Stochastic Calculus, Courant Institute, Fall 2019
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2019/StochCalc.html
Jonathan Goodman, September, 2019

Class 6, linear problems, more

1 Linear diffusions

Continuing from last week, a linear diffusion is one that satisfies an SDE of the form

$$dX_t = AX_t dt + BdW_t \; . \tag{1}$$

The qualitative behavior is determined by the eigenvalues of A. We will use eigenvalue and eigenvector analysis to study the qualitative behavior of the PDF, $X_t \sim p(\cdot, t)$ and of the auto-covariance function.

While talking about eigenvectors and eigenvalues, we discuss the special case of a symmetric matrix, which will be called C below, for "covariance". One purpose of this discussion is to explain *PCA*, which is *principal component analysis*. PCA is related to *variational principles* that describe eigenvalues and eigenvectors of symmetric matrices. These motivate PCA. The orthogonality of eigenvectors is one example of the fact that orthogonality can be a consequence of minimizing or maximizing. Another purpose is to explain how to use the basis of eivenvectors. The eigenvectors of a symmetric matrix "are" (usually are and always may be taken to be) orthogonal to each other. The eigenvectors of a general matrix (not necessarily symmetric) do not have to be orthogonal. In that case, "expansion" in the basis of right eigenvectors requires us to know left eigenvectors too.

Review, eigenvalues, eigenvectors, expansion

Here is a review of eigenvalue and eigenvector analysis. If this material is completely new to you, then look in a good linear algebra book. I recommend the book by Strang and the book by Lax. I do not or explain why everything is true.

The eigenvalues λ_j and right eigenvectors u_j satisfy

$$Au_j = \lambda_j u_j \; . \tag{2}$$

Even if A is real, the eigenvectors u_j and the eigenvalues λ_j can be complex. If an eigenvalue is not real, its complex conjugate also is an eigenvalue. The complex conjugate of $\lambda = \mu + i\omega$ is $\overline{\lambda} = \mu - i\omega$. Similarly, the complex conjugate of a complex column vector is the complex column vector whose entries are the complex conjugates

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \quad \Longleftrightarrow \quad \overline{u} = \begin{pmatrix} \overline{u}_1 \\ \vdots \\ \overline{u}_n \end{pmatrix} \; .$$

This is not the same as the conjugate transpose, which is the row vector

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \quad \Longleftrightarrow \quad u^* = \begin{pmatrix} \overline{u}_1 & \cdots & \overline{u}_n \end{pmatrix} .$$

If M is any matrix with complex entries, the conjugate transpose is M^*, which has entries $(M^*)_{ij} = \overline{M_{ji}}$. You take the transpose (take ij instead of ji) and then the complex conjugate. If M is real (has real entries), then $M^* = M^t$.

What about uniqueness? If $Au = \lambda u$ and $Au' = \lambda u'$ (two different column vectors, same λ), do we know that $u = u'$?

Complex conjugation "commutes with multiplication". You can conjugate first and then multiply, or you can multiply first and then conjugate. The answer is the same. If z and w are two complex numbers, then $\overline{zw} = \overline{z}\,\overline{w}$. This relation, applied to the eigenvalue relation (2) leads to

$$A\overline{u}_j = \overline{\lambda}_j \overline{u}_j .$$

This justifies the statement earlier, that if λ is an eigenvalue, then $\overline{\lambda}$ is an eigenvector. If λ is not real, then $\overline{\lambda}$ is a different number. Eigenvalues and eigenvectors come in *complex conjugate pairs*. The matrix A is *diagonalizable* if there are n distinct right eigenvectors u_1, ..., u_n that are linearly independent. The right eigenvector matrix is the matrix whose columns are the right eigenvectors:

$$U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} .$$

Saying that the u_j are linearly independent is the same as saying that $V = U^{-1}$ exists. The eigenvalue matrix is a diagonal matrix with eigenvalues on the diagonal:

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

The eigenvector/eigenvalue relations (2) may be written in matrix form as

$$AU = U\Lambda . \tag{3}$$

The columns of the matrix on the left side are Au_j, since

$$A \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ Au_1 & \cdots & Au_n \\ | & & | \end{pmatrix} .$$

The columns of the matrix on the right side are $\lambda_j u_j$, since (check this)

$$\begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = \begin{pmatrix} | & & | \\ \lambda_1 u_1 & \cdots & \lambda_n u_n \\ | & & | \end{pmatrix} .$$

The columns being equal are the eigenvalue/eigenvector relations (2).

