Bottom to top explanation of the Mahalanobis distance?

I'm studying pattern recognition and statistics and almost every book I open on the subject I bump into the concept of **Mahalanobis distance**. The books give sort of intuitive explanations, but still not good enough ones for me to actually really understand what is going on. If someone would ask me "What is the Mahalanobis distance?" I could only answer: "It's this nice thing, which measures distance of some kind" :)

The definitions usually also contain eigenvectors and eigenvalues, which I have a little trouble connecting to the Mahalanobis distance. I understand the definition of eigenvectors and eigenvalues, but how are they related to the Mahalanobis distance? Does it have something to do with changing the base in Linear Algebra etc.?

I have also read these former questions on the subject:

- What is Mahalanobis distance, & how is it used in pattern recognition?
- Intuitive explanations for Gaussian distribution function and mahalanobis distance (Math.SE)

I have also read this explanation.

The answers are good and pictures nice, but still I don't **really** get it...I have an idea but it's still in the dark. Can someone give a "How would you explain it to your grandma"-explanation so that I could finally wrap this up and never again wonder what the heck is a Mahalanobis distance? :) Where does it come from, what, why?

UPDATE:

Here is something which helps understanding the Mahalanobis formula:

https://math.stackexchange.com/questions/428064/distance-of-a-test-point-from-the-center-of-an-ellipsoid

normal-distribution | mathematical-statistics | distance | pattern-recognition | intuition

edited Apr 13 '17 at 12:44
Community ♦
1

asked Jun 19 '13 at 12:41

jjepsuomi
1,690 8 23 36

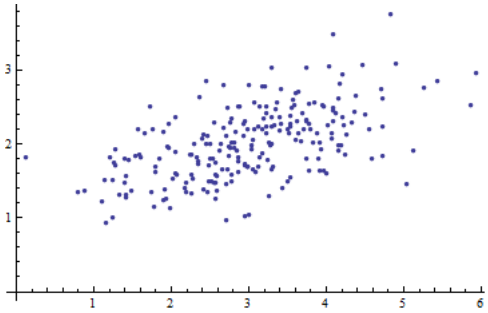
5 Please do not cross-post questions *within* SE sites. – whuber ♦ Jun 19 '13 at 14:10

Roger, sorry about that :) – jjepsuomi Jun 20 '13 at 5:49

4 NB spelling is Mahalanobis en.wikipedia.org/wiki/Prasanta_Chandra_Mahalanobis – Nick Cox Jun 25 '13 at 13:12

9 Answers

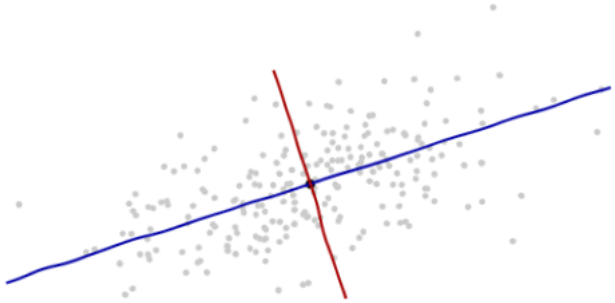
Here is a scatterplot of some multivariate data (in two dimensions):



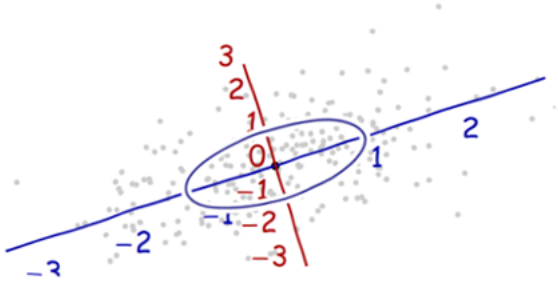
What can we make of it when the axes are left out?

Introduce coordinates that are suggested by the data themselves.

