Mean

The *mean* of a random variable (also called its *expected value* or *first moment*) is defined as

$$E(X) = \sum_x x\, P(X = x)$$

For example, if X is the number of spots that show up on the roll of a fair die, then

$$E(X) = \frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6 = 3\frac{1}{2}$$

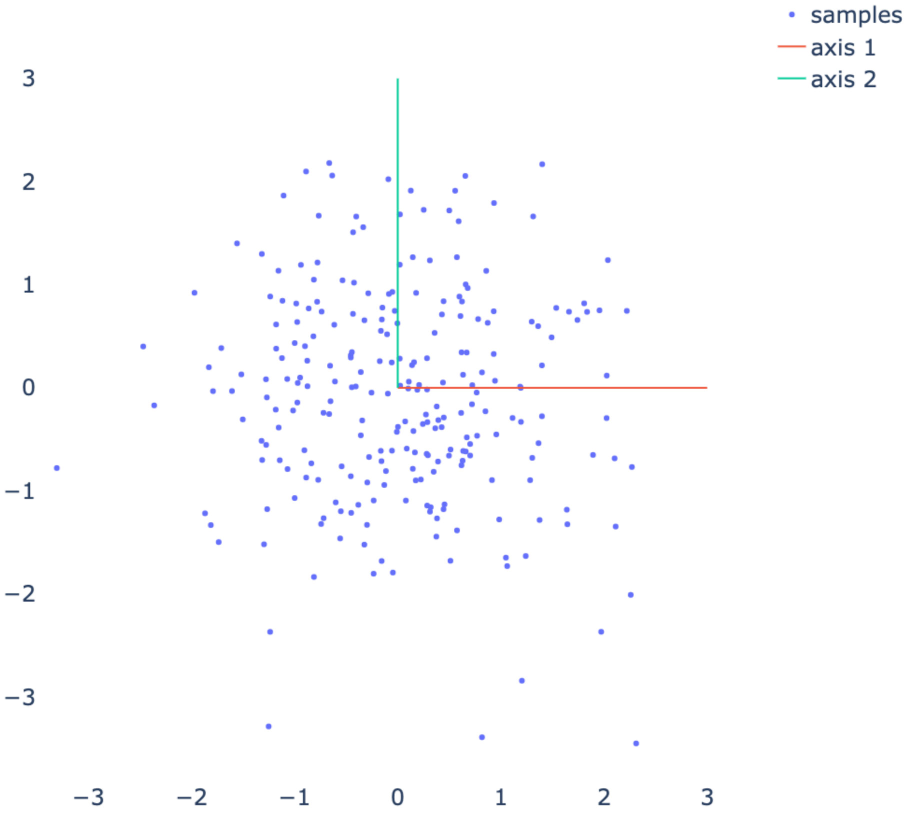
The mean tells us what the value of X will be on average: sometimes the value will be larger, sometimes it will be smaller, but the total positive and negative fluctuations will tend to cancel out over time.

An alternate notation for the mean is $E(X) = \bar{X}$. (We've also used a bar to denote the complement of a Boolean random variable, but the meaning will typically be clear from context.)

The mean of X has the same units as X itself: e.g., if X is the height of a person in cm, then $E(X)$ is measured in cm as well.

Scatter plot

In low dimensions we can visualize our data with a *scatter plot*: we take a sample from $P(X)$ and plot each of the samples $X_1, X_2, \ldots$ as a single point.



The mean is the center of mass of the scatter plot. While there might not be any samples exactly on the mean, the displacements of the individual samples away from the mean tend to add up to zero. (The average displacement would be exactly zero in an infinitely large sample.)

Linearity of expectation

One of the most useful properties for working with means is *linearity of expectation*: for a real-valued random variable X and constants a, b,

$$E(aX + b) = a\,E(X) + b$$

> *That is, expectation is a linear function from random variables to real numbers, so that we can pass linear functions into or out of expectations.*

For example, since the expected number of spots on a die roll is 3.5, the expected value of three times the number of spots is 10.5.

