Working Paper No. 13-09

Orthogonalization of Categorical Data:

Orthogonalization of Categorical Data: How to Fix a Measurement ro lem in Statistical Distance Metrics

Ross nippenberg[y]

November 10, 2013

Poli y m kers de end on e onomists, st tisti i ns, nd other so i l s ientists to m ke - ur te o serv tions nd dr w solid on lusions from qu ntit tive n lysis. E onometri s, for ex m le, h s ome long w y in the st entury nd guides m ny de isions m de tod y. On the other h nd, some st tisti l ro edures h ve not h d signifi nt dv n es, ut re inste d lied nd their origin l ssum tions re forgotten. The ro ri teness of m ny of these me surements h s ome into question, nd while riti ism is often e ted, little is done to orre t them. In re lity, there is rolifi me surement ro lem eing ommitted everyd y. This ro lem involves the use of st tisti l dist n e metri s to me sure so i l henomen . For ex m le, me surements whi h would routinely e used to nswer questions like: y how mu h h ve the im orts of the United St tes h nged in the st ye r? By how mu h h s r i l diversity h nged in the st de de? Does gre ter ethno-linguisti diversity le d to ivil onfli t? These nd simil r questions rely on ur te multi-v ri te dist n e metri s. However ll dis-

"M ny multiv ri te st tistic l methods c n be reg rded s techniques for investig ting s mple sp ce in which e ch s mple member is represented by point." John C. Go er (19 7), pg 13.

"Me surement is big p rt of mobilizing for imp ct. You set go l, nd then you use d t to m ke sure you're m king progress tow rd it. This is cruci l in business- nd it's just s import nt in the ght g inst poverty nd dise se" Bill H. Gates (2013), pg 52.

1 Introduction: The Problem

Before introducing an formal mathematics, consider the five follo ing measurement puzzles:

***Puzzle 1*:** Consider a t o-countr orld here Countr C exports half corn and half corn meal. Countr D exports half corn and half computers. Which one has the most diverse exports? Measures of export diversification indicate that both countries are exactl equall diverse.

***Puzzle 2*:** In Cit A exactl 5 percent of the labor force are Economics Professors. In Cit B, exactl 5 percent of the labor force are Research Economists. The Location Quotient doesn't recognize cross-discipline similarities, so bet een the t o cities, Cit A is classified as being relativel sparse in Research Economists and Cit B is classified as being relativel dense in Research Economists.

***Puzzle 3*:**

still maintaining the original structure of the data. That procedure is the subject of this paper and is detailed in the Methodolog section.

Ho big is this problem and the corresponding bias? That depends on the data, but a rough estimate is given b finding the average value of the data's similarit matrix, here ${}_{i;j}$ is the $(i;j)$ of a similarit matrix :

$$bias = \frac{\sum_{i=1}\sum_{j=1} {}_{i;j} \quad n}{n^2 \quad n} \qquad (1)$$

Using 4-digit SITC international trade data for the ear 2000, this number is $\frac{481801{:}6 \quad 772}{772^2 \quad 772} = 0{:}808$. In other ords, the average export product x_i is, on average 80.8 percent like product x_j. Ho ever, all current distance metrics, and hence all standard trade metrics, implicitl assume that similarit is zero bet een all categories. This is clearl not true, and ithout zero similarit bet een categories, the standard multi-dimensional metrics are not valid.

The question naturall arises: ho ide-spread is the problem? Well it exists in ever branch of procedure and so ma be unable to use this procedure until a similarit -calculating procedure is found. This ould be an ideal subject for future research.

To previe the proposed orthogonalization procedure, one can see it as a change of coordinate s stems. I take as given a set of data vectors and the measure of similarit bet een ever pair of its dimensions. The basic idea is that the similarit bet een dimensions can increase hich reduces distance bet een individual dimensions in a vector. This is best seen in the spherical coordinate s stem (see Spiegel 1959, Munkres 1991). The measure of the angle from the vector to an axis is given b   or . the orthogonalization procedure then uses the change of coordinates to find the length of this vector along each axis. The rectangular coordinate s stem is hat most empirical measures are based upon, at least those ith concepts like angle and distance. So in order for a quantitiative measure to be valid, e must change the coordinate s stem to hat the measure is assuming. This is the basic idea of the paper.

2 Literature

The aforementioned problem of heterogeneit bet een dimensions is, as far this author can tell, completel unackno ledged hen orking ith shares data. That said, the problem is recognized hen orking ith quantitative variables hich are not in the form of shares, and has been the focus of substantial research. I can identif 9 distinct orthogonalization procedures each of hich are based on t o basic methods, of hich there are undoubtedl more. The first method, found over helming in Statisitics and Econometrics, involves the use of a correlation or covariance matrix to find orthogonal dimensions. The second method, found in Mathematics and applied in Computer Science and Ph sics, involves kno ing exactl ho the s stem behaves in a non-stochastic fashion and having perfect measurements.

T o ideas distinguish m problem and solution from the rest of the literature. The first is that the data hich I am examing al a s exists on a unit simplex. Thus the range of possible values that variables ma take is relativel limited, and no negative values are allo ed. This eliminates the use of correlation and covariance matrices since these procedures commonl produce negative values. Secondl , m over-arching argument rests on the idea that the true coordinates of these observations are not kno n, but detailed information exists in the form of similarit values hich can be used to find the true location.

idea behind principal coordinates is that it takes a distance measure bet een all pairs of *observ tions* and gives them coordinates; according to Go er (19 7, pg.19), “We can ask ho the coordinates of points ith the given distances be found.”. On the other hand, the orthogonalization method that I discuss in this paper adjusts for a similarit measure bet een *v ri bles*. The idea of this adjustment bet een variables is that, after ard, similarit bet een observations (or other measures) can be measured, based on the adjusted variables. Consider that principle coordinates anal sis takes a matrix of similarities bet een observations as given, and then adjusts the variables to fit those similarities. In contrast, the orthogonalization procedure described herein is quite the opposite in that it seeks to create a similarit matrix bet een observations based on the given similarit bet een the variables.

Third, and also similar to principal components anal sis, is factor anal sis. In factor anal sis a researcher attempts to identif unobserved, latent categorical variables. In this case the covariance bet een dimensions leads to recognition of a previousl unidentified latent variable. So this statisexports.

