DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 13-09

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

Ross Knippenberg
University of Colorado at Boulder

November 2013

Department of Economics

University of Colorado at Boulder
Boulder, Colorado 80309

© November 2013 Ross Knippenberg

Orthogonalization of Categorical Data: How to Fix a Measurement Problem in Statistical Distance Metrics [*]

Ross Knippenberg [†]

November 10, 2013

Abstract

Policy makers depend on economists, statisticians, and other social scientists to make accurate observations and draw solid conclusions from quantitative analysis. Econometrics, for example, has come a long way in the past century and guides many decisions made today. On the other hand, some statistical procedures have not had significant advances, but are instead applied and their original assumptions are forgotten. The appropriateness of many of these measurements has come into question, and while criticism is often accepted, little is done to correct them. In reality, there is a prolific measurement problem being committed everyday. This problem involves the use of statistical distance metrics to measure social phenomena. For example, measurements which would routinely be used to answer questions like: by how much have the imports of the United States changed in the past year? By how much has racial diversity changed in the past decade? Does greater ethno-linguistic diversity lead to civil conflict? These and similar questions rely on accurate multi-variate distance metrics. However all distance metrics suffer from a common calculation problem. No one can deny that the math is correct, rather, the problem lies with an overlooked implicit assumption: that all categories are mutually orthogonal (right angles). This is a bold assumption in any context. In this paper I first show that this assumption is rarely valid, and second I suggest an orthogonalization procedure: measure the similarity or angle between categories, and then apply a transformation from spherical to rectangular coordinates. I illustrate the effect of the methodology using a simulation, a collection of potential applications, and two examples from international trade.

Keywords: Index Number Theory, International Trade, Orthogonalization, Principal Coordinates, Law of Cosines, Distance Metrics, Minkowski Metric, Euclidean Distance, Hirschman-Herfindahl Index, business analytics

JEL No. C43, C18, F10

Acknowledgements

I would like to thank my advisor, Murat Iyigun for guidance throughout this project, members of my dissertation committee: Carlos Martins-Filho, Lee Alston, Carol Shiue, Joanne Belknap, as well as Christian Dippel, Keith Maskus, Donald Waldman, Jack Hou, and participants of the 2013 WSSA conference for helpful comments. Also a special thanks to Jeanne Duflot for technical help. Lastly I want to thank to Mario Marazzi and Ken Younge for initially inspiring thought on this subject. All mistakes are my own.

[*] Persistent link to the latest draft: `http://bit.ly/1bqvKT4`

[†] Department of Economics, University of Colorado Boulder. Email: Ross.Knippenberg@Colorado.edu, Website: `https://sites.google.com/site/rossknippenberg/`.

"Many multivariate statistical methods can be regarded as techniques for investigating a sample space in which each sample member is represented by a point." John C. Gower (1967), pg 13.

"Measurement is a big part of mobilizing for impact. You set a goal, and then you use data to make sure you're making progress toward it. This is crucial in business-and it's just as important in the fight against poverty and disease" Bill H. Gates (2013), pg 52.

1 Introduction: The Problem

Before introducing any formal mathematics, consider the five following measurement puzzles:

***Puzzle 1*:** Consider a two-country world where Country C exports half corn and half corn meal. Country D exports half corn and half computers. Which one has the most diverse exports? Measures of export diversification indicate that both countries are exactly equally diverse.

***Puzzle 2*:** In City A exactly 5 percent of the labor force are Economics Professors. In City B, exactly 5 percent of the labor force are Research Economists. The Location Quotient doesn't recognize cross-discipline similarities, so between the two cities, City A is classified as being relatively sparse in Research Economists and City B is classified as being relatively dense in Research Economists.

***Puzzle 3*:** The Census Bureau send out a variety of questionnaires that include questions on race and ethnicity. The index of qualitative variation increases as the number of categories increases, so simply increasing the range of definitions of race and ethnicity lead to an increase in the measurement of diversity in the USA.

***Puzzle 4*:** In Index Number Theory, researchers have proposed hundreds of weighting structures for price indices, but, as far as this author knows, no consensus has been reached on a universally applicable price index, except, perhaps the Divisia Index (Malaney, 1996).

