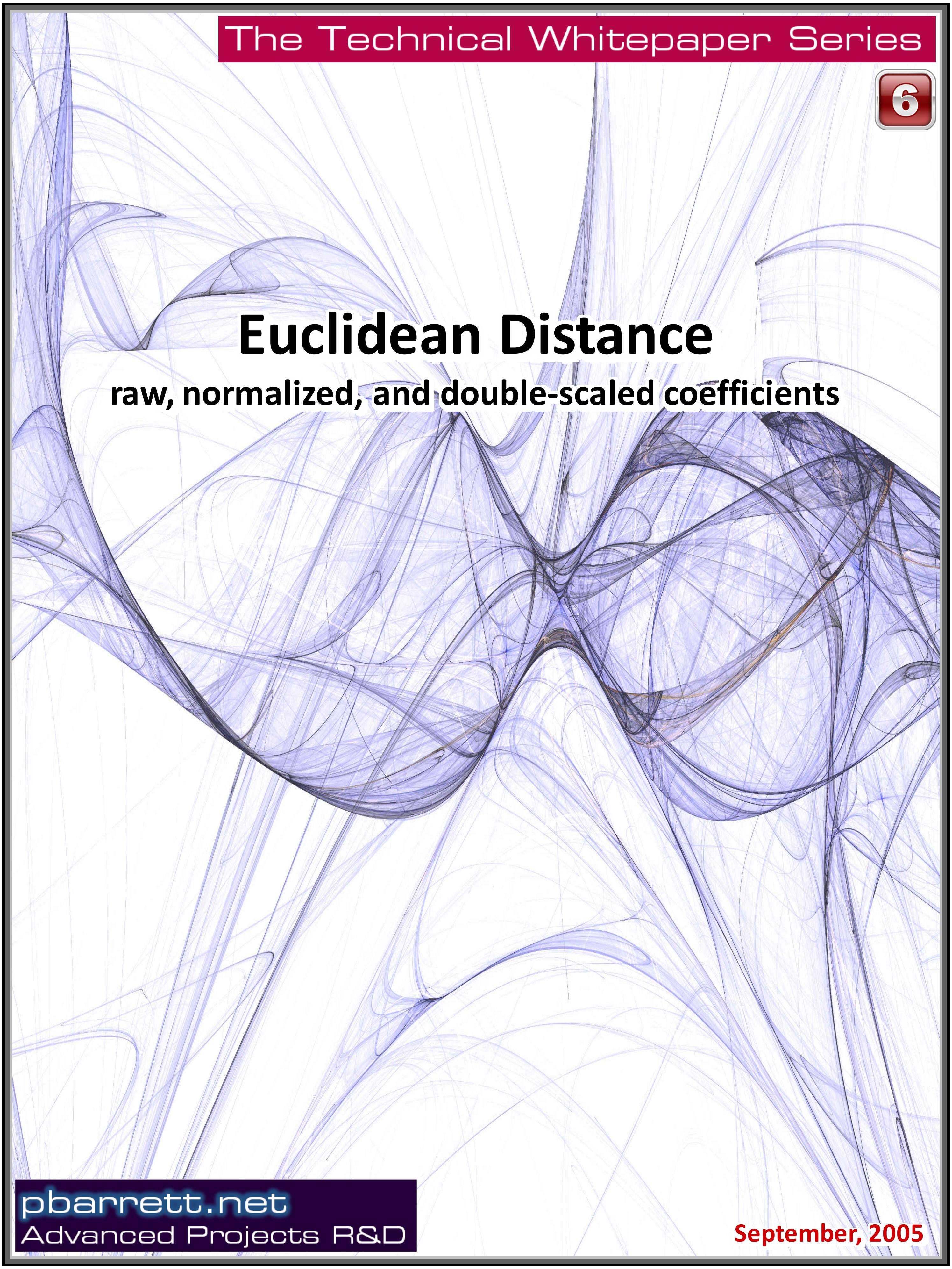
The Technical Whitepaper Series

6



Euclidean Distance

raw, normalized, and double-scaled coefficients

pbarrett.net
Advanced Projects R&D

September, 2005

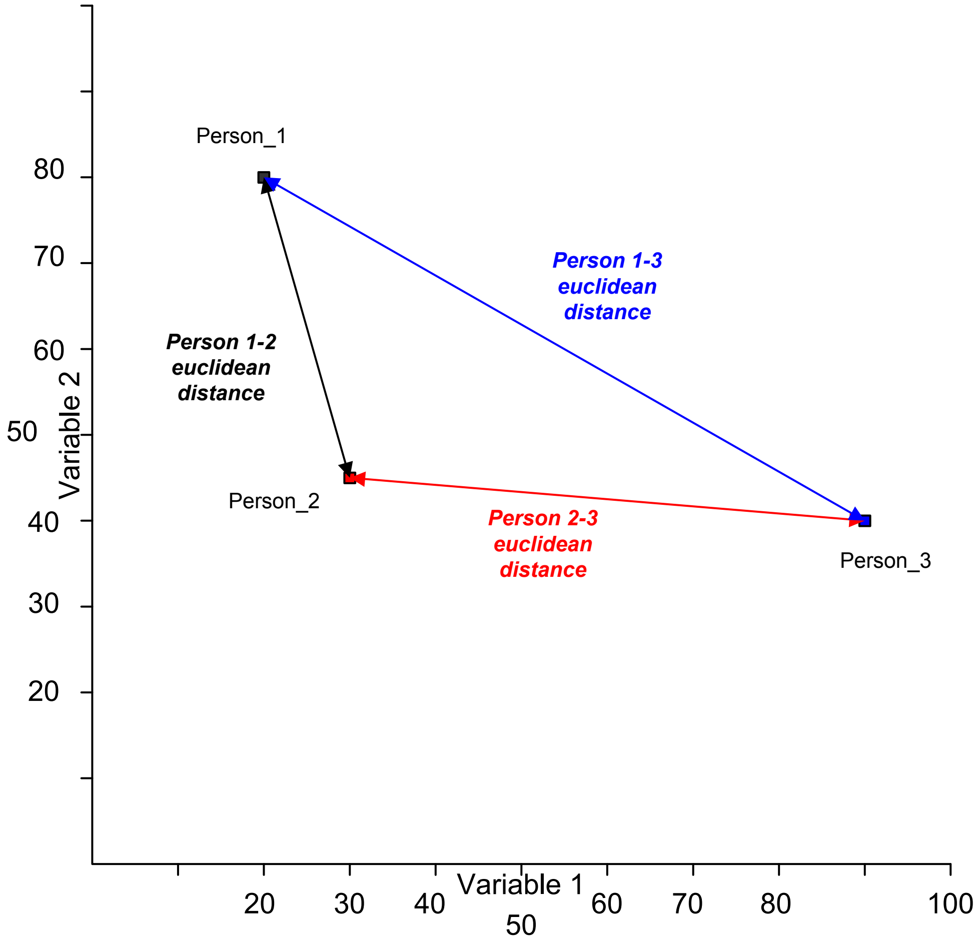
Euclidean Distance – Raw, Normalised, and Double-Scaled Coefficients

Having been fiddling around with distance measures for some time – especially with regard to profile comparison methodologies, I thought it was time I provided a brief and simple overview of Euclidean Distance – and why so many programs give so many completely different estimates of it. This is not because the concept itself changes (that of linear distance), but is due to the way programs/ investigators either transform the data prior to computing the difference, normalise constituent distances via a constant, or re-scale the coefficient into a unit metric. However, few actually make absolutely explicit what they do, and the consequences of whatever transformation they undertake. Given that I always use a double-scaling of distance into a unit metric for the coefficient, and never transform the raw data, I thought it time I explained the logic of this, and why I feel some of the coefficients used within some popular statistical programs are sometimes less than optimal (i.e. using "normal z-score" transformations).

