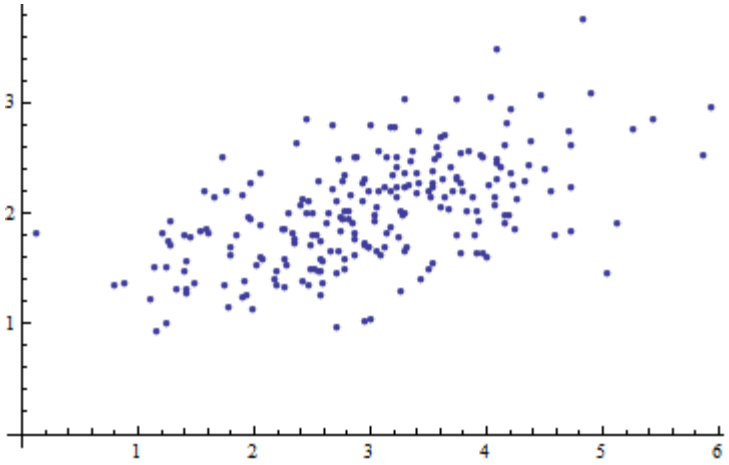
Mahalanobis Distance

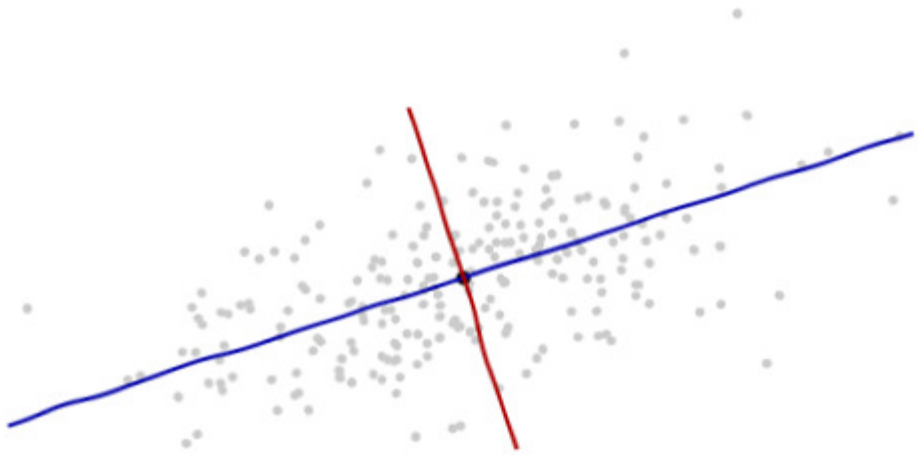
Here is a scatterplot of some multivariate data (in two dimensions):



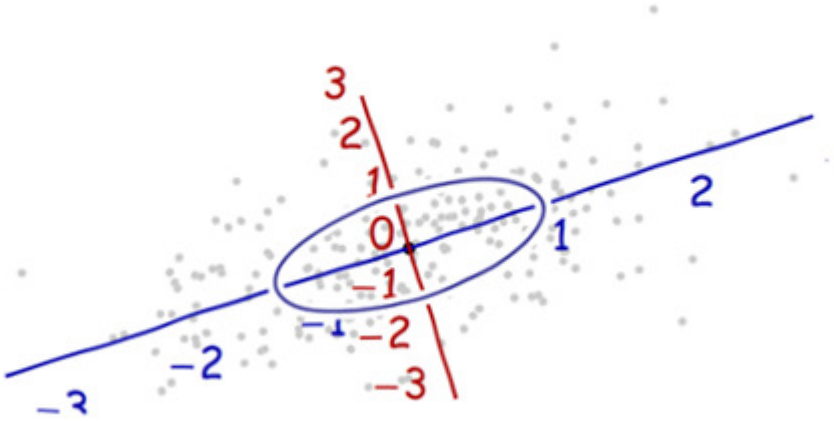
What can we make of it when the axes are left out?

Introduce coordinates that are suggested by the data themselves.