***Puzzle 5*:** Similar to *Puzzle 1*; add Country E to the world, where E exports half corn and half high fructose corn syrup. Are E's exports more like that of C (corn and corn meal) or D (corn and computers)? Measures of export similarity indicate all three countries are exactly equally alike because each exports the same proportion of corn.

Of course I don't agree, and neither should the reader, with the conclusions of these puzzles.

These puzzles are not the result of inadequate data, classifications, qualitative descriptions, unlying technological progression, preferences, supply, or demand. Rather these puzzles are purely the result of mis-applied statistical measurements. In this paper I outline the problem and suggest a geometrically-consistent solution which will solve the measurement problem.

In 1966, John C. Gower lamented to practitioners of multivariate statistical analysis: "The method of principal components analysis is often used, and misused, by statisticians. When unordered qualitative variates occur it is not applicable, except possibly for the special case of (0,1) data" (p.327). Indeed, I take this quote as the guiding motivation for this paper: that data on arbitrary categories must be treated appropriately, and not have statistical methods blindly applied to them.

Let me be clear about the type of data to which I am referring in this paper. I am looking at proportion or shares data: data whose shares sum to unity across categories. Shares data is unique in the way that it combines both a qualitative variable, the categorical label, and a quantitative variable, the ratio of the value of that category to the total value of all categories. For example weights in a price index index may be composed of food, entertainment and housing. Or within food could be the categories of meat, grains, fruits, vegetables and desserts. Similarly, an industrial production index may be composed of heavy manufacturing, light manufacturing, and precision manufacturing, and sub-cateogories thereof. Another example would be the shares of export categories of manufactured goods, primary products, and intermediate products in exports. Some of these categories are broken down by incredibly complex classification systems. For instance the third revision of Standard International Trade Classification system at the 5-digit level comprises over 3200 export product categories. The Harmonized System details trade at the 10-digit level, with many thousands more categories. A final example is to think about the ethnic origins of the population of the United States: you can break it down by race and then break down each of those by ethnicity or nationality. Either way there is a proportion, or share, of data in each category and no matter what the aggregation level, the sum of the shares is always 1.

Thousands of academic papers use measures which simplify vast amounts of categorical data into an index, and I agree with John Gower that these are misapplications. Geometrically, all of these measures are equivalent to plotting points in n-space and then finding the distance between them, using one of several common distance measures. [1] However, these categories of data may be arbitrarily chosen. For example, the Harmonized System (HS) of international trade data is broken down into industrial sectors and subsectors, always in a logical manner, but with no consistent relationships between categories. For another example, consider psychological questionnaire data which attempts to decipher various aspects of a person's personality. The problem is that this data is measured in certain arbitrary ways, methods which may be obvious to the initial observer, but does not follow the constraints of any particular theory to which a researcher may wish to analyze it. When the arbitrary categories are in some way aggregated into an index, each category is treated as if it were a separate dimension. In reality, these measures proxy for an underlying quantity: more specifically, a personality test measures perhaps four different aspects of personality, and no two questions will measure those differences perfectly symmetrically. Some examples of statistical methods used to correct for this arbitrariness of categories, and to find the underlying latent variables, include factor analysis, principal components analysis, principal coordinates analysis, and canonical correlation analysis.

Consider a concrete example using international trade categories. The following are 4-digit SITC Revision 3 categories: 8412 Men's suits and ensembles, 8413 Men's blazers and jackets, and 7832 Road Tractors for Semi-Trailers. Categories 8412 and 8413 are clearly quite similar since blazers and jackets make up one of the components of a suit. On the other hand, both 8412 and 8413 are probably not very similar to 7832: Road Tractors for Semi-Trailers. The problem is that all current distance metrics treat each of these categories (and every other category) as a separate dimension, orthogonal to every other dimension. However, in reality, categories are quite heterogeneous in their similarity to one another, so it would be fallacious to assume mutual orthogonality. A distance metric which accounts for this heterogeneity is described in a previous paper, Knippenberg (2012). However, a more useful procedure would enable a researcher to calculate any distance metric [2] while

[1] I use the terms "distance measure" and "distance metric" interchangeably.