Raw Euclidean Distance

The Euclidean metric (and distance magnitude) is that which corresponds to everyday experience and perceptions. That is, the kind of 1, 2, and 3-Dimensional linear metric world where the distance between any two points in space corresponds to the length of a straight line drawn between them. Figure 1 shows the scores of three individuals on two variables (Variable 1 is the x-axis, Variable 2 the y-axis) –

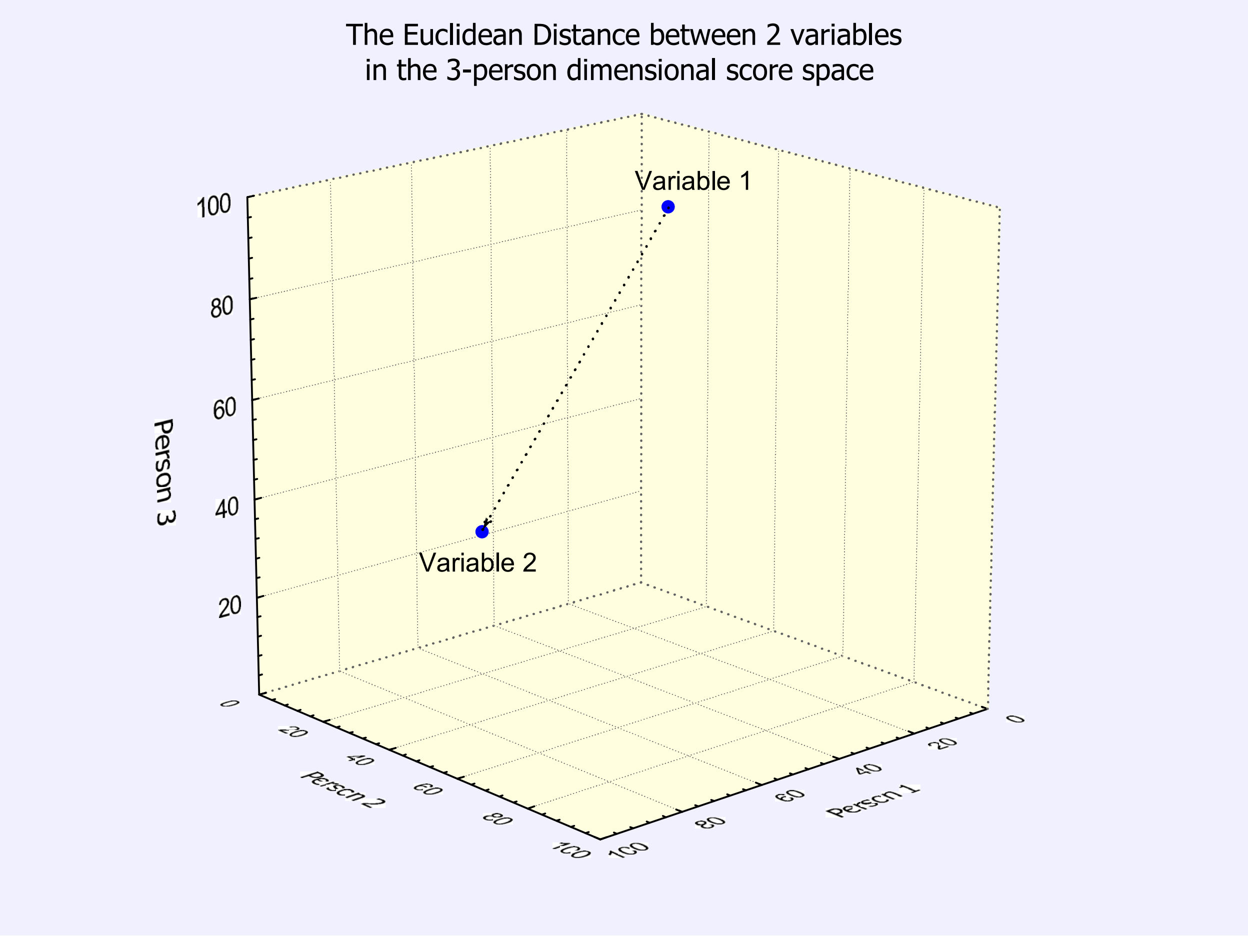
Figure 1



The straight line between each "Person" is the Euclidean distance. There would this be three such distances to compute, one for each person-to-person distance.

However, we could also calculate the Euclidean distance between the two variables, given the three person scores on each – as shown in Figure 2 …

Figure 2



The formula for calculating the distance between each of the three individuals as shown in Figure 1 is:

Eq. 1

$$d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$$

where the difference between two persons' scores is taken, and squared, and summed for v variables (in our example v=2). Three such distances would be calculated, for p1 – p2, p1 – p3, and p2 - p3.

The formula for calculating the distance between the two variables, given three persons scoring on each as shown in Figure 1 is:

Eq. 2

$$d = \sqrt{\sum_{i=1}^{p} (v_{1i} - v_{2i})^2}$$

where the difference between two variables' values is taken, and squared, and summed for p persons (in our example p=3). Only one distance would be computed – between v1 and v2.
Let's do the calculations for finding the Euclidean distances between the three persons, given their scores on two variables. The data are provided in Table 1 below …

Table 1

	1 Var1	2 Var2
Person 1	20	80
Person 2	30	44
Person 3	90	40

Using equation 1 …

$$d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$$

For the distance between person **1** and **2**, the calculation is:

$$d = \sqrt{(20-30)^2 + (80-44)^2} = 37.36$$

For the distance between person **1** and **3**, the calculation is:

$$d = \sqrt{(20-90)^2 + (80-40)^2} = 80.62$$

For the distance between person **2** and **3**, the calculation is:

$$d = \sqrt{(30-90)^2 + (44-40)^2} = 60.13$$

Using equation 2, we can also calculate the distance between the two variables …

$$d = \sqrt{\sum_{i=1}^{p} (v_{1i} - v_{2i})^2}$$

$$d = \sqrt{(20-80)^2 + (30-44)^2 + (90-40)^2} = 79.35$$

Equation 1 is used where say we are comparing two “objects” across a range of variables – and trying to determine how “dissimilar” the objects are (the Euclidean distance between the two objects taking into account their magnitudes on the range of variables. These objects might be two person’s profiles, a person and a target profile, in fact basically any two vectors taken across the same variables.

Equation 2 is used where we are comparing two variables to one another – given a sample of paired observations on each (as we might with a pearson correlation), In our case above, the sample was three persons.

In both equations, **Raw Euclidean Distance** is being computed.