If you multiply the matrix eigenvalue relation (3) from the left and from the right by $V = U^{-1}$, the result is

$$VA = \Lambda V . \tag{4}$$

The rows of V are the *left eigenvectors* of A, just as the columns of U are the right eigenvectors. Denote row j of V by v_j. Then

$$V = \begin{pmatrix} -- & v_1 & -- \\ & \vdots & \\ -- & v_n & -- \end{pmatrix}$$

The matrix equation (4) is

$$\begin{pmatrix} -- & v_1 & -- \\ & \vdots & \\ -- & v_n & -- \end{pmatrix} A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} \begin{pmatrix} -- & v_1 & -- \\ & \vdots & \\ -- & v_n & -- \end{pmatrix} .$$

Multiplying the two sides as before, we get

$$\begin{pmatrix} -- & v_1 A & -- \\ & \vdots & \\ -- & v_n A & -- \end{pmatrix} = \begin{pmatrix} -- & \lambda_1 v_1 & -- \\ & \vdots & \\ -- & \lambda_n v_n & -- \end{pmatrix} .$$

The fact that row j on the right and left are equal is the left eigenvector/eigenvalue relation

$$v_j A = \lambda_j v_j . \tag{5}$$

The left eigenvector is a row vector that goes on the left of A.

The matrix equation that $V = U^{-1}$ may be written as $VU = I$. We write

this in more detail as

$$\begin{pmatrix} -- & v_1 & -- \\ & \vdots & \\ -- & v_n & -- \end{pmatrix} \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} = \begin{pmatrix} v_1u_1 & v_1u_2 & \cdots & v_1u_n \\ v_2u_1 & v_2u_2 & & \vdots \\ \vdots & & \ddots & \\ v_nu_1 & \cdots & & v_nu_n \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{pmatrix}$$

Compare elements and you see that

$$v_j u_k = \delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases} . \tag{6}$$

These are *biorthogonality relations*. A left eigenvector v_j is orthogonal to every right eigenvector except u_j. Different left or right eigenvectors do not have to be orthogonal to each other.

If there is a basis of right eigenvectors (A is diagonalizable), then any vector may be represented as a linear combination (weighted sum) of right eigenvectors. For any $x \in \mathbb{R}^d$, there are weights w_k so that

$$x = \sum_{k=1}^{d} w_k u_k . \tag{7}$$

This may be described by saying that x has been *expanded* in the basis u_k with *expansion coefficients* w_k. The term *expansion* in this context refers to representing a general object, x in this case, in terms of a fixed collection of objects, u_k in this case. Taylor series is another kind of expansion, where a function $f(t)$ is represented as a sum of powers of t:

$$f(t) = \sum_{k=0}^{\infty} w_k t^k .$$

In this case, the expansion coefficients w_k are given in terms of the derivatives $f^{(k)}(0)$. There also are Fourier expansions (representing a function in terms of functions like $u_k(t) = e^{ikt}$) and others.

The expansion coefficients w_k in the right eigenvector expansion (7) are found using the left eigenvectors v_k:

$$w_k = l_k x . \tag{8}$$

You can find this formula using the bi-orthogonality relations (6). Basic linear algebra tells us that if there are d linearly independent vectors u_k, then they

are a basis, and therefore any x may be represented in terms of them as in (7). We multiply both sides of (7) by v_j and use the bi-orthogonality relations:

$$v_j x = v_j \left(\sum_{k=1}^{d} w_k u_k \right) \ .$$

On the right, we take v_j inside the sum. We move it past the coefficient w_k, which is just a number. The result is

$$v_j x = \sum_{k=1}^{d} w_k \left(v_j u_k \right) \ .$$

All the terms in this sum are zero, except the term $k = j$, which is 1. Therefore $v_j x = w_j$, which is the expansion coefficient formula (8).

The eigenvector expansion (7) is useful for understanding solutions of the differential equation

$$\frac{d}{dt} x = Ax \ . \tag{9}$$

Let x_t be a solution. Then the expansion coefficients are "time dependent" (are functions of t):

$$x_t = \sum_{k=1}^{d} w_{k,t} u_k \ , \quad w_{k,t} = v_k x_t \ .$$

The coefficients satisfy differential equations, which we find by substituting the expansion into the differential equation (9). On one side, we have

$$\frac{d}{dt} x_t = \sum_{k=1}^{d} \frac{dw_{k,t}}{dt} u_k \ .$$

On the other side, we have

$$\begin{aligned} Ax &= A \left(\sum_{k=1}^{d} w_{k,t} u_k \right) \\ &= \sum_{k=1}^{d} w_{k,t} A u_k \\ Ax &= \sum_{k=1}^{d} \lambda_k w_{k,t} u_k \ . \end{aligned}$$

Therefore

$$\sum_{k=1}^{d} \frac{dw_{k,t}}{dt} u_k = \sum_{k=1}^{d} \lambda_k w_{k,t} u_k \ .$$

Both sides represent expansions of the same vector in the same eigenvector basis, therefore the expansion coefficients are the same (expansion coefficients are unique). This gives, for each $k = 1, \ldots, d$,

$$\frac{dw_{k,t}}{dt} = \lambda_k w_{k,t} \ .$$

Of course, the solution is

$$w_{k,t} = e^{\lambda_k t} w_{k,0} \ . \tag{10}$$

The solution to the linear differential equation (9) is represented in terms of exponentials involving the possibly complex eigenvalues λ_k and possibly complex initial expansion coefficients $w_{k,0}$. An example of this is in Assignment 7.

Symmetric matrices, principal components

The eigenvalues and eigenvectors for symmetric matrices have more structure than for general matrices. They may be used for linear SDE if A is symmetric (which happens in certain applications). They also may used for *principal component analysis* of covariance matrices, as we will see.

