmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & f(\lambda_d) \end{pmatrix} U^T$$

This works because $U^T U = I$. For example, if $f(a) = a^2$, then you want $f(A) = A^2$. And that's what this definition gives

$$\begin{aligned} A^2 &= (U \Lambda U^T) (U \Lambda U^T) \\ &= U \Lambda (U^T U) \Lambda U^T \\ &= U \Lambda \Lambda U^T \\ &= U \Lambda^2 U^T . \end{aligned}$$

The square of the diagonal matrix Λ is the matrix with λ_j^2 on the diagonal. This also applies to the matrix inverse. If $f(a) = a^{-1}$ then $f(A) = A^{-1}$, which may be found by taking λ_j^{-1} .

One thing that makes the eigenvalue/eigenvector decomposition more useful for symmetric matrices than it might be for other matrices is the relationship between eigenvalues and the matrix norm. If $\|x\|$ is any vector norm, the corresponding matrix norm is

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} .$$

²This term is more commonly used in the advanced context of functional analysis, which is an infinite dimensional version of linear algebra.

Different vector norms lead to different matrix norms. We write $\|A\|_2$ for the matrix norm *induced* by the vector 2–norm. This definition is simple but abstract. It can be challenging to solve the optimization problem on the right and find $\|A\|$ explicitly. For the 2–norm, it’s easy (see below) to see that

$$\|\Lambda\| = \max |\lambda_j| . \quad (13)$$

You can see this from the Rayleigh quotient approach to eigenvalues/eigenvectors. You also can see it (as you normally do with the 2–norm) by taking the square:

$$\|A\|_2^2 = \max_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} .$$

For the diagonal matrix Λ , this becomes

$$\|\Lambda\|_2^2 = \max_{x \neq 0} \frac{\sum (\lambda_j x_j)^2}{\sum x_j^2} = \max_{x \neq 0} \frac{\sum \lambda_j^2 x_j^2}{\sum x_j^2} .$$

You maximize the fraction on the right by putting all the “weight” x_j^2 on the largest square eigenvalue.

$$\max_{x \neq 0} \frac{\sum \lambda_j^2 x_j^2}{\sum x_j^2} = \lambda_{\max}^2 .$$

Taking the square root gives the norm formula (13). Eigenvalues determine the 2–norm for any symmetric matrix, because

$$\|A\|_2 = \|U\Lambda U^T\|_2 = \|\Lambda U^T\|_2 = \|\Lambda\|_2 = \max_j |\lambda_j| .$$

We will often apply this together with the functional calculus:

$$\|f(A)\|_2 = \max_j |f(\lambda_j)| . \quad (14)$$

A matrix is called *normal* if it has d linearly independent eigenvectors that are orthogonal to each other. The Rayleigh quotient procedure shows that any symmetric matrix is normal. There are other categories of normal matrices that are important in other applications. Any normal matrix has a functional calculus that satisfies the norm formula (14). What makes this powerful is that it applies for any matrix size, d , and for “any” function f (any continuous function, also $f(a) = a^{-1}$, etc.). The norm of $f(A)$ can be harder to determine if it is not the 2–norm or if A is not normal. A possible inequality,

$$???? \quad \|f(A)\|_* \stackrel{??}{\leq} C(d, f, *) \max_j |f(\lambda_j)| \quad ????$$

might not be true at all, or the constant C might be very large when d is large, depending on which norm $*$ represents, which function, and “how non-normal” A might be. If A is normal but not symmetric ($A^T = -A$ or $A^T = A^{-1}$, etc.), then at least one of the eigenvalues of A is not real.

An important thing about the functional calculus is that $f(A)$ has the same eigenvectors as A . Only the eigenvalues change. It is unusual for two matrices to commute, but any two functions of A commute:

$$f(A)g(A) = g(A)f(A) .$$

It goes the other way too, with a technical proviso. If A has distinct eigenvalues and if $AB = BA$, then B is a function of A . There is some function $f(a)$ so that $B = f(A)$.