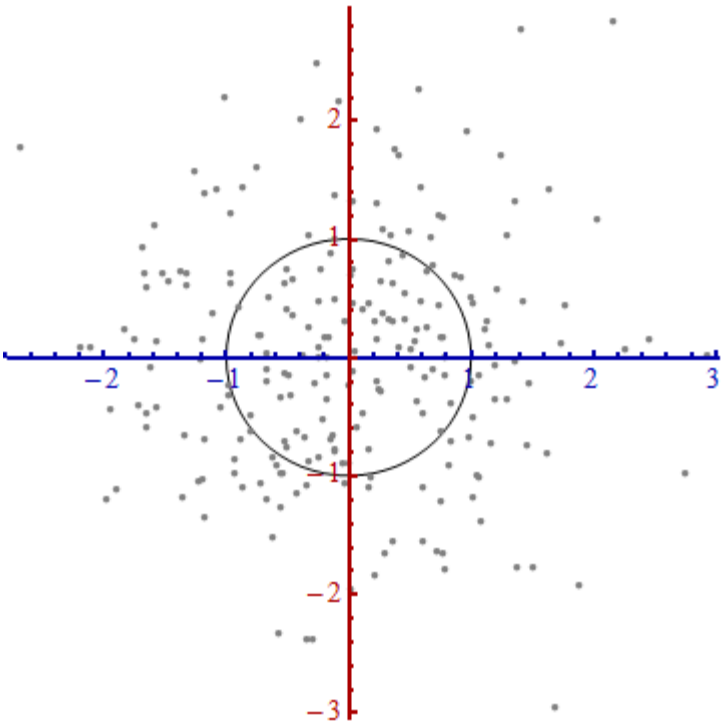
The **origin** will be at the centroid of the points (the point of their averages). The **first coordinate axis** (blue in the next figure) will extend along the "spine" of the points, which (by definition) is any direction in which the variance is the greatest. The **second coordinate axis** (red in the figure) will extend perpendicularly to the first one. (In more than two dimensions, it will be chosen in that perpendicular direction in which the variance is as large as possible, and so on.)



We need a **scale**. The standard deviation along each axis will do nicely to establish the units along the axes. Remember the 68-95-99.7 rule: about two-thirds (68%) of the points should be within one unit of the origin (along the axis); about 95% should be within two units. That makes it easy to eyeball the correct units. For reference, this figure includes the unit circle in these units:



That doesn't really look like a circle, does it? That's because this picture is *distorted* (as evidenced by the different spacings among the numbers on the two axes). Let's redraw it with the axes in their proper orientations--left to right and bottom to top--and with a unit aspect ratio so that one unit horizontally really does equal one unit vertically:

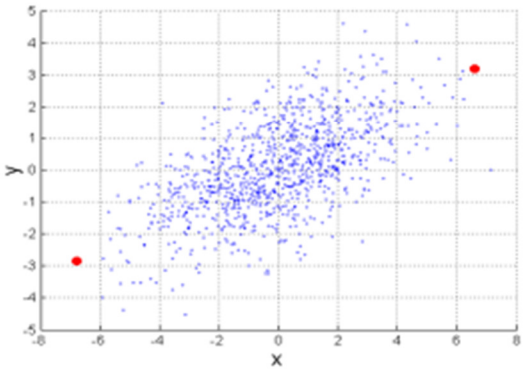


You measure the Mahalanobis distance in this picture rather than in the original.

What happened here? *We let the data tell us how to construct a coordinate system for making measurements in the scatterplot.* That's all it is. Although we had a few choices to make along the way (we could always reverse either or both axes; and in rare situations the directions along the "spines"--the *principal directions*--are not unique), *they do not change the distances in the final plot.*

Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$



Σ is the covariance matrix of the input data *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

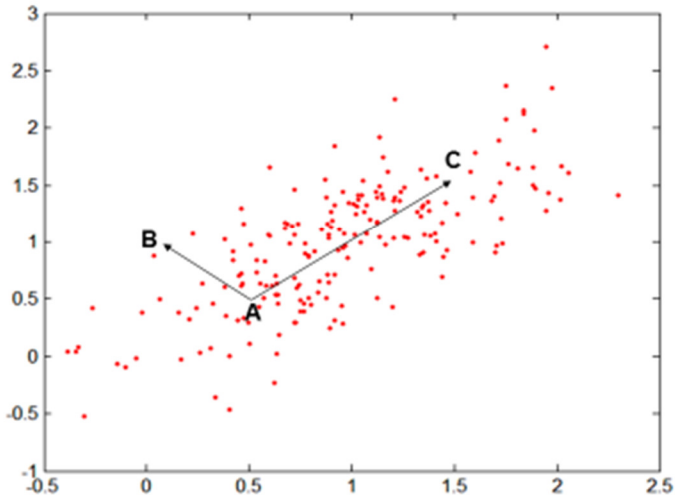
$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

