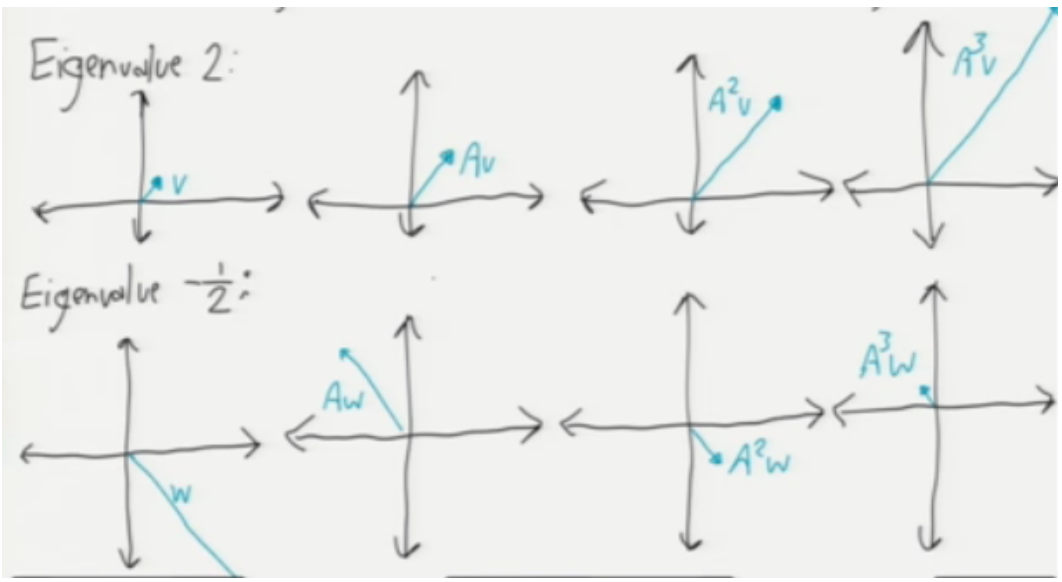
8 Eigenvectors and the Anisotropic Multivariate Normal Distribution

EIGENVECTORS

[I don't know if you were properly taught about eigenvectors here at Berkeley, but I sure don't like the way they're taught in most linear algebra books. So I'll start with a review. You all know the definition of an eigenvector:]

Given square matrix A, if $Av = \lambda v$ for some vector $v \neq 0$, scalar λ, then
v is an eigenvector of A and λ is the eigenvalue of A associated w/v.

[But what does that mean? It means that v is a magical vector that, after being multiplied by A, still points in the *same direction*, or in exactly the *opposite direction*.]



Draw this figure by hand (eigenvectors.png)

[For most matrices, most vectors don't have this property. So the ones that do are special, and we call them eigenvectors.]
[Clearly, when you scale an eigenvector, it's still an eigenvector. Only the direction matters, not the length. Let's look at a few consequences.]

Theorem: if v is eigenvector of A w/eigenvalue λ,
then v is eigenvector of A^k w/eigenvalue λ^k [k is a +ve integer; we will use Theorem later]

Proof: $A^2 v = A(\lambda v) = \lambda^2 v$, etc.

Theorem: moreover, if A is invertible,
then v is eigenvector of A^{-1} w/eigenvalue $1/\lambda$

Proof: $A^{-1} v = A^{-1}(\frac{1}{\lambda} A v) = \frac{1}{\lambda} v$ [look at the figures above, but go from right to left.]

[Stated simply: When you invert a matrix, the eigenvectors don't change, but the eigenvalues get inverted. When you square a matrix, the eigenvectors don't change, but the eigenvalues get squared.]

[Those theorems are pretty obvious. The next theorem is not obvious at all.]