2.1.4 Determinants

Many formulas involving Gaussian random variables involve determinants. A good linear algebra class will have a lot to say about determinants. They are defined only for square matrices. The multiplicative property is important

$$\det(AB) = \det(A) \det(B) .$$

One consequence is that $\det(AB) = \det(BA)$ even if $AB \neq BA$. A trivial consequence is $\det(I) = 1$, because if $\det(A) \neq 0$, then $\det(A) = \det(IA) = \det(I)\det(A)$ implies $\det(I) = 1$. Another important property of determinants, for us, is $\det(A) = \det(A^T)$. This implies that if U is orthogonal, then $\det(U) = \pm 1$. This is because $1 = \det(I) = \det(UU^T) = \det(U)\det(U^T) = \det(U)^2$ (if $\det(U)^2 = 1$ then $\det(U) = \pm 1$). Next, the determinant of a diagonal matrix is the product of the diagonal entries. That is,

$$\det(\Lambda) = \prod_{j=1}^d \lambda_j . \quad (15)$$

A *block matrix* is a matrix whose entries are matrices rather than numbers. For example, a 2×2 block matrix may be written

$$A = \begin{pmatrix} B & C \\ D & E \end{pmatrix} .$$

The sizes of the blocks has to work with this pattern. Suppose A is $r \times s$ (r rows and s columns, not necessarily square). If the “diagonal” blocks have sizes $r_1 \times s_1$ (for B) and $r_2 \times s_2$ (for E), then the other blocks must be $r_1 \times s_2$ (for C) and $r_2 \times s_1$ (for D). The number of rows of B plus the number of rows of D is the number of rows in all: $r_1 + r_2 = r$, and similarly for columns. Clearly, there could be block matrices larger than 2×2 . The matrix A is *block diagonal* if the off diagonal blocks (C and D in this case) are zero.

Most of our block matrices will have diagonal blocks that are square, which is $r_1 = s_1$ and $r_2 = s_2$ in the 2×2 case. Many of our matrices will have diagonal blocks of different sizes, which is $r_1 \neq r_2$ for 2×2 . In that case the off diagonal blocks are not square. A block diagonal square matrix with square diagonal blocks has determinant (the corresponding formula for more blocks is similar) is as it would be if the diagonal blocks were numbers:

$$\det \begin{pmatrix} B & 0 \\ 0 & E \end{pmatrix} = \det(B) \det(E) .$$

This is analogous to the scalar (1×1 blocks) determinant formula (15). The determinant of a 1×1 matrix (a number) is that number. There is (as far as I know) no general block determinant formula analogous to the scalar formula

$$\det \begin{pmatrix} b & c \\ d & e \end{pmatrix} = be - cd .$$

2.2 Gaussian integrals

Many properties of multi-variate normals may be seen using integrals involving the normal probability density function (1). It is possible to understand many integrals involving this density using “normal” coordinates aligned with the eigenvectors of H . There are other ways, possibly more elementary, but eigenvector coordinates are a general approach that you can try if you’re not feeling clever. In this section we always take the mean, μ , to be zero. You can always add in a non-zero mean once you’ve understood the $\mu = 0$ case.

2.2.1 Eigenvector coordinates

The information matrix, H , is symmetric and has d linearly independent eigenvectors u_j . We can take the u_j as a basis and write any x as a linear combination of the eigenvectors with weights y_j :

$$x = \sum_{j=1}^d y_j u_j . \quad (16)$$

This can be expressed in matrix/vector notation in terms of the column vector

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix}$$

and the eigenvector matrix U of (10):

$$x = Uy .$$

In multi-variate integration, there is a change of variables formula for linear change of variables that is informally written as

$$dy = \det(A) dx , \quad \text{if } y = Ax .$$

For $d = 1$, this is $dy = a dx$ if $y = ax$. If the change of variables involves a matrix, the *Jacobian factor* is the determinant. If the change of variables matrix A is orthogonal, then, as explained above, $\det(A) = \pm 1$.

We can always take $\det(U) = 1$. If we accidentally had eigenvectors u_j so that the eigenvector matrix had $\det(U) = -1$, we could do a “reflection” to reverse the sign of the determinant. One way to “reflect” is to replace the first eigenvector u_1 with its opposite $-u_1$. The resulting vector is still normalized, $\|-u_1\|_2 = 1$, and orthogonal to all the other eigenvectors. Thus, the resulting matrix, $\tilde{U}$, is still orthogonal. You can get $\tilde{U}$ by multiplying U by the diagonal matrix with all ones except -1 in the first spot (check this):

$$\tilde{U} = \begin{pmatrix} | & | & \cdots & | \\ -u_1 & u_2 & \cdots & u_d \\ | & | & \cdots & | \end{pmatrix} = \begin{pmatrix} | & | & \cdots & | \\ u_1 & u_2 & \cdots & u_d \\ | & | & \cdots & | \end{pmatrix} \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & 0 & \ddots & 0 \\ 0 & \cdots & & 1 \end{pmatrix} .$$

The last matrix on the right is diagonal with determinant -1 . Thus $\det(\tilde{U}) = \det(U)(-1)$. If $\det(U) = -1$, then $\det(\tilde{U}) = 1$.

