When does X3 contribute to increasing the discrimination power of (X1,X2)?

Let us consider a classification problem with two 3-dimensional normals centered respectively on (1,1,1) and (-1,-1,-1), both with covariance matrices equal to

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}.$$

Then there is a threshold n* beyond which the presence of feature X3 becomes advantageous, if the two other features X1 and X2 are already present.

Question: Given matrix Σ, characterized by a particular combination of the 6 parameters $\sigma_1, \sigma_2, \sigma_3, \rho_{12}, \rho_{13}, \rho_{23}$, how to evaluate the potential additional contribution of X3 to the discrimination between the two groups, whenever X1 and X2 are already present?

To analyze this question, 4 scenarios were created, and in each one of them, only 3 of the 6 parameters can be freely chosen.

In each of these 4 scenarios, given matrix Σ, the quotient Risk3/Risk2 can be calculated, where Risk3 and Risk2 are the Bayes risks relative, respectively:

a) to the situation in which all 3 features are present;
b) to the situation in which only X1 and X2 are present.

Supposedly, the higher Risk3/Risk2 ratio, the higher must be n*.

However, in order get a clearer understanding of the problem, besides that ratio, it could be interesting to have an indicator with a geometric flavor. How could we build such an indicator?

We know the level surfaces corresponding to a given 3D-normal distribution are ellipsoids centered on the respective centroid. Now, since the two centroids are (1,1,1) and (-1,-1,-1), the more the principal axis direction of these ellipsoids approaches the direction of the line joining the two centroids, the greater will be the intersection between the point clouds relative to the two groups, that is, the greater will be the expected error probability of the classifier, if all 3 features are present.

And, in order to standardize the discussion, let us consider an ellipsoid in R^3 (say, centered at the origin (0,0,0)) whose equation is $\mathrm{x}^{\mathrm{T}}\Sigma^{-1}\mathrm{x} = 1$. The line joining the two centroids is the set of all (c,c,c) vectors in R^3, whose three coordinates are the same. This line crosses the ellipsoid at some point whose three coordinates are equal and positive. Let d3 be the distance from this point to the origin. The greater this distance, the lower the classifier discrimination power, with the 3 features present.

Analogously, suppose only X1 and X2 are present. Let us consider an ellipse in R^2 (say, centered at the origin (0,0)) whose equation is $\mathrm{x}^{\mathrm{T}}\Sigma^{-1}\mathrm{x} = 1$. The line joining the two centroids (1,1) and (-1,-1) is the set of all (c,c) vectors of R^2, whose two coordinates are the same. This line intersects the ellipse at some point whose two coordinates are equal and positive. Let d2 be the distance from this point to the origin. The greater this distance, the lower the classifier discrimination power, with X1 and X2 present.

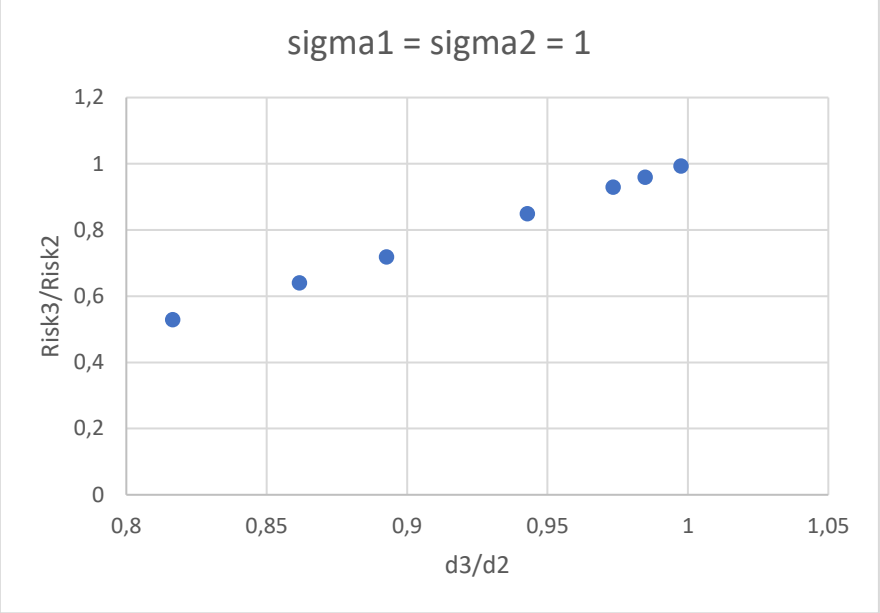
How to use d3 and d2 to create an X3 discrimination power indicator, given that X1 and X2 are already present?

d3/d2 is the indicator we propose. For all 4 scenarios, it is possible to see that, fixing the values of some parameters, Risk3/Risk2 is an increasing and smooth function of that indicator. Let us look at what happens in each scenario.