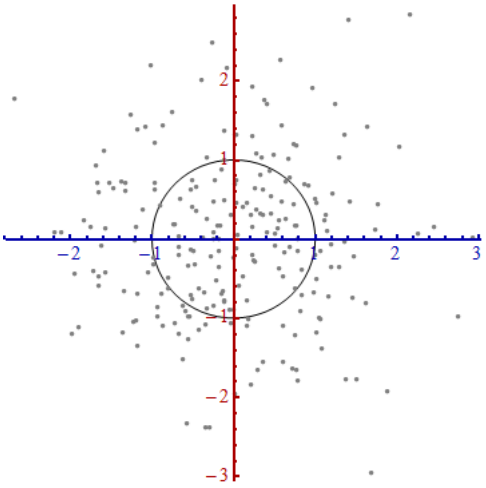
The **origin** will be at the centroid of the points (the point of their averages). The **first coordinate axis** (blue in the next figure) will extend along the "spine" of the points, which (by definition) is any direction in which the variance is the greatest. The **second coordinate axis** (red in 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.*

Technical comments

(Not for grandma, who probably started to lose interest as soon as numbers reappeared on the plots, but to address the remaining questions that were posed.)

- 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 stand for the covariance function, the new (Mahalanobis) distance between two points x and y is the distance from x to y divided by the square root of $C(x-y, x-y)$. The corresponding algebraic operations, thinking now of C in terms of its representation as a matrix and x and y in terms of their representations as vectors, are written $\sqrt{(x-y)'C^{-1}(x-y)}$. 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" in the expression $\exp(-\frac{1}{2}x^2)$ 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.)

edited Apr 13 '17 at 12:44
Community ♦
1

answered Jun 19 '13 at 21:26
whuber ♦
193k 31 406 766

3 Should anyone be curious, an affine transformation is " is a transformation which preserves straight lines... and ratios of distances between points lying on a straight line". (@whuber, I don't know if you might want to add something like this in the bulleted point.) – gung ♦ Jun 20 '13 at 2:11

@gung My mention of affine transformations is followed immediately by a characterization of them: a translation followed by a change of basis. I chose this language because it is the same used in the question. (We have to take "change of basis" somewhat liberally to encompass non-invertible linear transformations: that's an issue important for PCA, which

effectively drops some of the basis elements.) – whuber ♦ Jun 20 '13 at 13:40

12 @whuber, your explanation is probably the best one i've ever seen. Typically, when this is explained, it is covered very abstractly when they mentioned ellipsoids and spheres, and they fail to show what they mean. Kudos to you for demonstrating how the axis transformation transforms the data distribution into a "sphere" so that the distance can be "seen" as multples of the sd of the data from the mean of the data, as is readily the case for one dimensional data. This visualization is in my opinion is key, and is unfortunately left out of most discussions on the topic. Good job---your explanati – user32515 Nov 8 '13 at 14:23

Is there a robust PCA? A variation that allows us to throw away outlier data points when looking at the size of the covariance matrix? – EngrStudent Sep 24 '14 at 19:06

@Engr Sure: any robust estimation of the covariance matrix would lead to a robust PCA. Other direct methods exist, as indicated by references to them in answers to questions about robust PCA. – whuber ♦ Sep 24 '14 at 19:10

My grandma cooks. Yours might too. Cooking is a delicious way to teach statistics.

Pumpkin Habanero cookies are awesome! Think about how wonderful cinnamon and ginger can be in Christmas treats, then realize how hot they are on their own.

The ingredients are:

- habanero peppers (10, seeded and finely minced)
- sugar (1.5 cups)
- butter (1 cup)
- vanilla extract (1 tsp)
- eggs (2 medium)
- flour (2.75 cups)
- baking soda (1 tsp)
- salt (1 tsp)

Imagine your coordinate axes for your domain being the ingredient volumes. Sugar. Flour. Salt. Baking Soda. Variation along those directions, all else being equal, doesn't have nearly the impact to the flavor quality as variation in count of habanero peppers. A 10% change in flour or butter is going to make it less great, but not killer. Adding just a small amount more habanero will knock you over a flavor cliff from addictive-dessert to testosterone based pain-contest.

Mahalanobis isn't as much a distance in "ingredient volumes" as it is distance away from "best taste". The really "potent" ingredients, ones very sensitive to variation, are the ones you must most carefully control.