To summarize, we do a change of variables $x = Uy$ with $dx = dy$ so that the information matrix is diagonal in the y coordinate system

$$\begin{aligned} x^T H x &= (Uy)^T U \Lambda U^T (Uy) \\ &= y^T U^T U \Lambda U^T U y \\ &= y^T (U^T U) \Lambda (U^T U) y \\ x^T H x &= y^T \Lambda y . \end{aligned} \quad (17)$$

2.2.2 Diagonal quadratic forms and product integrals

Eigenvector coordinates are useful because they turn the general multi-variate integral into a product of one dimensional integrals. For a Gaussian, these one dimensional integrals are known in closed form. The multi-variable integral is evaluated as the product of these one variable formulas.

We illustrate the method on the problem of evaluating the normalization constant in (1),

$$Z = \int_{\mathbb{R}^d} e^{-\frac{1}{2}x^T H x} dx .$$

In eigenvector coordinates, this becomes (see (17))

$$Z = \int_{\mathbb{R}^d} e^{-\frac{1}{2}y^T \Lambda y} dy .$$

The exponential factor is

$$\begin{aligned} e^{-\frac{1}{2}y^T \Lambda y} &= e^{-\frac{1}{2}\lambda_1 y_1^2 - \dots - \frac{1}{2}\lambda_d y_d^2} \\ &= \left(e^{-\frac{1}{2}\lambda_1 y_1^2} \right) \dots \left(e^{-\frac{1}{2}\lambda_d y_d^2} \right) . \end{aligned} \quad (18)$$

The normalization factor is (step by step explanations below)

$$\begin{aligned} Z &= \int_{\mathbb{R}^d} e^{-\frac{1}{2}\lambda_1 y_1^2 - \dots - \frac{1}{2}\lambda_d y_d^2} dy_1 \dots dy_d \\ &= \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_1 y_1^2} dy_1 \right) \dots \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_d y_d^2} dy_d \right) \\ &= \left(\sqrt{\frac{2\pi}{\lambda_1}} \right) \dots \left(\sqrt{\frac{2\pi}{\lambda_d}} \right) \\ &= (2\pi)^{\frac{d}{2}} (\lambda_1 \dots \lambda_d)^{-\frac{1}{2}} \\ Z &= (2\pi)^{\frac{d}{2}} \det(H)^{-\frac{1}{2}} . \end{aligned} \quad (19)$$

In the first step, we recognize the integral of a product function as the product of integrals, as in

$$\int \int_{\mathbb{R}^2} f(x_1)g(x_2) dx_1 dx_2 = \left(\int_{-\infty}^{\infty} f(x_1) dx_1 \right) \left(\int_{-\infty}^{\infty} g(x_2) dx_2 \right) .$$

Putting the quadratic form $x^T H x$ in the form $y^T \Lambda y$ made the exponential into a product which made this possible. The basic Gaussian integral formula, which everyone doing this kind of thing should memorize, is

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \sqrt{2\pi} .$$

If we have

$$\int_{-\infty}^{\infty} e^{-\lambda \frac{1}{2}y^2} dy ,$$

and if $\lambda > 0$, we use the change of variables $\lambda y^2 = z^2$, which is $z = \sqrt{\lambda} y$, $dy = \frac{1}{\sqrt{\lambda}} dz$, so the integral is

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} \frac{dz}{\sqrt{\lambda}} = \frac{\sqrt{2\pi}}{\sqrt{\lambda}} ,$$

This makes the d fold integral a product of d such factors. Finally, the determinant formula is $\det(H) = \lambda_1 \dots \lambda_d$. We can use the explicit formula for the normalization constant into (1) to get

$$p(x) = \frac{\sqrt{\det(H)}}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}x^T H x} . \quad (20)$$

The simpler formula (1) is often enough.