Spectral Theorem: every real, symmetric $n \times n$ matrix has real eigenvalues and
n eigenvectors that are mutually orthogonal, i.e., $v_i^\top v_j = 0 \quad$ for all $i \neq j$

[This takes about a page of math to prove.
One minor detail is that a matrix can have more than n eigenvector directions. If two eigenvectors happen to have the same eigenvalue, then every linear combination of those eigenvectors is also an eigenvector. Then you have infinitely many eigenvector directions, but they all span the same plane. So you just arbitrarily pick two vectors in that plane that are orthogonal to each other. By contrast, the set of eigenvalues is always uniquely determined by a matrix, including the multiplicity of the eigenvalues.]

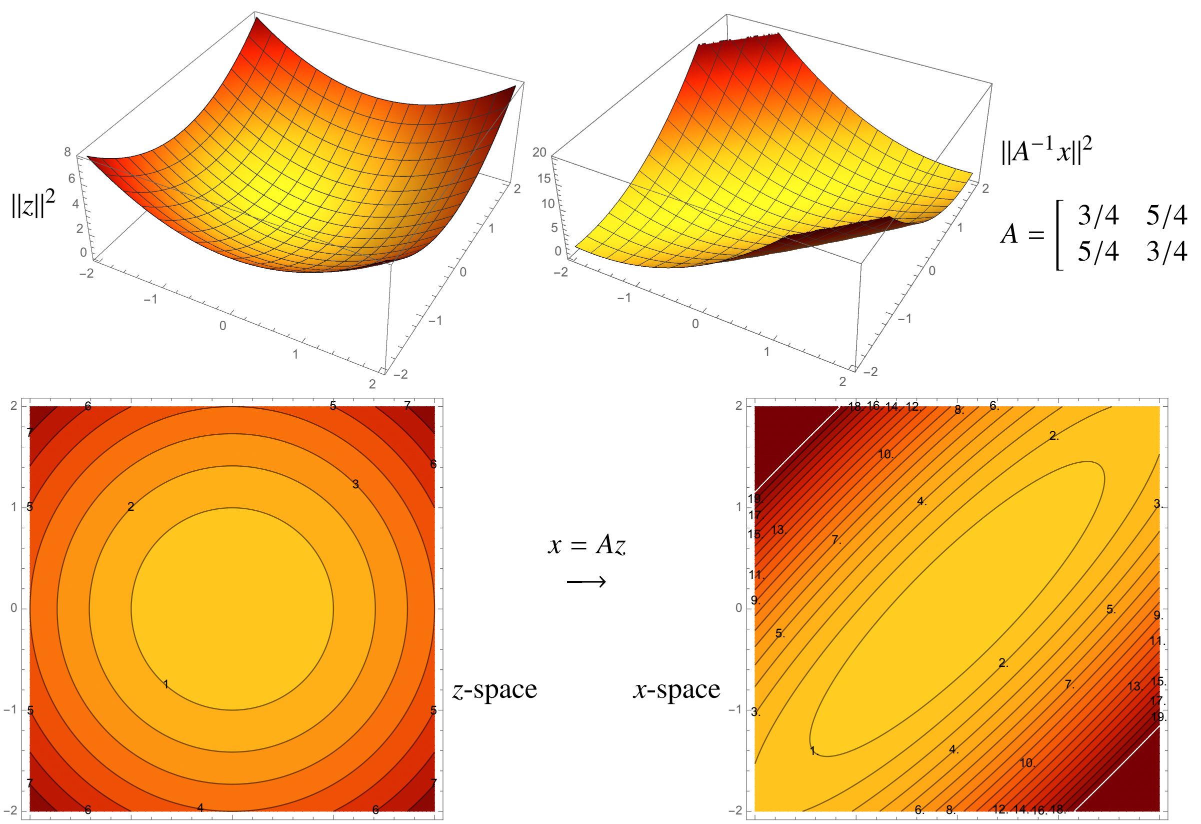
We can use them as a basis for $\mathbb{R}^n$.

Quadratic Forms

[My favorite way to visualize a symmetric matrix is to graph something called *the quadratic form*, which shows how applying the matrix affects the length of a vector. The following example uses the same two eigenvectors and eigenvalues as above.]