We can do this with more complicated expressions too. For example, the solution to the normal equations $XX^Tw = Xy^T$ depends linearly on y when X is fixed. So if y is a random variable, then the solution w is also a random variable, and we can write $XX^T\,E(w) = X\,E(y^T)$.

Variance

The mean tells us what value our random variable will take on average, but it ignores how much the variable fluctuates around this average. There are a number of ways to quantify this fluctuation — that is, to answer how far a random variable will typically be from its mean. The most widely used ways are a pair of related measurements called the *variance* and the *standard deviation*.

The variance is defined as

$$V(X) = E((X - E(X))^2)$$

In words, the variance is the expected squared difference between a random variable and its mean.

For example, suppose X is a fair coin flip: it takes values 0 and 1 with equal probability. The mean is $E(X) = \frac{1}{2}$. The difference between X and $E(X)$ is either $+\frac{1}{2}$ or $-\frac{1}{2}$, with equal probability; squaring and averaging tells us that $V(X) = \frac{1}{4}$.

The units of variance are the *square* of the units of the original random variable. For example, the variance of a height might be measured in cm^2. This is not very intuitive. So, it's common to report the *standard deviation* instead: this is the square root of the variance, and it is often written $\sigma(X)$. The standard deviation has the same units as X and $E(X)$.

The standard deviation is a pretty good match for our intuitive idea of how far a random variable tends to be from its mean. Its main flaw is that it is sensitive to *outliers*: that is, if there is a small probability of encountering a measurement that is very far from the mean, that will have an outsized effect on the standard deviation.

For example, let's make a low-probability change to our fair coin flip from above. Suppose that X takes values 0 or 1 with probability 49.95% each; but with the remaining probability of 0.1%, it takes value 100. The low-probability event causes the mean of X to move somewhat away from 0.5, to around 0.6. But the variance is now around 10 instead of $\frac{1}{4}$ — a much larger change than the mean.

This sort of sensitivity means that it can be difficult to estimate the variance accurately. In the example above, if we take fewer than 1000 samples of X, we have a good chance of not even seeing the large value, and thinking that X is an ordinary fair coin.

Exercise: if the variance of X is 1, what is the variance of $aX + b$?

Exercise: if the standard deviation of X is 1, what is the standard deviation of $aX + b$?

Normalizing a random variable

When working with a random variable, it's common to (approximately) subtract out its mean, so that its mean is (approximately) zero. This is called *centralizing* the random variable.

After centralizing, it's also common to (approximately) divide by the standard deviation, so that the standard deviation is (approximately) 1. This is called *standardizing* or *z-scoring* the random variable.

Both of these transformations are sometimes called *normalizing* the random variable; this is a less-specific name that encompasses any processing that's intended to remove some kind of idiosyncracy in a random variable.

Normalizing affects different downstream machine learning procedures in different ways. So, it is sometimes but not always a good idea to normalize; and different learning procedures need different kinds of normalization.

Moments

We can measure lots of additional properties of a random variable by looking at $E(f(X))$ for different functions f. Such expectations are called *moments* of X.

For example, if

$$Q(X) = \begin{cases} 1 & \text{if } X \in [3,4] \\ 0 & \text{o/w} \end{cases}$$

then $E(Q(X))$ is the probability that $X \in [3,4]$. Or, if

$$Q(X) = e^{2\pi i t X}$$

for some value of t, then $E(Q(X))$ tells us whether X has a periodicity of length $2\pi/t$: this moment will far from zero if there is some value x such that X tends to take values close to $x, x \pm 2\pi/t, x \pm 4\pi/t, \ldots$, and close to zero if not (i.e., if X is approximately aperiodic).

The polynomials are a common and useful source of functions to use in defining moments. The expectations of the monomials X, X^2, X^3, ... are common enough to have special names: they are called the first, second, third, ... moments.

We can recognize from the definitions that the mean is the first moment. It's common to remove the mean (centralize the random variable) before taking second, third, and higher moments, so that we get $E((X - E(X))^2)$, $E((X - E(X))^3)$, and so forth. These are called *central moments*, and we can recognize from the definitions that the variance is the second central moment.