2.2.3 Principal components

The eigenvector coordinates (16) are more than a trick to calculate integrals. They are a way to understand multivariate Gaussians. If $X \sim p$ (multi-variate normal), the corresponding eigenvector coordinates form a random vector Y with components defined by the representation relation

$$X = \sum_{j=1}^d Y_j u_j . \quad (21)$$

The Y_j are sometimes called *principal components*. The representation (21) is sometimes called *principal component analysis*, or , *PCA*. I say “sometimes” because the term “PCA” is applied by different people to different aspects of the eigenvalue/eigenvector analysis. Sometimes the eigenvectors u_j are called principal components. We saw that the scalar formula (21) may be written in vector form as

$$X = UY .$$

You can multiply both sides from the left by U^T and get a formula for the principal component vector,

$$Y = U^T X .$$

The crucial point, which is unique for Gaussians, is that the Y_j are independent random variables. This is essentially a consequence of the factorization (18) that we used to evaluate the Gaussian integral. The components of a multi-component random variable are independent when the joint PDF is a product of one variable density functions.

The principal component vector is random and it has a probability density that we call $p_Y(\cdot)$. This is related to the PDF of X , which we call $p_X(\cdot)$ (This kind of notation is common in theoretical statistics.) If A is any invertible $D \times d$ matrix, and random variables X and Y are related by A as $Y = AX$, $X = A^{-1}Y$, then the PDF of X and Y are related by the same linear transformation, up to a scaling factor. Informally, $p_Y(y) = \frac{1}{Z} p_X(x)$, understanding that $y = Ax$. This can be made formal in several ways, including

$$p_Y(Ax) = \frac{1}{Z} p_X(x) , \quad p_Y(y) = \frac{1}{Z} p_X(A^{-1}y) . \quad (22)$$

It is traditional to use the same letter Z to represent any normalization factor. In particular, the Z here is not the same as Z in (1). We apply the second formula to (1), with $A^{-1} = U^T$ and $\mu = 0$. As before $y = U^T x$ and $x = Uy$ implies that $x^T H x = y^T \Lambda y$. Thus,

$$p_Y(y) = \frac{1}{Z} e^{-\frac{1}{2} y^T \Lambda y} .$$

As before, this factors as

$$p_Y(y_1, \dots, y_d) = \left(\frac{1}{Z_1} e^{-\frac{1}{2} \lambda_1 y_1^2} \right) \dots \left(\frac{1}{Z_d} e^{-\frac{1}{2} \lambda_d y_d^2} \right) .$$

The individual normalization factors Z_j are chosen so that the one variable density functions are normalized

$$1 = \int_{-\infty}^{\infty} p_j(y_j) dy_j = \frac{1}{Z_j} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \lambda_j y_j^2} dy_j$$

The scalar normalization factors are (as we’ve already seen)

$$Z_j = \sqrt{\frac{2\pi}{\lambda_j}} .$$

This gives

$$p_j(y_j) = \sqrt{\frac{\lambda_j}{2\pi}} e^{-\frac{1}{2}\lambda_j y_j^2} .$$

This looks more familiar if we substitute (possible because $\lambda_j > 0$)

$$\sigma_j^2 = \frac{1}{\lambda_j} .$$

The Y_j PDF becomes

$$Y_j \sim p_j(\cdot) , \quad p_j(y_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{y_j^2}{2\sigma_j^2}} .$$

The conclusion is that the principal components Y_j are independent scalar Gaussians with mean zero and variance $\sigma_j^2 = \lambda_j^{-1}$.

This partly explains calling H the *information* matrix. More “information” means that the random variable is more narrowly constrained. The standard deviations of the principal components go down when the eigenvalues λ_j go up.

2.2.4 Covariance matrix

If X is a d component random variable, the variance/covariance matrix is Σ with entries

$$\sigma_{jk} = \text{cov}(X_j, X_k) .$$

We usually call Σ the covariance matrix. It has variances on the diagonal and covariances off the diagonal. When X has mean zero, the covariance formula simplifies to

$$\sigma_{jk} = \text{E}[X_j X_k] .$$

A matrix/vector version of this formula is simpler and easier to do linear algebra with. If y is a d component column vector, then y^T is a d component row vector and ordinary matrix multiplication makes yy^T the $d \times d$ matrix whose jk entry is $y_j y_k$:

$$yy^T = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \cdots & y_d \end{pmatrix} = \begin{pmatrix} y_1 y_1 & y_1 y_2 & \cdots & y_1 y_d \\ y_2 y_1 & y_2 y_2 & & \vdots \\ \vdots & & & \\ y_d y_1 & & y_j y_k & y_d y_d \end{pmatrix} .$$