[2] Be it an absolute distance like Hirschman-Herfindahl, or a relative distance from another object like Euclidean,

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

How big is this problem and the corresponding bias? That depends on the data, but a rough estimate is given by finding the average value of the data's similarity matrix, where $\phi_{i,j}$ is the (i, j) of a similarity matrix Φ:

$$bias = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{i,j} - n}{n^2 - n} \tag{1}$$

Using 4-digit SITC international trade data for the year 2000, this number is $\frac{481801.6-772}{772^2-772} = 0.808$. In other words, the average export product x_i is, on average 80.8 percent like product x_j. However, all current distance metrics, and hence all standard trade metrics, implicitly assume that similarity is zero between all categories. This is clearly not true, and without zero similarity between categories, the standard multi-dimensional metrics are not valid.

The question naturally arises: how wide-spread is the problem? Well it exists in every branch of science where each variable is treated as a separate dimension, and a distance metric is computed. This includes Economics, Statistics, Biology, Psychology, Sociology, Political Science, and Computer Science, to name a few. In Statistics the problem is most prevalent in Multi-dimensional scaling, Cluster Analysis, Correspondence Analysis, and Procrustes Analysis where distances between multi-dimensional objects is the chief concern. In this paper I will provide guidance on how to correct metrics in these fields. Second, how can one create useful and meaningful statistics based on high dimensional data? I suggest a methodological solution and show how it changes the accepted results of international trade statistics. Social science fields in particular could benefit greatly from this procedure because of the extensive use of indexes of arbitary categories and weights. The main limitation of this procedure is that it assumes a known similarity (or distance) matrix of its components, which is not always well-defined. For my largest example I use international trade data because a well-known procedure for calculating similarity has been introducted by Hidalgo, Klinger, Barabási, and Hausmann (2007), however most sample spaces do not have an established

Canberra, or Manhattan.

procedure and so may be unable to use this procedure until a similarity-calculating procedure is found. This would be an ideal subject for future research.

To preview the proposed orthogonalization procedure, one can see it as a change of coordinate systems. I take as given a set of data vectors and the measure of similarity between every pair of its dimensions. The basic idea is that the similarity between dimensions can increase which reduces distance between individual dimensions in a vector. This is best seen in the spherical coordinate system (see Spiegel 1959, Munkres 1991). The measure of the angle from the vector to an axis is given by θ or ϕ. the orthogonalization procedure then uses the change of coordinates to find the length of this vector along each axis. The rectangular coordinate system is what most empirical measures are based upon, at least those with concepts like angle and distance. So in order for a quantitiative measure to be valid, we must change the coordinate system to what the measure is assuming. This is the basic idea of the paper.

2 Literature

The aforementioned problem of heterogeneity between dimensions is, as far this author can tell, completely unacknowledged when working with shares data. That said, the problem is recognized when working with quantitative variables which are not in the form of shares, and has been the focus of substantial research. I can identify 9 distinct orthogonalization procedures each of which are based on two basic methods, of which there are undoubtedly more. The first method, found overwhelming 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 Physics, involves knowing exactly how the system behaves in a non-stochastic fashion and having perfect measurements.

Two ideas distinguish my problem and solution from the rest of the literature. The first is that the data which I am examing always exists on a unit simplex. Thus the range of possible values that variables may take is relatively limited, and no negative values are allowed. This eliminates the use of correlation and covariance matrices since these procedures commonly produce negative values. Secondly, my over-arching argument rests on the idea that the true coordinates of these observations are not known, but detailed information exists in the form of similarity values which can be used to find the true location.

2.1 Stochastic Methods of Orthogonalization

First I examine statistical methods in which the data is understood to be stochastic. That is, every observation is multi-dimensional and has error, so notions of variance, covariance, and correlation are applicable. Many orthogonalization methods take a latent variable (eigenvalue-eigenvector) approach. These methods includes multidimensional scaling, principal components analysis, principle coordinates analysis and factor analysis, which I discuss in detail below. Furthermore regression analysis is a statistical method of orthogonalization but does not use latent variables.