The third central moment is called *skew*, and it measures the symmetry of a distribution: if X has positive skew it means that large positive values of X are more likely than large negative ones, while negative skew means the reverse. The fourth central moment is called *kurtosis*, and it measures the balance between small and large values of X: if we keep the variance fixed, a high kurtosis means that we put higher probability on very large values of X while compensating by making small values smaller.

High order polynomial moments are even more sensitive to outliers than the variance is. That makes it even harder to estimate them from data. But, if we know lots of moments, that tells us a lot about a random variable.

Exercise: what is the variance of a biased coin flip, in terms of the probability p of seeing 1? What about skew or kurtosis?

Covariance and correlation

Suppose that we have two random variables X and Y. The covariance between them is defined as

$$\mathrm{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

If the covariance is positive, it means that positive values of X tend to co-occur with positive values of Y, and negative values of X tend to co-occur with negative values of Y. If the covariance is negative, it means the reverse: positive values of X are seen with negative values of Y on average, and vice versa.

If we standardize X and Y, then their covariance is also called the *correlation* of X and Y:

$$\operatorname{Corr}(X, Y) = E\left(\frac{X - E(X)}{\sigma(X)} \frac{Y - E(Y)}{\sigma(Y)}\right)$$

Correlation is always in the range $[-1, 1]$: a correlation of $+1$ means that Y is a linear function of X with positive slope, while a correlation of -1 means that Y is a linear function of X with negative slope.

Correlation and covariance are not the same; for example, covariance can be bigger than $+1$ or smaller than -1. We haven't defined independence yet, but we will see later that it is separate from both covariance and correlation. If two variables are independent, they will have zero covariance and zero correlation, but the reverse is not true.

Exercise: given an example of two random variables that are perfectly dependent (i.e., one is a function of the other) but have zero correlation.

Conditional mean and variance

Suppose we have a random variable X with some mean $E(X)$ and variance $V(X)$. If we condition on an event A, the distribution of X can change; this new distribution will have its own mean and variance, which we write $E(X \mid A)$ and $V(X \mid A)$. These are called the *conditional* mean and variance of X given A.

If we have a family of events $Y = 1, Y = 2, \ldots$, then $E(X \mid Y)$ represents a table: it shows $E(X \mid Y = y)$ for all possible values of y.

Vector-valued random variables

If X is a random variable that takes values in some vector space V, then we define $E(X)$ exactly as before:

$$E(X) = \sum_x x \, P(X = x)$$

If our vector space is $\mathbb{R}^n$, this is equivalent to taking the mean *componentwise*. (This follows from linearity of expectation.)

For $X \in \mathbb{R}^n$, we generalize the definition of the variance of X: we define the *covariance matrix* of X to be

$$V(X) = E[(X - E(X))(X - E(X))^T]$$

or, expanding,

$$V(X) = \sum_x (x - E(X))(x - E(X))^T P(X = x)$$

By comparing this formula to the scalar case, we can see that the diagonal elements of $V(X)$ are the variances of the individual components of X, and the off-diagonal elements are the covariances of pairs of components of X. We'll also sometimes refer to $V(X)$ as just the variance of X, with the understanding that $V(X)$ is a scalar if X is a scalar, and a matrix if X is a vector.

Note that the covariance matrix is always symmetric — we can see this from the above expression, or by noting that the covariance of X_i and X_j doesn't depend on which way we order them.

Also note that the covariance matrix is always positive semidefinite: from the definition of variance and linearity of expectation, for any x we have