The *variational formulation* is one way to understand the special features of the symmetric eigenvalue/eigenvector problem. A variational formulation of a math problem is a formulation that gives the answer as a maximizer or minimizer of something. Variational principles not only explain some properties of solutions, they also explain why solutions exist at all. It is easier to show that one function has a minimum or a maximum than it is to show that a system of equations in many variables has a solution. The variational formulation of the symmetric eigenvalue problem explains the fact that eigenvalues are real and the eigenvectors for different eigenvalues are orthogonal. It explains that there is a full basis of orthogonal eigenvectors even when there are "multiple eigenvalues" (more than one eigenvalue being equal).

The *Rayleigh quotient* is the function that is maximized or minimized to find eigenvalues and eigenvectors. Let C be a symmetric $d \times d$ matrix and $x \neq 0$ a $d-$component real column vector. The Rayleigh quotient is

$$R(x) = \frac{x^t C x}{x^t x} \ . \tag{11}$$

[This is named for Lord Rayleigh, an English lord, whose actual name was John William Strutt. He was a Cambridge professor in the late 1800's and early 1900's. His book *Theory of Sound* (where you can read about Rayleigh quotients and many other things) is still in print.]

The numerator $Q(x) = x^t C x$ is the *quadratic form* based on C. For $d = 1$, we could write $Q(x) = Cx^2$. For higher dimensions, $x^t C x$ is a number. It satisfies the binomial theorem in the sense that

$$Q(x + y) = (x + y)^t C (x + y) = x^t C x + 2x^t C y + y^t C y \ .$$

This is true only if C is symmetric. Otherwise, the middle term would be $x^tCy + y^tCx$. The middle term involves, $B(x,y) = x^tCy$, which is the *bilinear form* related from C. *Bilinear* means that it is linear as a function of x for every y and linear in y for every x. That is,

$$B(ax_1 + bx_2, y) = aB(x_1, y) + bB(x_2, y) \ , \quad \text{for every } y \ , \quad \text{and}$$
$$B(x, ay_1 + by_2) = aB(x, y_1) + bB(x, y_2) \ , \quad \text{for every } x \ .$$

If C is a symmetric matrix then $B(x,y)$ is a symmetric function in that $B(x,y) = B(y,x)$.

Suppose $Z = (Z_1, \ldots, Z_d)$ is a $d-$component random variable and $S = x_1Z_1 + \cdots + x_d, Z_d$ is a *linear combination*. Let C cov(Z) be the $d \times d$ covariance matrix of Z. Then the variance of S is the quadratic form involving C:

$$\text{var}(S) = x^tCx \ . \tag{12}$$

The Rayleigh quotient analysis of the variance of linear combinations is part of *principal component analysis*. The eigenvectors of C (or possibly the eigenvalues) are *principal components* of C. If $S_1 = x_1^tZ$ and $S_2 = x_2^tZ$, then

$$\text{cov}(S_1, S_2) = B(x_1, x_2) = x_1^tCx_2 \ .$$

If x_1 and x_2 are eigenvectors of C, then $\text{cov}(S_1, S_2) = 0$. Different eigenvectors of C correspond to different linear functionals S that are uncorrelated.

Here's the math relating *stationary values* of the Rayleigh quotient to eigenvalues. If $f(x)$ is any function, then a point x_* where $\nabla f(x_*) = 0$ is a *stationary point*. The corresponding value $f(x_*)$ is the stationary *value*. A stationary point can be a *local maximum* (if $f(x_*) \geq f(y)$ for y near x_*), a *local minimum* (if $f(x_*) \geq f(y)$ for y near x, or a *saddle point*. A local maximum is a *global maximum* if $f(x_*) \geq f(y)$ for all y. A local maximum is *strict* if $f(x_*) > f(y)$ for y near x_* and $y \neq x_*$. A global maximum is strict if $f(x_*) > f(y)$ for all $y \neq x_*$.

Stationary points of the Rayleigh quotient $R(x)$ are eigenvectors. The corresponding values of R are eigenvalues. The maximum of R is the largest eigenvalue and the minimum is the smallest. All eigenvalues between ($\lambda_{\min} < \lambda < \lambda_{\max}$ are saddle points.

Before getting to the good stuff about eigenvalues, we have to ask why the Rayleigh quotient has stationary points at all. The function $f(x) = x^3 + x$ has no stationary points. There is a basic principle that applies here: if $f(x)$ is continuous and defined on a subset of $\mathcal{R}^d$ that is closed and bounded, then there is some x where $f(x)$ is a maximum. A subset is $\mathcal{S}$ is *bounded* if there is an r so that $|x| \leq r$ for all $x \in \mathcal{S}$. Such an r is a *upper bound* for the sizes of $|x|$ for $x \in \mathcal{S}$. The subset $\mathcal{S} \subseteq \mathbb{R}^n$ is *closed* if it contains all its limit points. That means that if $x_n \in \mathcal{S}$ and $x_n \to x$ as $n \to \infty$, then $x \in \mathcal{S}$. An example in $d = 1$ is the interval $[0,1]$ which includes the endpoints 0 and 1. If $0 \leq x_n \leq 1$ and $x_n \to x$ as $n \to \infty$, then $0 \leq x \leq 1$. If you leave out the endpoint $x = 0$ to get $\mathcal{S} = (0,1]$, then the set is no longer closed. The points $x_n = \frac{1}{n}$ have $x_n \to 0$.