If you think about any Gaussian distribution vs. the Standard Normal distribution, what is the difference? Center and scale based on central tendency (mean) and variation tendency (standard deviation). One is the coordinate transform of the other. Mahalanobis is that transform. It shows you what the world looks like if your distribution of interest was re-cast as a standard normal instead of a Gaussian.

edited Jun 19 '13 at 21:58

answered Jun 19 '13 at 14:04

EngrStudent
5,389 1 18 59

4 Gaussian distributions *are* Normal distributions, so what distinction are you trying to make in your last paragraph? – whuber ♦ Jun 19 '13 at 20:21

1 @Whuber - standard. I meant standard. Thought I said it. Should check edit history. Following sentences repeat the main thought. – EngrStudent Jun 19 '13 at 21:58

1 What then do you mean by "**the** Gaussian distribution"? – whuber ♦ Jun 19 '13 at 21:59

1 Better? It could be a Gaussian distribution with any mean and variance - but the transform maps to the standard normal by subtracting the mean and scaling by the standard deviation. – EngrStudent Jun 19 '13 at 22:02

4 Yes, now it's clearer. I'm puzzled why you use two terms (Gaussian and normal) to refer to the same thing, though, but that's OK now that you have explained it. I am also a little confused about your last claim, which seems to say that every multivariate distribution can be turned into a standard Normal (which according to the definition you link to is *univariate*): I think you mean it can be made to look standard Normal *in each component.* Regardless, the analogy you start off with is nice. – whuber ♦ Jun 19 '13 at 22:07

As a starting point, I would see the Mahalanobis distance as a suitable deformation of the usual Euclidean distance $d(x,y) = \sqrt{\langle x,y \rangle}$ between vectors x and y in $\mathbb{R}^n$. The extra piece of information here is that x and y are actually *random* vectors, i.e. 2 different realizations of a vector X of random variables, lying in the background of our discussion. The question that the Mahalanobis tries to address is the following:

"how can I measure the "dissimilarity" between x and y, knowing that they are realization of the same multivariate random variable?"

Clearly the dissimilarity of any realization x with itself should be equal to 0; moreover, the dissimilarity should be a symmetric function of the realizations and should reflect the existence of a random process in the background. This last aspect is taken into consideration by introducing the covariance matrix C of the multivariate random variable.

Collecting the above ideas we arrive quite naturally at

$$D(x,y) = \sqrt{(x-y)\,C^{-1}(x-y)}$$

If the components X_i of the multivariate random variable $X = (X_1, \ldots, X_n)$ are uncorrelated, with, for example $C_{ij} = \delta_{ij}$ (we "normalized" the X_i's in order to have $Var(X_i) = 1$), then the Mahalanobis distance $D(x,y)$ *is* the Euclidean distance between x and y. In presence non trivial correlations, the (estimated) correlation matrix $C(x,y)$ "deforms" the Euclidean distance.

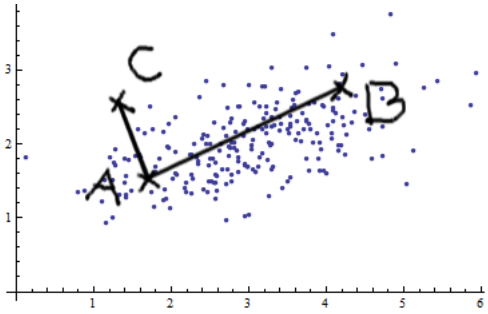
edited Jun 25 '13 at 16:52

StasK
24.9k 57 135

answered Jun 19 '13 at 14:25

Avitus
385 5 15

Let's consider the two variables case. Seeing this picture of bivariate normal (thanks @whuber), you cannot simply claim that AB is larger than AC. There is a positive covariance; the two variables are related to each other.



You can apply simple Euclidean measurements (straight lines like AB and AC) only if the variables are

1. independent
2. have variances equal to 1.

Essentially, Mahalanobis distance measure does the following: it transforms the variables into uncorrelated variables with variances equal to 1, and then calculates simple Euclidean distance.