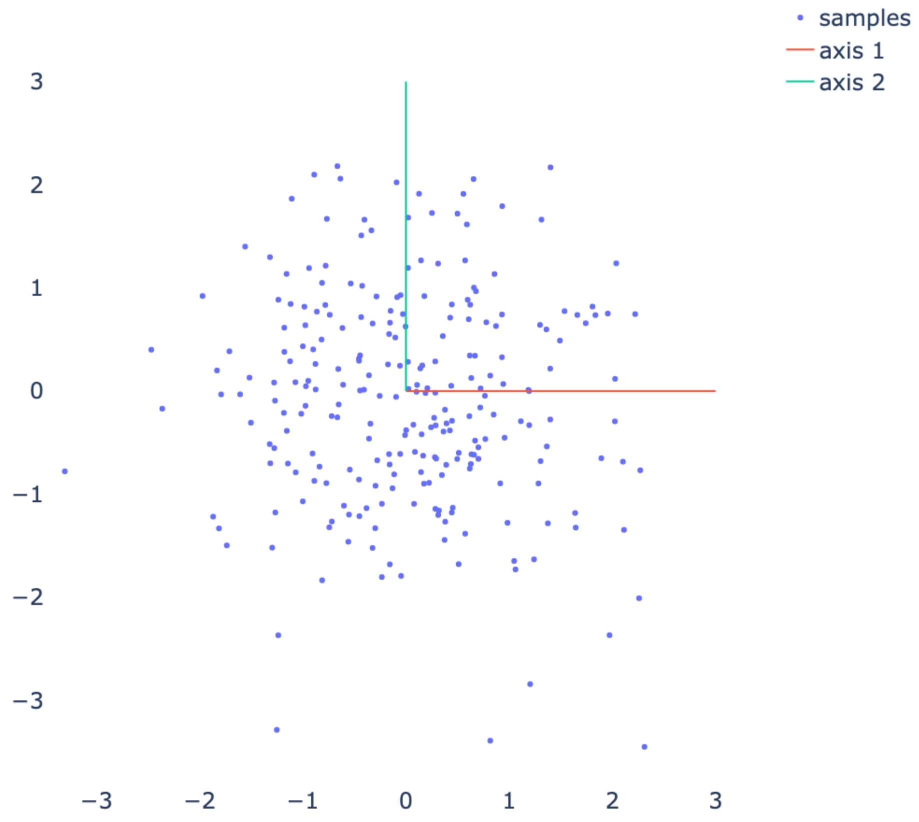
$$x^T V(X) x = E[x^T (X - E(X))(X - E(X))^T x] = E[(x^T (X - E(X))^2] \geq 0$$

> *If X takes values in some inner product space V instead of $\mathbb{R}^n$, we define the variance to be a linear operator on V: let $Y = X - E(X)$ and set $V(X)\,z = E(Y\langle Y, z\rangle)$ for each $z \in V$. This reduces to the above definition if $V = \mathbb{R}^n$.*

The simplest possible covariance matrix is the identity, $V(X) = I$. This means that each individual coordinate has variance 1, and any two coordinates are uncorrelated. We can make a sample with variance I by sampling each coordinate independently and standardizing. NumPy provides a convenient method for doing so: `numpy.random.randn(m, n)` makes an $m \times n$ matrix of independent random variables with mean zero and variance 1. If we take each column of this matrix as one of our samples, we will have the variance matrix equal to I as desired.

Interpreting covariance matrices

The covariance matrix $\Sigma = V(X)$ tells us something about the shape of the distribution of X. In $\mathbb{R}^2$ or $\mathbb{R}^3$ we can visualize this shape with a scatter plot.