But $0 \neq \mathcal{S}$. The function $f(x) = \frac{1}{x}$ is continuous on this set $(0, 1]$, but it has no maximum value. Also, the function $f(x) = 1 - x$ has maximum value $f(0) = 1$, but $0 \notin \mathcal{S}$ and $f(x) < 1$ for $x \in \mathcal{S}$. This, again, $f(x)$ does not have a maximum value in $\mathcal{S}$.

These things do not happen for closed and bounded sets. The function $f(x) = \frac{1}{x}$ is not continuous on the closed set $[0, 1]$. It does not even have a value (you might call it ∞, which is not a number) for $x = 0$. The function $f(x) = 1 - x$ has maximum value $f(0) = 1$. A closed and bounded $\mathcal{S} \subseteq \mathbb{R}^d$ is called *compact*. But it might seem that $(0, 1]$ (the point zero left out) is just as "compact" as $[0, 1]$ (zero included). In $d > 1$ dimensions, the *unit sphere* of points $|x| = (x^t x)^{\frac{1}{2}}$ is compact. It is clear bounded (with $r = 1$). The limit of a sequence with $|x_n| = 1$ has $|x| = 1$. This unit sphere is called S_{d-1}. It is a theorem that a continuous function on a compact set has a maximum and a minimum. If you have not seen a proof, I hope the examples make the theorem seem reasonable. This is an "existence theorem" that says the maximum and minimum exist. But it does not say how to find them. An existence theorem that is not "constructive" (contain a recipe for finding it) may seem more like philosophy than mathematics.

The Rayleigh quotient $R(x)$ defined in (11) is defined for all $x \in \mathbb{R}^d$ with $x \neq 0$. This is not a compact set, so the theorem does not directly apply. The trick is to realize that $R(x)$ is *scale invariant*, which means that $R(ax) = R(x)$ for any number a. The scale invariance is a consequence of the fact that R is a quotient of two quadratic forms, each of which gets a factor of a^2 when scaled. That is $Q(ax) = (ax)^t C(ax) = a^2 x^t C x = a^2 Q(x)$, and $(ax)^t(ax) = a^2 x^t x$. These a^2 factors cancel in the ratio (11). If $x \neq 0$ is any vector, then there is an a so that $|ax| = 1$. Therefore, the maximum value of R is the same as the maximum value of R for $|x| = 1$. The set of vectors with $|x| = 1$ is called the *unit sphere*, and it is compact. But you don't have to restrict to the unit sphere when minimizing. Any minimizer (an $x_* \in \mathbb{R}^d$ that minimizes $R(x)$) has $\nabla R = 0$.

The relation between R and eigenvalues comes from the formula for ∇R. The quotient rule from calculus is one part of this calculation:

$$\begin{aligned} \nabla R(x) &= \nabla \frac{Q(x)}{x^t x} \\ &= \frac{\nabla Q(x)}{x^t x} - \frac{Q(x)}{x^t x} \frac{\nabla x^t x}{x^t x} \\ &= \frac{\nabla Q(x)}{x^t x} - R(x) \frac{\nabla x^t x}{x^t x} \ . \end{aligned}$$

We need a formula for $\nabla Q(x) = \nabla x^t C x$, starting with a formula for $x^t C x$. First,

$$x^t C x = \sum_{j=1}^{d} x_j \left(Cx \right)_j \ .$$

The matrix vector formula gives

$$(Cx)_j = \sum_{k=1}^{d} C_{jk} x_k \ .$$

We put these together to get

$$\begin{aligned} x^t C x &= \sum_{j=1}^{d} \left(\sum_{k=1}^{d} C_{jk} x_k \right) \\ &= \sum_{j=1}^{d} \sum_{k=1}^{d} x_j x_k C_{jk} \ . \end{aligned}$$

Component i of $\nabla x^t C x$ is

$$\partial_{x_i} \sum_{j=1}^{d} \sum_{k=1}^{d} x_j x_k C_{jk} = \sum_{j=1}^{d} \sum_{k=1}^{d} \left(\partial_{x_i} x_j x_k \right) C_{jk} \ .$$

There are three ways the term $\partial_{x_i} x_j x_k$ can be different from zero:

1. $i = j$ and $j \neq k$. Then the result is x_k. The sum of these terms is

   $$\sum_{k \neq i} C_{ik} x_k \ .$$

   It's C_{ik} because $j = i$.

2. $i = k$ and $j \neq k$. Then the result is x_j. The sum of these terms is

   $$\sum_{j \neq i} C_{ji} x_j \ .$$

   Note that i is the second index here (C_{ji}) and the first index above. This does not matter if C is symmetric, so $C_{ji} = C_{ij}$. We can call the summation variable k instead of j. Adding this sum to the previous one gives simply

   $$2 \sum_{k \neq i} C_{ik} x_k \ .$$

3. $i = j = k$. Then the result is $2x_i$. The term is $2C_{ii}x_i$. This is the term missing in the sum $\sum_{k \neq i}$.