Normalised Euclidean Distance

The problem with the raw distance coefficient is that it has no obvious bound value for the maximum distance, merely one that says 0 = absolute identity. Its range of values vary from 0 (absolute identity) to some maximum possible discrepancy value which remains unknown until specifically computed. Raw Euclidean distance varies as a function of the magnitudes of the observations. Basically, you don't know from its size whether a coefficient indicates a small or large distance.

If I divided every person's score by 10 in Table 1, and recomputed the euclidean distance between the persons, I would now obtain distance values of 3.736 for person 1 compared to 2, instead of 37.36. Likewise, 8.06 for person 1 and 3, and 6.01 for persons 2 and 3. The raw distance conveys little information about absolute dissimilarity.

So, raw euclidean distance is acceptable only if relative ordering amongst a fixed set of profile attributes is required. But, even here, what does a figure of 37.36 actually convey. If the maximum possible observable distance is 38, then we know that the persons being compared are about as different as they can be. But, if the maximum observable distance is 1000, then suddenly a value of 37.36 seems to indicate a pretty good degree of agreement between two persons.

The fact of the matter is that unless we know the maximum possible values for a euclidean distance, we can do little more than rank dissimilarities, without ever knowing whether any or them are actually similar or not to one another in any absolute sense.

A further problem is that raw Euclidean distance is sensitive to the scaling of each constituent variable. For example, comparing persons across variables whose score ranges are dramatically different. Likewise, when developing a matrix of Euclidean coefficients by comparing multiple variables to one another, and where those variables' magnitude ranges are quite different.

For example, say we have 10 variables and are comparing two person's scores on them … the variable scores might look like …

Table 2

	1 Person 1	2 Person 2
Var 1	1	2
Var 2	1	1
Var 3	4	5
Var4	6	6
Var5	1200	1300
Var6	3	3
Var7	2	2
Var8	3	5
Var9	2	3
Var10	8	8

The two persons' scores are virtually identical except for variable 5. The raw Euclidean distance for these data is: **100.03**. If we had expressed the scores for variable 5 in the same metric as the other scores (on a 1-10 metric scale), we would have scores of 1.2 and 1.3 respectively for each individual. The raw Euclidean distance is now: **2.65**.

Obviously, the question "is 2.65 good or bad" still exists – given we have no idea what the maximum possible Euclidean distance might be for these data.

This is where SYSTAT, Primer 5, and SPSS provide Standardization/Normalization options for the data so as to permit an investigator to compute a distance coefficient which is essentially "scale free".

Systat 10.2's normalised Euclidean distance produces its "normalisation" by dividing each squared discrepancy between attributes or persons by the total number of squared discrepancies (or sample size).

Eq. 3

$$d = \sqrt{\sum_{i=1}^{v} \left(\frac{(p_{1i} - p_{2i})^2}{v} \right)}$$

So, comparing two persons across their magnitudes on 10 variables, as in the Table 3 below,

Table 3

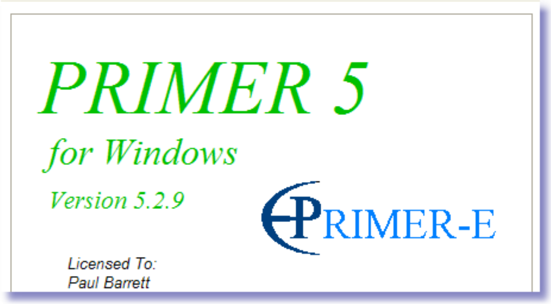
	1 Person 1	2 Person 2
Var 1	1	2
Var 2	1	1
Var 3	4	5
Var4	6	6
Var5	1.2	1.3
Var6	3	3
Var7	2	2
Var8	3	5
Var9	2	3
Var10	8	8

We calculate …

$$d = \sqrt{\left(\frac{(1-2)^2}{10}\right) + \left(\frac{(1-1)^2}{10}\right) + \left(\frac{(4-5)^2}{10}\right) \ldots + \left(\frac{(9-8)^2}{10}\right)} = 0.837$$

For the data in Table 2, the SYSTAT normalized Euclidean distance would be **31.634**

Frankly, I can see little point in this standardization – as the final coefficient still remains scale-sensitive. That is, it is impossible to know whether the value indicates high or low dissimilarity from the coefficient value alone.



Primer 5 – an ecological/marine biology software package allows the calculation of raw Euclidean distance as well as a normalized Euclidean distance. But, this normalization is problematic when just two variables or persons are to be compared to one another and these are the only two persons or variables in the dataset.

An immediate problem is encountered when trying to analyse the data in Tables 2 or 3 causes an error message

This is due to the fact that Primer 5 is actually standardizing each row of data in the file … hence, when two values are equal, as for variables 2, 4, 6, 7 etc., there is no variance, no standard deviation or it is set to zero, which then causes a division by zero in the standardization formula.

I modified the data in Table 3 – to allow unequal values on each pair of variable scores for the two persons …

Table 4

	1 Person 1	2 Person 2	3 Person 1 - Row Standardized	4 Person 2 - Row Standardized
Var 1	1	2	-0.707106781	0.707106781
Var 2	1	2	-0.707106781	0.707106781
Var 3	4	5	-0.707106781	0.707106781
Var4	6	5	0.707106781	-0.707106781
Var5	1.2	1.3	-0.707106781	0.707106781
Var6	3	4	-0.707106781	0.707106781
Var7	2	3	-0.707106781	0.707106781
Var8	3	5	-0.707106781	0.707106781
Var9	2	3	-0.707106781	0.707106781
Var10	8	7	0.707106781	-0.707106781

What we see in columns 3 and 4 is what Primer 5 does with the data (by standardizing rows) …

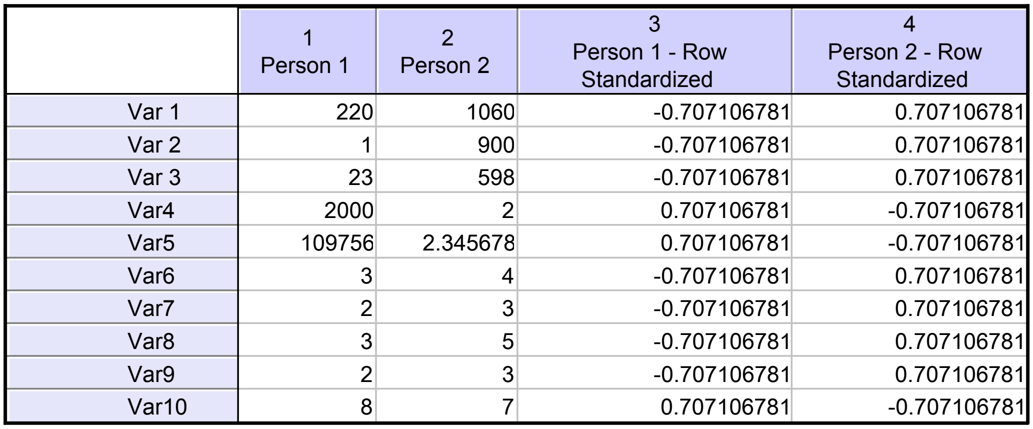
It produces a normalized Euclidean distance calculation of **4.4721** for the data in columns 1 and 2. The raw Euclidean distance is **3.4655**

If we change variable 5 to reflect the 1200 and 1300 values as in Table 2, the normalized Euclidean distance remains as **4.4721**, whilst the raw coefficient is: **100.06**. So, its normalization certainly ensures stability of coefficient scaling given unequal metrics of the constituent variables, but the value itself is now a function of the number of variables. For example, if we had made the calculation over 500 variables, the normalized Euclidean distance would be **31.627**.