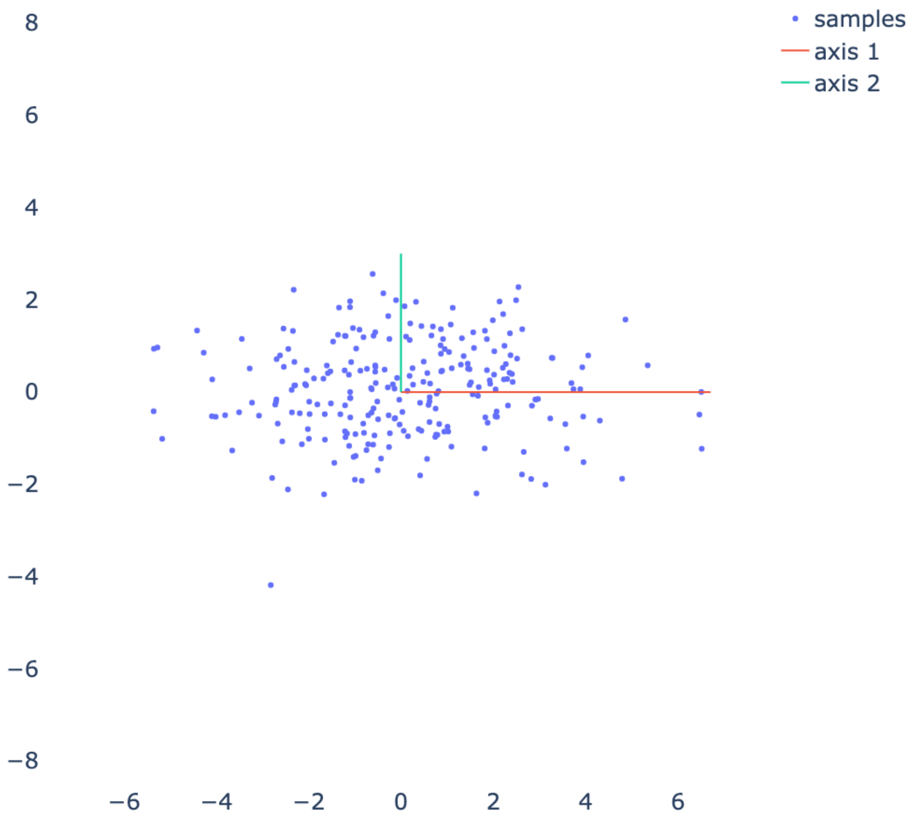


By looking at the plot, we can see some information about how the samples tend to vary. This information shows up as well in the covariance matrix. In the sample shown above, the covariance matrix is the identity, which results in a perfectly symmetric blob of points around the origin.

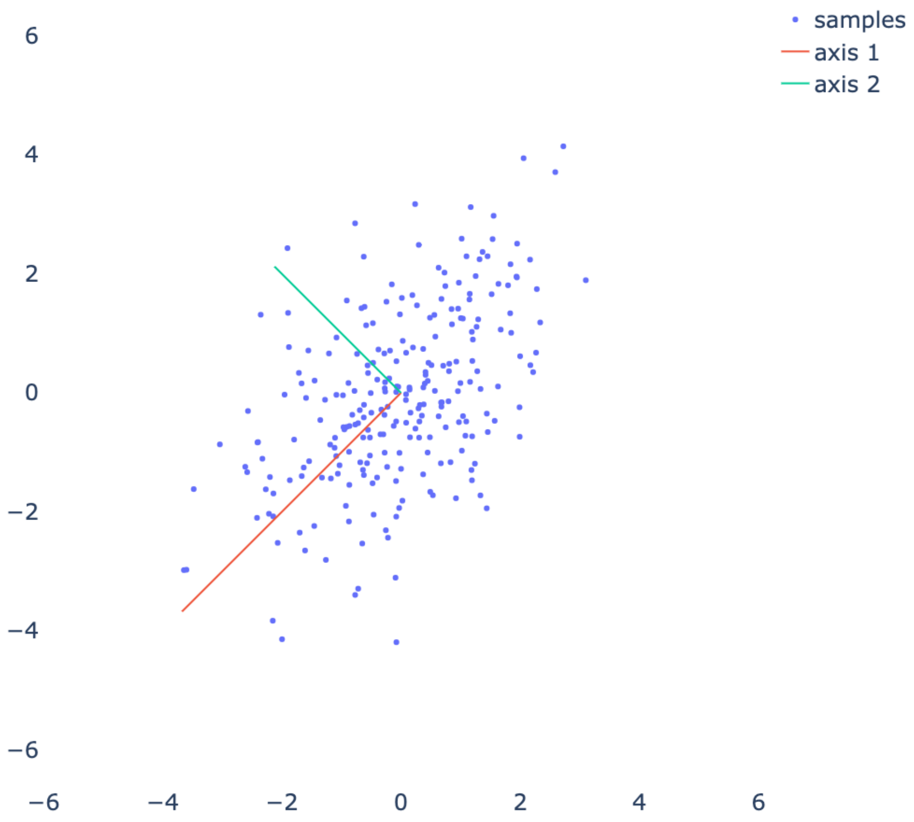
> *In this section we're showing scatter plots for samples that have zero mean. If the mean were nonzero they would look the same except for being shifted.*