We put it in and get

$$\partial_{x_i} x^t C x = 2 \sum_{k=1}^{d} C_{ik} x_k = 2 \left(Cx \right)_i \ .$$

On the right, we recognize the sum as component i of Cx. This shows that

$$\nabla Q(x) = \nabla x^t C x = 2Cx \; . \tag{13}$$

In dimension $d = 1$, this becomes $\frac{d}{dx} Cx^2 = 2Cx$, which is the formula from ordinary calculus.

If C is not symmetric, the formula is (you can check this)

$$\nabla x^t C x = Cx + C^t x = 2 \left[\frac{1}{2} \left(C + C^t \right) x \right] \; .$$

The matrix $\frac{1}{2}\left(C + C^t\right)$ is called the *symmetric part* of C. The *anti-symmertic part* is $\frac{1}{2}\left(C - C^t\right)$. The symmetric part is a symmetric matrix and the anti-symmetric part is anti-symmetric (check this). You can see that C is the sum of its symmetric and anti-symmetric parts. If C is symmetric then C is equal to its symmetric part and its anti-symmetric part is zero. The gradient of $x^t C x$ involves only the symmetric part of C. In fact, the quadratic form $Q(x) = x^t C x$ only "sees" the quadratic part of C. Since Q is a number, it is equal to its transpose. Therefore

$$Q = Q^t = \left(x^t C x \right)^t = x^t C^t \left(x^t \right)^t = x^t C^t x \; .$$

Thus the quadratic form based on C and the one based on C^t are the same, they produce the same values. The average of them also produces this value. The average is

$$\frac{1}{2} \left(x^t C x + x^t C^t x \right) = x^t \left[\frac{1}{2} \left(C + C^t \right) \right] x \; .$$

This is the quadratic form based on the symmetric part of C. The conclusion is: if you have a quadratic form involving a matrix C, you may as well use the symmetric part of C instead. In other words, any quadratic form is equivalent to a quadratic form with a symmetric C.

This fact may seem interesting, but it is not necessary to symmetrize (take the symmetric part) in most applications. If you have a quadratic form that came from some problem, the matrix C probably was symmetric already. For example, a covariance matrix is symmetric. Other quadratic forms arise naturally in other problems. I don't know any of them that isn't naturally symmetric.

Back to the Rayleigh quotient, our gradient formula (13) allows us to compute

$$\nabla R = \frac{2}{x^t x} Cx - \frac{2}{x^t x} R(x) x \; . \tag{14}$$

If we set the gradient to zero to find a stationary point, the result (cancelling common factors) is

$$Cx_* = R(x_*) x_* \; . \tag{15}$$

This implies that x_* is an eigenvector and $\lambda = R(x_*)$ is the corresponding eigenvalue. To summarize: let $u_1 \in \mathbb{R}^d$ be the vector that minimizes $R(x)$ with $\|u_1\| = 1$. Then this u_1 is an eigenvector of C with eigenvalue $\lambda_1 = R(u_1)$. If x

is any other eigenvector of C, which satisfies $Cx = \lambda x$, then $\lambda \geq \lambda_1$. This last claim true because you can calculate the Rayleigh quotient for an eigenvector:

$$R(x) = \frac{x^t C x}{x^t x} = \frac{x^t \lambda x}{x^t x} = \lambda \, .$$

Since u_1 minimized R, the value λ cannot be lower than the minimum, which is λ_1.

The eigenvalues $\lambda_2 \geq \lambda_1$ and $\lambda_3 \geq \lambda_2$ also have variational principles. For these we need to talk about *orthogonal complements* and *invariant subspaces*, as well as Rayleigh quotients. If $\mathcal{T} \subseteq \mathbb{R}^d$ is a subspace, the *orthogonal complement* is the set of vectors $x \in \mathbb{R}^d$ that are orthogonal to $\mathcal{T}$ (orthogonal to every vector in $\mathcal{T}$. It is written $\mathcal{T}^\perp$. We will call it $\mathcal{S}$. It is defined by

$$x \in \mathcal{T}^\perp \iff x^t y = 0 \text{ for all } y \in \mathcal{T} \, .$$

["Complement" with an "e" is something that makes something complete. The subspace $\mathcal{S} \subseteq \mathbb{R}^d$ completes $\mathcal{T}$ in the sense that $\mathcal{S}$ and $\mathcal{T}$ together span all of $\mathbb{R}^d$. "Compliment" with an "i" means to say something nice about someone. Subspaces don't do that.]

