Variable scaling in cluster analysis of linguistic data

Hermann Moisl
School of English Literature, Language, and Linguistics
University of Newcastle, UK

hermann.moisl@ncl.ac.uk
http://www.ncl.ac.uk/elll/staff/profile/hermann.moisl
http://www.staff.ncl.ac.uk/hermann.moisl/

Abstract

Where the variables selected for cluster analysis of linguistic data are measured on different numerical scales, those whose scales permit relatively larger values can have a greater influence on clustering than those whose scales restrict them to relatively smaller ones, and this can compromise the reliability of the analysis. The first part of this discussion describes the nature of that compromise. The second part argues that a widely used method for removing disparity of variable scale, Z-standardization, is unsatisfactory for cluster analysis because it eliminates differences in variability among variables, thereby distorting the intrinsic cluster structure of the unstandardized data, and instead proposes a standardization method based on variable means which preserves these differences. The proposed mean-based method is compared to several other alternatives to Z-standardization, and is found to be superior to them in cluster analysis applications.

Keywords

Linguistic data analysis; cluster analysis; variable scaling; variable standardization

Introduction

The variables selected for a linguistics research project involving cluster analysis may require measurement on different scales. In sociolinguistics, for example, speakers might be described by a set of variables one of which represents the frequency of usage of some phonetic segment, another one age, and a third income. Because these variables represent different kinds of thing in the world, they are measured in numerical units and ranges appropriate to them: phonetic frequency in the integer range, say, 1..1000, age in the integer range 20..100, and income in some currency in the real-valued range, again say, 0..50000.00. Humans understand that one can't compare apples and oranges and, faced with different scales, use the variable semantics to interpret their values sensibly. But cluster analysis methods don't have common sense. Given an *m* x *n* data matrix M in which the *m* rows represent the *m* objects to be clustered, the *n* columns represent the *n*

variables, and the entry at M_{ij} (for i = 1..m, j = 1..n) represents a numerical measure of object i in terms of variable j, a clustering method has no idea what the values in the matrix mean and calculates the degrees of similarity between the row vectors purely on the basis of the relative numerical magnitudes of the variable values. As a consequence, variables whose scales permit relatively larger magnitudes can have a greater influence on the cluster analysis than those whose scales restrict them to relatively smaller ones, and this can compromise the reliability of the analysis, as has often been noted (for example Romesburg 1984: Ch.7; Kaufman and Rousseeuw 1990: 4-11; Gnanadesikan 1997: 102-105; Hair et al. 2006: Ch.8; Kettenring 2006; Tan et al. 2006: 64-65; Chu et al. 2009). The first part of this discussion examines the nature of this compromise, and the second proposes a resolution; the discussion relates to numerical data only, and does not consider categorical data.

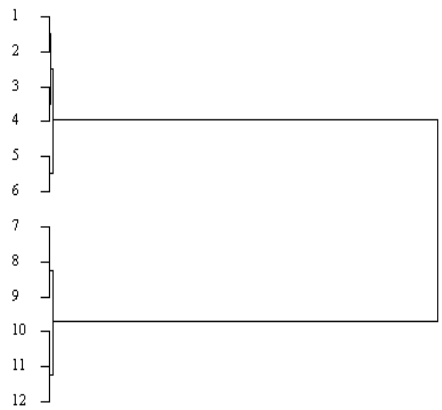
1. The problem

Table 1 shows variants (a)-(c) of a matrix that describes a dozen speakers in terms of three variables and, for each variant, a cluster analysis of the matrix rows using squared Euclidean distance and Ward's Method (Everitt et al. 2001).

Table 1: Versions of a data matrix with different variable scales and corresponding cluster analyses of their row vectors

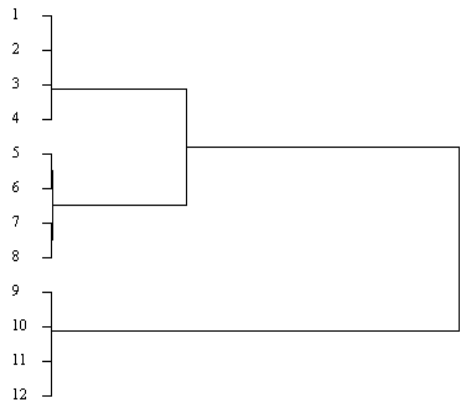
a

	Frequency	**Age (years)**	**Income (Euros)**
1	100	20	30000
2	105	21	30250
3	110	22	30500
4	115	30	30750
5	200	31	31000
6	205	32	31250
7	210	40	35000
8	215	41	35250
9	300	42	35500
10	305	50	35750
11	310	51	36000
12	315	52	36250



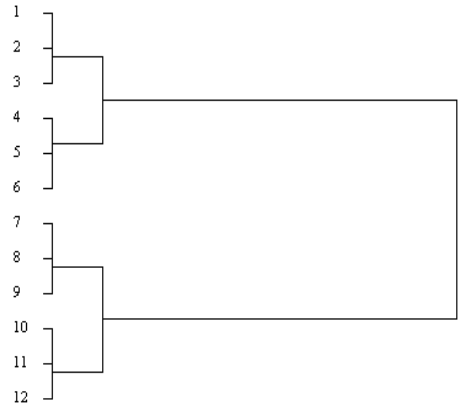
b

	Frequency	Age (years)	Income (K Euros)
1	100	20	30.00
2	105	21	30.25
3	110	22	30.50
4	115	30	30.75
5	200	31	31.00
6	205	32	31.25
7	210	40	35.00
8	215	41	35.25
9	300	42	35.50
10	305	50	35.75
11	310	51	36.00
12	315	52	36.25



c

	Frequency	Age (days)	Income (K Euros)
1	100	7300	30.00
2	105	7665	30.25
3	110	8030	30.50
4	115	10950	30.75
5	200	11315	31.00
6	205	11680	31.25
7	210	14600	35.00
8	215	14965	35.25
9	300	15330	35.50
10	305	18250	35.75
11	310	18615	36.00
12	315	18980	36.25



In Table 1a the first variable represents the frequency of speakers' usage of some phonetic segment of interest, the second age in years, and the third annual income in Euros. In Table 1b *Frequency* and *Age* are as in 1a but *Income* is now expressed as the number of thousands of Euros (K), and Table 1c both retains the *Income* scale of 1b and also expresses *Age* in terms of days rather than years. Using the variable semantics, a human interpreter would see from direct inspection of the matrices that, regardless of variation in scale, the descriptions of the speakers are in fact equivalent and that they fall into three phonetic frequency groups, four age groups, and two income groups. That same interpreter would expect cluster analysis to use these groupings as the basis for a result that is consistent across all three matrices and independent of the variation in scaling, but it does not. The trees in Table 1 differ substantially, and they cluster the speakers according to the relative magnitude of values in the matrix columns. In Table 1a the largest values are those in the *Income* column and the corresponding cluster tree divides the speakers into two main groups, those with incomes in the range 30000-31250 and those with incomes in the range 35000-36250; in Table 1b the largest

values are those in the *Frequency* column, and the corresponding cluster tree classifies the speakers into three main groups (100-115), (200-215), and (300-315) by frequency; in Table 1c the *Age* column is the one with the largest values, and, predictably, the speakers are now divided into four main groups (7300-8030), (10950-11680), (14600-15330), and (18250-18980) by age.

That the result of cluster analysis should be contingent on the vagaries of scale selection is self-evidently unsatisfactory both in the present case and in general. Some way of eliminating scale as a factor is required; Section 2 proposes one.