The reason for this is because whatever the values of the variables for each individual, the standardized values are always equal to 0.707106781 !

Look at the following data in Table 5 below …

Table 5

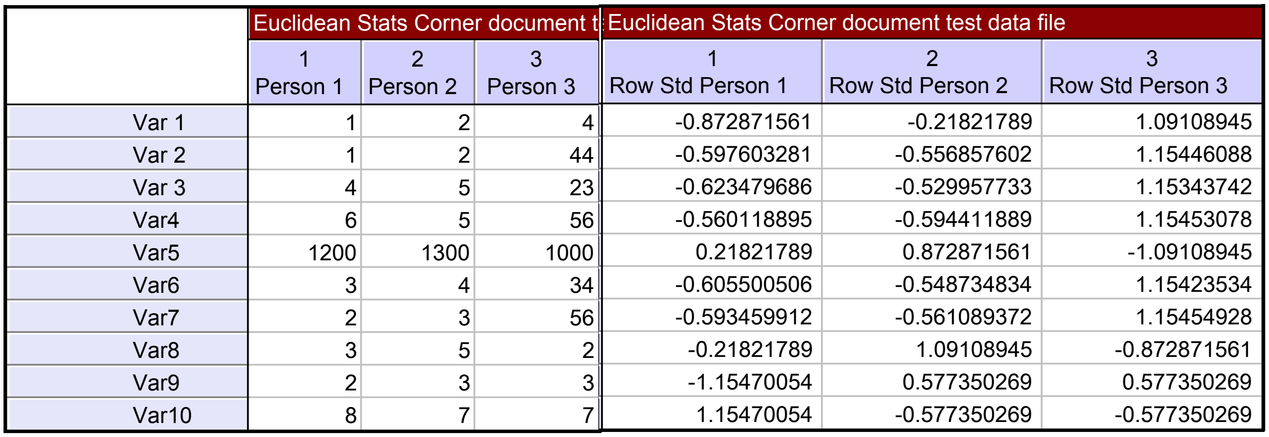


	1 Person 1	2 Person 2	3 Person 1 - Row Standardized	4 Person 2 - Row Standardized
Var 1	220	1060	-0.707106781	0.707106781
Var 2	1	900	-0.707106781	0.707106781
Var 3	23	598	-0.707106781	0.707106781
Var4	2000	2	0.707106781	-0.707106781
Var5	109756	2.345678	0.707106781	-0.707106781
Var6	3	4	-0.707106781	0.707106781
Var7	2	3	-0.707106781	0.707106781
Var8	3	5	-0.707106781	0.707106781
Var9	2	3	-0.707106781	0.707106781
Var10	8	7	0.707106781	-0.707106781

The raw euclidean distance is **109780.23**, the Primer 5 normalized coefficient remains at **4.4721**.

It's clear that Primer 5 cannot provide a normalized Euclidean distance where just two objects are being compared across a range of attributes or samples. It seems to work only where more than two objects exist in a data matrix, and more than two variables or samples are present. Then the standardization permits differentiation of values for samples or variables such that coefficients may be calculated. As a double-check – I added a 3rd person to the data of Table 5, shown in Table 6 …

Table 6



	Euclidean Stats Corner document t			Euclidean Stats Corner document test data file		
	1 Person 1	2 Person 2	3 Person 3	1 Row Std Person 1	2 Row Std Person 2	3 Row Std Person 3
Var 1	1	2	4	-0.872871561	-0.21821789	1.09108945
Var 2	1	2	44	-0.597603281	-0.556857602	1.15446088
Var 3	4	5	23	-0.623479686	-0.529957733	1.15343742
Var4	6	5	56	-0.560118895	-0.594411889	1.15453078
Var5	1200	1300	1000	0.21821789	0.872871561	-1.09108945
Var6	3	4	34	-0.605500506	-0.548734834	1.15423534
Var7	2	3	56	-0.593459912	-0.561089372	1.15454928
Var8	3	5	2	-0.21821789	1.09108945	-0.872871561
Var9	2	3	3	-1.15470054	0.577350269	0.577350269
Var10	8	7	7	1.15470054	-0.577350269	-0.577350269

The Row Standardized values for each variable are shown as the last 3 variables.

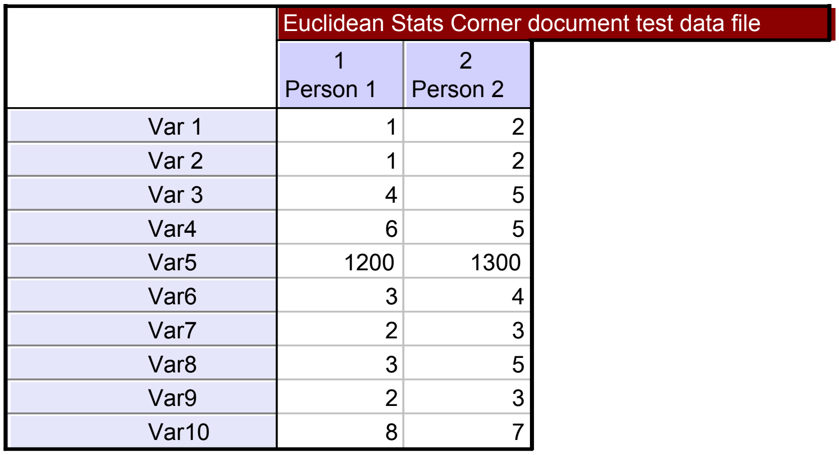
The normalised Euclidean distance between:
Persons 1 and 2 is now **2.9304** (was **4.4721** when just two persons in the file)
Persons 1 and 3 is 5.2268
Persons 2 and 3 is 4.9085

The actual raw data never changed – but the distance measure does, solely as a function of the transformation.

So, since on many occasions we might be trying to produce pair-wise coefficients for ranking, direct comparisons, or profiling etc., it seems Primer 5 is just not built for these kinds of applications. Also, the normalised distance measure itself will change as a function of how many objects there are to be compared in a data file – even though the actual data may remain completely the same for a subset of that file. The normalization by row is, at first glance, a sensible way of ensuring that each variable is expressed in the same metric. But, it must not be forgotten that this is a **data** transformation. That is, you are no longer expressing Euclidean distance between the data at hand, but of derived, transformed values of that data.