The jj diagonal entry is y_j^2 . This matrix is symmetric, which you can see directly or by calculating that $(yy^T)^T = yy^T$. The jk entry of XX^T is $X_j X_k$. Its expected value is $\sigma_{jk} = \text{E}[X_j X_k]$. Therefore, the expected value of the matrix XX^T is

$$\Sigma = \text{E}[XX^T] . \tag{23}$$

If $\text{E}[X] = \mu$, the formula becomes

$$\Sigma = \text{E}[(X - \mu)(X - \mu)^T] .$$

The covariance formula (23) makes it easy to track what happens to covariances under linear transformations. If X and Y are distinct random variables, we write Σ_X and Σ_Y for the corresponding

covariance matrices. If $Y = AX$ then

$$\begin{aligned}
\Sigma_Y &= E[YY^T] \\
&= E[AX(AX)^T] \\
&= E[AXX^T A^T] \\
&= A E[XX^T] A^T \\
\Sigma_Y &= A \Sigma_X A^T .
\end{aligned} \tag{24}$$

You should check that if M is a random matrix and A and B are fixed matrices, then

$$E[AMB] = A E[M] B .$$

The expected value is an integral and constants “come out of” integrals, but matrix multiplication is not commutative, so A comes out on the left and B on the right to preserve the order of the matrices.

We can now calculate the covariance of a multi-variate normal with information matrix H . One way to do this is to use the eigenvector coordinates for X , which means writing $X = UY$. Our linear transformation covariance formula (24) wants us to write Y in terms of X , which is $Y = U^T X$ (the inverse is the transpose for orthogonal matrices). Therefore (24), with $A = U^T$, gives

$$\Sigma_X = U^T \Sigma_Y U .$$

We need to figure out Σ_Y , which is the covariance matrix of the eigenvector coordinates. But we already did. The diagonal entries are $\sigma_{Y_j}^2 = \lambda_j^{-1}$ and the off-diagonal entries are zero because the principle components are uncorrelated. Therefore

$$\Sigma_Y = \begin{pmatrix} \sigma_{Y_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{Y_2}^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{Y_d}^2 \end{pmatrix} = \begin{pmatrix} \lambda_1^{-1} & 0 & \cdots & 0 \\ 0 & \lambda_2^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_d^{-1} \end{pmatrix} = \Lambda^{-1}$$

Therefore,

$$\Sigma_X = U^T \Lambda^{-1} U = H^{-1} . \tag{25}$$

You can see this by quoting the functional calculus above, or you can check that $U^T \Lambda^{-1} U H = U^T \Lambda^{-1} U U^T \Lambda U = I$, which makes $U^T \Lambda^{-1} U = H^{-1}$. The conclusion is:

The covariance matrix and the information matrix are inverses of each other.

This allows us to write the Gaussian density function (1) in terms of its covariance matrix. You can leave the normalization constant as Z or write it in terms of Σ_X . The relation between determinants is

$$\Sigma_X = H^{-1} \implies \det(H) = \frac{1}{\det(\Sigma_X)} .$$

Therefore,

$$p(x) = \frac{1}{Z} e^{-\frac{1}{2}(x-\mu)^T \Sigma_X^{-1} (x-\mu)} = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{\sqrt{\det(\Sigma_X)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma_X^{-1} (x-\mu)} . \tag{26}$$

This formula justifies the statement that the multi-variate normal is determined by its mean and covariance matrix.

Conversely, any positive definite matrix Σ determines a mean zero multi-variate normal. Any positive definite matrix is the covariance matrix of some multi-component random variable. This makes the multi-variate normal useful for modeling correlated phenomena. For example finance people

model defaults on loans as random events. A financial institution will hold many loans (a “basket” of loans) to different counter-parties. The default risk for the basket depends on the default risk for the individual loans and also on the correlation between defaults. If the defaults are independent and each one has 10% default probability then a large basket is unlikely to lose significantly more than 10% to defaults. However, if there is one event (a recession) that causes all the loans to default, and the default probability is 10%, then there is a 10% probability of losing the whole basket. A typical strategy for modeling correlated defaults is to start with a multivariate normal with given correlations and map the components to defaults. This is the *copula* model.³

Finally, a piece of terminology. The *correlation coefficient* between one component (scalar) random variables X and Y is

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\mathbb{E}[(X - \mu_X)^2] \mathbb{E}[(Y - \mu_Y)^2]}} . \quad (27)$$

This is a dimensionless measure of the relation between X and Y ; the units of X and Y on the top and bottom cancel. Some facts about ρ are (see a probability book):

- ρ between -1 and 1 .
- If $\rho = \pm 1$, then $Y = aX + b$ for some constants a and b .
- It can happen that $|\rho| < 1$ even though Y is completely determined by X . For example, if $Y = X^3$ and X is standard normal (or basically any distribution).