2. Proposed solution

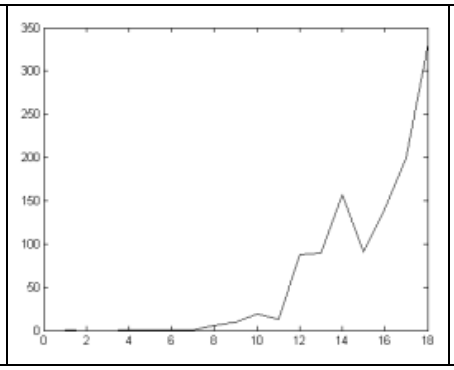
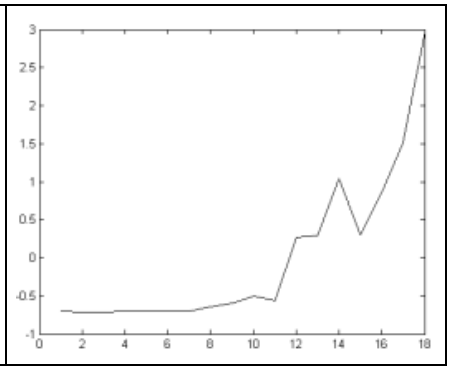
Relative to an *m* x *n* matrix M, the solution to the above problem is to standardize the variables by transforming the values in the column vectors of M in such a way that variation in scale among them is removed; if all the variables are measured on the same scale, none can dominate. The textbook method for doing this is via standard scores, also known as Z-scores and autoscaling (for example Romesburg 1984: 11; Kaufman and Rousseeuw 1990: 6-11; Everitt 2001: 51; Hair et al. 2006: Ch.8; Kettenring 2006; Boslaugh and Watters 2008: 369-370; Chu et al. 2009), which transforms the original values in any column vector M_j (for $j = 1..n$) into ones which say how many standard deviations those original values are from the vector mean. In what follows, this method is referred to as *Z-standardization*. Expression 1 gives the formula for Z-standardizing the *i*'th value (for $i = 1..m$) in any given column vector M_j.

$$zscore(M_{ij}) = \frac{(M_{ij} - \mu(M_j))}{\sigma(M_j)} \qquad 1$$

where $\mu(M_j)$ is the column vector mean and $\sigma(M_j)$ its standard deviation. The Z-standardization of an arbitrary vector *x* is shown in Table 2.

Table 2: Z-standardization of an arbitrary vector *x*

	Original *x*	**Z-standardized *x***
Values	[1 0 0 1 1 1 1 6 10 19 13 88 90 157 91 141 199 331]	[-0.7006 -0.7118 -0.7118 -0.7006 -0.7006 -0.7006 -0.7006 -0.6449 -0.6004 -0.5001 -0.5669 0.2686 0.2909 1.0373 0.3020 0.8591 1.5052 2.9759]
Mean	63.89.	0
Standard deviation	89.76	1

Plot		

Z-standardization transforms any vector into one having a mean of 0 and a standard deviation of 1, and, because division by a constant is a linear operation, the shape of the distribution of the original values is preserved, as is shown by the pre- and post-standardization plots in Table 2. Only the scale changes: 0..331 for original *x*, and -0.7006..2.9759 for transformed *x*.

When Z-standardization is applied to each of the column vectors of a matrix, any variation in scale across those variables disappears because all the variables are now expressed in terms of the number of standard deviations from their respective means. Table 3 shows, for example, Z-standardization of the data matrices in Table 1.

Table 3: Comparison of matrices in Table 1 and their Z-standardized versions

a: Unstandardized matrices from Table 1a				**b: Corresponding Z-standardized matrices**			
	Frequency	**Age (years)**	**Income (Euros)**		**Frequency**	**Age (years)**	**Income (Euros)**
1	100	20	30000	**1**	-1.3135	-1.4273	-1.2322
2	105	21	30250	**2**	-1.2524	-1.3381	-1.1336
3	110	22	30500	**3**	-1.1913	-1.2489	-1.035
4	115	30	30750	**4**	-1.1302	-0.53523	-0.93644
5	200	31	31000	**5**	-0.091641	-0.44603	-0.83787
6	205	32	31250	**6**	-0.030547	-0.35682	-0.7393
7	210	40	35000	**7**	0.030547	0.35682	0.7393
8	215	41	35250	**8**	0.091641	0.44603	0.83787
9	300	42	35500	**9**	1.1302	0.53523	0.93644
10	305	50	35750	**10**	1.1913	1.2489	1.035
11	310	51	36000	**11**	1.2524	1.3381	1.1336
12	315	52	36250	**12**	1.3135	1.4273	1.2322
Mean	207.5	36	33125	**Mean**	0	0	0
Std dev	81.84	11.21	2536.20	**Std dev**	1	1	1
	Frequency	**Age (years)**	**Income (K Euros)**		**Frequency**	**Age (years)**	**Income (K Euros)**
1	100	20	30.00	**1**	-1.3135	-1.4273	-1.2322
2	105	21	30.25	**2**	-1.2524	-1.3381	-1.1336
3	110	22	30.50	**3**	-1.1913	-1.2489	-1.035

4	115	30	30.75	**4**	-1.1302	-0.53523	-0.93644
5	200	31	31.00	**5**	-0.091641	-0.44603	-0.83787
6	205	32	31.25	**6**	-0.030547	-0.35682	-0.7393
7	210	40	35.00	**7**	0.030547	0.35682	0.7393
8	215	41	35.25	**8**	0.091641	0.44603	0.83787
9	300	42	35.50	**9**	1.1302	0.53523	0.93644
10	305	50	35.75	**10**	1.1913	1.2489	1.035
11	310	51	36.00	**11**	1.2524	1.3381	1.1336
12	315	52	36.25	**12**	1.3135	1.4273	1.2322
Mean	207.5	26	33.13	**Mean**	0	0	0
Std dev	81.84	11.21	2.54	**Std dev**	1	1	1

	Frequency	**Age (days)**	**Income (K Euros)**		**Frequency**	**Age (days)**	**Income (K Euros)**
1	100	7300	30.00	**1**	-1.3135	-1.4273	-1.2322
2	105	7665	30.25	**2**	-1.2524	-1.3381	-1.1336
3	110	8030	30.50	**3**	-1.1913	-1.2489	-1.035
4	115	10950	30.75	**4**	-1.1302	-0.53523	-0.93644
5	200	11315	31.00	**5**	-0.091641	-0.44603	-0.83787
6	205	11680	31.25	**6**	-0.030547	-0.35682	-0.7393
7	210	14600	35.00	**7**	0.030547	0.35682	0.7393
8	215	14965	35.25	**8**	0.091641	0.44603	0.83787
9	300	15330	35.50	**9**	1.1302	0.53523	0.93644
10	305	18250	35.75	**10**	1.1913	1.2489	1.035
11	310	18615	36.00	**11**	1.2524	1.3381	1.1336
12	315	18980	36.25	**12**	1.3135	1.4273	1.2322
Mean	207.5	13140	33.13	**Mean**	0	0	0
Std dev	81.84	4091.69	2.54	**Std dev**	1	1	1

The Z-standardized versions of the matrices in Table 3b are identical despite the variations of scale in those of 3a. Cluster analysis of the rows of this standardized matrix, moreover, generates a tree, shown in figure 1, that differs from any of those in Table 1. It was generated using squared Euclidean distance and Ward's method, as before, and this combination is used throughout the remainder of the discussion to maintain comparability among analyses.

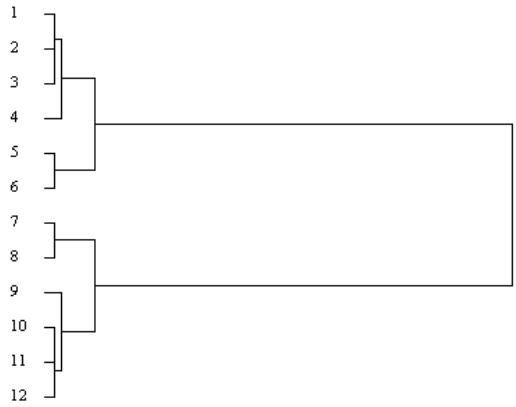


Figure 1: Cluster analysis of matrix rows in Table 3b

No one variable is dominant by virtue of the magnitudes of its values relative to the magnitudes of the others. Instead, all three play an equal part in determining the cluster structure, resulting in a symmetrical tree which reflects the symmetry of the standardized matrix's row vectors in Table 3b: vectors 1 and 12 are numerically identical but with opposite signs, as are 2 and 11, 3 and 10, and so on.

Variants of Z-standardization based on the mean absolute deviation or the median absolute deviation rather than on the standard deviation are often used because these are less sensitive to distortion by outliers (Kaufman and Rousseeuw 1990: 5-9), but the difference between them and Z-standardization is not significant for present purposes, and they are not further considered here.