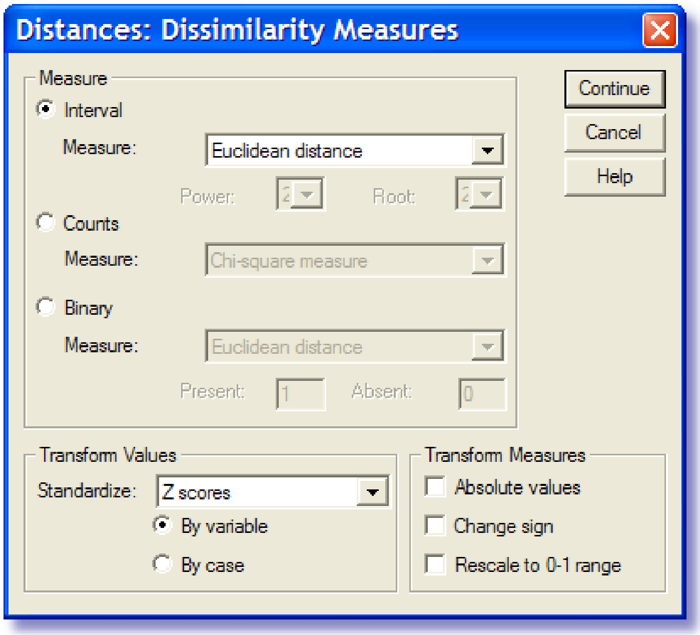
SPSS permits a number of transformations of simple Euclidean distance – and allows both row "standardization" (normalisation in Primer 5) as well as "column" standardization. Not only that, but it allows for several other transformations of the data prior to computing a Euclidean distance. No formula or examples are given of each in any SPSS documentation; the effect on a coefficient value is relative to the particular transformation sought. For example, for the data file in Table 7,

Table 7



	Euclidean Stats Corner document test data file	
	1 Person 1	2 Person 2
Var 1	1	2
Var 2	1	2
Var 3	4	5
Var4	6	5
Var5	1200	1300
Var6	3	4
Var7	2	3
Var8	3	5
Var9	2	3
Var10	8	7

using the following SPSS Correlate-Distances settings … where the "variables" denote person 1 and person 2 …



Distances: Dissimilarity Measures

Measure

● Interval

Measure: Euclidean distance

Power: 2 Root: 2

○ Counts

Measure: Chi-square measure

○ Binary

Measure: Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: Z scores

● By variable

○ By case

Transform Measures

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

we obtain a “Z-score standardized data” Euclidean distance of **0. 008016.** Here, the data have been “column standardized” prior to the calculation being made. If we instead seek a “by case” row standardization, we obtain a value of **4.4721** as per Primer 5.

Other transformations produce other values … with little rationale.

Double-Scaled Euclidean Distance

When comparing two variables/persons, what we can do is to calculate the Euclidean distance from data which is transformed into a 0-1 metric using a strictly linear method (rather than non-linear normalisation-standardisation), then re-scale the resultant Euclidean distance measure itself into a 0-1 range, scaling it into a range defined by 0 through to the maximum possible distance observable between the two variables/persons.

Case 1: Comparing two variables or persons, where observations exists across many persons/cases. In the case of two variables whose minimum and maximum possible values are fixed (either by scale bounds or physical property characteristics) or when comparing say persons across variables, each of whose metric range differs, then each variable's maximum observable discrepancy will need to be calculated and each constituent squared discrepancy of the Euclidean distance calculation will need to be initially normalised using the particular maximum observable discrepancy. The resulting square root of the sum of squared values represents the raw Euclidean distance of these normalised values, which in turn is then scaled to a metric between 0 and 1.0 (where 1.0 represents the maximum discrepancy between the two variables' scaled scores). Computationally, each squared discrepancy is transformed into a 0-1 range, using a simple linear conversion – we keep our raw data "as is" – and simply scale the squared discrepancies for the variables into a 0 to 1.0 range.

Computational steps … Comparing two persons/objects across many variables

Step 1: Determine the maximum possible squared discrepancy for each variable comparison using the minimum and maximum values as specified. Call these values *md.* Each variable will possess a minimum and maximum so the md for each variable is just:

$md_i = (\text{Maximum for variable } i - \text{Minimum for variable } i)^2$

Step 2. Compute the sum of squared discrepancies per variable, dividing through the squared discrepancy (across persons) for each variable by the maximum possible discrepancy for that variable. Then take the square root of the sum to produce the scaled variable Euclidean distance.

Given the formula in equation 1: $d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$

it is now modified to reflect the operations in step 2 …

Eq. 3 $$d_1 = \sqrt{\sum_{i=1}^{v} \left(\frac{(p_{1i} - p_{2i})^2}{md_i} \right)}$$

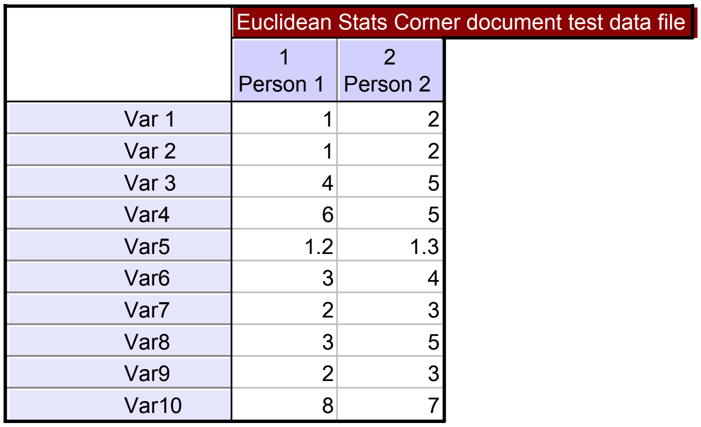
where d_1 = the "scaled variable" Euclidean distance

md_i = the maximum possible squared discrepancy per variable i of v variables.

Step 3. Compute the scaled value from step 3 by dividing it by $\sqrt{v}$, where v = the number of variables.

Eq. 4 $$d_2 = \frac{\sqrt{\sum_{i=1}^{v}\left(\frac{(p_{1i}-p_{2i})^2}{md_i}\right)}}{\sqrt{v}}$$

Example 1:
Assume we have two persons on which we observe the following observations (Table 4)…

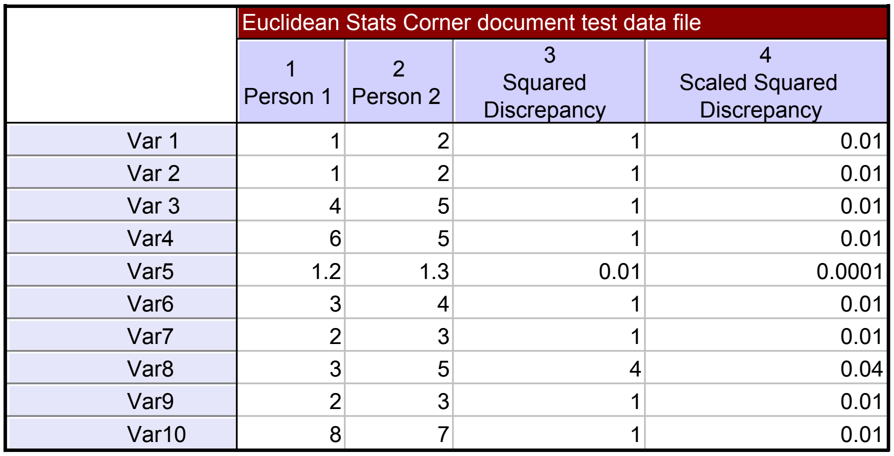


	Euclidean Stats Corner document test data file	
	1 Person 1	2 Person 2
Var 1	1	2
Var 2	1	2
Var 3	4	5
Var4	6	5
Var5	1.2	1.3
Var6	3	4
Var7	2	3
Var8	3	5
Var9	2	3
Var10	8	7

Given we have prior information which tells us that each variable possesses a fixed minimum of 0 and a fixed maximum of 10, then the calculations are as follows:

Step 1:
Given any variable's observations can possess 0 as the minimum, and 10 as the maximum, the maximum possible observable squared discrepancy is thus (0-10)^2 = 100 for each variable. These are our *md*'s.