First consider principal components analysis. In principal components analysis, for a given set of multidimensional observations, there are assumed to be underlying latent variables which are correlated, and can be combined into fewer dimensions. Using the eigenvalues of the covariance or correlation matrix, the corresponding eigenvectors tell how the data should be rotated and scaled, and judgment may be exercised to identify the most important of the resulting dimensions. So principal components analysis is similar to what I am proposing because of the rotation. The resulting data produced from principal components has the identity matrix as its correlation matrix. Similarly, my proposed orthogonalization procedure also produces the identity matrix as its similarity matrix. Now it is worth noting that my orthogonalization procedure differs from principal components analysis in several ways. For one, I am only using shares data, for which the analysis is not the same as with correlation data. With shares data similarity matrices are more relevant than covariance matrices, because the dimensions can only be positively related. Third, reduction and identification of dimensions is not at all a priority in this paper, contrary to principal components analysis. Lastly, the big advantage over principal components analysis is that this orthogonalization process can have more variables than observations.

Similar to principal components analysis is principal coordinates analysis. This process allows a researcher to create a dataset of orthogonal variables, much like this paper suggests. However, the idea behind principal coordinates is that it takes a distance measure between all pairs of *observations* and gives them coordinates; according to Gower (1967, pg.19), "We can ask how the coordinates of points with the given distances be found.". On the other hand, the orthogonalization method that I discuss in this paper adjusts for a similarity measure between *variables*. The idea of this adjustment between variables is that, afterward, similarity between observations (or other measures) can be measured, based on the adjusted variables. Consider that principle coordinates analysis takes a matrix of similarities between 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 similarity matrix between observations based on the given similarity between the variables.

Third, and also similar to principal components analysis, is factor analysis. In factor analysis a researcher attempts to identify unobserved, latent categorical variables. In this case the covariance between dimensions leads to recognition of a previously unidentified latent variable. So this statistical method successfully deals with the problem of heterogeneous similarity between categories.

Fourth, this paper is related to the multidimensional scaling literature which is a modern, simplified version of principal coordinates analysis and which makes very wide use of similarity matrices in order to represent large multivariate data sets in several smaller dimensions. In terms of economics, this is the same as breaking downs factors into wide categories like high- or low-skill labor, capital, land, etc. Or in consumer preferences as demand for different foods like grains versus fruits, vegetables, meat, etc. What differs from multidimensional scaling? Consider this: a researcher is often not interested in differences between *attributes* and how to categorize them, but rather is interested in how different attributes affect *agents*. In other words, consider three goods: corn, wheat, and computers. This paper is concerned not with the attributes of corn, flour, and computers, but rather with the reasons why one person would prefer one bundle of these goods over another. Or in the context of international trade, why one country which produces a given bundle would trade with a country that produces another bundle. In such an analysis, the similarities between any two goods is trivial and only matters because it affects how I calculate production or exports.

Fifth, regression analysis and analysis of variance, are chiefly concerned with accounting for covariance. Each variable is treated as a dimension, and the covariance between dimensions can greatly affect the estimate of the mean value of a regressor on the regressand. In an abstract way this is similar because the practitioner realizes that variables are not completely independent of one another, and so the covariance, or angle between dimensions, is included in the process by design. Not including important covariates leads to omitted variable bias: the magnitude of a parameter is inaccurate. This is geometrically equivalent to a parameter value being projected onto an n-plane but not parallel to its coordinate axis, with the angle between its proposed axis and the actual projection proportional to the correlation with the omitted variable. This is exactly the argument that I am making for distance measures.

2.2 Exact Methods of Orthogonalization

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

2.3 Other Literature

Sixth, this paper has ties to Measure Theory. A main point of measure theory concerns distinguishing a measurement of an attribute from the attribute itself. Consider common commodities like wheat, corn, and computers. How different are these things? As economists, we don't particularly care about wheat, corn, or computers in themselves, but rather about the implied underlying productive or demand structures of each of which determine the composition of output. So what we are able to measure is not necessarily the same as what we need to measure to form general theoretical statements.