Z-standardization appears, therefore, to be a good general solution to the problem of variation in scaling among data variables, and it is in fact widely used for that purpose. It is, however, arguable that, for cluster analysis, Z-standardization should be used with caution or not at all, again as others have observed (Romesburg 1984: Ch.7; Milligan and Cooper 1988; Gnanandesikan et al. 1995; Kettenring 2006; Chu et al. 2009). The remainder of this section is in three parts: the first presents the argument against Z-standardization, the second proposes an alternative standardization method, and the third assesses the alternative method relative to some others proposed in the literature.

2.1. *The argument against Z-standardization*

The argument depends on making a distinction between three properties of a variable.

- The absolute magnitude of values of a variable is the numerical size of its values, and can for present purposes be taken as the absolute maximum of those values. For *Frequency* in Table 3a, for example, it is 315 on that criterion.

- The absolute magnitude of variability is the amount of variation in the values of a variable expressed in terms of the scale of those values, and is measured by the standard deviation. In Table 3a the absolute magnitude of variability of the *Frequency* column is 81.84.

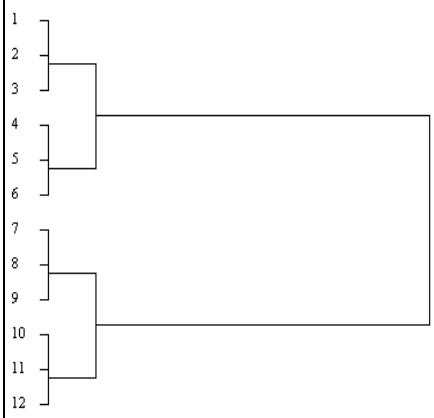
- The intrinsic variability is the amount of variability in the values of a variable expressed independently of the scale of those values. This is measured in statistics by the coefficient of variation (see for example Boslaugh and Watters 2008: 62), which is defined with respect to a variable *v* as the ratio of *v*'s standard deviation to its mean, as in Expression 2.

$$CoefficientOfVariation(v) = \frac{\sigma_v}{\mu_v} \qquad 2$$

The intuition gained from direct inspection of the matrices in Table 3a is that there is much more variability in the values of the *Frequency* and *Age* columns than there is for those in the *Income* column regardless of the variation in their absolute magnitudes and absolute magnitudes of variability. The coefficient of variation captures this intuition: for *Frequency* it is 0.383, for *Age* almost as much at 0.311, and for *Income* much less at 0.141. Because the coefficient of variation is scale-independent it can be used as a general way of comparing the degrees of variability of variables measured on different scales.

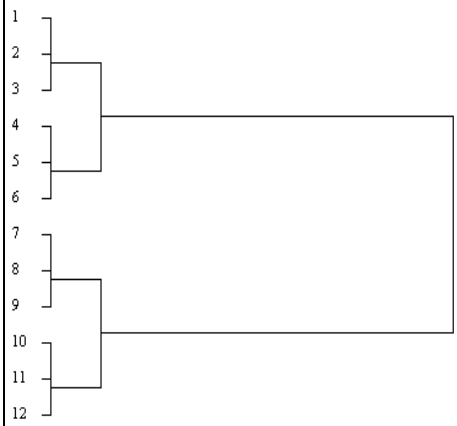
Table 4 exemplifies the interrelationship of these three properties. Each row (a)-(d) shows a two-dimensional matrix with the standard deviations and coefficients of variation of its column vectors together with a cluster analysis of the row vectors; the values of *v1* are altered in various ways in the (a)-(d) sequence, and those of *v2* are held constant.

Table 4: Interrelationship of absolute magnitude, absolute magnitude of variability, and intrinsic variability

	Data matrix			Cluster analysis
a		**v1**	**v2**	
	1	1000	100	
	2	1000	100	
	3	1000	100	
	4	1000	200	
	5	1000	200	
	6	1000	200	
	7	1000	300	
	8	1000	300	
	9	1000	300	
	10	1000	400	
	11	1000	400	
	12	1000	400	
	Standard deviation	0	111.80	
	Coefficient of variation	0	0.447	

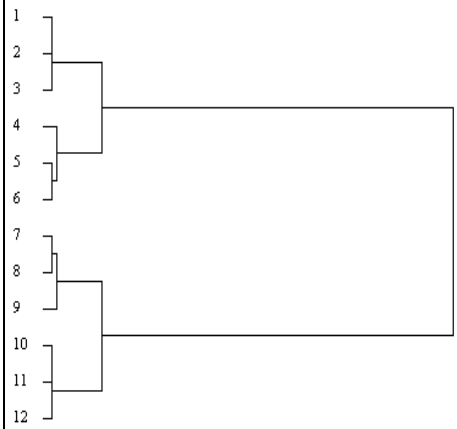
b

	v1	v2
1	10000	100
2	10000	100
3	10000	100
4	10000	200
5	10000	200
6	10000	200
7	10000	300
8	10000	300
9	10000	300
10	10000	400
11	10000	400
12	10000	400
Standard deviation	0	111.80
Coefficient of variation	0	0.447



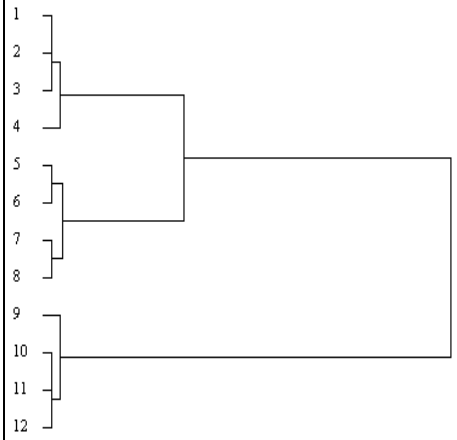
c

	v1	v2
1	1000	100
2	1000	100
3	1000	100
4	1000	200
5	1050	200
6	1050	200
7	1050	300
8	1050	300
9	1100	300
10	1100	400
11	1100	400
12	1100	400
Standard deviation	40.82	111.80
Coefficient of variation	0.039	0.447



d

	v1	v2
1	1000	100
2	1000	100
3	1000	100
4	1000	200
5	1200	200
6	1200	200
7	1200	300
8	1200	300
9	1400	300
10	1400	400
11	1400	400
12	1400	400
Standard deviation	151.84	111.80
Coefficient of variation	0.136	0.447



In Table 4a there is no variability in the values of *v1*, the coefficient of variation and standard deviation are commensurately 0, and, even though the absolute magnitude of the values in *v1* is much greater than that in *v2*, clustering is determined entirely by the variation in the values of *v2*: there are four primary clusters corresponding to the 100-400 value-groups. In Table 4b the absolute magnitude of *v1* is substantially increased but the increase is uniform so that there is still no variability and the standard deviation and coefficient of variation remain 0; the cluster analysis is again determined by the variability in *V2* and is identical to the one in 4a. In Table 4c a relatively small amount of variability is introduced into the values of *v1*, which results in nonzero standard deviation and coefficient of variation, though both of these are smaller than those of *v2*; the cluster tree differs from the one in 4a / 4b in that the same four primary clusters remain, but the pattern of variability across rows 4-6 and 7-9 is now different from that in rows 1-3 and 10-12, and this is expressed in the internal structures of the corresponding clusters. Finally, the amount of variability in *v1* is increased still further in Table 4d, and this is reflected in a higher standard deviation and coefficient of variation. For the first time, however, the standard deviation of *v1* is greater than that of *v2*, and, even though the coefficient of variation is still smaller than *v2*'s, there are now three rather than the previous four primary clusters corresponding to the *v1* value groups 1000, 1200, and 1400. It is neither the absolute magnitude nor the intrinsic variability of a variable's values that determine clustering, but its absolute magnitude of variability: the larger the standard deviation of a variable, the greater its effect on clustering.

How does this relate to the use of Z-standardization of data for cluster analysis? It is a general property of every Z-standardized vector, noted above, that its standard deviation is 1. Z-standardization of multiple columns of a matrix therefore imposes a uniform absolute magnitude of variability on them. This is shown in Table 5; the coefficient of variation cannot be shown for the Z-standardized variables in 5b because the formula for the coefficient of variation involves division by the mean and, for a Z-standardized vector, this is always 0.