The diagonal entries of Σ are the variances of individual coordinates of X. A large variance Σ_{ii} means that the distribution of X_i is spread out: the scatter plot will stretch out along dimension i. The spread is proportional to the standard deviation, the square root of Σ_{ii}. A small variance means that the corresponding coordinate is tightly packed together. In the limit, a variance of 0 means that the value is always the same, making the scatter plot into a line in 2D or a pancake in 3D.

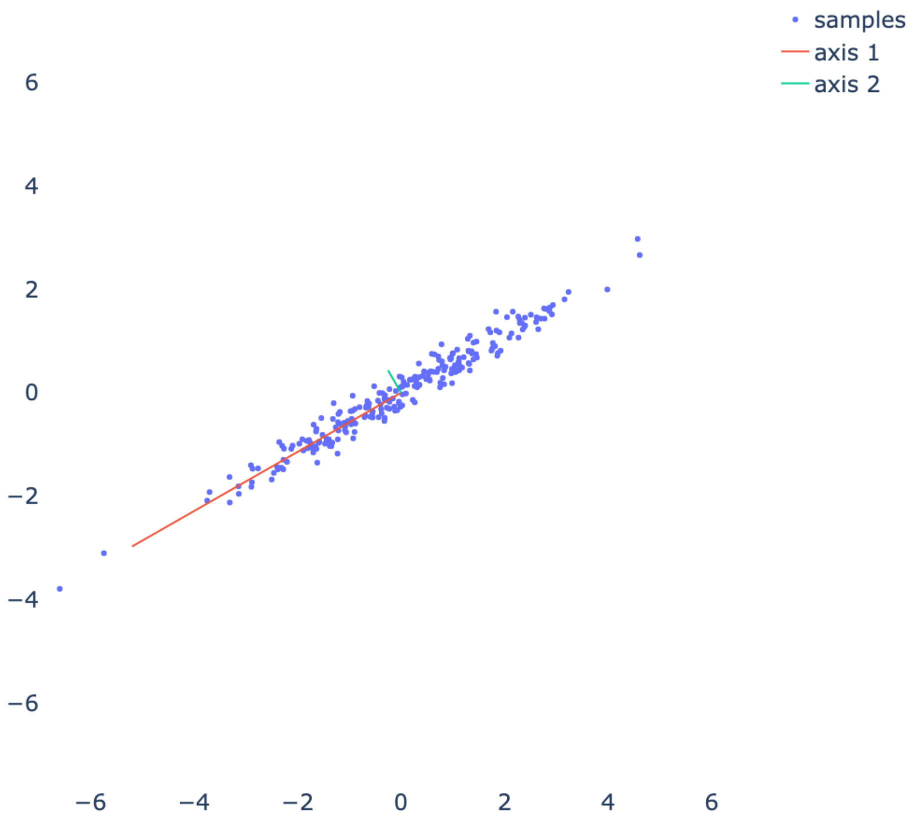
If $V(X)$ is a diagonal matrix, then this sort of stretching or squashing of the individual axes means that our data distribution will look basically like an ellipsoid, with the axes of the ellipsoid pointing along the coordinate axes. The diameter of the ellipsoid in each direction will be proportional to the standard deviation in that direction (the square root of the variance). Here is a plot where the horizontal axis has variance 5 and the vertical axis has variance 1.