Seventh, this paper closely relates to Index Number Theory, however, this paper has nothing direct to say about prices. In Index Number Theory, one can typically identify two distinct approaches: the Axiomatic Approach versus the Economic Approach. What is the point to having these two different approaches? The point is that the data does not line up exactly with theory because the theory is coordinate-free. So the Axiomatic Approach is more like a theoretical approach in that it carries over weights from one period to another, or from one country to another, whereas the economic approach is more like an empirical approach in that it allows comparison of consumers revealed preferences. The reason for these two approaches is that data is not exact enough to specify exactly what the researcher is trying to show. The problem is that product categories are pair-wise arbitrarily determined, so that no pair of goods can be assumed to have any particular relationship to any other pair of goods. This is implicitly understood by economists, and the two approaches take different steps to ameliorating this problem. This paper suggests another method: to make the pair-wise relationships between goods the same in any data vector. This process will make both approaches equally valid because it will eliminate the problem of assigning weights.

No papers directly address or recognize the non-orthogonality problem in relation to distance metrics. I can find only two exceptions. The first is the textbook Gentle (2007) which dedicates an entire section of the book to angular representations of correlation, covariance, similarity, and distance matrices. And second are pages 25-26 of Gower (1967). Gower writes at length about how the angle between two vectors is given by a cosine representation, which is exactly the same problem that I discuss in this paper. The difference here is that Gower is assuming these angles are the subject of interest, whereas I assume that these angles are given and are not of any particular interest. Furthermore, he assumes that the angles between observations is given, not the angle between variates (see Table 1). So, in a sense, I am assuming and proving the contra-positive of Gower's passage.

3 Methodology

The problem, stated in yet another way, is that the categories in which much data is classified is ad hoc, with some categories more alike than others. To fix this problem, one first needs a measure of similarity between all dimensions, to which I will defer to other papers. For example in International Trade see Hidalgo, et al (2007) or for a more general treatment see Dauxois and Nkiet (2002). Second, accoding to Gentle (2007), this similarity data is best viewed as representing the angle between 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 system, Euclidean n-space. Then one can apply any number of distance metrics. This projection is best viewed as a change from hyperspherical [3] to rectangular coordinates for each individual dimension.

3.1 Similarity Matrices and Angle Between Dimensions

Similarity matrices are rarely studied in the Economics literature. For one, the term "similarity" is ambiguous in that it could refer to a host of attributes. Additionally, similarity does not come from any specific theoretical model. In relation to international trade data, two interpretations of similarity should immediately come to mind. First on the demand side, similarity could be a measure of the elasticity of substitution between goods A and B: more similar goods are better substitutes for one another. Second, on the production side, similarity could refer to how similar are two production processes or to the similarity in factor content. In this paper I am abstracting away from this discussion and leaving it to other scholars to debate. I simply assume that a similarity matrix is given.

Now for some notation. The most important distinction here from most textbooks is that I will be treating each entry $x_{c,i}$ of a shares vector as a vector in and of itself. This is perfectly valid because each entry represents a magnitude and, together with the similarity matrix, represents a direction.

According to Gentle (2007), a similarity matrix gives information about the orientation of a set

[3] Known as polar coordinates when $n = 2$, spherical coordinates when $n = 3$, and hyperspherical coordinates when $n \geq 4$.

of vectors: “The cosine of the angle between two vectors is related to the correlation between the vectors, so a matrix of the cosine of the angle between the columns of a given matrix would also be a similarity matrix” (pg 298). This is the most important fact: the values in a similarity matrix are interpretable as the angle between dimensions. Furthermore, from pg. 26 of the same book:

$$angle(x, y) = cos^{-1}\left(\frac{< x, y >}{||x||||y||}\right) \tag{2}$$

Let x denote a vector of elements $x_1, ..., x_n$ of arbitrary category shares such that $\sum_{i=1}^{n} x_i = 1$. Let ϕ_{ij} denote the measure of similarity between any two dimensions i and j (equivalently, let $1 - \phi_{ij}$ denote the measure of dissimilarity), where $\phi_{ij} = 0$ denotes completely dissimilar (orthogonal) categories and $\phi_{i,j} = 0$ denotes categories which would otherwise be identical if not for the arbitrary misclassification.