The subspace $\mathcal{S}$ is *invariant* under the "action" of a matrix C if $x \in \mathcal{S} \Longrightarrow Cx \in \mathcal{S}$. The word "invariant" means that C does not change $\mathcal{S}$. Vectors in $\mathcal{S}$ "go to" other vectors in $\mathcal{S}$ under the action of C, so the vectors are not invariant. But the subspace as a whole stays the same. Alas, this is not quite true. Define the set $C\mathcal{S}$ to be "image" of $\mathcal{S}$ under the action of C. That is

$$C\mathcal{S} = \{Cx \mid x \in \mathcal{S}\} \, .$$

Invariance maybe should mean that $C\mathcal{S} = \mathcal{S}$. But our definition is only that $\mathcal{S}$ is invariant if $C\mathcal{S} \subseteq \mathcal{S}$. It can happen that this inclusion is *strict*, which means that there is some $x \in \mathcal{S}$ so that $x \notin C\mathcal{S}$. If $\mathcal{S}$ is invariant under the action of C, then C is a linear transformation on $\mathcal{S}$. The inclusion $C\mathcal{S} \subset \mathcal{S}$ is strict if and only C is singular (in the sense of singular matrices or singular linear transformations) as a linear transformation of $\mathcal{S}$.

The variational principles for λ_2, is based on the fact that the orthogonal complement of u_1 is invariant under the action of C. [This is an "abuse of terminology" because the subspace $\mathcal{T}$ is replaced by the vector u_1, but being orthogonal to u_1 is the same as being orthogonal to the vector space "generated by u_1.] The orthogonal complement of u_1 is the set of vectors with $x^t u_1 = 0$. Call this subspace $\mathcal{S}_2$. We will see that u_2 with $Cu_2 = \lambda_2 u_2$ is in $\mathcal{S}_2$. The definition is

$$x \in \mathcal{S}_2 \iff x^t u_1 = 0 \, .$$

We will show that $\mathcal{S}_2$ is an invariant subspace for C. This is the same as saying that if $x^t u_1 = 0$, then $(Cx)^t u_1 = 0$, so $Cx \in \mathcal{S}_2$ also.

This is a proof by contradiction. We will see that if $\mathcal{S}_2$ is not invariant, then u_1 is not the minimizer. More concretely, if there is an x with $x^t u_1 = 0$ but $(Cx)^t u_1 \neq 0$, then $\nabla R(u_1) \neq 0$. In fact, not only is the gradient of R

at u_1 not equal to zero, but the directional derivative in the x direction is not zero. That means that you can find a lower value of R by moving in the x direction (possibly in the $-x$ direction). Roughly speaking, if $x^t u_1 = 0$ (x is perpendicular to u_1), then (draw a picture) $\|u_1 + \epsilon x\| \approx \|u_1\|$. More precisely, $\|u_1 + \epsilon x\| = \|u_1\| + O(\epsilon^2)$ (Calculations below). On the other hand, if $(Cx)^t u_1 \neq 0$, then $(u_1 + \epsilon x)^t C(u_1 + \epsilon x) = u_1^t C u_1 + O(\epsilon)$. If ϵ is small enough, the $O(\epsilon)$ is bigger than $O(\epsilon^2)$.

The directional derivative/gradient calculation is motivated by the geometry of the previous paragraph but it skips the $O(\cdot)$ stuff. The gradient formula (14) leads to (using $u_1^t u_1 = 1$)

$$\nabla R(u_1) = 2Cu_1 - 2R(u_1)u_1 \ .$$

Suppose $x^t u_u = 0$ but $x^t C u_1 \neq 0$. Then the directional derivative in the x_1 direction at u_1 is

$$x^t \nabla R(u_1) = 2x^t C u_1 - 2R(u_1) x^t u_1 \neq 0 \ .$$

That's it. If $u_1^{\perp} = \mathcal{S}_2$ is not an invariant subspace, then u_1 does not minimize R.

This happens in other places – the minimizer has extra orthogonality properties. A very simple example: take $x \in \mathbb{R}^d$ and let $\mathcal{T}$ be a subspace. Let $y \in \mathcal{T}$ be the point in $\mathcal{T}$ closest to x in the sense that $\|x - y\| = [(x-y)^t(x-y)]^{\frac{1}{2}}$ is minimized over all $y \in \mathcal{T}$. Then $(x - y)$ is perpendicular to $\mathcal{T}$ (draw a picture in the plane with $\mathcal{T}$ being a line that does not go through x). That is, if $z \in \mathcal{T}$, then $(x-y)^t z = 0$. It's a similar proof by contradiction. If $(x-y)^t z \neq 0$, then $\|x - (y + \epsilon z)\|$ can be made smaller than $\|x - y\|$ by taking ϵ small enough and with the correct sign.

Now we know that $\mathcal{S}_2$ is an invariant subspace for C. That is, C acts on the vector space $\mathcal{S}_2$. We can minimize the Rayleigh quotient over $\mathcal{S}_2$ (leave out $x = 0$) and find a number λ_2 as the minimum value:

$$\lambda_2 = \min_{x \in \mathcal{S}_2, x \neq 0} R(x) \ .$$

The value is $\lambda_2 \geq \lambda_1$. This is because λ_1 was the overall minimum and λ_2 is the minimum only considering "people" in $\mathcal{S}_2$. We also saw that the minimizer is an eigenvector. Since $\mathcal{S}_2$ is an invariant subspace, if u_2 is an eigenvector in $\mathcal{S}_2$, then u_2 is an eigenvector in $\mathbb{R}^d$. More simply, the equation $Cu_2 = \lambda_2 u_2$ holds in $\mathcal{S}_2$, so it also is true in $\mathbb{R}^d$. It is the same C and matrix multiplication means the same thing.