Step 2:
Compute the sum of squared discrepancies per variable, dividing through the squared discrepancy (across persons) for each variable by the maximum possible discrepancy for that variable…. these values are:



	Euclidean Stats Corner document test data file			
	1 Person 1	2 Person 2	3 Squared Discrepancy	4 Scaled Squared Discrepancy
Var 1	1	2	1	0.01
Var 2	1	2	1	0.01
Var 3	4	5	1	0.01
Var4	6	5	1	0.01
Var5	1.2	1.3	0.01	0.0001
Var6	3	4	1	0.01
Var7	2	3	1	0.01
Var8	3	5	4	0.04
Var9	2	3	1	0.01
Var10	8	7	1	0.01

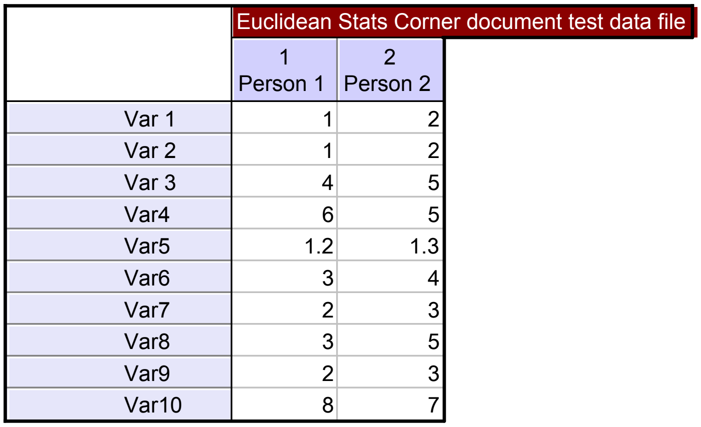
Step 3:
Sum the Scaled Squared Discrepancies, take the square root of the sum, and then scale this coefficient into a 0-1 metric by dividing the scaled Euclidean distance by $\sqrt{10}$ … =

$$d_2 = \frac{\sqrt{0.1201}}{\sqrt{10}} = 0.10959$$

Note that if the minimum and maximum for each variable was between 0 – 25, then the double-scaled Euclidean distance d_2 = 0.043838. This is a reminder that the Euclidean distance is relative the maximum possible distance between variables. This point is taken up again in the next example.

Example 2:

Assume we have two persons on which we observe the following observations (Table 4)…

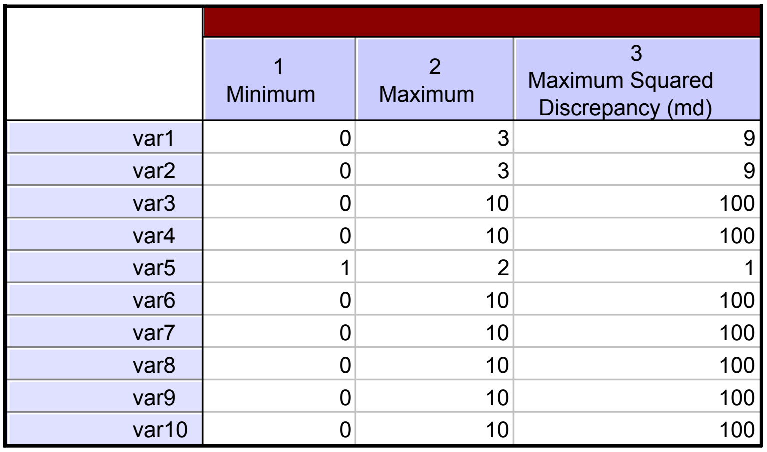


Euclidean Stats Corner document test data file

	1 Person 1	2 Person 2
Var 1	1	2
Var 2	1	2
Var 3	4	5
Var4	6	5
Var5	1.2	1.3
Var6	3	4
Var7	2	3
Var8	3	5
Var9	2	3
Var10	8	7

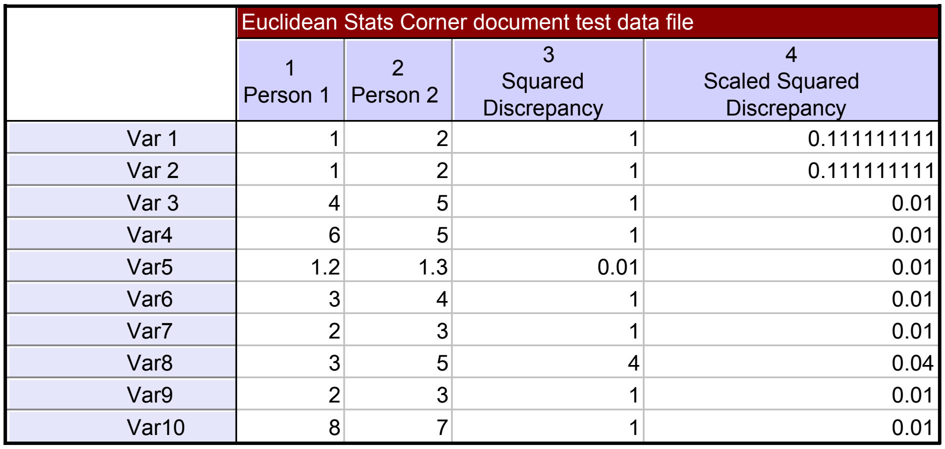
However, variables 1 and 2 have minimum and maximum values of 0 and 3, whilst variable 5 has a minimum of 1 and a maximum of 2. The remainder have a minimum of 0 and a maximum of 10.

Step 1: Determine the maximum possible squared discrepancy for each variable comparison using the minimum and maximum values as specified.



	1 Minimum	2 Maximum	3 Maximum Squared Discrepancy (md)
var1	0	3	9
var2	0	3	9
var3	0	10	100
var4	0	10	100
var5	1	2	1
var6	0	10	100
var7	0	10	100
var8	0	10	100
var9	0	10	100
var10	0	10	100

Step 2:
Compute the sum of squared discrepancies per variable, dividing through the squared discrepancy (across persons) for each variable by the maximum possible discrepancy for that variable…. these values are:



	Euclidean Stats Corner document test data file			
	1 Person 1	2 Person 2	3 Squared Discrepancy	4 Scaled Squared Discrepancy
Var 1	1	2	1	0.111111111
Var 2	1	2	1	0.111111111
Var 3	4	5	1	0.01
Var4	6	5	1	0.01
Var5	1.2	1.3	0.01	0.01
Var6	3	4	1	0.01
Var7	2	3	1	0.01
Var8	3	5	4	0.04
Var9	2	3	1	0.01
Var10	8	7	1	0.01

Step 3:
Sum the Scaled Squared Discrepancies, take the square root of the sum, and then scale this coefficient into a 0-1 metric by dividing the scaled Euclidean distance by $\sqrt{10}$ … =

$$d_2 = \frac{\sqrt{0.332222}}{\sqrt{10}} = 0.18227$$

Note, the distance is now larger than in example 1 because the ranges of variables 1, 2, and 5 are now far less than 10, hence a discrepancy of 0.1 within a range of 1.0 is far greater than a 0.1 discrepancy within a 0-10 potential range.