Scenario 1 ($\rho_{12} = \rho_{13} = \rho_{23} = 0$):

Setting $\sigma_1 = \sigma_2 = 1$, and varying σ_3, we get:

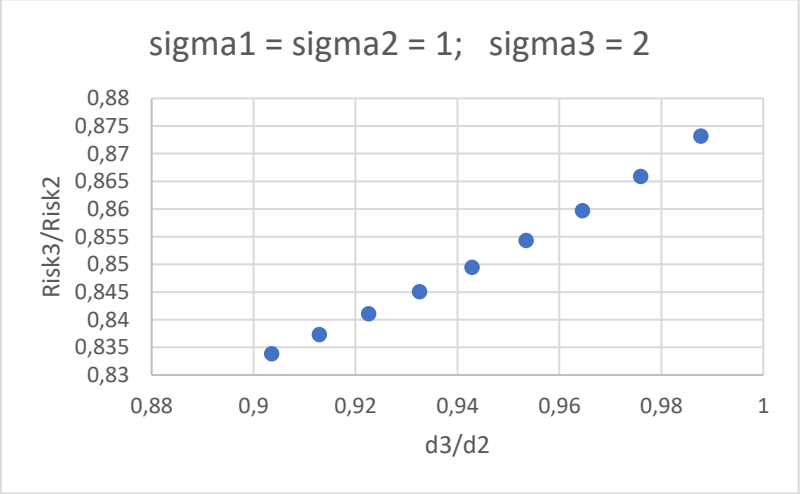
σ_3	d3/d2	Risk3/Risk2
1	0.816497	0.529338439
1.2	0.86155	0.640172221
1.4	0.892607	0.719083617
2	0.942809	0.849428329
3	0.973328	0.929649359
4	0.984731	0.959714436
10	0.997509	0.993427246



Scenario 2 ($\boldsymbol{\sigma_1 = \sigma_2,\ \rho_{13} = \rho_{23} = 0}$):

Setting $\sigma_1 = \sigma_2 = 1$, $\sigma_3 = 2$, and varying ρ_{12} from -1 to +1, we get:

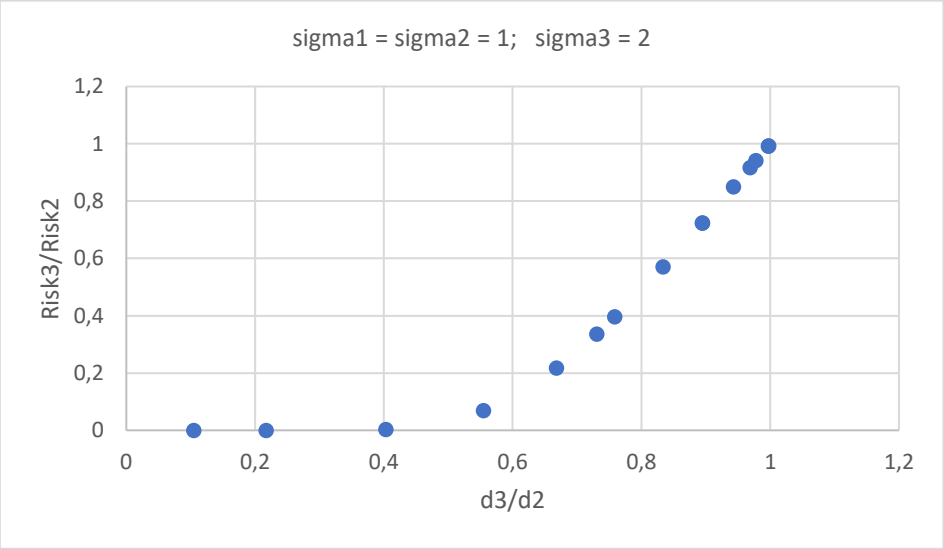
ρ_{12}	d3/d2	Risk3/Risk2
-0.8	0.987729	0.873159598
-0.6	0.9759	0.865841917
-0.4	0.964485	0.859665032
-0.2	0.953463	0.85426079
0	0.942809	0.849428329
0.2	0.932505	0.845042485
0.4	0.922531	0.841018529
0.6	0.912871	0.837295723
0.8	0.903508	0.833828655



Scenario 3 ($\boldsymbol{\sigma_1 = \sigma_2}$, $\boldsymbol{\rho_{12} = 0}$, $\boldsymbol{\rho_{13} = \rho_{23}}$): (Restriction: $|\rho_{13}| < \frac{\sqrt{2}}{2}$)

Setting $\sigma_1 = \sigma_2 = 1$, $\sigma_3 = 2$, and varying $\rho_{13} = \rho_{23}$ from -0.7 to 0.7, we get:

$\rho_{13} = \rho_{23}$	d3/d2	Risk3/Risk2
-0.7	0.104685	0
-0.6	0.402887	0.002846727
-0.5	0.554699	0.068579171
-0.4	0.667757	0.217341613
-0.3	0.758575	0.395931843
-0.2	0.833269	0.570008019
-0.1	0.894425	0.723756338
0	0.942807	0.849428329
0.1	0.9778	0.941441075
0.2	0.997292	0.992858026
0.3	0.996963	0.991991025
0.4	0.968466	0.916846491
0.5	0.894425	0.723756338
0.6	0.730295	0.33571378
0.7	0.21693	4.49348E-10



Scenario 4 ($\boldsymbol{\sigma_1 = \sigma_2 = \sigma_3}$, $\boldsymbol{\rho_{13} = \rho_{23}}$): (Restriction: $|\rho_{13}| < \sqrt{\frac{1+\rho_{12}}{2}}$)

Setting $\sigma_1 = \sigma_2 = \sigma_3 = 1, \rho_{12} = -0.1$, and varying $\rho_{13} = \rho_{23}$ from -0.6 to 0.6:

$\rho_{13} = \rho_{23}$	d3/d2	Risk3/Risk2
-0.6	0.2747211	4.22993E-07
-0.5	0.4259177	0.003420082
-0.4	0.535182	0.03929572
-0.3	0.6246951	0.12510334
-0.2	0.7017782	0.247388519
-0.1	0.7698004	0.38818455
0	0.8304548	0.534008242
0.1	0.8844333	0.675493612
0.2	0.9315174	0.805154984
0.3	0.9701425	0.914408902
0.4	0.9957173	0.987679476
0.5	0.993808	0.982190703
0.6	0.8944272	0.702607486