We can continue in this way. Define $\mathcal{S}_3$ to be the orthogonal complement of the vector space spanned by u_1 and u_2. Or (this is equivalent), define $\mathcal{S}_3$ to be the orthogonal complement to u_2 in $\mathcal{S}_2$. Then $\mathcal{S}_3$ also is an invariant subspace for C. Minimizing R over $\mathcal{S}_3$ gives u_3 and $\lambda_3 = R(u_3) \geq \lambda_2 \geq \lambda_1$.

This process stops when we run out of dimensions. The orthogonal complement to u_1 is defined by one constraint, so it has dimension $d-1$. The dimension

of $\mathcal{S}_3$ is $\dim(\mathcal{S}_2) - 1 = d - 2$, and so on. The result u_1, ..., u_d, all normalized, all orthogonal, all eigenvectors.

The maximum of $R(x)$ is given by $\lambda_n = R(u_n)$. This is because $\lambda_1 \leq \cdots \leq \lambda_n = R(u_n)$. Therefore, we could have found the eigenvalues and eigenvectors starting from the top and working down rather than starting at the bottom and working up. This would have been

$$\begin{aligned} \lambda_n &= \max_{x \neq 0} R(x) \\ \lambda_{n-1} &= \max_{x \neq 0, x^t u_n = 0} R(x) \\ &\text{etc.} \end{aligned}$$

The arguments about orthogonal complements and invariant subspaces apply equally well if we do it this way (check if you don't believe this). It may have been better to do this, given that PCA uses the eigenvalues and eigenvectors starting from the top. But it's traditional to take $\lambda_1 \leq \lambda_2 \leq \cdots$, so we would have had to call the first (largest) eigenvalue λ_n.

Principal component analysis

Principal component analysis, usually called *PCA*, refers to using the eigenvectors and eigenvalues of C to find components of the random variable X that have the largest variance. Let X be a d component random variable with covariance C. The components of X is the eigenvector directions are

$$Y_k = u_k^t X \ .$$

The Y_k may be called *principal components*, or the eigenvectors u_k may be called that. The Y_k may have a stronger claim to the name because they are components of X, in the sense of linear algebra. We will call the eigenvectors principal *directions*, but that is not a standard term. It is common that a multi-component random variable is highly variable in some directions and has less variation in other directions. If $v \in \mathbb{R}^d$, the variance corresponding to v might be $\text{var}(v^t X)$. This can be increased or decreased by scaling v, which is cheating. Therefore, we require that $\|v\| = 1$. If we maximize the variance with this constraint, the problem is

$$\max_{\|v\|=1} \text{var}(v^t X) \ .$$

But we have the variance formula (12), so this is λ_n. The corresponding v is u_n. For this reason, $Y_n = u_n^t X$ is called the largest principal component of X.

The second largest principal component is a random variable $v^t X$ that is uncorrelated with Y_n that has the largest variance. The covariance of $Z = v^t X$ with Y_n is $v^t C u_n = \lambda_n v^t u_n$. The random variables are uncorrelated if v is orthogonal to u_n, which is $v^t u_n = 0$. If we maximize the variance of Z with the constraint $\|v\| = 1$ and $v^t u_n = 0$, the answer is (we saw above) $v = u_{n-1}$. This makes $Y_{n-1} = u_{n-1}^t X$ the second largest principal component of X.

Statisticians talk about *explained* and *unexplained* variance. When talking about data, this is the explained and unexplained sum of squares. Suppose Z and W are correlated random variables both with mean zero. A *linear prediction* of Z using W is a formula $\widehat{Z} = \beta W$. The variance of Z is (given that Z has mean zero)

$$\sigma_Z^2 = \mathrm{E}\big[\, Z^2\big] \ .$$

The explained variance is the variance of the predictor

$$\sigma_{\widehat{Z}}^2 = \mathrm{E}\Big[\, \widehat{Z}^2\Big] \ .$$

[Coefficients like β may be called *regression coefficients* and are often denoted by β.] The *mean square* prediction error is

$$R^2 = \mathrm{E}\left[\left(Z - \widehat{Z}\right)^2\right] \ .$$

[Here, "mean" means "average" or "expected value". R^2 is the expected value of the square of the prediction error.] The optimal β is the value of *beta* that minimizes the mean square prediction error. The next paragraph has a formula for it. The optimal mean square prediction error is also the unexplained variance, because of the formula

$$\sigma_Z^2 = \sigma_{\widehat{Z}}^2 + R^2 \ . \tag{16}$$

This formula looks like the Pythagorean theorem. Like the Pythagorean theorem, this formula is related to orthogonality. Minimizing the prediction error, it turns out, is equivalent to making the prediction error "orthogonal to" (uncorrelated to) the estimator.