Fifth, regression anal sis and anal sis of variance, are chiefl concerned ith accounting for covariance. Each variable is treated as a dimension, and the covariance bet een dimensions can greatl affect the estimate of the mean value of a regressor on the regressand. In an abstract a this is similar because the practitioner realizes that variables are not completel independent of one another, and so the covariance, or angle bet een dimensions, is included in the process b design. ot including important covariates leads to omitted variable bias: the magnitude of a parameter is inaccurate. This is geometricall equivalent to a parameter value being projected onto an n-plane but not parallel to its coordinate axis, ith the angle bet een its proposed axis and the actual projection proportional to the correlation ith the omitted variable. This is exactl the argument that I am making for distance measures.

2.2 Exac Me h ds f Or h g naliza i n

The second class of orthogonalizatin procedures is based on mathematical procedures for rotation, have no stochastic assumption, and the underl ing data generating process has no latent variables. The ever-present implicit assumption that I am tr ing to upend here is that n-space coordinates are al a s kno n. For this reason the Gram-Schmidt process, the Householder Transformation, and the Givens Rotation can all be ruled out as potential orthogonalization techniques because the all make this assumption. I am not going to detail each method, because none of them can ork due to this assumption. Again, each assumes that the coordinates of a vector are kno n, hereas I onl assume partial information about the coordinates is kno n.

2. O her Li era ure

Sixth, this paper has ties to Measure Theor . A main point of measure theor concerns distinguishing a measurement of an attribute from the attribute itself. Consider common commodities like heat, corn, and computers. Ho different are these things? As economists, e don't particularl care about heat, corn, or computers in themselves, but rather about the implied underl ing proable to measure is not necessaril the same as hat e need to measure to form general theoretical statements.

Seventh, this paper closel relates to Index umber Theor , ho ever, this paper has nothing direct to sa about prices. In Index umber Theor , one can t picall identif t o distinct approaches: the Axiomatic Approach versus the Economic Approach. What is the point to having these t o different approaches? The point is that the data does not line up exactl ith theor be-

3 ethodology

The problem, stated in et another a , is that the categories in hich much data is classified is ad hoc, ith some categories more alike than others. To fix this problem, one first needs a measure of similarit bet een all dimensions, to hich I ill defer to other papers. For example in International Trade see Hidalgo, et al (2007) or for a more general treatment see Dauxois and kiet (2002). Second, accoding to Gentle (2007), this similarit data is best vie ed as representing the angle bet een dimensions. With this in mind, the orthogonalization procedure is then to take each data share $x_{c;i}$ and project it onto an orthogonal coordinate s stem, Euclidean n-space. Then one can appl an number of distance metrics. This projection is best vie ed as a change from h perspherical [3] to rectangular coordinates for each individual dimension.

.1 imilari y Ma rices and Angle Be ween Dimensi ns

of vectors: “The cosine of the angle bet een t o vectors is related to the correlation bet een the vectors, so a matrix of the cosine of the angle bet een the columns of a given matrix ould also be bet een the z-axis and the x- plane, in radians. Let be the angle bet een the x-axis and the z- plane. Then given the values for spherical coordinates $(r; \; ; \;)$, the corresponding rectangular coordinates $(x_1; y_1; z_1)$ can be found b :

$$x = r \sin \quad \cos \tag{3}$$

$$y = r \sin \quad \sin \tag{4}$$

$$z = r \cos \tag{5}$$

The above equations are a projection of a vector in spherical coordinates into the rectangular coordinate s stem. These should be familiar to the reader and are t picall first encountered in multvariate Calculus.

. The Or h g naliza i n Pr cedure: Change f C rdina es

The most promising method to obtain an orthogonal coordinate s stem is to use a change from h perspherical to rectangular coordinates. I use the algorithm described in Lin (1995) [4]. The basic idea here is to treat each dimension of a vector as its o n vector. Then, because the angle of each dimension is kno n in regards to ever other dimension, and using a trigonometric-based algorithm, one can project the length of the vector onto each dimension, repeat for each entr in the vector, and sum them up at the end.

Define a vector of shares data b x hich has n ro s indexed b i. The associated n b n similarit matrix is , ith elements $_{i;j}$ here ro s are indexed b i and columns indexed b j. Redefine in terms of degrees and convert it from a similarit matrix to a distance matrix:

$$d_{i;j} = (1 \quad _{i;j})90 \qquad (\;)$$

For ever i, define each entr in the vector x_i as a radius. Define each column entr j in ro i of matrix as the angle formed b the vector i to dimension j. To convert to rectangular coordinates,

[4] I thank Professor Jeanne uflot for roviding me with an equivalent algorithm.

align the numeraire good as the first good. This represents a simple rotation of the coordinate s stem [5].

$$xq_{1;i} = x_1 \cos(\ _{1;1}) \tag{7}$$

o , similarl , find the projection of the second good onto each axis. Do this for each of the n goods using the follo ing algorithm.

$$
\begin{aligned}
xq_{2;i} &= x_1 \cos \ _{1;i} \\
xq_{3;i} &= x_1 \sin \ _{1;i} \cos \ _{2;i} \\
xq_{4;i} &= x_1 \sin \ _{1;i} \sin \ _{2;i} \cos \ _{3;i} \\
& \qquad\qquad (8) \\
x_n \backslash_{2;i} &= r \sin \ _{1;i} \sin \ _{2;i} \sin \ _{3;i} \quad \sin \ _{n \ 3;i} \cos \ _{n \ 2;i} \\
x_n \backslash_{1;i} &= r \sin \ _{1;i} \sin \ _{2;i} \sin \ _{3;i} \quad \sin \ _{n \ 2;i} \cos \ _{n \ 1;i} \\
xq_{n;i} &= r \sin \ _{1;i} \sin \ _{2;i} \sin \ _{3;i} \quad \sin \ _{n \ 1;i} \sin \ _{n;i}
\end{aligned}
$$

The previous algorithm is adapted from Lin (1995).[6] ote that the pattern of the h perspherical algorithm is such that the vast majorit of terms are sine and each line ends ith cosine except for the ver last line hich ends in sine.

Repeat the above producedure for all i and then define for all j:

$$xb_j =$$

And finall , because this is shares data and exists on a unit simplex, the sum of the entries must add to 1. Define total unadjusted shares (TUS)as:

$$TUS = \sum_{j=1} x_j \tag{11}$$

And normalize each entr using TUS:

$$x_i^{ca} = \frac{x_i}{TUS} \tag{12}$$

The above equations outline the orthogonalization procedure for a single data vector. Likel a researcher ould be comparing man different data vectors and ould need to complete this procedure for each vector. This is the end of the orthogonalization procedure.