$\|z\|^2 \quad = z^\top z \qquad \Leftarrow$ quadratic; isotropic; isosurfaces are spheres

$\|A^{-1}x\|^2 \quad = x^\top A^{-2} x \qquad \Leftarrow$ quadratic form of the matrix $A^{-2} \qquad$ (A symmetric)
anisotropic; isosurfaces are ellipsoids



circlebowl.pdf, ellipsebowl.pdf, circles.pdf, ellipses.pdf

[Both figures at left are plots of $\|z\|^2$, and both figures at right are plots of $\|A^{-1}x\|^2$.
(Draw the stretch direction $(1, 1)$ with eigenvalue 2 and the shrink direction $(1, -1)$ with eigenvalue $-\frac{1}{2}$ on the ellipses at bottom right.)]

[The matrix A maps the circles on the left to the ellipses on the right. They're stretching along the direction with eigenvalue 2, and shrinking along the direction with eigenvalue $-1/2$. Let's verify that.]

$\|A^{-1}x\|^2 = 1$ is an ellipsoid with axes $v_1, v_2, \ldots, v_n$ and
radii $\lambda_1, \lambda_2, \ldots, \lambda_n$

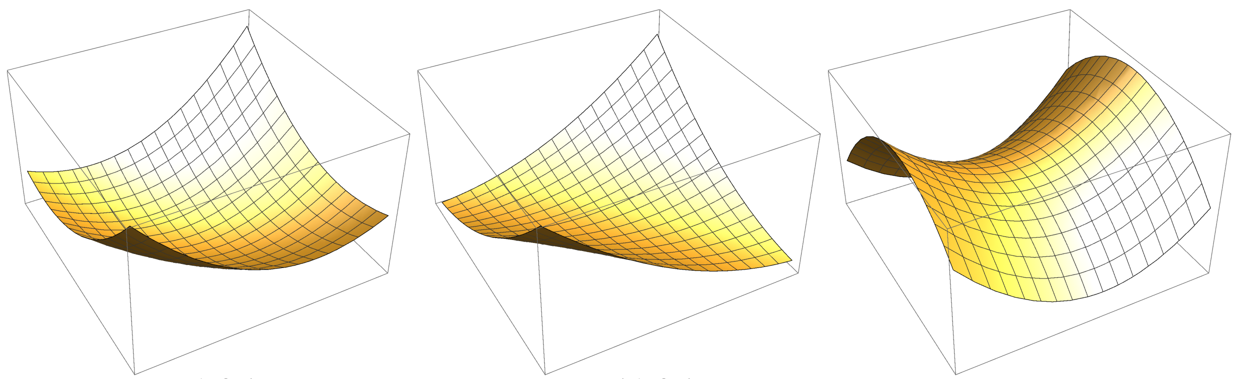
because if $A^{-1}x = v_i$ has length 1 (v_i lies on circle), $x = Av_i$ has length λ_i (Av_i lies on the ellipsoid).

Special case: A is diagonal $\Leftrightarrow$ eigenvectors are coordinate axes
$\Leftrightarrow$ ellipsoids are axis-aligned

[Draw axis-aligned isocontours for a diagonal metric.]

A symmetric matrix M is

- positive definite if $w^\top M w > 0$ for all $w \neq 0$. $\Leftrightarrow$ all eigenvalues positive
- positive semidefinite if $w^\top M w \geq 0$ for all w. $\Leftrightarrow$ all eigenvalues nonnegative
- indefinite if +ve eigenvalue & $-$ve eigenvalue
- invertible if no zero eigenvalue



pos definite   pos semidefinite   indefinite

$\boxed{\text{posdef.pdf, possemi.pdf, indef.pdf}}$

[Examples of quadratic forms for positive definite, positive semidefinite, and indefinite matrices. Positive eigenvalues correspond to axes where the curvature goes up; negative eigenvalues correspond to axes where the curvature goes down. (Draw the eigenvector directions, and draw the flat trough in the positive semidefinite bowl.)]