The calculation of the optimal regression coefficient β will be familiar to many. You write the mean square prediction error as a quadratic in β, then optimize the quadratic. That is:

$$\begin{aligned}
\mathrm{E}\left[\left(Z - \widehat{Z}\right)^2\right] &= \mathrm{E}\Big[(Z - \beta W)^2\Big] \\
&= \mathrm{E}\big[Z^2 - \beta ZW + \beta W^2\big] \\
&= \sigma_Z^2 - \beta \mathrm{cov}(Z,W) + \beta^2 \sigma_W^2 \ .
\end{aligned}$$

The maximum of this satisfies

$$2\mathrm{cov}(Z,W) = 2\beta\sigma_W^2 \implies \beta = \frac{\mathrm{cov}(Z,W)}{\sigma_W^2} \ .$$

The covariance of the prediction error with the prediction variable is

$$\begin{aligned}
\mathrm{cov}(Z - \widehat{Z}, W) &= \mathrm{cov}(Z,W) - \beta\sigma_W^2 \\
&= \mathrm{cov}(Z,W) - \frac{\mathrm{cov}(Z,W)}{\sigma_W^2}\sigma_W^2 \\
&= 0 \ .
\end{aligned}$$

This leads to the orthogonality formula (16), starting with the representation of Z as the sum of the prediction and the prediction error:

$$Z = \widehat{Z} + \left(Z - \widehat{Z} \right) \ .$$

The expected square is

$$\begin{aligned} \mathrm{E}\big[Z^2\big] &= \mathrm{E}\left[\widehat{Z}^2 + 2\widehat{Z}\left(Z - \widehat{Z} \right) + \left(Z - \widehat{Z} \right)^2 \right] \\ &= \sigma^2_{\widehat{Z}} + 0 + R^2 \ . \end{aligned}$$

This is the formula (16) writing the total variance as the sum of the explained variance and the unexplained variance. We review it here because it is important and because it motivates a similar discussion involving principal components.

Suppose we "predict" the d component random variable X using the k largest principal components. The predictor is

$$\widehat{X}_k = Y_n u_n + Y_{n-1} u_{n-1} + \cdots + Y_{n-k+1} u_{n-k+1} \ .$$

For example, you can check that $k = 1$, only the one largest principal component, leads to one term on the right. We want something like the variance to measure the size of X. Since X is a vector, we can use the 2 norm, which is

$$\|X\|^2 = X^t X = \sum_{j=1}^{n} X_j^2 \ .$$

But this is random, we use its expected value

$$S_X^2 = \mathrm{E}\big[\, X^t X \big] \ .$$

The size of the prediction is

$$S^2_{\widehat{X}_k} = \mathrm{E}\Big[\, \widehat{X}_k^t \widehat{X}_k \Big] \ .$$

The size of the error, measured this way, is

$$R_k^2 = \mathrm{E}\left[\left(X - \widehat{X}_k \right)^t \left(X - \widehat{X}_k \right) \right] \ .$$

The orthogonality built into the u_j leads to approximation formulas like for linear regression above. Analogous to (16), we have a multi-component version

$$S_X^2 = S^2_{\widehat{X}_k} + R_k^2 \ . \tag{17}$$

Each of these quantities may be expressed in terms of the eigenvalues of C (also called principal components, wrongly I would argue, but still). The

formulas are

$$S_X^2 = \sum_{j=1}^{n} \lambda_j \tag{18}$$

$$S_{\widehat{X}_k}^2 = \sum_{j=n-k+1}^{n} \lambda_j \tag{19}$$

$$R_k^2 = \sum_{j=1}^{n-k} \lambda_j \tag{20}$$

Recall that $\lambda_j = \mathrm{var}(Y_j)$. The "explained sum" involving the principal components used (19) is equal to

$$\sum_{j=n-k+1}^{n} \mathrm{var}(Y_j) \ .$$

You explain the most of X by using the principal components with the largest variances. The proof is a calculation that uses orthogonality of the eigenvectors. One of these calculations will explain how to do the other ones and make the identities (18) (19) and (20) clear. You can see that our explained/unexplained thing (17) follows from these. Here is the verification of (20). It uses the trick that represents the square of a sum as a double sum

$$\left(\sum_j a_j \right)^2 = \sum_i \sum_j a_i a_j \ .$$

$$\begin{aligned}
R_k^2 &= \mathrm{E}\left[\left(X - \widehat{X}_k \right)^t \left(X - \widehat{X}_k \right) \right] \\
&= \mathrm{E}\left[\left(\sum_{j=1}^{n=k} Y_j u_j \right)^t \left(\sum_{j=1}^{n=k} Y_j u_j \right) \right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n=k} \sum_{j=1}^{n=k} Y_i Y_j u_i^t u_j \right] \\
&= \sum_{i=1}^{n=k} \sum_{j=1}^{n=k} \mathrm{E}[\, Y_i Y_j] \, u_i^t u_j \ .
\end{aligned}$$

The u_j are ortho-normal, so $u_i^t u_j = 0$ unless $i = j$, and then $u_j^t u_j = 1$. Also, $\mathrm{E}[\, Y_j Y_j] = \lambda_j$. Therefore,

$$R_k^2 = \sum_{j=1}^{k} \lambda_j \ .$$

This is it.