.4 Dis ance Me rics

The follo ing is an introduction to a subset of common distance metrics used in man different statistical and social science fields. Man more distance metrics exist, and as ith the literature revie , this list is b no means exhaustive. In various fields these distance metrics go b specific

Where onl the positive root is used. When $p = 1$, the distance metric is kno n as either Manhattan, or Cit -Block Distance:

$$D_{Manhattan} = \sum_{i=1}^{n} |x_{c;i} - x_{d;i}| \tag{14}$$

Cit -block distance gets its name from the fact that to get from one point to another in a cit grid one must follo the streets. Particularl in Manhattan, streets intersect at right angles, so the absolute value in the differe nManha1ach(Manhastreet)-384(dimensio0(Manhais)-anhathe)-384(total)-384(area)-3

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{n} (x_{c;i} - x_{d;i})^2} \tag{15}$$

When p approaches , one Cheb shev's Distance:

$$D_{Chebyshev} = \max_{i=1}^{n} |x_{c;i} - x_{d;i}| \tag{1 }$$

A second class of distance estimators scales the coordinate values. Canberra Distance t pifies this class:

$$D_{Canberra} = \sum_{i=1}^{n} \frac{|x_{c;i} - x_{d;i}|}{|x_{c;i}| + |x_{d;i}|} \tag{17}$$

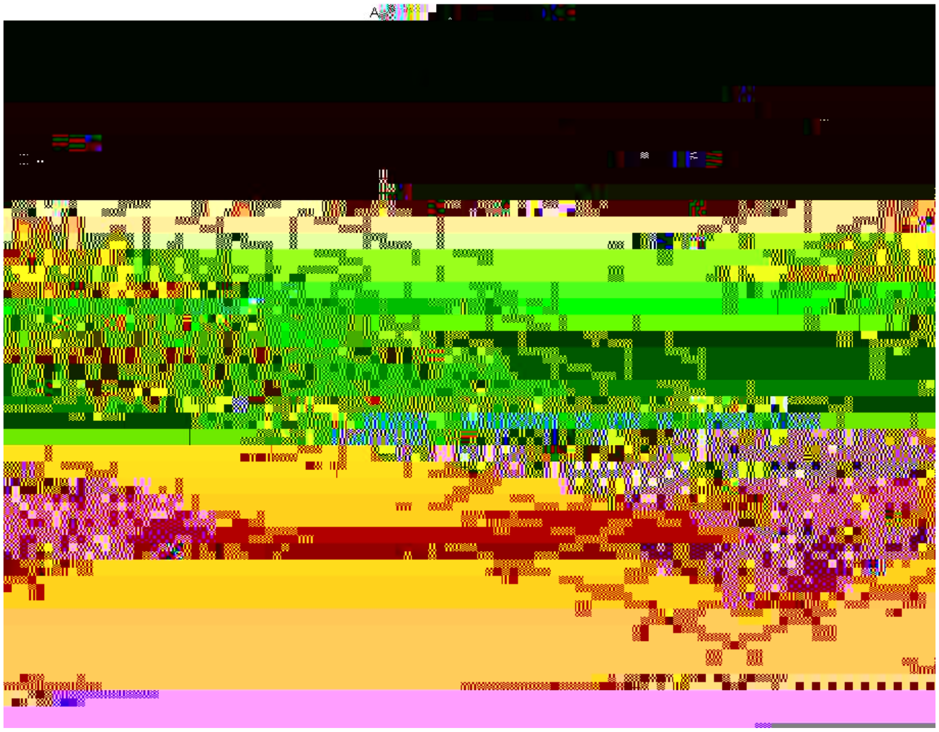
A third class of distance include the Czekano ski Coefficient, hich goes b a m riad

4 Simulation

M above qualitative argument for the need for an orthogonalization procedure is hopefull persuasive. Ho ever, I find it useful to present a ver general example using a series of simple simulations. I ill consider a three-dimensional orld here a single observation x_i is composed of k share attributes, here the sum of k attributes is one. Using a random number generator, I ill assign values to the k attributes as ell as to the $\frac{k^2}{2}$ k similarit bet een attributes. This is equivalent to finding a random point in a random k-space. I ill then calculate the Euclidean distance to the origin first ignoring the similarities, and then compare this to the Euclidean distance using the orthogonalization procedure. I repeat this for var ing values of n and k, ith the results displa ed in Figures 2 and 1. Here the number of observations are $n = 1; 2; :::120$ [7] hich are plotted along the x-axis, and the number of dimensions is $k = 2; 3; :::; 160$, [8] plotted along the -axis. The z-axis (vertical) represents the measured distance on k dimensions bet een a point and the origin, averaged for n observations. Figure 2 ignores the similarit bet een dimensions and computes Euclidean distance in the normal a . Compare these average values to those in Figure 1 hich do take into account the similarit bet een dimensions and thus compute the true average distances. Figure 3 plots the simple difference bet een the t o surfaces.

B definition, the distances using the La of Cosines are correct, it is the Euclidean distances,

Figure 1: Actual Distance sing the Law of Cosines



pletel invariant to the number of observations, and are completel dependent on the number of dimensions used. So these measures depend more on the number of dimensions used rather than the actual values in the shares data.

Applications

The follo ing section outlines a fe examples in the literature here the orthgonalization method can potentiall ield great benefit. I plan to academicall pursue these topics in the near future. I have drafted or am orking on proposals on all of the follo ing topics.

5.1 Applica i n: Price Indices

The computation of index numbers suffers from three primar challenges. The first is that the data is in the form of categories, hich naturall do not obe the la s of arithmetic. The second is that the eights of categories change over time. These first t o challenges are commonl referred to as the “Index umber Problem.” The third is that classification and categorical ambiguit