Table 5: Unstandardized and Z-standardized versions of a matrix

	v1	**v2**	**v3**
1	100	20	1000.10
2	110	21	1000.11
3	120	22	1000.08
4	130	40	1000.01
5	200	41	1000.20
6	210	42	1000.07
7	220	60	1000.23
8	230	61	1000.30
9	300	62	1000.03
10	310	80	1000.14
11	320	81	1000.13
12	330	82	1000.68
Standard deviation	82.411	22.376	0.171
Coefficient of	0.383	0.439	0.00017

	v1	**v2**	**v3**
1	-1.395	-1.385	-0.426
2	-1.274	-1.341	-0.368
3	-1.153	-1.296	-0.542
4	-1.031	0.492	-0.949
5	-0.182	0.447	0.155
6	-0.061	0.402	-0.600
7	0.061	0.402	0.329
8	0.182	0.447	0.736
9	1.031	0.492	-0.833
10	1.153	1.296	-0.194
11	1.274	1.341	-0.252
12	1.395	1.385	2.943
Standard deviation	1	1	1
Coefficient	-	-	-

variation					of variation				
a: Unstandardized					b: Z-standardized				

Because the absolute magnitude of variability determines the degree of a variable's effect on clustering, the implication is that all the column vectors in a Z-standardized matrix have an equal influence. This obviously eliminates any possibility of dominance by variables with relatively high absolute magnitudes of variability, but there is a price, and that price might be felt to be too high in any given research application. Intuitively, real-world objects can be distinguished from one another in proportion to the degree to which they differ: identical objects cannot be distinguished, objects that differ moderately from one another are moderately easy to distinguish, and so on. Data variables used to describe real-world objects to be clustered are therefore useful in proportion to the variability in their values: a variable with no variability says that the objects are identical with respect to the characteristic it describes and can therefore contribute nothing as a clustering criterion, a variable with moderate variability says that the corresponding objects are moderately distinguishable with respect to the associated characteristic and is therefore moderately useful as a clustering criterion, and again so on. Variables *v1* and *v2* in Table 5 have high intrinsic variabilities relative to *v3* and are therefore more useful clustering criteria than *v3*. In fact, the variability of *v3* is so small that it could be the result of random observational noise with respect to a characteristic that is constant across the objects to be clustered. To equate *v3* with *v1* and *v2* in terms of its influence on clustering, as Z-standardization does, cannot be right. Rescaling data values so that all variables have an identical absolute magnitude of variability diminishes the distinguishing power of high-variability variables and enhances the power of low-variability ones relative to what is warranted by observed reality. In other words, Z-standardization can distort the validity of data as an accurate description of reality, and this is the reason why it should be used with caution or not at all in data preparation for cluster analysis.

For multivariate data whose variables are measured on different scales, what is required is a standardization method that, like Z-standardization, eliminates the distorting effect of disparity of variable scale on clustering but, unlike Z-standardization, also preserves the relativities of size of the pre-standardization intrinsic variabilities in the post-standardization absolute magnitudes of variation, or, in other words, generates standardized variable vectors such that the ratios of their absolute magnitudes of variability are identical to the ratios of the intrinsic variabilities of the unstandardized ones. In this way the standardized variables can influence the clustering in proportion to the real-world distinguishability of the objects they describe.

2.2 *An alternative to Z-standardization*

The literature (Milligan and Cooper 1998; Gnanandesikan et al. 1995; Chu et al. 2009) contains a variety of alternatives to Z-standardization, but, relative to the desiderata just stated, one of them seems the obvious choice: MEAN-standardization (first proposed by Anderberg 1973). As its name indicates,

this standardization involves division of the values of a numerical vector *v* by their mean μ_v, as in Expression 3.

$$v_{stdMEAN} = \frac{v}{\mu_v} \qquad 3$$

Table 6 shows the application of MEAN-standardization to the column vectors of the unstandardized matrix of Table 5a.

Table 6: MEAN-standardization of the matrix in Table 5a

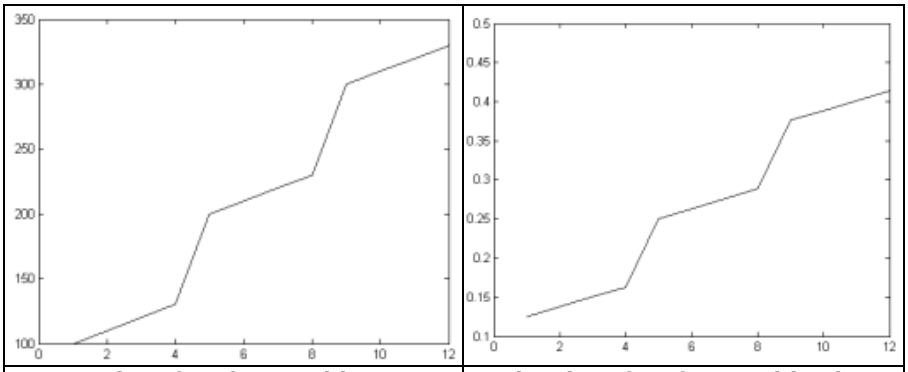
	v1	**v2**	**v3**
1	100	20	1000.10
2	110	21	1000.11
3	120	22	1000.08
4	130	40	1000.01
5	200	41	1000.20
6	210	42	1000.07
7	220	60	1000.23
8	230	61	1000.30
9	300	62	1000.03
10	310	80	1000.14
11	320	81	1000.13
12	330	82	1000.68
Standard deviation	82.411	22.376	0.171
Coefficient of variation	0.383	0.439	0.00018

a: Unstandardized

	v1	**v2**	**v3**
1	0.46512	0.39216	0.99993
2	0.51163	0.41176	0.99994
3	0.55814	0.43137	0.99991
4	0.60465	0.78431	0.99984
5	0.93023	0.80392	1
6	0.97674	0.82353	0.9999
7	1.0233	1.1765	1.0001
8	1.0698	1.1961	1.0001
9	1.3953	1.2157	0.99986
10	1.4419	1.5686	0.99997
11	1.4884	1.5882	0.99996
12	1.5349	1.6078	1.0005
Standard deviation	0.383	0.439	0.00018
Coefficient of variation	0.383	0.439	0.00018

b: MEAN-standardized

Note that MEAN-standardization has preserved the coefficients of variation of the unstandardized variables. This is because division by a scalar, here the column vector mean, is a linear operation that alters the scale while preserving the shape of the original value distribution, as shown for the general case in Figure 1 and for *v1* in the present case in Figure 2.



a: Plot of v1 from Table 6a | **b: Plot of v1 from Table 6b**

Figure 2: Preservation of distribution shape by linear transformation

Note also that the standard deviations of *v1-v3* in Table 6b are identical to the corresponding coefficients of variation. This is because, for any vector *v*, it is always the case that its coefficient of variation is identical to the standard deviation of the MEAN-standardized version of *v*, as shown in Table 7.

Table 7: Calculation of coefficient of variation and standard deviation of a vector *v*

Coefficient of variation of *v*	$CoeffVar(v) = \frac{1}{\mu_v} StdDev(v)$
Standard deviation of $v_{MEANstd}$	$StdDev(v_{MEANstd}) = StdDev(\frac{1}{\mu_v} v)$

Table 7 shows that for the coefficient of variation of *v* the standard deviation is calculated first and then multiplied by the inverse of the mean, and for the standard deviation of the MEAN-standardized version of *v*, *v* is first divided by its mean and the standard deviation of the result then calculated. But one of the properties of the standard deviation is that, for a vector *v* and a constant *c*, stddev*(cv)* = *c*stddev*(v)*, that is, the two are mathematically equivalent. Since, therefore, (i) the coefficient of variation is a scale-independent measure of variability, and (ii) the standard deviation of a mean-standardized variable is always identical to the coefficient of variation of the unstandardized variable, and (iii) the standard deviation of a variable is what measures its absolute magnitude of variability, MEAN-standardization fulfils the above-stated requirements for a general standardization method: that it eliminate the distorting effect of disparity of variable scale on clustering while preserving the ratios of the intrinsic variabilities of the unstandardized variables in the ratios of the absolute magnitudes of variability of the standardized ones — the absolute magnitudes of variability of MEAN-standardized variables are identical to the intrinsic variabilities of the unstandardized ones, and hence so are the ratios.