What does this tell us about $x^\top A^{-2} x$?

[We've been visualizing the quadratic form of a matrix A^{-2}. The eigenvalues of A^{-2} are the inverse squares of the eigenvalues of A, so A^{-2} cannot have a negative eigenvalue. Moreover, A^{-2} cannot have a zero eigenvalue, because A cannot have an infinite eigenvalue (and A^{-2} does not exist if A has a zero eigenvalue.) Therefore, A^{-2} is positive definite (if it exists).]

What about the isosurfaces of $x^\top M x$ for a +ve definite M?

[If M is positive definite, the contour plot of M's quadratic form has ellipsoidal isosurfaces whose radii are determined by the eigenvalues of $M^{-1/2}$, which are the inverse square roots of the eigenvalues of M. We will use these ideas to define Gaussian distributions, and for that, we'll need a strictly positive definite matrix.]

[If M is only positive semidefinite, but not positive definite, the isosurfaces are cylinders instead of ellipsoids. These cylinders have ellipsoidal cross sections spanning the directions with nonzero eigenvalues, but they run in straight lines along the directions with zero eigenvalues.]

Building a Quadratic

[There are a lot of applications where you're given a matrix, and you want to extract the eigenvectors and eigenvalues. But when you're learning the math, I think it's more intuitive to go in the opposite direction. Suppose you have an ellipsoid isosurface in mind. Suppose you pick the ellipsoid axes and the radius along each axis, and you want to create the matrix whose quadratic form will have isosurfaces matching the ellipsoid of your dreams.]

Choose n mutually orthogonal **unit** n-vectors $v_1, \ldots, v_n$ [so they specify an orthonormal coordinate system]
Let $V = [v_1 \quad v_2 \quad \ldots \quad v_n]$ $\quad \Leftarrow n \times n$ matrix
Observe: $V^\top V = I$ [off-diagonal 0's because the vectors are orthogonal]
[diagonal 1's because they're unit vectors]

$$\Rightarrow \quad V^\top = V^{-1} \quad \Rightarrow \quad VV^\top = I$$

V is orthonormal matrix: acts like rotation (or reflection)

Choose some radii λ_i:

$$\text{Let } \Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$$ [diagonal matrix of eigenvalues]

Defn. of "eigenvector": $AV = V\Lambda$
[This is the same definition of eigenvector I gave you at the start of the lecture—$Av = \lambda v$—but this is how we express it in matrix form, so we can cover all the eigenvectors in one statement.]

$\Rightarrow AVV^\top = V\Lambda V^\top$ [which leads us to . . .]

Theorem: $A = V\Lambda V^\top = \sum_{i=1}^{n} \lambda_i \underbrace{v_i v_i^\top}_{\text{outer product: } n \times n \text{ matrix, rank 1}}$ has chosen eigenvectors/values

This is a matrix factorization called the eigendecomposition. [every real, symmetric matrix has one]
Λ is a diagonalized version of A.
$V^\top$ rotates the ellipsoid to be axis-aligned.
Paraboloid with specified axes and radii: $\|A^{-1}x\| = 1$ or $x^\top A^{-2} x = 1$.
[This completes our task of specifying a paraboloid whose isosurfaces are ellipsoids with specified axes and radii.]

Observe: $A^2 = V\Lambda V^\top V\Lambda V^\top = V\Lambda^2 V^\top$ $\qquad A^{-2} = V\Lambda^{-2}V^\top$
[This is another way to see that squaring a matrix squares its eigenvalues without changing its eigenvectors. It also suggests a way to define a matrix square root.]
Given a symmetric PSD matrix Σ, we can find a symmetric square root $A = \Sigma^{1/2}$:

compute eigenvectors/values of Σ
take square roots of Σ's eigenvalues
reassemble matrix A [with the same eigenvectors as Σ but changed eigenvalues]

[The first step of this algorithm—computing the eigenvectors and eigenvalues of a matrix—is much harder than the remaining two steps.]