The off-diagonal entries of Σ are the covariances of pairs of coordinates of X. A positive covariance means that the two coordinates tend to have the same sign; so, the scatter plot stretches out to the top right and bottom left. The larger the covariance, the more the scatter plot stretches. A negative covariance, on the other hand, means that the two coordinates tend to have opposite signs. In the following plot, each coordinate of X has variance 2, and the covariance is $+1$:



If we fix the variances of two coordinates while increasing their covariance, their relationship will become closer and closer to deterministic. In 2D, the scatter plot will approach a diagonal line, so that we can perfectly predict one coordinate from the other. The slope of the line will depend on the standard deviations of the individual coordinates. The largest possible covariance will be the *geometric mean* of the individual variances, $\Sigma_{ij} = \pm\sqrt{\Sigma_{ii}\Sigma_{jj}}$. At this value, the scatter plot will be a perfect line, and the correlation between the two coordinates will be ± 1. (The sign depends on whether the slope of the line is positive or negative.) In the following plot, the covariance is $\begin{pmatrix} 3 & 1.7 \\ 1.7 & 1 \end{pmatrix}$; note that 1.7 is just slightly smaller than $\sqrt{3}$, the geometric mean of 1 and 3.



Variance of a linear transform

If we apply a linear transform A to a random variable X, we have

$$V(AX) = A\,V(X)A^T$$

We can prove this fact using linearity of expectation.

Exercise: do so. (Hint: assume without loss of generality that $E(X) = 0$; then use the definition of variance $V(X) = E(XX^T)$.)

There are a lot of uses for this identity. For example, we can use it to generate samples with any desired covariance matrix Σ, given a factorization of Σ. Say we have the Cholesky factorization $\Sigma = LL^T$: we start by generating samples X with covariance I. If we then transform them to $Y = LX$, we have

$$V(Y) = V(LX) = L\,V(X)L^T = LL^T = \Sigma$$

SVD

Putting all of the above together, we can now describe what a scatter plot looks like for a general covariance matrix. To do so, we'll use the singular value decomposition of $V(X)$:

$$V(X) = USU^T$$

Here U is square and orthonormal, and S is diagonal and nonnegative. Note that we've used the form of the SVD for symmetric positive semidefinite matrices, where we have the same orthonormal matrix on both the left and the right.

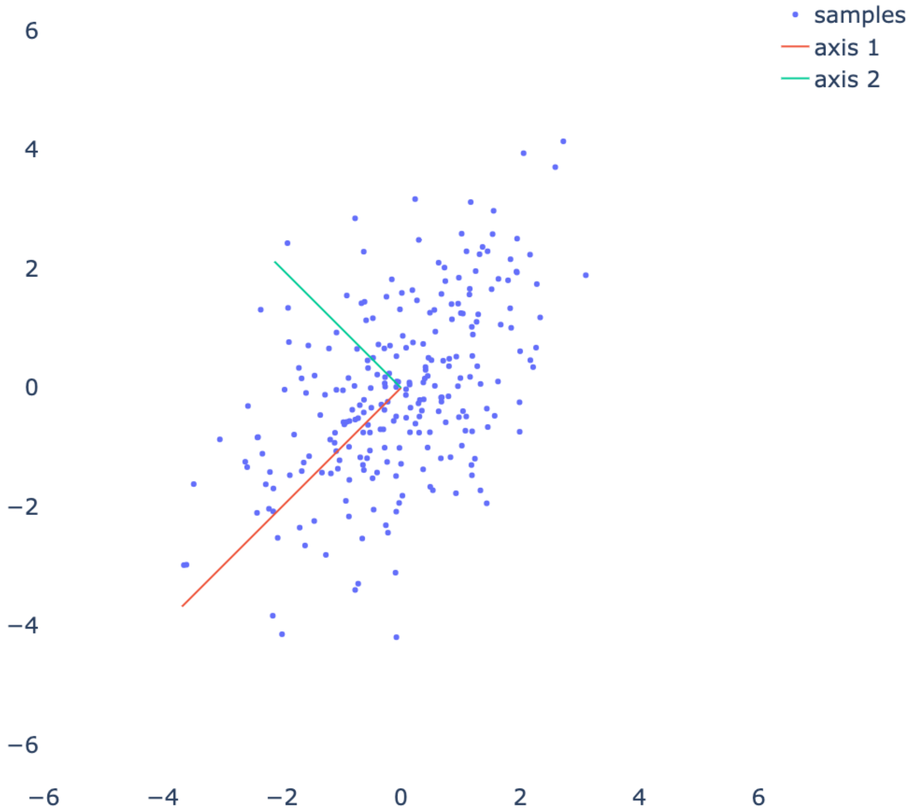
Using the identity for covariance of a linear transform, we know that