Figure 3 compares the cluster trees for the unstandardized matrix in Table 5a, the Z-standardized version in Table 5b, and the MEAN-standardized version in Table 6b.

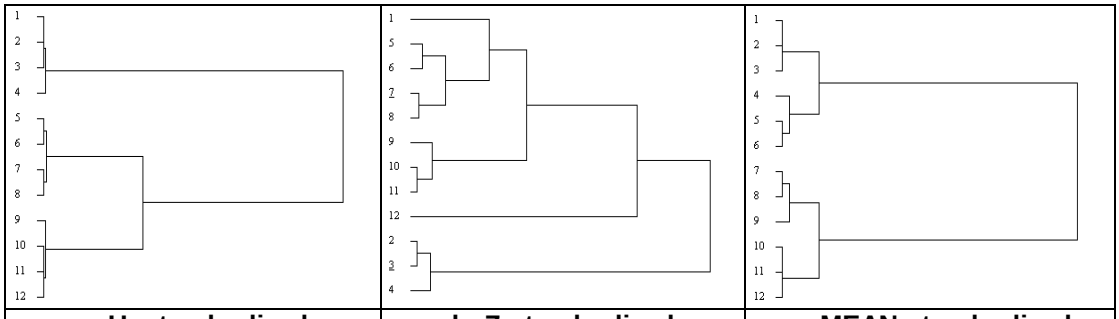


Figure 3: Cluster analyses of unstandardized, Z-standardized, and MEAN-standardized versions of the matrix in Table 5a

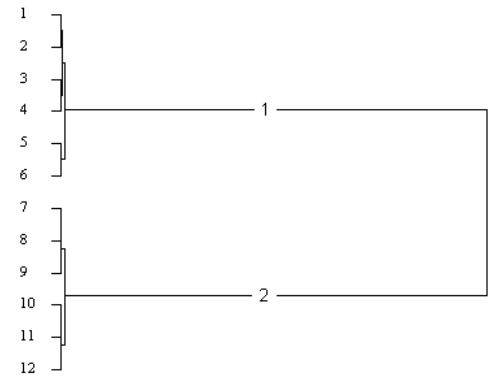
Direct inspection of the unstandardized matrix in Table 5a reveals three value-groups for *v1*, four groups for *v2*, and small random variations on a constant for *v3*. The primary clustering in Figure 3a is by *v1* because it has the highest absolute magnitude of variability and subclustering within the three primary clusters is by *v2*, with the effect of *v3* invisible, all as expected. The cluster tree for the Z-standardized matrix is much more complex, and any sense of the groups observable in *v1* and *v2* in the unstandardized matrix is lost as the clustering algorithm takes account of the numerically much-enhanced random variation in *v3* generated by Z-standardization; the tree in Figure 3b bears no obvious relationship to any reasonable intuition about structure in the unstandardized matrix. The tree corresponding to the MEAN-standardized version of the matrix, however, captures these intuitions very well: there are four primary clusters corresponding to the four numerical groups in *v2* of Table 6b, which has the highest intrinsic variability and therefore represents the characteristic that most strongly distinguishes objects 1-12 from one another in the real world; the effect of *v2*, the variable with the next-highest intrinsic variability, is seen in the internal structures of the primary clusters, so that, for example, the flat subtree for objects 1-3 corresponds to very similar row vectors in the unstandardized matrix, the segregation of 4 from 5 and 6 in the subtree corresponds to the anomalously-low value of 130 in row 4 of the unstandardized matrix, and similarly for the remaining two groups 7-9 and 10-12; the influence of *v3*, with its very low intrinsic variability, is invisible.

Returning now to the matrix of Table 1a with which the discussion began, how do MEAN and Z-standardization of it compare with respect to cluster analysis? An answer requires a basis for comparison. The basis used here is how well analyses of the respective standardized matrices capture the intuition gained from direct examination of the unstandardized data. The values in Table 1a were selected so that a specific speaker structure could readily be observed by inspection, and one consequently knows what to expect from cluster analysis: younger, less-well-off speakers use the phonetic feature of interest less frequently than older, better-off ones. Though the matrix is much smaller than the data one would want to analyze in actual sociolinguistic research or in corpus-based linguistic research more generally, the variable values it contains reflect those one might expect in data abstracted from the real world, and as such cluster analysis of standardized versions of it gives a good idea of how effective these standardizations would be in practice. Table 8 juxtaposes the cluster trees for the unstandardized, Z-standardized, and MEAN-standardized versions of the matrix in Table 1a.

Table 8: Cluster analyses of (a) unstandardized, (b) Z-standardized, and (c) MEAN-standardized versions of the matrix in Table 1a

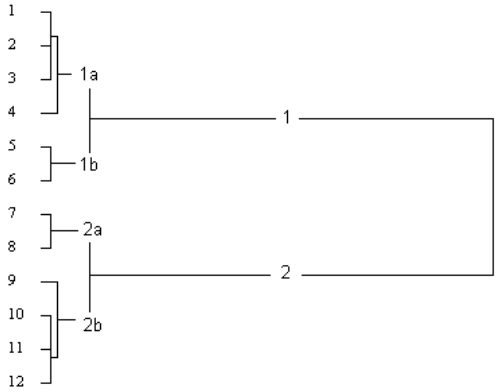
a

	Frequency	Age (years)	Income (Euros)
1	100	20	30000
2	105	21	30250
3	110	22	30500
4	115	30	30750
5	200	31	31000
6	205	32	31250
7	210	40	35000
8	215	41	35250
9	300	42	35500
10	305	50	35750
11	310	51	36000
12	315	52	36250
Standard deviation	81.84	11.21	2536.20
Coefficient of variation	0.394	0.311	0.077



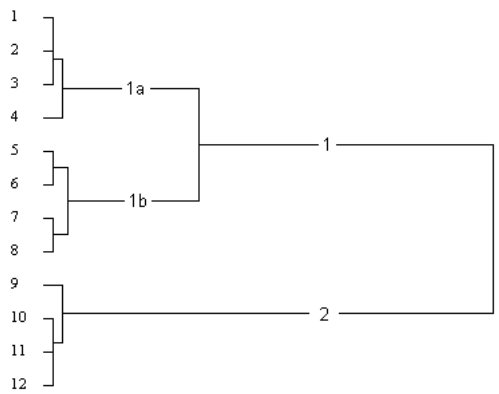
b

	Frequency	Age (years)	Income (Euros)
1	-1.3135	-1.4273	-1.2322
2	-1.2524	-1.3381	-1.1336
3	-1.1913	-1.2489	-1.035
4	-1.1302	-0.5352	-0.9364
5	-0.0916	-0.4460	-0.8379
6	-0.0305	-0.3568	-0.7393
7	0.0305	0.35682	0.7393
8	0.0916	0.44603	0.8379
9	1.1302	0.53523	0.9364
10	1.1913	1.2489	1.0350
11	1.2524	1.3381	1.1336
12	1.3135	1.4273	1.2322
Standard deviation	1	1	1
Coefficient of variation	-	-	-



c

	Frequency	Age (years)	Income (Euros)
1	0.4819	0.5556	0.9057
2	0.5060	0.5833	0.9132
3	0.5301	0.6111	0.9208
4	0.5542	0.8333	0.9283
5	0.9638	0.8611	0.9359
6	0.9879	0.8889	0.9434
7	1.0120	1.1111	1.0566
8	1.0361	1.1389	1.0642
9	1.4458	1.1667	1.0717
10	1.4699	1.3889	1.0792
11	1.4940	1.4167	1.0868
12	1.5181	1.4444	1.0943
Standard deviation	0.394	0.311	0.077
Coefficient of variation	0.394	0.311	0.077



The cluster tree based on the unstandardized matrix accords poorly with intuition. It says that the speakers fall into two main clusters, and that, within these clusters, there is very little variation among them.

The tree based on the Z-standardized data accords well with intuition. It says (i) that the speakers fall into two main clusters, where cluster 1 consists of younger, less well paid ones who use the phonetic feature in question relatively infrequently and cluster 2 of older, better paid ones who use it more frequently, (ii) that the speakers in cluster 1 are subclustered such that 1a consists of the youngest, least well paid ones who use the segment least frequently and 1b of those who are somewhat older, somewhat better paid, and somewhat more frequent users, and (iii) that the speakers in cluster 2 subcluster in a way analogous to those of 1.