ANISOTROPIC GAUSSIANS

[Let's revisit the multivariate Gaussian distribution, with different variances along different directions.]

$X \sim \mathcal{N}(\mu, \Sigma)$ [X and μ are d-vectors. X is a random variable with mean μ.]

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right)$$

$\uparrow$ determinant of Σ

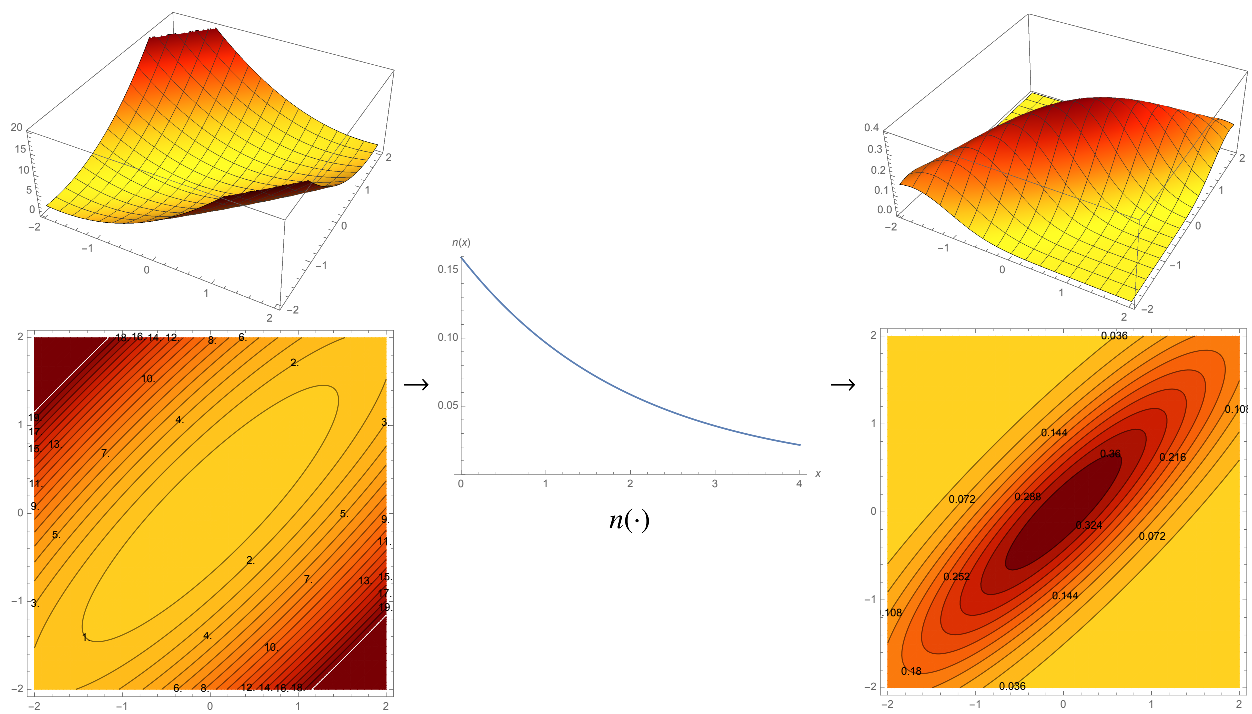
Σ is the $d \times d$ SPD covariance matrix.
Σ^{-1} is the $d \times d$ SPD precision matrix.

Write $f(x) = n(q(x))$, where $q(x) = (x-\mu)^\top \Sigma^{-1} (x-\mu)$

$\uparrow$ $\mathbb{R} \to \mathbb{R}$, exponential $\quad$ $\uparrow$ $\mathbb{R}^d \to \mathbb{R}$, quadratic

[Now $q(x)$ is a function we understand—it's just a quadratic bowl centered at μ, the quadratic form of the precision matrix Σ^{-1}. The other function $n(\cdot)$ is a simple, monotonic, convex function, an exponential of the negation of its argument. This mapping $n(\cdot)$ does not change the isosurfaces.]