A special note here on the “Almost Orthogonal” property. The Almost Orthogonal property, as described in Gentle (2007) on page 38 does not apply here. A simple proof shows that, for any arbitrary vector with a 45° angle to each axis, the angle with any particular axis approaches 90° as the number of dimensions increases. However, this does not apply to heterogeneous spaces such as the product space. See Appendix C for further explanation.

3.2 The Spherical Coordinate System

Hyperspherical coordinate systems have found widespread use in Quantum Chemistry and Quantum Physics, as the movement of atoms and molecules relative to one-another can be more parsimoniously described. Other than the physical sciences, I can find only one paper that applies hyperspherical coordinates. This is a Computer Science paper written about distance functions in search indexing by Panda, Chang, and Qamra (2006). The only paper in the Statistics literature that is remotely related to this study is that of Marsaglia (1972) which suggests that Monte Carlo simulation suffers from sampling problems and the way to fix this problem is to treat the data as if it were from a sphere.

The following borrows heavily from Spiegel (1959). Denote a vector in three-dimensional Euclidean space as (x_1, y_1, z_1). Let r be the norm (Euclidean length) of a vector. Let θ be the angle between the z-axis and the x-y plane, in radians. Let ϕ be the angle between the x-axis and the z-y plane. Then given the values for spherical coordinates (r, θ, ϕ), the corresponding rectangular coordinates (x_1, y_1, z_1) can be found by:

$$x = r \sin\theta \cos\phi \tag{3}$$

$$y = r \sin\theta \sin\phi \tag{4}$$

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

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

3.3 The Orthogonalization Procedure: Change of Coordinates

The most promising method to obtain an orthogonal coordinate system is to use a change from hyperspherical 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 own vector. Then, because the angle of each dimension is known in regards to every other dimension, and using a trigonometric-based algorithm, one can project the length of the vector onto each dimension, repeat for each entry in the vector, and sum them up at the end.

Define a vector of shares data by x which has n rows indexed by i. The associated n by n similarity matrix is Φ, with elements $\phi_{i,j}$ where rows are indexed by i and columns indexed by j. Redefine Φ in terms of degrees and convert it from a similarity matrix to a distance matrix:

$$\widehat{\phi_{i,j}} = (1 - \phi_{i,j})90 \tag{6}$$

For every i, define each entry in the vector x_i as a radius. Define each column entry j in row i of matrix Φ as the angle formed by the vector i to dimension j. To convert to rectangular coordinates,

[4] I thank Professor Jeanne Duflot for providing me with an equivalent algorithm.

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

$$\widehat{x_{1,i}} = x_1 \cos(\phi_{1,1}) \tag{7}$$

Now, similarly, find the projection of the second good onto each axis. Do this for each of the n goods using the following algorithm.

$$\begin{aligned}
\widehat{x_{2,i}} &= x_1 \cos \phi_{1,i} \\
\widehat{x_{3,i}} &= x_1 \sin \phi_{1,i} \cos \phi_{2,i} \\
\widehat{x_{4,i}} &= x_1 \sin \phi_{1,i} \sin \phi_{2,i} \cos \phi_{3,i} \\
&\cdots \\
\widehat{x_{n-2,i}} &= r \sin \phi_{1,i} \sin \phi_{2,i} \sin \phi_{3,i} \cdots \ \sin \phi_{n-3,i} \cos \phi_{n-2,i} \\
\widehat{x_{n-1,i}} &= r \sin \phi_{1,i} \sin \phi_{2,i} \sin \phi_{3,i} \cdots \ \sin \phi_{n-2,i} \cos \phi_{n-1,i} \\
\widehat{x_{n,i}} &= r \sin \phi_{1,i} \sin \phi_{2,i} \sin \phi_{3,i} \cdots \ \sin \phi_{n-1,i} \sin \phi_{n,i}
\end{aligned} \tag{8}$$

The previous algorithm is adapted from Lin (1995).[6] Note that the pattern of the hyperspherical algorithm is such that the vast majority of terms are sine and each line ends with cosine except for the very last line which ends in sine.