The tree based on the MEAN-standardized data also accords well with intuition. Like the preceding one, it divides the speakers into two main clusters where, as with the Z-standardized analysis, cluster 1 consists of younger, less well paid speakers who use the phonetic feature in question relatively infrequently and cluster 2 of older, better paid speakers who use it more frequently, but it allocates them differently so that cluster 2 contains only the oldest, best paid, most frequent users, and cluster 1 the remainder. Cluster 1, moreover, contains two quite strongly defined and differently-structured subclusters: 1a consists of the youngest, least well paid speakers, and 2 of somewhat older, somewhat better paid, somewhat more frequent ones.

The first of these analyses is an artefact of variable scale selection, as we have seen, and can be dismissed. Given that the other two both give intuitively plausible results, however, which one should be preferred? The answer is implicit in what was said earlier about the relationship between a variable and the aspect of reality it describes — essentially that, if it is to be a

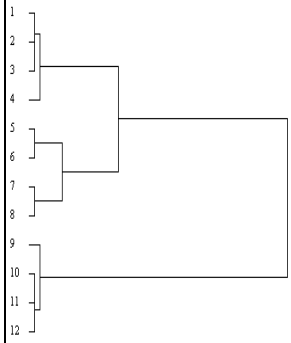
good representation of reality, its variability must reflect variation in the real world. Z-standardization imposes a uniform absolute magnitude of variability on all variables, whereas MEAN-standardization preserves their intrinsic variability ratios in those of their absolute magnitudes of variability. In the present case, Z-standardization has diminished the absolute magnitude of variability of the *Frequency* and *Age* variables and enhanced it for the *Income* variable, and, because it was identical across all three, cluster analysis treated all the variables equally. MEAN-standardization has on the other hand retained the three variables' intrinsic variabilities in their absolute magnitudes of variability, and the analysis consequently clustered the speakers first by the variable with the largest magnitude, *Frequency*, and then by the variable with the second largest magnitude, *Age*; the effect of *Income,* with its relatively low variability, is invisible. In the present case, therefore, it seems clear that the analysis based on MEAN-standardization should be preferred, and, given the foregoing discussion, this generalizes.

It is implicit in the foregoing discussion that, the closer the intrinsic variabilities of the variables in a data matrix are to uniformity, the more similar the clustering results based on Z-standardization and MEAN-standardization will be. This is exemplified in Table 9. The values of the *Income* variable in Table 1a were changed so that its intrinsic variability is closer to that of *Frequency* and *Age*. The matrix was then Z- and MEAN-standardized, and the corresponding nearly-identical cluster trees are shown below.

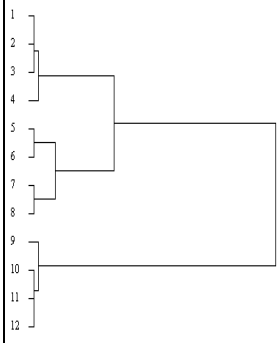
Table 9: Cluster analyses of matrix in Table 1a with *Income* given greater intrinsic variability

	Frequency	**Age (years)**	**Income (Euros)**
1	100	20	30000
2	105	21	32000
3	110	22	34000
4	115	30	36000
5	200	31	38000
6	205	32	40000
7	210	40	60000
8	215	41	62000
9	300	42	64000
10	305	50	66000
11	310	51	68000
12	315	52	70000
Standard deviation	81.84	11.21	15383.97
Coefficient of variation	0.394	0.311	0.308

Emended Table 1a matrix



Clustering based on Z-standardized matrix



Clustering based on MEAN-standardized matrix

Z-standardization is inadvisable only when there is more or less substantial variation in intrinsic variability among variables, in which case MEAN-standardization should be used. But, since MEAN-standardization is a more generally-applicable method, it's difficult to see why one would want to bother with Z-standardization at all, assuming the research application requires preservation of intrinsic variability.

2.3 *Comparison of mean standardization to other methods*

Most general statistics, data processing, and cluster analysis textbooks say something about standardization. Z-standardization is always mentioned and, when different methods are cited or proposed, there is typically little discussion of the relative merits of the alternatives, though, as noted earlier, quite a few express reservations about Z-standardization for the same reason as in Section 2.1 above. The relatively few studies that are devoted specifically to the issue (Milligan and Cooper 1988; Gnanandesikan et al. 1995; Chu et al. 2009) are empirical, that is, they assess various methods' effectiveness in allowing clustering algorithms to recover clusters known *a priori* to exist in specific data sets. Their conclusions are inconsistent with one another and with the results of the present study. Milligan and Cooper compared eight methods and concluded that standardization using the range of variables works best; Gnanandesikan et al. proposed and favoured one that uses estimates of within-cluster and between-cluster variability, though also noted that "much more research is needed before one attempts to cull out the best approaches"; Chu et al. concluded that, for the data they used, "there is no consistent performance benefit that is likely to be obtained from the use of any particular standardization method". This section takes a principled rather than an empirical approach to comparison of various standardization methods. Its criterion is the degree to which the methods preserve the pre-standardization intrinsic variabilities of variables in post-standardization absolute magnitudes of variability. To avoid prolonging the discussion overly, only a selection of methods cited in the above three studies is considered together with one, cosine normalization, that is extensively used in the Information Retrieval literature (for example Singhal et al. 1996). These are listed in Table 10.

Table 10: Standardization methods

SUM standardization divides the values of v by their sum, which rescales the values to the interval 0..1 such that their sum in standardized v is 1.	$v_{std} = \frac{v}{\sum_{i=1..m} v_i}$
COSINE standardization divides the values of v by the norm or length of v, as a result of which the standardized vector is always of length 1.	$v_{std} = \frac{v}{\lvert v \rvert}$
MAX standardization divides the values of v by the largest of the values in v, which rescales all the values to the interval 0..1.	$v_{std} = \frac{v}{\max(v)}$

RANGE standardization divides the values of *v* by their range, that is, by the difference between the maximum and minimum values in *v*.	$v_{std} = \frac{v}{(\max(v) - \min(v))}$

All the methods in Table 10 involve division of a vector by a scalar, that is, a linear transformation, and as such all preserve the shape of the unstandardized distribution together with its intrinsic variabilitiy, just as MEAN-standardization does. The degree to which they preserve the ratios of the intrinsic variabilities of the unstandardized variables in the ratios of their standardized absolute magnitudes of variability differs from method to method, however, and is determined by the nature of the divisor in any particular case. Table 11 compares the ratios of the intrinsic variabilities of the unstandardized variables in Table 1a to the ratios of the absolute magnitudes of variability of those variables standardized by the methods in Table 10. The methods are arranged in descending order of closeness to the benchmark MEAN-standardization which, as already noted, preserves the ratios perfectly.

Table 11: Ratios of pre- and post-standardization variabilities of the Table 1a matrix

	v1/v2	**v1/v3**	**v2/v3**
Ratios of intrinsic variabilities of unstandardized variables in Table 1a	1.27	5.15	4.07
Ratios of absolute magnitudes of variation of standardized variables			
MEAN (benchmark)	1.27	5.15	4.07
SUM	1.27	5.15	4.07
COSINE	1.23	4.81	3.89
MAX	1.21	3.71	3.08
RANGE	1.09	0.94	0.86

The ratios for SUM are identical to those for MEAN. This is because the column vectors of a given matrix M SUM-standardized and MEAN-standardized are just linear variants of one another via division of the SUM version by a constant *n* or multiplication of the MEAN version by *n*, where *n* is the number of rows in M. That SUM and MEAN standardizations are equivalent self-evidently applies not only in the present case but in general.

The ratios for COSINE-standardization are close but not identical to those for MEAN and SUM. This is a consequence of the nonlinearity that COSINE introduces into standardization. To see where this nonlinearity comes from and its effect, it is first necessary to understand the general relationship between vector variability and vector length. This relationship is examined with reference to the list of vectors in Table 12a. These vectors are two-dimensional for graphical display, but the discussion based on them extends straightforwardly to higher dimensions. They all contain values in the range 0..10, though the choice of range is arbitrary and could have been anything, and all have different value distributions within that. The value distributions were selected such that they all sum to the range maximum 10, and go from one distribution extreme [0 10] to the other [10 0] in constant 0.5 increments;

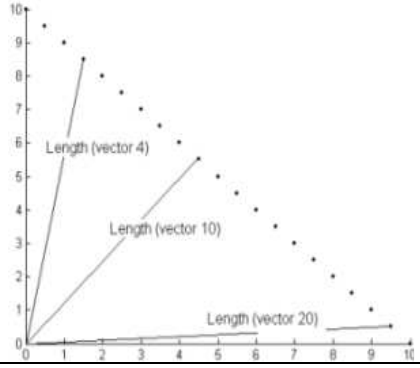
these are simplifications for clarity of exposition, and do not affect the generality of the discussion to follow.