Figure 2: Estimate sing Euclidean Distance

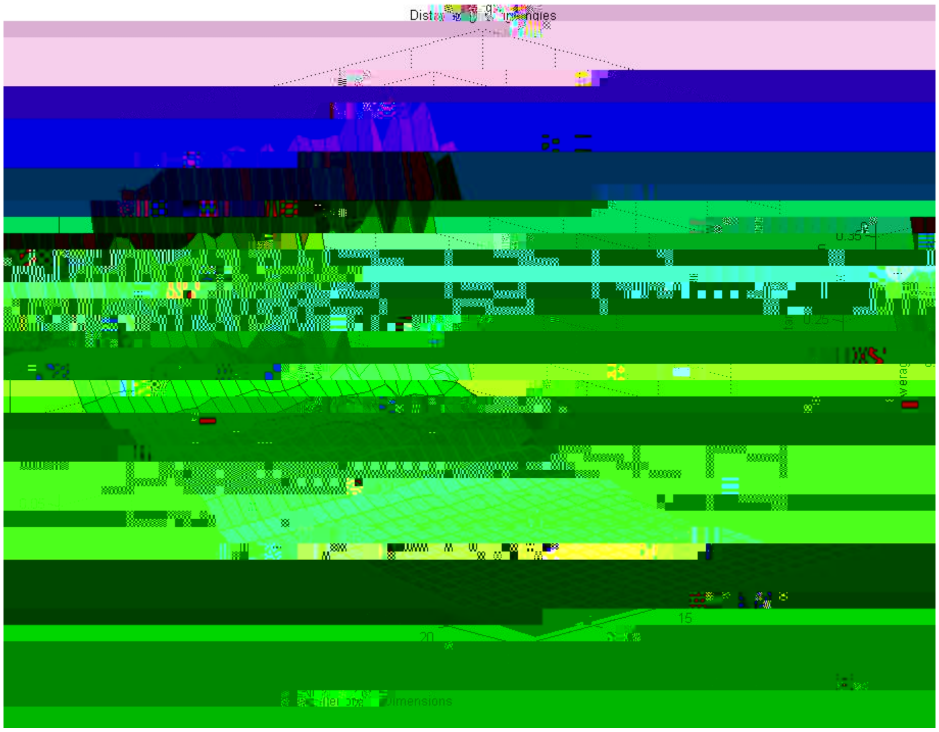
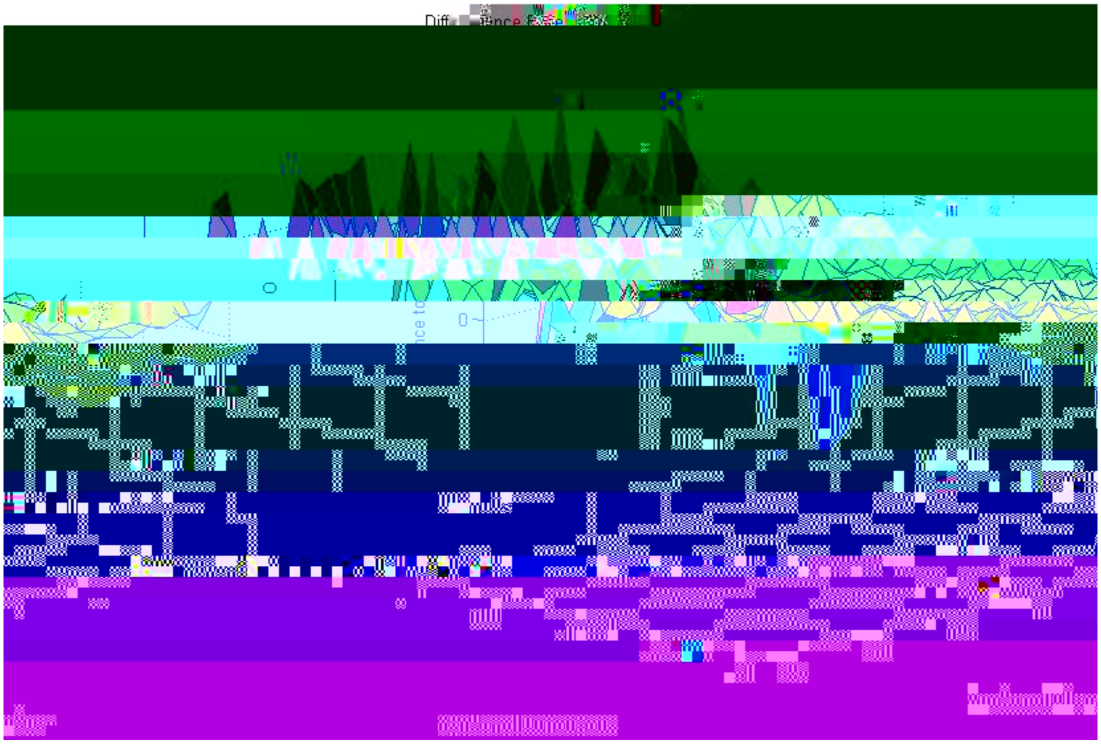


Figure 3: Difference Between Actual and Estimated



egorizing people as red, ello , bro n, black, and hite (Funderburg 2013, 83). The simplicit of this categorization has, quite understandabl , been contentiousl opposed. The offense is likel not so much in the names used, as it is in the broadness of each categor . When being given a label, most individuals ould likel ant to be recognized as closel as possible to the categor in hich the self-identif . To this end, a researcher ma feel compelled to divide the categories into smaller subcategories. The onl problem is that the index monotonicall increases ith the number of categories. While it's possible that this measure is correct at an aggregation level, the point is that it's not clear hich level of aggregation is appropriate. In particular, think about multi-racial people. In computing the IQV, most researchers treat a bi- or multi-racial person as being in a completel different categor . Ho ever, a multi-racial person is reall , b definition, a combination

$$IQV = \frac{1 \quad \text{P} \; (p_k)^2}{n \quad 1} \tag{22}$$

Where p_k is the share of group k in the total population. This is a normalized version of Euclidean distance from the origin. So ith cross-categor observations, the bi- or multi-racial observations can be partiall grouped into categories, changing the computation in a a that is not immediatel clear.

5. Applica i n: Devel pmen and P li ical Ins i u i ns: The Index f E hn - linguisi ic Frac i naliza i n

Incredibl similar to the IQV is a measure kno n as the Index of Ethno-linguistic Fractionalization (ELF), hich is applied extensivel in the literature in the fields of political science and economic development. The equation is given b :

$$ELF = 1 \quad \sum_{i=1}^{k} p_k^2 \tag{23}$$

Where $k \quad 2$ and p_k^2 is the share of ehtnic group k in the total population. This is a version of non-normalized Euclidean distance from the origin.

The basic idea behind the Index is, just like the IQV, to measureiversit . The problem, as clearl defined b Laitin and Posner (2001) is t o-fold. First, a researcher needs to be careful about the level of aggregation used in defining ethnic groups. Second, not all ethnic groups are equall unalike. To this end, Bossert, D'Ambrosio and Ferrara (2005) define a Herfindahl Index that the coin the Generalized Index of Ethno-linguistic Fractionalization Index hich accounts for similarties bet een categories. Indeed the almost define the La of Cosines distance metric in Knippenberg (2013), but stop short of taking a geometric interpretation of distance metrics. So I am pleased that this problem of non-zero similarit bet een categories has been recognized before in this literature, but has not had an orthogonalization procedure applied.