We say X and Y are *uncorrelated* if $\rho_{xy} = 0$. Since the denominator terms in (27) are positive except in trivial cases, X and Y are uncorrelated when $\sigma_{XY} = 0$ (the covariance). It is common to say X and Y are uncorrelated rather than “uncovarianced” when the covariance is zero. We uncorrelated instead of uncovarianced when covariance matrix Σ_{XY} is zero.

2.2.5 Linear prediction and regression

The notation in this subsection will be different from above. Here, there will be a multi-variate random variable with two blocks, called X and Y . The notation is

$$\begin{aligned} X &= \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} , & X &\in \mathbb{R}^n . \\ Y &= \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} , & Y &\in \mathbb{R}^m . \\ W &= \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Y_1 \\ \vdots \\ Y_m \end{pmatrix} & W &\in \mathbb{R}^d , \quad d = n + m . \end{aligned}$$

³The copula model was blamed for banks under-estimating the default risk for baskets of loans that led to a large crisis in 2007. In retrospect it wasn’t the copula method itself, but the willingness of modelers to put in artificially low correlations.

The covariance matrix of W will be Σ_W . It is a 2×2 block matrix

$$\Sigma_W = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{pmatrix}$$

This is a symmetric matrix (like any covariance matrix) so the top right block is the transpose of the bottom left block. You can understand the block structure by

$$\Sigma_W = E[WW^T] = E\left[\begin{pmatrix} X \\ Y \end{pmatrix} (X^T \ Y^T)\right] = \begin{pmatrix} E[XX^T] & E[XY^T] \\ E[YX^T] & E[YY^T] \end{pmatrix}$$

This identifies the blocks as

$$\begin{aligned} \Sigma_X &= E[XX^T] \\ \Sigma_{XY} &= E[XY^T] \\ \Sigma_{YX} &= \Sigma_{XY}^T = E[YX^T] \\ \Sigma_Y &= E[YY^T] \end{aligned}$$

The prediction problem assumes you know Y and want to predict X . Call the prediction function

$$\hat{X} = \hat{X}(Y) .$$

The notation $\hat{X}$ refers both to the prediction and to the prediction function. The prediction function is a non-random function, but its output is random because it is a function of the random input, Y . The prediction residual (prediction error) will be called $R = X - \hat{X}$. The optimal prediction problem is to design the best prediction function.

This linear Gaussian prediction problem has a solution that is optimal in many senses. Most prediction problems have predictors that depend on the prediction loss criterion, but not this one. We will see that this optimal predictor is linear. There is a prediction matrix K and⁴

$$\hat{X} = KY . \tag{28}$$

We choose K (this is the main idea) so that the prediction residual is uncorrelated to the data. That is, the covariance between $X - \hat{X}$ and Y is zero

$$\begin{aligned} \Sigma_{RY} &= E[RY^T] \\ &= E[(X - \hat{X})Y^T] \\ &= E[XY^T] - E[\hat{X}Y^T] \\ &= \Sigma_{XY} - E[KYY^T] \\ \Sigma_{RY} &= \Sigma_{XY} - K\Sigma_Y . \end{aligned} \tag{29}$$

This is the difference of two term, the first being Σ_{XY} . The second term is

$$E[\hat{X}Y^T] = E[KYY^T] = K\Sigma_Y .$$

We set $\Sigma_{RY} = 0$ and solve for K :

$$\Sigma_{XY} - K\Sigma_Y = 0 \implies K = \Sigma_{XY}\Sigma_Y^{-1} . \tag{30}$$

⁴The “K” is for Kalman, who used these ideas to design optimal filters for linear noisy dynamical systems. Those are called *Kalman filters* today and are used in applications ranging from airplane auto-pilots to weather forecasting to high frequency stock trading.