Table 12: The relationship between vector variability and vector length

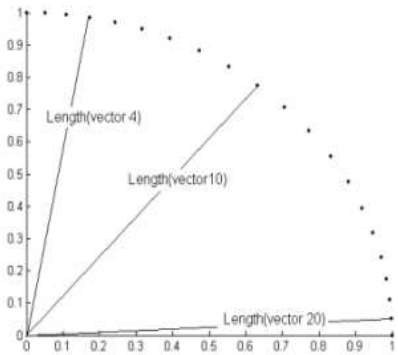
	Vectors	Std dev	Coeff var	Length
1	[0 10]	5	1	10
2	[0.5 9.5]	4.5	0.9	9.513
3	[1 9]	4	0.8	9.055
4	[1.5 8.5]	3.5	0.7	8.631
5	[2 8]	3	0.6	8.246
6	[2.5 7.5]	2.5	0.5	7.906
7	[3 7]	2	0.4	7.106
8	[3.5 6.5]	1.5	0.3	7.211
9	[4 6]	1	0.2	7.382
10	[4.5 5.5]	0.5	0.1	7.616
11	[5 5]	0	0	7.071
12	[5.5 4.5]	0.5	0.1	7.106
13	[6 4]	1	0.2	7.211
14	[6.5 3.5]	1.5	0.3	7.382
15	[7 3]	2	0.4	7.616
16	[7.5 2.5]	2.5	0.5	7.906
17	[8 2]	3	0.6	8.246
18	[8.5 1.5]	3.5	0.7	8.631
19	[9 1]	4	0.8	9.055
20	[9.5 0.5]	4.5	0.9	9.513
21	[10 0]	5	1	10

	Vectors	Std dev	Coeff var	Length
1	[0 1]	0.5	1	1
2	[0.053 0.999]	0.473	0.9	1
3	[0.110 0.994]	0.442	0.8	1
4	[0.174 0.985]	0.406	0.7	1
5	[0.243 0.970]	0.364	0.6	1
6	[0.316 0.949]	0.316	0.5	1
7	[0.394 0.919]	0.263	0.4	1
8	[0.474 0.880]	0.203	0.3	1
9	[0.555 0.832]	0.139	0.2	1
10	[0.633 0.774]	0.070	0.1	1
11	[0.707 0.707]	0	0	1
12	[0.774 0.633]	0.070	0.1	1
13	[0.832 0.555]	0.139	0.2	1
14	[0.880 0.474]	0.203	0.3	1
15	[0.919 0.394]	0.263	0.4	1
16	[0.949 0.316]	0.316	0.5	1
17	[0.970 0.243]	0.364	0.6	1
18	[0.985 0.174]	0.406	0.7	1
19	[0.994 0.110]	0.442	0.8	1
20	[0.999 0.053]	0.473	0.9	1
21	[1 0]	0.500	1	1

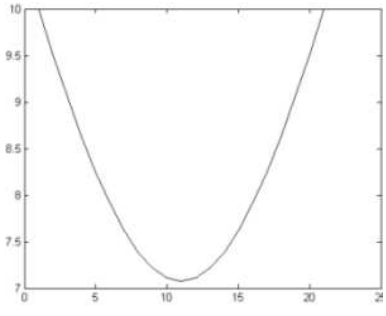
a: Pre-standardization vectors



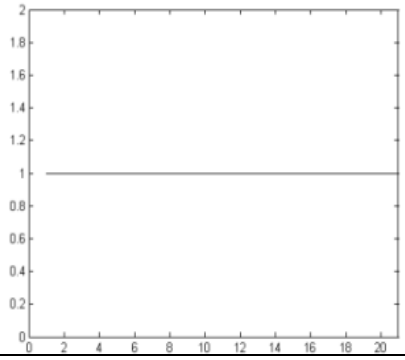
e: Post-standardization vectors



b: Locations of pre-standardization vectors in 2-dimensional space



f: Locations of post-standardization vectors in 2-dimensional space



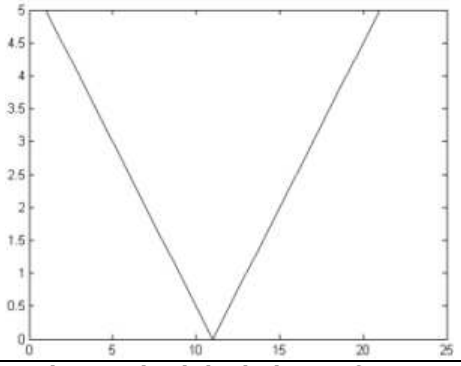
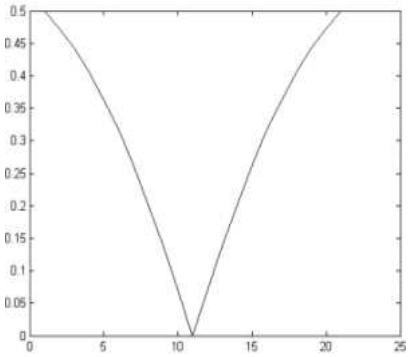
c: Lengths of pre-standardization vectors	g: Lengths of post-standardization vectors
	
d: Standard deviations of pre-standardization vectors	**h: Standard deviations of post-standardization vectors**

Table 12b shows the locations of the vectors 1-21 in 12a in two-dimensional space together with lines representing the lengths of arbitrarily selected ones. The plot shows that the vectors vary in length and that there is a systematic relationship between the lengths and the distribution of vector values. This is confirmed by reference to the *length* column in 12a: vector length is at its maximum at the distribution extremes, that is, at [0 10] and [10 0], and at its minimum for the uniform vector [5 5]. Table 12c is a plot of these vector lengths, and it shows that the relationship between value distribution and length is nonlinear, with most of the nonlinearity at and near the minimum length. Reference to the standard deviations of the vectors in 12a and the corresponding plot in 12d allows this to be restated in terms of variability: a vector's length is nonlinearly related to its standard deviation, that is, to its absolute magnitude of variability.

The implications of the nonlinear relationship of vector length to absolute magnitude of variability are that, when COSINE-standardization is applied to a collection of vectors, any differences in the variabilities of the vectors will be nonlinearly reflected in their lengths, and, because the lengths are the divisors in the COSINE formula, that the post-standardization absolute magnitudes of variability will be nonlinearly related to the pre-standardization ones. This is exemplified in 12e-12h. Table 12e shows the vectors of 12a COSINE-standardized together with the corresponding standard deviations, coefficients of variation, and lengths. COSINE achieves a uniform vector length of 1, shown in the *length* column of 12e and in 12g, by transforming the values in each vector such that it lies on a curve of constant radius 1 in two-dimensional space, as in 12f. A consequence of this transformation is that the absolute magnitudes of variability of the standardized vectors are nonlinearly related to the absolute magnitudes of variability of the unstandardized ones, as can be seen by comparing 12d and 12h.

Recall, moreover, that COSINE standardization preserves intrinsic variability, as shown in the *coeff var* columns of 12a and 12e, and further note that the intrinsic variabilities are linear, having the same shape as the standard deviations in 12d. Since the absolute magnitudes of variability of the standardized vectors are nonlinear and the intrinsic variabilities are linear, the post-standardization absolute magnitudes of variability are nonlinearly related to the intrinsic variabilities, and it follows that the ratios of the variables'

intrinsic variabilities are not preserved in the ratios of the absolute magnitudes of variability; this failure to preserve the ratios can be demonstrated by calculating and comparing a small sample of them from 12e, shown in Table 13.