Principle: given monotonic $n : \mathbb{R} \to \mathbb{R}$, isosurfaces of $n(q(x))$ are same as $q(x)$ (different isovalues).



$q(x)$ $\quad$ $f(x) = n(q(x))$

ellipsebowl.pdf, ellipses.pdf, exp.pdf, gauss3d.pdf, gausscontour.pdf

[(Show this figure on a separate "whiteboard" for easy reuse next lecture.) A paraboloid (left) becomes a bivariate Gaussian (right) after you compose it with the univariate Gaussian (center).]

[One of the main ideas is that if you understand the isosurfaces of a quadratic function, then you understand the isosurfaces of a Gaussian, because they're the same. The differences are in the isovalues—in particular, the Gaussian achieves its maximum at the mean, and decreases to zero as you move infinitely far away from the mean.]

$q(x)$ is the squared distance from $\Sigma^{-1/2}x$ to $\Sigma^{-1/2}\mu$. Consider the metric

$$d(x,\mu) = \left\|\Sigma^{-1/2}x - \Sigma^{-1/2}\mu\right\| = \sqrt{(x-\mu)^\top\Sigma^{-1}(x-\mu)} = \sqrt{q(x)}.$$

[So we think of the precision matrix as a "metric tensor" which defines a metric, a sort of warped distance from x to the mean μ.]

covariance: Let R, S be random variables—column vectors or scalars
$\text{Cov}(R,S) = \text{E}[(R - \text{E}[R])\,(S - \text{E}[S])^\top] = \text{E}[RS^\top] - \mu_\text{R}\,\mu_\text{S}^\top$
$\text{Var}(R) = \text{Cov}(R,R)$
If R is a vector, covariance matrix for R is

$$\text{Var}(R) = \begin{bmatrix} \text{Var}(R_1) & \text{Cov}(R_1,R_2) & \dots & \text{Cov}(R_1,R_d) \\ \text{Cov}(R_2,R_1) & \text{Var}(R_2) & & \text{Cov}(R_2,R_d) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(R_d,R_1) & \text{Cov}(R_d,R_2) & \dots & \text{Var}(R_d) \end{bmatrix} \qquad \text{[symmetric; each } R_i \text{ is scalar]}$$

For a Gaussian $R \sim \mathcal{N}(\mu,\Sigma)$, one can show $\text{Var}(R) = \Sigma$.
[. . . by integrating the expectation in anisotropic spherical coordinates. It's a painful integral.]
[An important point is that statisticians didn't just arbitrarily decide to call Σ a covariance matrix. Rather, statisticians discovered that if you find the covariance of the normal distribution by integration, it turns out that the covariance is Σ. This is a happy fact; it's rather elegant.]

R_i, R_j independent $\Rightarrow$ $\text{Cov}(R_i,R_j) = 0$ [the reverse implication is not generally true, but . . .]
$\text{Cov}(R_i,R_j) = 0$ AND multivariate normal dist. $\Rightarrow$ R_i, R_j independent
all features pairwise independent $\Rightarrow$ $\text{Var}(R)$ is diagonal [the reverse is not generally true, but . . .]
$\text{Var}(R)$ is diagonal AND joint normal
$\Leftrightarrow$ axis-aligned Gaussian; squared radii on diagonal of $\Sigma = \text{Var}(R)$
$$\Leftrightarrow \quad \underbrace{f(x)}_{\text{multivariate}} = \underbrace{f(x_1)\,f(x_2)\cdots f(x_d)}_{\text{univariate Gaussians}}$$

[So when the features are independent, you can write the multivariate Gaussian PDF as a product of univariate Gaussian PDFs. When they aren't, you can do a change of coordinates to the eigenvector coordinate system, and write it as a product of univariate Gaussian PDFs in eigenvector coordinates.]