$$V(U^T X) = U^T V(X) U = (U^T U) S (U^T U) = S$$

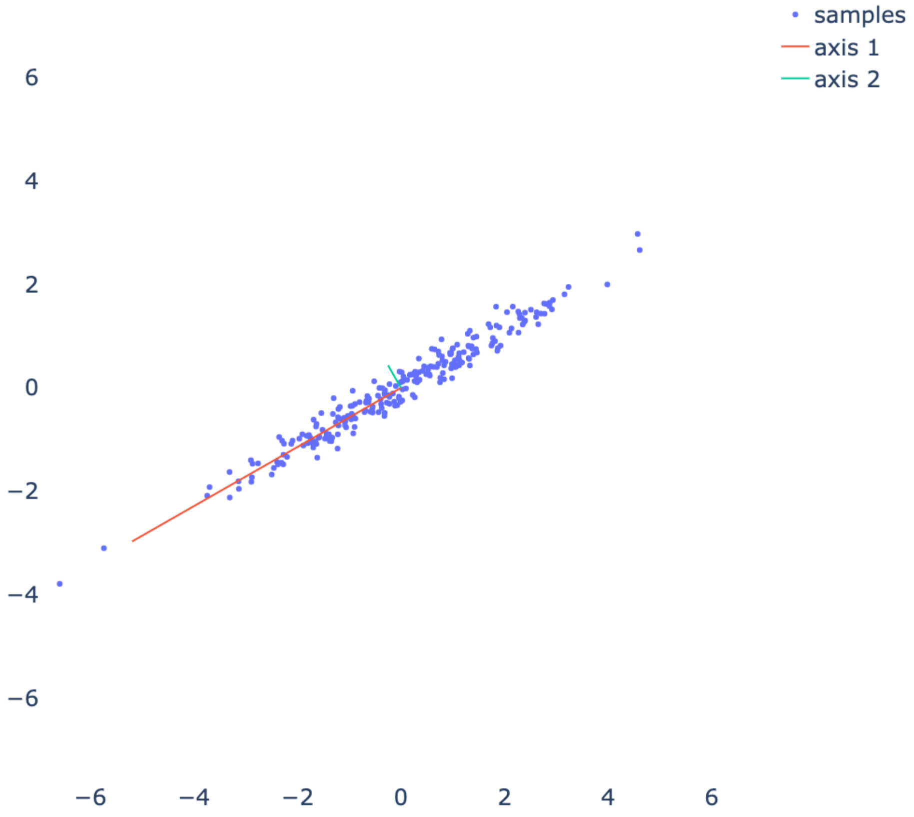
That is, the covariance of $Y = U^T X$ is diagonal. That means we understand the shape of the distribution of Y: its scatter plot will look like an axis-parallel ellipsoid, with the diameter along axis i proportional to $\sqrt{S_{ii}}$.

But now we can go back to $X = UY$. We can think of the columns of U as an orthonormal basis. Multiplying Y by U means transforming each coordinate axis of Y into one of the basis vectors. Since the basis vectors are orthonormal, the overall effect will be a combination of rigid rotations and reflections.

So, the scatter plot of X will look like a rotated (and maybe reflected) version of the scatter plot of Y: that is, it will be an ellipsoid, but not necessarily axis-parallel any more. Each column U_i will point along one axis of the ellipsoid, and the spread of the ellipsoid along the direction U_i will be proportional to $\sqrt{S_{ii}}$.



● samples
axis 1
axis 2
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6



The vectors U_i are called the singular vectors of the covariance matrix. The variances S_{ii} are called the singular values. For a symmetric PSD matrix, the singular values and singular vectors are also called the eigenvalues and eigenvectors — though eigenvectors and singular vectors are different for general matrices. So you will often see a distribution described in terms of the eigenvectors of its covariance matrix.