Which brings home the issue of “relativity” of the distance to a predefined metric space. The distance is always relative to the maximum possible distance for two variables to differ. So, a distance of just 2 relative to a maximum possible distance of 4 will be larger than the same distance between two variables whose maximum discrepancy is 400. This linear scaling embodies precisely the concept of Euclidean distance, but relative to absolute maximum distance.

What if you don't know the minimum or maximum values for a variable?

Good question. All you can do is use your best estimate. One option is simply to use the observed minimum and maximum for each variable in your dataset – but if using several data files or blocks of data on the same variables with the aim of making comparisons between them (in terms of distance/similarity analysis), then you should really set a fixed minimum and maximum for each variable which will permit a unified metric scale to be applied to all such variables.

For example, in a marine biology study using variables such as "Depth" or "temperature" of water at which numbers of fish are found; you would have to decide what might be the lowest valid depth or temperature at which you might make any observations (say 0 meters or 0 deg C, through to 40 meters and 25 degrees C). The validity of these figures will be provided by logic, theory, and common-sense.

The choice you make will determine the scaling constant for the variables.

Still within **Case 1: Comparing two variables or persons, where observations exists across many persons/cases,** let's look at Comparing two variables across many persons/observations.

Computational steps … Comparing two variables across many persons/observations

Step 1: Determine the maximum possible squared discrepancy for each variable comparison using the minimum and maximum values as specified. Call these values *md.* Because each variable's minimum and maximum values will be different, and neither minimum need necessarily be 0.0, a simple decision algorithm is required to determine the maximum discrepancy. Given …

minv1 = minimum value for variable 1 maxv1 = maximum value for variable 1
minv2 = minimum value for variable 2 maxv2 = maximum value for variable 2

Then the maximum discrepancy is given by:

```
if (abs(maxv2-minv1)) <= (abs(maxv1-minv2)) then
  md = $(\text{maxv1-minv2})^2$
else
  md= $(\text{maxv2-minv1})^2$
```

* Note that *md* acts as a constant here.

Step 2. Compute the sum of squared discrepancies per observation, dividing through the squared discrepancy for each pair of observations by the maximum possible discrepancy observable given these two variables. Then take the square root of the sum to produce the scaled variable Euclidean distance.

Given the formula in equation 2: $d = \sqrt{\sum_{i=1}^{p}(v_{1i} - v_{2i})^2}$

where p = the number of paired observations $i = 1$ to p

it is now modified to reflect the operations in step 2 …

Eq. 3 $$d_1 = \sqrt{\sum_{i=1}^{p}\left(\frac{(v_{1i} - v_{2i})^2}{md}\right)}$$

where d_1 = the "scaled variable" Euclidean distance
md = the maximum possible squared discrepancy between these two variables

Step 3. Compute the scaled value from step 3 by dividing it by $\sqrt{p}$, where p = the number of paired observations

Eq. 4 $$d_2 = \frac{\sqrt{\sum_{i=1}^{p}\left(\frac{(v_{1i}-v_{2i})^2}{md}\right)}}{\sqrt{p}}$$

Example
For two variables, with 9 observations on each variable, and where the minimum value on each variable is 0, with the maximum on variable 1 of 20, and 40 for variable 2.

Step 1: Determine the maximum possible squared discrepancy for each variable comparison using the minimum and maximum values as specified.

```
if (abs(maxv2-minv1)) <= (abs(maxv1-minv2)) then
  md = (maxv1-minv2)^2
else
  md= (maxv2-minv1)^2
```

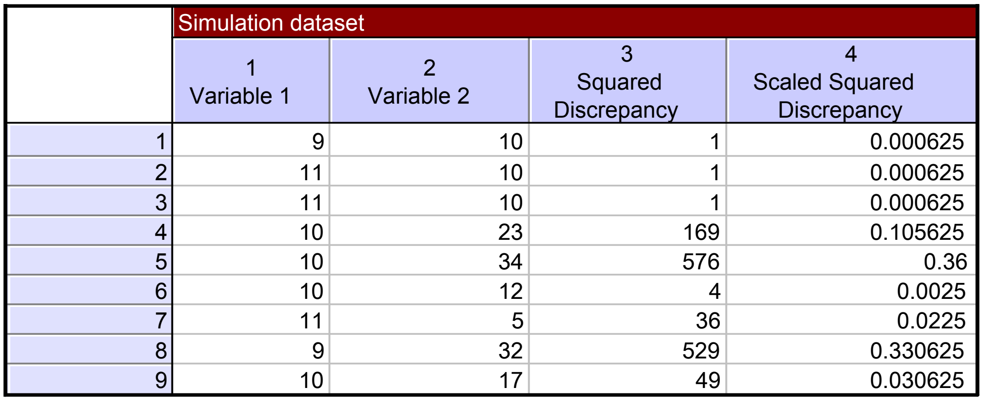
Which, when we use our actual values …

```
if (abs(40-0)) <= (abs(20-0)) then
  md = (20-0)^2
else
  md= (40-0)^2
```

which means that *md* in this example is **1600**.

Step 2. Compute the sum of squared discrepancies per observation, dividing through the squared discrepancy for each pair of observations by the maximum possible discrepancy observable given these two variables. Then take the square root of the sum to produce the scaled variable Euclidean distance.

The relevant data are



	Simulation dataset			
	1 Variable 1	2 Variable 2	3 Squared Discrepancy	4 Scaled Squared Discrepancy
1	9	10	1	0.000625
2	11	10	1	0.000625
3	11	10	1	0.000625
4	10	23	169	0.105625
5	10	34	576	0.36
6	10	12	4	0.0025
7	11	5	36	0.0225
8	9	32	529	0.330625
9	10	17	49	0.030625

Step 3. Compute the scaled value from step 3 by dividing it by $\sqrt{p}$, where p = the number of paired observations

Sum the Scaled Squared Discrepancies, take the square root of the sum, and then scale this coefficient into a 0-1 metric by dividing the scaled Euclidean distance by $\sqrt{9}$ … =

$$d_2 = \frac{\sqrt{0.85375}}{\sqrt{9}} = 0.307995$$

Case 2: Comparing observations to a target profile, where a fixed array of values are used as the target profile.

When comparing persons to say a fixed target of variable values (as in person-target profiling), then the maximum discrepancy per variable is a function of the target value, as well as the upper and lower bounds. As above, each variable's permissible range determines the scaling factor – but because the target values are fixed, and because the comparison is thus constrained by these values, the potential combinations of values on both variables is actually constrained by the fixed target. In effect, the distance used to express score/value discrepancy for each variable is adjusted relative to the target profile variable values.