Bill Gates (2013, pg 52).

the same notation as above, denote total exports of countr c as X_c, here $X_c = \sum_{i=1} X_{c;i}$. Define $x_{c;i}$ to be the share of good i in total exports of countr c, here $x_{c;i} = \frac{X_{c;i}}{X_c}$, and, consequentl , $\sum_{i=1} x_{c;i} = 1$. Equivalentl for a second, but ith the subscript d, and for the orld, ith the subscript w. All measurements except the last t o are assumed to be taken in the same time period, so time subscripts are other ise suppressed.

The Hirschman-Herfindahl Index:

$$HHI_{c;d} = \sqrt{\frac{1}{\left(\sum_{i=1} x_{c;i}\right)^2}} \tag{24}$$

The export similarit Index, Finger and Kreinin (1979):

$$FK_{c;d} = \sum_{i=1} \min(x_{c;i}; x_{d;i}) \tag{25}$$

The Grubel-Llo d Index, Grubel and Llo d (1971):

$$GL_{c;d} = 1 \quad \frac{\sum_{i=1} |X_{c;i} \quad X_{d;i}|}{\sum_{i=1} (X_{c;i} \quad X_{d;i})} \tag{2 }$$

T o definitions are common for the Export Diversification Index. The first follo s directl from Finger and Kreinin (1979):

$$DX1_c = \sum_{i=1} \min(x_{c;i}; x_{w;i}) \tag{27}$$

Where the subscript w stands for " orld". The more common definition is exactl the same as the Hirschman Index:

$$DX2_c = \sqrt{\frac{1}{\left(\sum_{i=1} x_{c;i}\right)^2}} \tag{28}$$

The Trade Compatibilit Index, Michael (199)[11]:

The Export Specialization Index:

$$ES_c = \frac{x_{c;i}}{m_{d;i}} \tag{30}$$

Changes in Global Demand for Major Exports:

$$CGD_c = \sum_{i=1}^{n} S_{i;0}\,(X_{i;t} - X_{i;0}) \tag{31}$$

Changes in Global Market Share for Major Exports:

$$CGMS_c = (S_{i;t} - S_{i;0})\,M_{g;t} \tag{32}$$

And lastly, the Thiel Index of export concentration:

$$T_c = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} x_i} \ln \frac{x_i}{\sum_{i=1}^{n} x_i} \tag{33}$$

As the reader can see, each trade statistic treats each export (or import) product as a separate dimension, and there is no system of weights or compensation for dimensions being more or less alike.

These trade statistics can be classified in several ways. The Hirschman Index is of the absolute type: they describe a country's export shares as some distance from the origin. All of the others are of the relative type. The export diversification (Finger and Kreinin) tells the manhattan distance between a country's export shares and the world export shares. The rest give the distance between two country's export shares: the Grubel-Lloyd gives the distance exactly in terms of Canberra distance, and the rest of the trade statistics are of the relative type: they tell the distance between two non-origin points. The export similarity and export diversification measures (both based on the work of Finger and Kreinin) are nearly identical to the Czekanowski Coefficient, except that they are already in terms of shares, whereas the Czekanowski Coefficient converts to shares after summing the values.

.7. Simple Example: x International Trade

Consider a two-country world with two goods: guns and butter. The first country, denoted by c, produces 20 percent butter and 80 percent guns, while the second country, denoted by d, produces

70 percent butter and 30 percent guns:

$$y_c = \begin{bmatrix} y_{c;b} \\ y_{c;g} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}; \quad y_d = \begin{bmatrix} y_{d;b} \\ y_{d;g} \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}.$$

Then suppose the production of guns and butter share some common attributes. For example, both need land: butter producers more so to raise dair co s and but guns producers also need land for placing factories. Both also need metal: butter producers need metal for producing churns and vats, and but gun producers need metal relativel more to produce stocks and barrels. B some external measurement process e kno the the similarit bet een guns and butter to be 0.8, or 80 percent of the inputs are alike. Then the similarit matrix, denoted b ith individual elements

$_{b;b}$, $_{b;g}$, $_{g;b}$, and $_{g;}$ d[]TJF6

Figure 4: nadjusted Shares

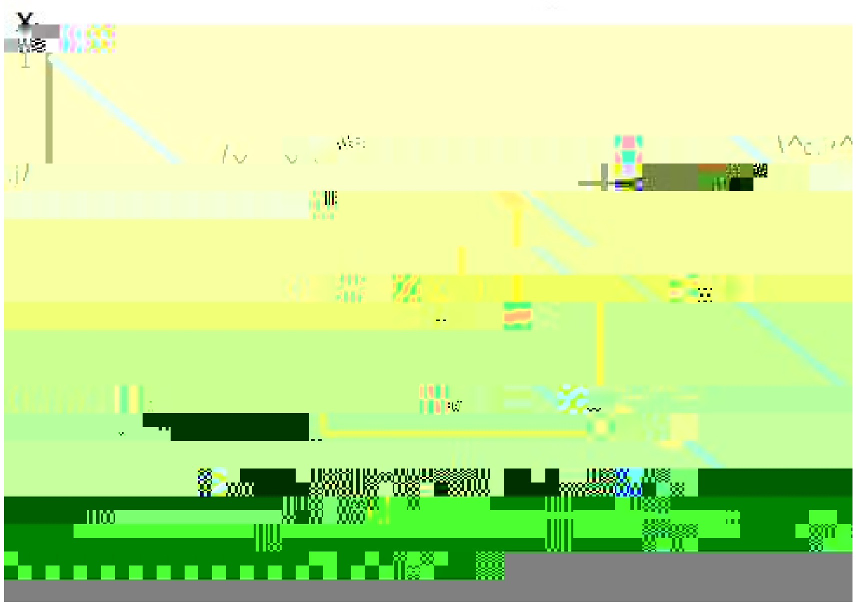
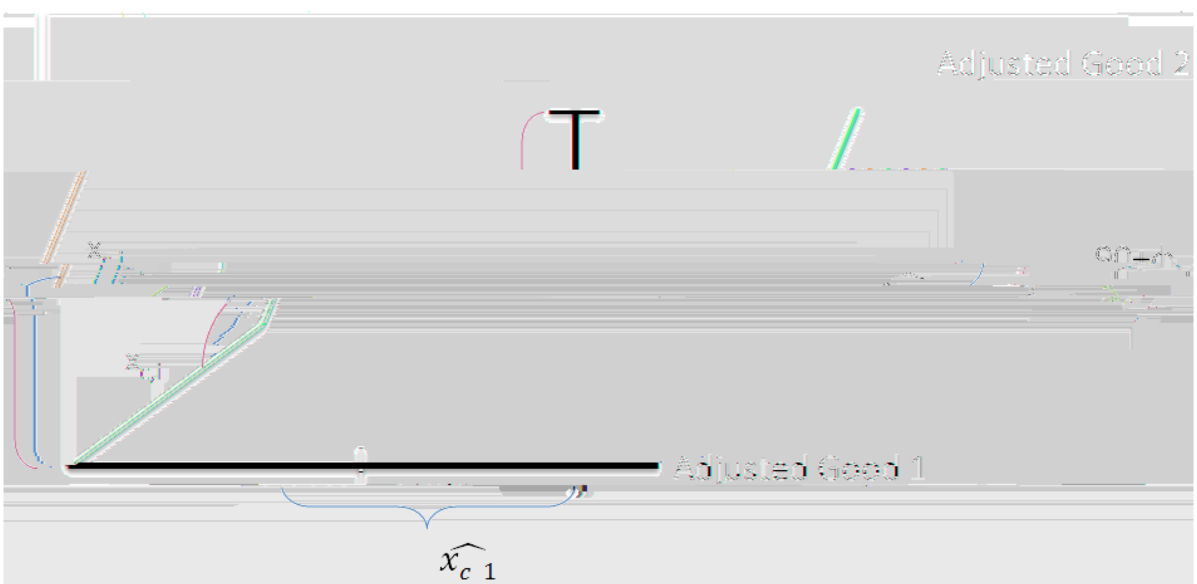