The linear prediction formula (28) is optimal in many senses because the prediction error is (see below) independent of the data. The data is (are?) Y and the prediction error is $R = X - \hat{X}$. The random variables Y and R are independent:

$$X = \text{prediction, function of the data} + \text{mean zero and independent of the data.}$$

The technical term for “mean zero and independent of the signal (data)” is *noise*. Thus, if you’re only want the most accurate possible prediction, (28) with gain matrix (30) is what you should do. However, if you’re going to take some action based on the prediction, it might be better not to use a linear predictor. For example, the consequences for overestimating ($\hat{A} > X$, assuming only one component) might be different from the consequence for underestimating. That might make a biased “prediction” more useful than an unbiased one. This is decision theory, which we will discuss more later.

The Gaussian fact uncorrelated $\implies$ independent is a consequence of the linear change of variables formula (22) and the relation between the covariance and information matrices. The linear change of variable formula implies that if W is Gaussian with information matrix H_W and $V = AW$ for some invertible matrix A , then V is Gaussian with information matrix

$$H_V = A^T H_W A .$$

Therefore, if the pair (X, Y) is jointly normal, then so is the pair $\hat{X}, R$. The linear relation between them is, in block matrix form (this encodes the two formulas $R = X - KY$ and $Y = Y$)

$$\begin{pmatrix} R \\ Y \end{pmatrix} = \begin{pmatrix} X - \hat{X} \\ Y \end{pmatrix} = \begin{pmatrix} I & -K \\ 0 & I \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

The inverse of the block matrix

$$A = \begin{pmatrix} I & -K \\ 0 & I \end{pmatrix}$$

is (multiply and check)

$$A^{-1} = \begin{pmatrix} I & K \\ 0 & I \end{pmatrix}$$

This encodes the formulas $X = R + KY$ and $Y = Y$.

Now, suppose $W = (X, Y)$ is a multivariate Gaussian with two block components. We saw that X and Y being uncorrelated means that the covariance matrix of X has the block form

$$\text{cov}(W) = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix} .$$

The inverse of a block diagonal matrix is block diagonal (as the inverse of a diagonal matrix is diagonal). In particular, the information matrix for W is

$$H_W = \Sigma_W^{-1} = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_X^{-1} & 0 \\ 0 & \Sigma_Y^{-1} \end{pmatrix} = \begin{pmatrix} H_X & 0 \\ 0 & H_Y \end{pmatrix} .$$

Therefore, the joint PDF of X and Y is

$$p_W(x, y) = \frac{1}{Z} e^{-\frac{1}{2}(x^T H_X x + y^T H_Y y)} = \left(\frac{1}{Z} e^{-\frac{1}{2}x^T H_X x} \right) \left(\frac{1}{Z} e^{-\frac{1}{2}y^T H_Y y} \right) = p_X(x) p_Y(y) .$$

This shows that X and Y are independent random variables (their density is a product). The three Z normalization factors in these formulas are all different.

What is usually called *linear regression* is a variant on this. Here, the data consists of n rows with $d + 1$ entries in each row. In regression notation (totally different from the above, alas), data row k might consist of the $d + 1$ numbers

$$(Y_k, X_{k,1}, \dots, X_{k,d}) .$$

The regression problem is to find regression coefficients $(\beta_1, \dots, \beta_d)$ to make a linear prediction of Y_k from the corresponding *regressors*

$$Y_k \approx \sum_{j=1}^d X_{k,j} \beta_j . \quad (31)$$

The idea is that if you get a new set of predictor variables $(X_1, \dots, X_d)$, you will be able to predict the corresponding Y value using the regression formula (31). Standard linear regression formulas are a consequence of this “desire” together with an error model that the prediction residuals are independent mean zero normals with variance σ^2 . This is traditionally (in discussing linear regression) as

$$\epsilon_k = Y_k - \sum_{j=1}^d X_{k,j} \beta_j , \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2) .$$

As we saw in earlier notes, the maximum likelihood estimate of the parameter vector $\beta = (\beta_1, \dots, \beta_d)$ is the least squares problem

$$\hat{\beta} = \arg \max_{\beta} \|Y - X\beta\|_2^2 . \quad (32)$$

This is the notation of linear regression (check a textbook to verify). Note that even though σ is not assumed known, the least squares problem (32) for the regression coefficients does not depend on σ . This is an oddity of the gaussian model that is not shared by other models.