Let's take the following dataset …

	Person-Target Profiling example - Double-scaled Euclidean			
	1 Target Profile	2 Comparison Profile	3 Minimum for var	4 Maximum for var
Extraversion	10	12	0	24
Conscientiousness	15	17	0	24
Days Absenteeism	1	3	0	10
Gallup Score	40	33	12	60
Communication Rating	4	3	1	5
Accident History	1	0	0	5
Performance Rating	70	92	0	100
Verbal Ability	15	17	0	32
Abstract Ability	12	19	0	24
Numerical Ability	10	13	0	28

What we see is a target profile taken over a mixture of 10 personal attributes, ratings, and behavioural criteria (you also immediately see the problem in using mixed-variable profiles – in that you actually need to combine distance metrics with decision/threshold/bound type statements).

Note that the minimum and maximum value is different for almost every variable.

What we do first is to compute the maximum squared discrepancy for each variable, taking into account that a comparison profile value can only differ from the target value, and not the minimum or maximum variable value. So, the maximum discrepancy is between the target value and the minimum or maximum value for a variable, whichever is the larger.

For example, from the above data, with our target value for variable 1 as 10, and a minimum and maximum value for that variable as 0 and 24, then the maximum discrepancy observable is the larger of

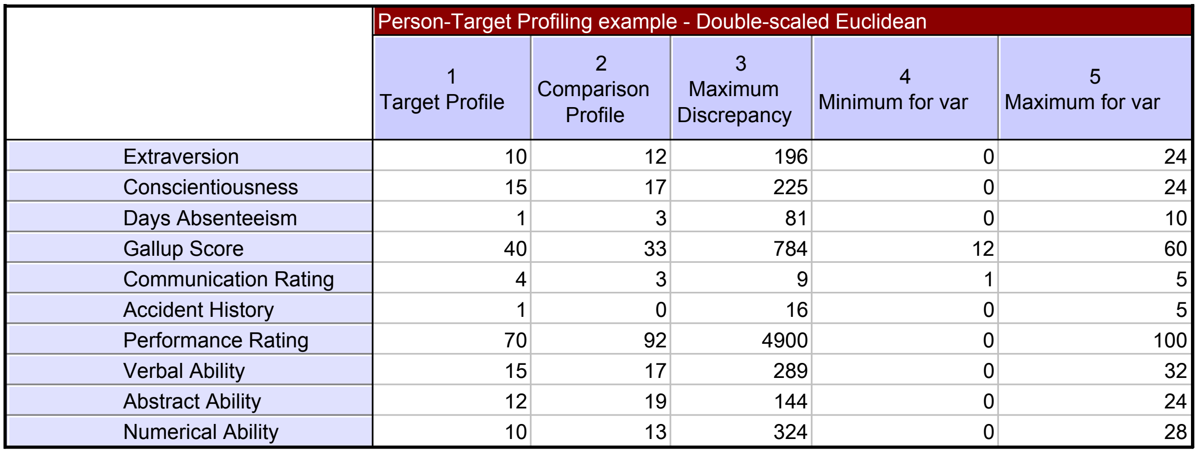
abs(target – minimum variable value)   or   *abs*(target-maximum variable value)

In our example, this translates to: *abs*(10-0)=10   or   *abs*(10-24)=14

The rule is to choose the larger of the two discrepancies. Thus, our maximum discrepancy is 14, which when squared is 196.

That is, our comparison profile value can vary between 0 and 24, but the maximum possible discrepancy which can be observed between the target and a comparison profile value is actually between 10 and 24 (because the target value is in fact fixed, whilst the comparison profiles with observations on this variable can vary between 0 and 24).

If we do the calculations for the maximum discrepancies we have …



	Person-Target Profiling example - Double-scaled Euclidean				
	1 Target Profile	2 Comparison Profile	3 Maximum Discrepancy	4 Minimum for var	5 Maximum for var
Extraversion	10	12	196	0	24
Conscientiousness	15	17	225	0	24
Days Absenteeism	1	3	81	0	10
Gallup Score	40	33	784	12	60
Communication Rating	4	3	9	1	5
Accident History	1	0	16	0	5
Performance Rating	70	92	4900	0	100
Verbal Ability	15	17	289	0	32
Abstract Ability	12	19	144	0	24
Numerical Ability	10	13	324	0	28

Now, we apply the usual scaling formulae to yield a double-scaled Euclidean distance …

Eq. 5 $$d_1 = \sqrt{\sum_{i=1}^{v}\left(\frac{(t_i - c_i)^2}{md_i}\right)}$$

where t = the target profile variable value

c = the comparison profile variable value

d_1 = the “scaled variable” Euclidean distance

md_i = the maximum possible squared discrepancy per variable i of v variables.

Step 3. Compute the scaled value from step 3 by dividing it by $\sqrt{v}$, where v = the number of variables.

Eq. 6 $$d_2 = \frac{\sqrt{\sum_{i=1}^{v}\left(\frac{(t_i - c_i)^2}{md_i}\right)}}{\sqrt{v}}$$

The data now look like ..

	Person-Target Profiling example - Double-scaled Euclidean						
	1 Target Profile	2 Comparison Profile	3 Maximum Discrepancy	4 Squared Discrepancy (Target-Comparison)	5 Scaled Squared Discrepancy	6 Minimum for var	7 Maximum for var
Extraversion	10	12	196	4	0.0204081633	0	24
Conscientiousness	15	17	225	4	0.0177777778	0	24
Days Absenteeism	1	3	81	4	0.049382716	0	10
Gallup Score	40	33	784	49	0.0625	12	60
Communication Rating	4	3	9	1	0.111111111	1	5
Accident History	1	0	16	1	0.0625	0	5
Performance Rating	70	92	4900	484	0.0987755102	0	100
Verbal Ability	15	17	289	4	0.0138408304	0	32
Abstract Ability	12	19	144	49	0.340277778	0	24
Numerical Ability	10	13	324	9	0.0277777778	0	28

With the final calculation as: $d_2 = \frac{\sqrt{0.804352}}{\sqrt{10}} = 0.283611$

The Raw Euclidean Distance is **24.6779.**

If we had just treated the data as per Case 1 above, comparing two persons across different variables, with unequal minimums and maximums, and not used Person 1 (column 1 of the data file) as a "fixed target" … we would have obtained a d_2 distance of **0.18606**. The lower d_2 distance is due to the fact that we have expanded the maximum possible discrepancies using the absolute minimums for each variable to define the maximum discrepancy, rather than the permissible minimums (as per the target values).

Again, this example serves to underline the importance of determining the actual "Euclidean space" within which distance calculations will be made.

To repeat my own words from above ..

> "Which brings home the issue of "relativity" of the distance to a predefined metric space. The distance is always relative to the maximum possible distance for two variables to differ. So, a distance of just 2 relative to a maximum possible distance of 4 will be larger than the same distance between two variables whose maximum discrepancy is 400. This linear scaling embodies precisely the concept of Euclidean distance, but relative to absolute maximum distance."

One final point … given the double-scaled metric coefficient, it is easy to turn it into a measure of similarity by subtracting it from 1.0, this in the above example, the dissimilarity coefficient is 0.283611. If we express this as a similarity coefficient, it becomes (1-0.283611) = 0.716389.

A useful feature for profile comparison/profile matching applications.