edited Jun 4 '14 at 16:34

gung ♦
101k 33 238 494

answered Jun 4 '14 at 16:13

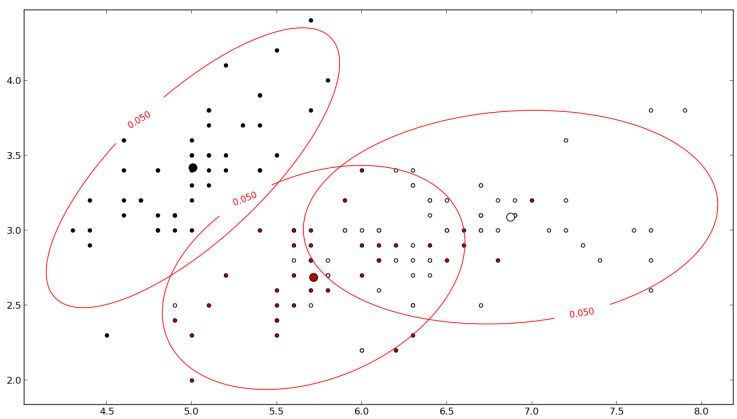
den2042
125 1 8

are you suggesting that every time I see a correlation in a graph as shown in your answer here, I should only think about calculating Mahalanobis rather than the Euclidean distance? What would tell me when to use which? – sandyp Jul 14 at 21:14

I'll try to explain you as simply as possible:

Mahalanobis distance measures the distance of a point x from a data distribution. The data distribution is characterized by a mean and the covariance matrix, thus is hypothesized as a multivariate gaussian.

It is used in pattern recognition as similarity measure between the pattern (data distribution of training example of a class) and the test example. The covariance matrix gives the shape of how data is distributed in the feature space.



The figure indicates three different classes and the red line indicates the same Mahalanobis distance for each class. All points lying on the red line have the same distance from the class mean, because it is used the covariance matrix.

The key feature is the use of covariance as a normalization factor.

answered Jun 19 '13 at 13:15

robbisg
115 8

Just to add to the excellent explanations above, the Mahalanobis distance arises naturally in (multivariate) linear regression. This is a simple consequence of some of the connections between the Mahalanobis distance and the Gaussian distribution discussed in the other answers, but I think it's worth spelling out anyway.

Suppose we have some data $(x_1, y_1), \ldots, (x_N, y_N)$, with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^m$. Let's assume that there exists a parameter vector $\beta_0 \in \mathbb{R}^m$ and a parameter matrix $\beta_1 \in \mathbb{R}^{m \times n}$ such that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_N$ are iid m-dimensional Gaussian random vectors with mean 0 and covariance C (and they are independent of the x_i). Then y_i given x_i is Gaussian with mean $\beta_0 + \beta_1 x_i$ and covariance C.

It follows that the negative log-likelihood of y_i given x_i (as a function of $\beta = (\beta_0, \beta_1)$) is given by

$$-\log p(y_i \mid x_i; \beta) = \frac{m}{2}\log(2\pi \det C) + \frac{1}{2}(y_i - (\beta_0 + \beta_1 x_i))^\top C^{-1}(y_i - (\beta_0 + \beta x_i)).$$

We are taking the covariance C to be constant, so

$$\operatorname{argmin}_\beta[-\log p(y_i \mid x_i; \beta)] = \operatorname{argmin}_\beta D_C(\beta_0 + \beta_1 x_i, y_i),$$

where

$$D_C(\hat{y}, y) = \sqrt{(y - \hat{y})^\top C^{-1}(y - \hat{y})}$$

is the Mahalanobis distance between $\hat{y}, y \in \mathbb{R}^m$.

By independence, the log-likelihood $\log p(\mathbf{y} \mid \mathbf{x}; \beta)$ of $\mathbf{y} = (y_1, \ldots, y_N)$ given $\mathbf{x} = (x_1, \ldots, x_N)$ is given by the sum

$$\log p(\mathbf{y} \mid \mathbf{x}; \beta) = \sum_{i=1}^N \log p(y_i \mid x_i; \beta)$$

Therefore,

$$\operatorname{argmin}_\beta[-\log p(\mathbf{y} \mid \mathbf{x}; \beta)] = \operatorname{argmin}_\beta \frac{1}{N}\sum_{i=1}^N D_C(\beta_0 + \beta_1 x_i, y_i),$$

where the factor $1/N$ does not affect the argmin.