Figure 5: Projection onto Principal Axes



$\hat{y}_{c,2} = y_{c,2,1} + y_{c,2,2} = 0 + 0.247 = 0.247$

Lastly, because this is shares data, the the sum of the shares must equal 1:

$\hat{y}_{c,1} + \hat{y}_{c,2} = 0.961 + 0.247 = 1.208$

And then:

$\hat{y}_{c,1} = \frac{0.961}{1.208} = 0.796$

$\hat{y}_{c,2} = \frac{0.247}{1.208} = 0.204$

Equivalently for country d: Project the $y_{d,b}$ vector onto the $y_{d,1}$ and $y_{d,2}$ axes:

$y_{d,[1,1} = y_{d,b} \cos(0^\circ) = 0.7(1) = 0.7$

$y_{d,[2,1} = y_{d,b} \sin(0^\circ) = 0.7(0) = 0$

Similarly, projecting the $y_{d,g}$ vector onto the $y_{d,1}$ and $y_{d,2}$ vector space yields:

$y_{d,[1,2} = y_{d,g} \cos(18^\circ) = 0.3(0.951) = 0.285$

$y_{d,[2,2} = y_{d,g} \sin(18^\circ) = 0.3(0.309) = 0.093$

And the last step for country c is to add together the results of the two projections:

$\hat{y}_{d,1} = y_{d,[1,1} + y_{d,[1,2} = 0.7 + 0.285 = 0.985$

$\hat{y}_{d,2} = y_{d,[2,1} + y_{d,[2,2} = 0 + 0.093 = 0.093$

Lastly, because this is shares data, the the sum of the shares must equal 1:

$\hat{y}_{d,1} + \hat{y}_{d,2} = 0.985 + 0.093 = 1.078$

And then:

$\hat{y}_{d,1} = \frac{0.985}{1.078} = 0.$

And notice that for the unadjusted vectors, the normalized Euclidean distance ould have been:

$$dist^{a}_{c,d} = \frac{1}{\sqrt{2}}\sqrt{(0.2 - 0.7)^2 + (0.8 - 0.3)^2} = \frac{0.707}{\sqrt{2}} = 0.5, \tag{35}$$

As an aside, the La of Cosines distance metric in Knippenberg (2012) finds the Euclidean distance bet een the unadjusted vectors hich is equivalent to the distance bet een the adjusted but non-normalized vectors, see Appendix B for the proof.

The reader can hopefull see that hen similarit is zero, then $\cos(0) = 1$, allo ing the orthogonalization process to return the original vectors of guns and butter. o , to take the anal sis a step further, assume that the export share vector of each countr is in exactl the same proportion as their production vectors:

$$x_c = y_c = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}; \quad x_d = y_d = \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix}.$$

To abstract from an confounding effects, assume that each countr has equal economic output, that these are the onl t o countries in the orld, and that each exports goods equal to 1 normalized unit of value. Abstracting a a from an theor on h the countries are trading or on their quantities of that trade, the empirical international trade literature suggests a number of measures.

Using the original, unadjusted trade vectors, the composition of bilateral trade is given b $GL^{un}_{c,d} = 0.5$. Contras1nal trade ls4d[(:)(ectorsis)-s1nal trade

1945, 19 4):

$$H_c = \sum_{i=1} \left(\frac{x_{c;i}}{X_c} \right)^2 \qquad (3\ )$$

Where $x_{c;i}/X_c$ is the share of good i in the export bundle of countr c. Using the original data, this comes out to be $H_c^{un} = 0.68$ and $H_d^{un} = 0.58$. And using the adjusted data vectors this comes out as: $H_c^{a} = 0.584$ and $H_d^{a} = 0.777$. Again, the economic significance of the differences bet een these t o measures is subjective, but hat is interesting is that the ordering has reversed. Where in the unadjusted index, countr c as the more concentrated countr , in the adjusted index, the more concentrated countr is no d.