Table 13: Ratios of intrinsic variabilities and post-standardization absolute magnitudes of variability for a selection of vectors in figure 12e

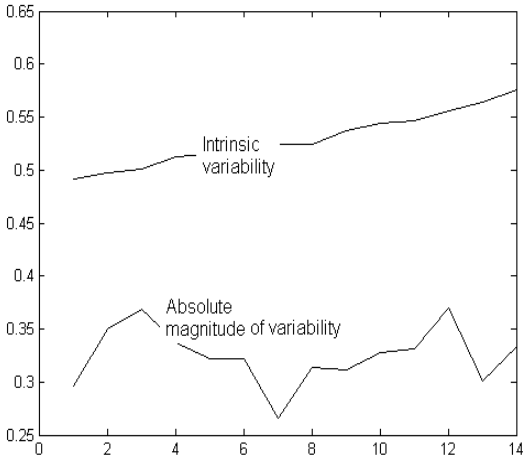
	v1/v2	**v1/v3**	**v1/v4**	**v1/v5**
Intrinsic variabilities	1 / 0.9 = 1.111	1 / 0.8 = 1.250	1 / 0.7 = 1.429	1 / 0.6 = 1.667
Post-std absolute magnitudes of variability	0.5 / 0.473 = 1.057	0.5 / 0.442 = 1.132	0.5 / 0.406 = 1.233	0.5 / 0.364 = 1.374

The conclusion must therefore be that, with respect to the collection of vectors in 12a, COSINE is inferior to MEAN and SUM on account of the nonlinearity it introduces into the standardization procedure. And, since nonlinearity is a general property of the relationship between vector variability and length, this conclusion can be extended to any collection of vectors with different variabilities such as those of the matrix in Table 1a.

The ratios for MAX in Table 11 are reasonably close but not identical to those for MEAN and SUM. The reason for this is easier to see than for COSINE: it is unpredictability in the distribution of maxima across the column vectors of the matrix being standardized. In general, for real-valued data having any given level of intrinsic variability, there is an arbitrary number of different vectors which have that variability, and the maxima across those vectors can differ; because the maximum is the divisor in the MAX formula, this unpredictability affects the post-standardization absolute magnitudes of variability and may cause those magnitudes to vary across the column vectors in a way that distorts the intrinsic variability ratios to greater or lesser degrees. To exemplify this, a collection of vectors with very similar intrinsic variabilities but with substantial variation in maxima was generated and MAX standardized. The unstandardized vectors are shown in Table 14a, sorted in ascending order of intrinsic variability, together with the corresponding intrinsic variabilities, maxima, and post-standardization absolute magnitudes of variability.

Table 14: MAX-standardization of vectors with similar intrinsic variabilities and substantially different maxima

		Vector	**Intrinsic variability**	**Max**	**Post-std absolute magnitude of variability**
a	1	[0.51 0.54 0.50 0.44 0.01]	0.4912	0.981	0.2962
	2	[0.07 0.93 0.54 0.95 0.84]	0.4968	0.951	0.3498
	3	[0.98 0.27 0.26 0.58 0.85]	0.5009	0.546	0.3683
	4	[0.31 0.02 0.44 0.46 0.28]	0.5124	0.468	0.3373
	5	[0.96 0.91 0.26 0.65 0.22]	0.5145	0.967	0.3220
	6	[0.93 0.32 0.92 0.21 0.49]	0.5203	0.930	0.3218
	7	[0.43 0.40 0.17 0.39 0.91]	0.5237	0.913	0.2660
	8	[0.09 0.73 0.97 0.71 0.40]	0.5238	0.974	0.3134
	9	[0.30 0.25 0.02 0.21 0.42]	0.5374	0.421	0.3112
	10	[0.46 0.04 0.46 0.74 0.85]	0.5446	0.857	0.3276

	11	[0.65 0.64 0.32 0.11 0.25]	0.5469	0.658	0.3316
	12	[0.65 0.32 0.62 0.70 0.02]	0.5561	0.703	0.3694
	13	[0.13 0.38 0.96 0.71 0.37]	0.5645	0.967	0.3007
	14	[0.93 0.64 0.02 0.70 0.40]	0.5753	0.931	0.3342
b					

From the co-plots of intrinsic variability and post-standardization absolute magnitude of variability in 14b, it is readily seen that there is no systematic relationship between the two: the intrinsic variability plot increases slowly and fairly smoothly whereas the other jumps around without any obvious pattern, reflecting the unpredictability of the maxima in the unstandardized vectors. The effect of this on the preservation of intrinsic variability ratios for a sample of vector pairs in 14a is shown in Table 15.

Table 15: Ratios of intrinsic variabilities and post-standardization absolute magnitudes of variability for a selection of vectors in figure 14a

	v1/v2	**v1/v3**	**v1/v4**	**v1/v7**
Intrinsic variabilities	0.989	0.981	0.959	0.938
Post-standardization absolute magnitudes of variability	0.847	0.804	0.878	1.114

The unpredictability of maxima means that MAX standardization cannot be relied on to preserve the intrinsic variability ratios in the post-standardization absolute magnitudes of variability consistently, and hence MAX must, like COSINE, be regarded as inferior to MEAN and SUM.

The ratios for RANGE differ greatly from those of the benchmark, and are in fact closer to the uniformity characteristic of Z-standardization. The reason for this is straightforward: like standard deviation, the divisor in Z-standardization, range is a measure of dispersion, though a cruder one, and therefore gives a similar result. RANGE is therefore subject to the same criticism as Z, and can be rejected as a viable data standardization method for cluster analysis.

Conclusion

This discussion has addressed a generally recognized problem in cluster analysis: that, where the data variables are measured on different scales, variables whose scales permit relatively large values tend to have an effect on clustering which is disproportionately large relative to their importance as

descriptors of the domain of interest, and the effect of those whose scales permit only relatively small values disproportionately small relative to their importance, making cluster results contingent on the vagaries of scaling choice. A widely used solution to that problem, Z-standardization, was found to be inadvisable at least under some circumstances because it imposes a uniform variability on all the data variables whatever the differences in their intrinsic variabilities, thereby distorting their validity as representations of the domain of interest and consequently the reliability of cluster analysis based on them. An alternative standardization method based on the variable mean was proposed and compared to several other methods, and, together with SUM-standardization, was found to be most effective in preserving intrinsic variability. In applications where preservation of the intrinsic variabilities of data variables is felt to be important for reliable cluster analysis, therefore, MEAN / SUM-standardization should be used. Corpus-based sociolinguistics provided one example of such an application in the foregoing discussion, though it is not difficult to think of others both within that subdiscipline and in corpus-based linguistics more generally.

References

Anderberg, Michael. 1973. *Cluster Analysis for Applications*. London: Academic Press.

Boslaugh, Sarah & Watters, Paul. 2008. *Statistics in a Nutshell*. Cambridge MA: O'Reilly.

Chu, Chia-Wei, Holliday, John & Willett, Peter. 2009. Effect of data standardization on chemical clustering and similarity searching. *Journal of Chemical Information and Modeling* 49.155-161.

Everitt, Brian, Landau, Sabine & Leese, Morven. 2001. *Cluster Analysis*, 4th edn. London: Arnold.

Gnanadesikan, R. 1997. *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd edn. New York: Wiley-Interscience.

Gnanandesikan, R., Tsao, S., Kettenring, John. 1995. Weighting and selection of variables for cluster analysis. *Journal of Classification* 12. 113-136.

Hair, Joseph, Black, William, Babin, Barry & Anderson, Ralph. 2006. *Multivariate data analysis*, 6th edn. Upper Saddle River NJ: Pearson Prentice Hall.

Kaufman, Leonard & Rousseeuw, Peter. 1990*. Finding Groups in Data. An Introduction to Cluster Analysis*. New York: Wiley-Interscience.

Kettenring, John. 2006. The practice of cluster analysis. *Journal of Classification* 23. 3-30.

Milligan, Glenn & Cooper, Martha. 1988. A study of standardization of variables in cluster analysis. *Journal of Classification* 5. 181-204.

Romesburg, H. 1984. *Cluster Analysis for Researchers*. Florence KY: Wadsworth.

Singhal, A., Salton, G., Mitra, M., Buckley, C. 1996. Document length normalization. *Information Processing and Management* 32. 619-633.

Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin. 2006. *Introduction to Data Mining*. Upper Saddle River NJ: Pearson Addison Welsley.