In summary, the coefficients β_0, β_1 that minimize the negative log-likelihood (i.e. maximize the likelihood) of the observed data also minimize the empirical risk of the data with loss function given by the Mahalanobis distance.

edited Dec 3 '16 at 20:23

answered Dec 3 '16 at 2:12

Ben CW
151 5

1 Well, not quite. That term corresponding to $\log \det C$ changes things quite a bit. And you seem to have focused on the other dimension: the Mahalanobis distance actually plays a much more important role in the n dimensional space spanned by the columns, because that is related to leverage. Readers will likely be confused by that, though, due to the reversal of the roles of x and β in your notation: x is the parameter vector and β the design matrix! – whuber ♦ Dec 3 '16 at 4:57

My intent was for (x, y) here to denote a single labelled training example (so no design matrix here); the reason y is a vector is that I am doing multivariate regression (otherwise the noise term ϵ would be a single-variable Gaussian, there would be no covariance matrix, and the example might seem too trivial). Perhaps my notation is non-standard, as my background is not in statistics. Regarding the presence of the $\log \det C$ term, what I meant is that $\operatorname{argmin}_\beta[-\log p(y \mid x; \beta)] = \operatorname{argmin}_\beta \sqrt{(y - \beta x)^\top C^{-1}(y - \beta x)}$. – Ben CW Dec 3 '16 at 5:11

It is important to explain what your symbols refer to rather than requiring readers to guess. Quite possibly your explanation is a good one, but without that explanation (which you have begun with that latest comment) I suspect most readers will have trouble understanding your meaning. – whuber ♦ Dec 3 '16 at 17:09

2 I see your point. I've edited the original answer to incorporate some of the ideas in these comments. – Ben CW Dec 3 '16 at 20:26

I might be a bit late for answering this question. This paper in here is a good start for understanding the Mahalanobis distance. They provides a complete example with numerical values. What I like about it is the geometric representation of the problem is presented.

answered Oct 3 '14 at 1:55

CroCo


168 2 7

1 It's not late :) thank you for your help! :) – jjepsuomi Oct 3 '14 at 6:13

I'd like to add a little technical information to Whuber's excellent answer. This information might not interest grandma, but perhaps her grandchild would find it helpful. The following is a bottom-to-top explanation of the relevant linear algebra.

Mahalanobis distance is defined as $d(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$, where Σ is an estimate of the covariance matrix for some data; this implies it is symmetric. If the columns used to estimate Σ are not linearly dependent, Σ is positive definite. Symmetric matrices are diagonalizable and their eigenvalues and eigenvectors are real. PD matrices have eigenvalues which are all positive. The eigenvectors can be chosen to have unit length, and are orthogonal (i.e. orthonormal) so we can write $\Sigma = Q^T D Q$ and $\Sigma^{-1} = Q D^{-\frac{1}{2}} D^{-\frac{1}{2}} Q^T$. Plugging that into the distance definition, $d(x,y) = \sqrt{[(x-y)^T Q] \, D^{-\frac{1}{2}} D^{-\frac{1}{2}} \, [Q^T (x-y)]} = \sqrt{z^T z}$. Clearly the products in the square brackets are transposes, and the effect of multiplication by Q is rotating the vector $(x-y)$ into an orthogonal basis. Finally, $D^{-\frac{1}{2}}$, which is diagonal, and formed by inverting each element on the diagonal, then taking the square root, is rescaling each element of each vector. In fact, $D^{-\frac{1}{2}}$ is precisely the inverse standard deviation of each feature in the orthogonal space (i.e. D^{-1} a precision matrix, and because the data are in an orthogonal basis, the matrix is diagonal). The effect is to transform what Whuber calls a rotated ellipse into a circle by "flattening" its axes. Clearly $z^T z$ is measured in the squared units, so taking the square root returns the distance into the original units.

answered Feb 11 '17 at 3:56


Sycorax
30.6k 5 85 137

Mahalanobis distance is an euclidian distance (natural distance) wich take into account the covariance of data. It give a bigger weight to noisy component and so is very usefull to check for similarity between two datasets.

As you can see in your exemple here when variables are correlated, the distribution is shifted into one direction. You may want to remove this effects. If you take correlation into account in your distance, you can remove the shift effect.

answered Jun 19 '13 at 13:12


Were_cat
1,071 11 26

2 I believe the Mahalanobis distance effectively *downweights* the large-covariance directions, rather than giving "bigger" weights there. – whuber ♦ Jun 19 '13 at 20:20

protected by kjetil b halvorsen Nov 21 '17 at 22:24

Thank you for your interest in this question. Because it has attracted low-quality or spam answers that had to be removed, posting an answer now requires 10 reputation on this site (the association bonus does not count).

Would you like to answer one of these unanswered questions instead?