.7. The Product Space

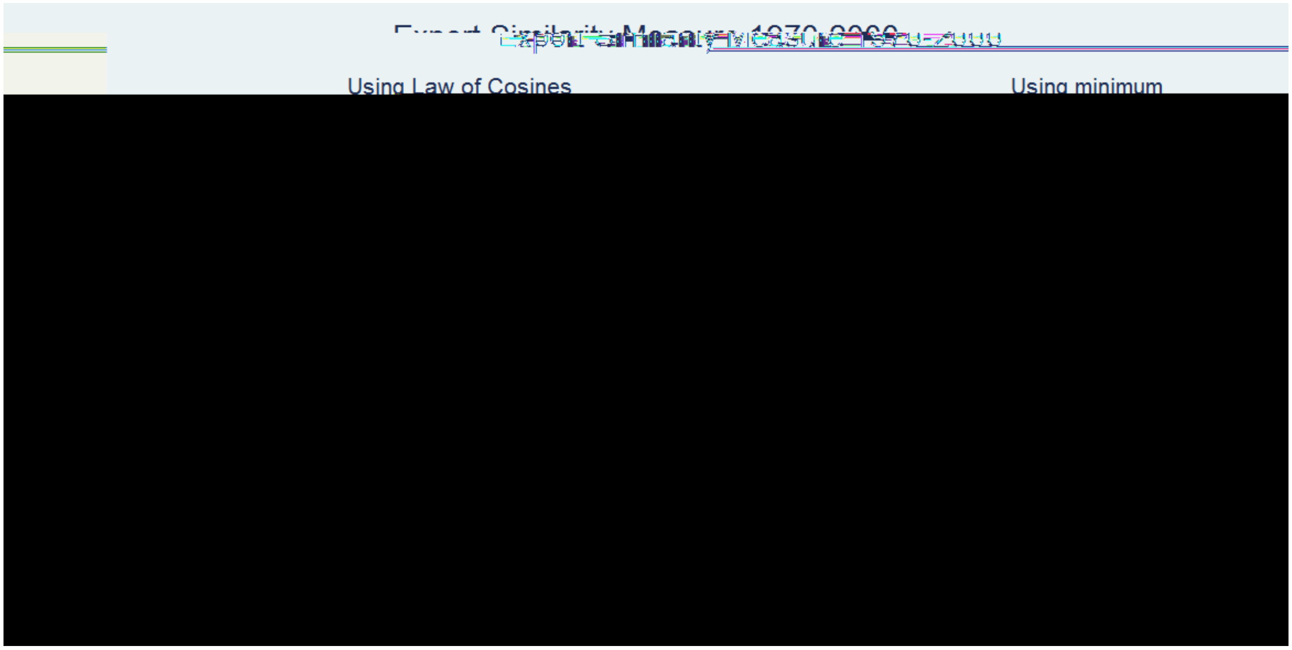
Here I demonstrate the change-of-coordinates orthogonalization procedure in a high-dimensional example: that of the product space of international trade. The product space is an idea conceived and visualized b Hidalgo, et al. (2007), ho use export shares to find a measure of similarit bet een export product categories and then map them using a net ork anal sis approach. I take their anal sis a step further b using the similarit measures to adjust the original countr export vectors, and I sho that the measurements, hile clearl correlated, are ver different. Because of the computational intensit of the orthogonalization producedure [12] , I have onl produced estimates on the export similarit measure. A detailed treatment of the consequences of changing the export similarit measure can be found in Knippenberg (2012), here I insert the ne export similarit measure into a gravit equation of international trade and find ver different results from previous studies.

Export similarit as first conceived b Finger and Kreinin (1979) as a simple measure for comparing export content across either countries or time. I denote this measure as $FK_{c;d}$ and it is defined in equation (22). I use a version of $FK_{c;d}$, hich is derived in Sun and g (2000), and is given in equation (19). The measure has been used in hundreds of academic papers on international trade.

[12] Com uting this variable for 47,653 observations took a roximately four weeks on a deskto com uter with a quad-core 3.3Ghz rocessor.

The steps taken to arrive at these export similarit indices are as follo s. First I do nloaded the export data from Feenstra's ebsite. The data is 4-digit SITC trade data ith 799 categories. I am using 5- ear intervals from 1970 to 2000 for 133 countries. Second, I transform export values into export shares. Third, I follo Hidalgo, et al. (2007) to calculate similarit bet een export categories. Fourth, using this similarit matrix and export shares, I appl the orthogonalization procedure to obtain the adjusted export vectors. Lastl , I appl a Euclidean Distance algorithm

Figure 6: Histograms of Export Similarity easures



to situations in hich similarit bet een variates or correlation is alread accounted for, such as regression anal sis or principal components anal sis. Given the nature of international trade shares data, this orthogonalizaiton procedure is clearl applicable. Furthermore, e Trade Theor models assume an equal marginal rate of substitution bet een varieties of a good. Ho ever, if t o varieties are more similar than either are to an third, then equal marginal rates of substitution cannot mathematicall hold. After appl ing this orthogonalization procedure, the marginal rates of substitution bet een the adjusted goods should be equal because the variables are orthogonal to one another. This ould make the data consistent ith the theor , and is a promising area for future research.

This procedure orks onl hen a bivariate notion of "similarit " or "distance" is computable, as these similarit measures directl feed into the equation. This procedure is not applicable here similarit is not defined or calculable. Finding a a to calculate this similarit in man different contexts is an area for future research here notions of covariance, correlation, ma be ver important. Furthermore, a simple lack of a a to calculate similarit doesn't make the previous distance metrics an more valid - the are still computed using the incorrect coordinate s stem.

Conclusion

I like the follo ing quote from a linear algebra textbook: "Ph sical La s must be independent of an particular coordinate s stem used in describing them mathematicall , if the are to be valid" Spiegel (1959 pg 1). It reminds me that just because ou can measure something doesn't mean that hat ou have measured must necessaril obe the la s of our theor : sometimes a researcher has to manipulate data to make sense of it. In the case of shares data, often it is in the rong coordinate s stem and must be converted to the proper s stem before familiar measures can be applied, like measures of distance in the rectangular coordinate s stem. I have argued throughout this paper that arbitrar classifications are not automaticall defined b the rectangular coordinate s stem. Ho ever the rectangular coordinate s stem is the onl requirement for appl ing familiar statistical distance metrics. In other ords, the principle axes of the coordinate s stem are rarel the same as the axes of the data, so distance metrics cannot be immediatel applied.

Besides the justification of the orthogonalization procedure, the previous paragraphs have also laid out areas for future research. The more mundane of these include re-estimating the effects of unbiased indices on outcomes. For example in trade, this ould include the effect of export similarit or diversification on bilateral trade (Knippenberg 2012), or like ise the effect of the Grubel-Llo d or Herfindahl Indices on various response variables. Theoretical research, on the other hand, holds even more promising avenues. As touched upon earlier, the continuum of goods assumption in International Trade can be re-visited: after normalizing the goods vectors, each adjusted good should have equal marginal rates of substitution, as each represents an orthogonal underl ing good. I have ritten this paper in an attempt to sta as general as possible about its applications: the extensive examples in international trade are merel a consequence of m o n experience. The concepts described herein have ide applicabilit in all areas of empirical research and I look for ard to conducting these applications in the near future.

eferences

[1] Axler, Sheldon Ja . (1997).

[18] Lin, Chii-Dong. (1995) "H perspherical Coordinate Approach to Atomic and Other Coulombic Three-Bod S stems." *Physics Reports* 257: 1-83.

[19] Malane , Pia. (199) "The Index umber Problem: A Differential Geometric Approach." *Ph.D. Thesis*, Harvard Universit , Department of Economics, December 199 .

[20] Marsaglia, George. (1972). "Picking a Point from the Surface of a Sphere". *The Ann ls of M them tic l St tistics* 43(2): 45- 4 .