Covariance Matrix: Sigma

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

Get Inverse of Covariance Matrix: Sigma^{-1} to calculate Mahalanobis distance

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Matrix = [a b

c d]

Matrix Inverse with determinant = (ad –bc)

Matrix $^{-1}$ = 1/ (ad –bc) [d -b

-c a]

When Covariance Matrix

Sigma = [0.3 0.2

0.2 0.3]

So Inverse of Covariance Matrix is

Sigma $^{-1}$ = 1/ ((0.3 * 0.3) – (-0.2 * -0.2) [0.3 -0.2

-0.2 0.3]

= 1/ (0.09 – 0.04) [0.3 -0.2

-0.2 0.3]

= 20 * [0.3 -0.2

-0.2 0.3]

= [6 -4

-4 6]

$$mahalanobis(p,q) = (p-q)\sum^{-1}(p-q)^T$$

A = (p1, p2) = (0.5, 0.5)

B = (q1, q2) = (0, 1)

So mahalanobis distance (A, B)

= [(0.5 – 0) (0.5 – 1)] * [6 -4

-4 6] * [(0.5 – 0)

(0.5 – 1)]

= [0.5 -0.5] * [6 -4

-4 6] * [0.5

-0.5]

= [(0.5 * 6) + (-0.5 * -4) (0.5 * -4) + (-0.5* 6)] * [0.5

-0.5]

= [(3 + 2) (-2-3)] * [0.5

-0.5]

= [5 -5] * [0.5

-0.5]

= 2.5 + 2.5 = 5

Technical comments

- Unit vectors along the new axes are the *eigenvectors* (of either the covariance matrix or its inverse).
- We noted that undistorting the ellipse to make a circle *divides* the distance along each eigenvector by the standard deviation: the square root of the covariance.

  Letting C C stand for the covariance function, the new (Mahalanobis) distance between two points x x and y y is the distance from x x to y y divided by the square root of $C(x-y,x-y)$ C(x–y,x–y) . The corresponding algebraic operations, thinking now of C C in terms of its representation as a matrix and x x and y y in terms of their representations as vectors, are written $(x-y)\prime C^{-1}(x-y)$ $\sqrt{(x-y)'C-1(x-y)}$. This works *regardless of what basis is used to represent vectors and matrices.* In particular, this is the correct formula for the Mahalanobis distance *in the original coordinates.*
- The amounts by which the axes are expanded in the last step are the (square roots of the) *eigenvalues* of the inverse covariance matrix. Equivalently, the axes are *shrunk* by the (roots of the) eigenvalues of the covariance matrix. Thus, the more the scatter, the more the shrinking needed to convert that ellipse into a circle.
- Although this procedure always works with any dataset, it looks this nice (the classical football-shaped cloud) for data that are approximately multivariate Normal. In other cases, the point of averages might not be a good representation of the center of the data or the "spines" (general trends in the data) will not be identified accurately using variance as a measure of spread.
- The shifting of the coordinate origin, rotation, and expansion of the axes collectively form an *affine transformation.* Apart from that initial shift, this is a change of basis from the original one (using unit vectors pointing in the positive coordinate directions) to the new one (using a choice of unit eigenvectors).
- There is a strong connection with Principal Components Analysis (PCA). That alone goes a long way towards explaining the "where does it come from" and "why" questions--if you weren't already convinced by the elegance and utility of letting the data determine the coordinates you use to describe them and measure their differences.
- For multivariate Normal distributions (where we can carry out the same construction using properties of the probability density instead of the analogous properties of the point cloud), the Mahalanobis distance (to the new origin) appears in place of the "x x " in the expression $\exp(-12\ x\ 2\ )$ exp (–12x2) that characterizes the probability density of the standard Normal distribution. Thus, in the new coordinates,

a multivariate Normal distribution looks *standard Normal* when projected onto any line through the origin. In particular, it is standard Normal in each of the new coordinates. From this point of view, the only substantial sense in which multivariate Normal distributions differ among one another is in terms of how many dimensions they use. (Note that this number of dimensions may be, and sometimes is, less than the nominal number of dimensions.)