[21] Michael , Michael. (199). "Trade Preferential Agreements in Latin America: An Ex-Ante Assessment " World Bank Polic Research Paper 1583, March 200 .

[22] Mikic, Mia. (2005). "Commonl Used Trade Indicators: A ote" ART eT Capacit Building Workshop on Trade Research.

[23] Munkres, James. (1991). *An lysis on M nifolds* Westvie Press, 1991, Chapter six.

[24] g, Francis. (2002). "Appendix B: Trade Indicators and Indices" in *Development, Tr de nd the WTO: A H ndbook* Worldbank Bank Publications, Washington, D.C, pg 585-588.

[25] Panda, avneet, Ed ard Y. Chang and Arun Qamra. (200)."H persphere Indexer". Lecture otes in Computer Science, Volume 4080, 200 .

[2] Spiegel, Murra . (1959). *Vector An lysis With n Introduction to Tensor An lysis* McGra -Hill, e York, e York. ISB 07-0 0228-X.

[27] Sun, Guang- hen and Ye K ang g. (2000). "The measurement of structural differences bet een economies: An axiomatic characteization." *Economic Theory* 1 : 313-321.

A Appendi : Proof to Equi alence of Finger-Kreinin and Sun-Ng Distance easures

This section provides a proof that the export similarit measures from Finger and Kreinin (1979) (FK) and Sun and g (2000) (S) are perfectl negativel correlated. Because of the minimum function in FK and the absolute function in S , this proof is not conducive to deduction, but an inductive argument is easier to sho . Define FK and S according to their authors:

$$FK = \sum_{i=1} \min\left(\frac{x_{c;i}}{X_c}; \frac{x_{d;i}}{X_d}\right); \tag{37}$$

and:

$$SN = \sum_{i=1} \frac{|x_{c;i} \quad x_{d;i}|}{2} \tag{38}$$

Pr p si i n:

Let n denote the number of export products. Let c and d be any two countries. Denote export share of good i in country c as $x_{c;i}$, where $i = 1; ::: ; n$. Because $x_{c;i}$ is an export share,

$$\sum_{i=1} x_{c;i} = 1; \tag{39}$$

is satisfied by the definition of a share. The same equation also holds for any other country d. Let the sums FK and SN be defined as above, then the following equality always holds:

$$SN = 1 \quad FK \tag{40}$$

Pr f:

A.1 Case 1.1

Let $n = 2$ and Let $x_{c;1} = x_{d;1}$, then because $x_{c;1} + x_{c;2} = 1$ and $x_{d;1} + x_{d;2} = 1$, it must also be true that $x_{c;2} = x_{d;2}$. In this case,

$$\begin{aligned} FK &= min(x_{c;1}; x_{d;1}) + min(+x_{c;2}; x_{d;2}) \\ &= x_{c;1} + x_{c;2} \\ &= 1 \end{aligned} \tag{41}$$

Similarly,

$$SN = \frac{x_{c;1} \quad x_{d;1}}{2} + \frac{x_{c;2} \quad x_{d;2}}{2} \tag{42}$$

By assumption, $x_{c;1} \quad x_{d;1} = 0$ and since $x_{c;2} = x_{d;2}$, then $x_{c;2}$ 1 STdiak

A.4 Case 2.2

Let $n \geq 2$ and $x_{c;i} > x_{d;i}$ for $i = 1; \ldots; j$. Let $x_{c;i} > x_{d;i}$ for $i = 1; \ldots; k$. Let $x_{c;i} > x_{d;i}$ for $i = 1; \ldots; l$. Where $j + k + l = n$, and $j; k; l \geq 0$. Then by defintion, Equation (37) implies:

$$FK = \sum_{i=1}^{j} x_{c;i} + \sum_{i=1}^{k} x_{d;i} + \sum_{i=1}^{l} x_{c;i}. \tag{56}$$

Or equivalently where the last summation is replaced by $x_{d;i}; i = 1; \ldots; l$. By the shares definition, Equation (39) implies for country c:

$$\sum_{i=1}^{j} x_{c;i} + \sum_{i=1}^{k} x_{c;i} + \sum_{i=1}^{l} x_{c;i} = 1; \tag{57}$$

as well as for country d:

$$\sum_{i=1}^{j} x_{d;i} + \sum_{i=1}^{k} x_{d;i} + \sum_{i=1}^{l} x_{d;i} = 1. \tag{58}$$

And by the definition SN (38):

$$SN = \frac{1}{2}\left[\sum_{i=1}^{j} (x_{c;i} - x_{d;i}) + \sum_{i=1}^{k} (x_{d;i} - x_{c;i}) + \sum_{i=1}^{l} (x_{c;i} - x_{d;i})\right]. \tag{59}$$

Distributing through the summations and rearranging yields:

$$SN = \frac{1}{2}\left[\sum_{i=1}^{j} x_{c;i} - \sum_{i=1}^{k} x_{c;i} + \sum_{i=1}^{l} x_{c;i} - \sum_{i=1}^{j} x_{d;i} + \sum_{i=1}^{k} x_{d;i} - \sum_{i=1}^{l} x_{d;i}\right]. \tag{60}$$

Rearranging (57) implies:

$$\sum_{i=1}^{j} x_{c;i} + \sum_{i=1}^{l} x$$

Simplifying:

$$SN = 1 \quad \left[\sum_{i=1}^{k} x_{c;i} \quad \sum_{i=1}^{j} x_{d;i} \quad \sum_{i=1}^{l} x_{d;i} \right] : \tag{65}$$

Then substituting in the definition of F , Equation (56), yields the desired result:

$$SN = 1 \quad FK: \tag{66}$$

Thus the relationship holds for both $n = 2$ and $n \quad 2$, proving the proposition by induction.

B Appendi : Proof of Equi alence Between Orthogonalization and the n-Dimesional Law of Cosines

Pr p si i n:

In an n-Hilbert space, the norm distance $\| < x_1, x_2 > \|$ with similarity matrix equals $\| < \hat{x}_1, \hat{x}_2 > \|$ with similarity matrix I, the identity matrix.

I will only prove this equivalence for the two-good case. The notation needed to prove

However, this property does not apply to a heterogeneous space such as the product space for two reasons. First of all, as evidenced in Hidalgo et al (2007), the Product Space is extremely heterogeneous: some areas of the product space are dense and others are disparate. In terms of the previous equation, this difference caencein-27(te(prtced)ce)-aises ex(o)-27orthars7h

fom: nor(anoms.)6oy7 orstri(o)-1s(in)-