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

$$\widehat{x_j} = \sum_{i=1}^{n} \widehat{x_{1,i}} \tag{9}$$

The algorithm can be simplified to the following for the i-th good and j-th adjusted good:

$$\widehat{x_i} = \sum_{j=1}^{n} x_j \left(\prod_{k=1}^{i-1} \sin \phi_{k,j} \right) \cos \phi_{i,j}, \quad \forall i < n. \tag{10}$$

And when $i = n$, replace the cosine term with a sine term.

[5] A strictly coordinate-less or good-less coordinate system is not possible to calculate in this way because the similarity matrix is symmetric with 1's down the diagonals, so the rank is at most (n-1), and declaring a numeraire good allows one to compute the rectangular coordinates for n dimensions. This simple rotation does not affect the magnitude of a distance measure, according to Borg and Groenen (1997, pg 281).

[6] As well as from a personal correspondence with my former undergraduate professor Jeanne Duflot.

And finally, 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:

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

And normalize each entry using TUS:

$$\widehat{x_i^a} = \frac{\widehat{x_i}}{TUS} \tag{12}$$

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

3.4 Distance Metrics

The following is an introduction to a subset of common distance metrics used in many different statistical and social science fields. Many more distance metrics exist, and as with the literature review, this list is by no means exhaustive. In various fields these distance metrics go by specific names. For example, the Hirschman-Herfindahl Index is computationaly equivalent to both the Simpson Index in Sociology as well as the Hunter-Gaston Index in Microbiology, which are all simply non-normalized Euclidean distance metrics of an observation's distance from the origin. In other words, all multivariate measures on ratio data are applications one of the following distance metrics.

Let X_c, $X_d \in \mathbb{R}$, where $X_c = \sum_{i=1}^{n} X_{c,i}$ and $X_d = \sum_{i=1}^{n} X_{d,i}$. And denote $x_{c,i} = \frac{X_{c,i}}{X_c}$ and $x_{d,i} = \frac{X_{d,i}}{X_d}$ so that $\sum_{i=1}^{n} x_{c,i} = 1$ and $\sum_{i=1}^{n} x_{d,i} = 1$. Let C and D be points in Euclidean n-space such that

$$C = \begin{bmatrix} x_{c,1} \\ x_{c,2} \\ \dots \\ x_{c,n} \end{bmatrix}, \quad D = \begin{bmatrix} x_{d,1} \\ x_{d,2} \\ \dots \\ x_{d,n} \end{bmatrix}.$$

. The distance between points C and D can be calculated in several ways. The first basic class of distance metrics is the Minkowski Metric:

$$D_{Minkowski} = \left(\sum_{i=1}^{n} |x_{c,i} - x_{d,i}|^p \right)^{\frac{1}{p}}, p \geq 1 \tag{13}$$

Where only the positive root is used. When $p = 1$, the distance metric is known as either Manhattan, or City-Block Distance:

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

City-block distance gets its name from the fact that to get from one point to another in a city grid one must follow the streets. Particularly in Manhattan, streets intersect at right angles, so the absolute value in the difference in each street dimension is the total area one must travel. When p=2, this is the straight-line distance, known as Euclidean Distance:

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

When p approaches infinity, one gets Chebyshev's Distance:

$$D_{Chebyshev} = \max_{i=1}^{n} |x_{c,i} - x_{d,i}| \tag{16}$$

A second class of distance estimators scales the coordinate values. Canberra Distance typifies 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 estimators include the Czekanowski Coefficient, which goes by a myriad of other names in other academic fields. These metrics measure the amount of Manhattan overlap between two multidimensional observations:

$$D_{Czekanowski} = 1 - \frac{2\sum_{i=1}^{n} \min\left(x_{c,i}, x_{d,i}\right)}{\sum_{i=1}^{n} (x_{c,i} + x_{d,i})} \tag{18}$$

Or with a slightly different weighting structure which is invariant to relative size:

$$D_{FK} = \sum_{i=1}^{n} \min\left(x_{c,i}, x_{d,i}\right) \tag{19}$$

$$D_{SN} = \sum_{i=1}^{n} \frac{|x_{c,i} - x_{d,i}|}{2} \tag{20}$$

Where I show in Appendix A that:

$$D_{FK} + D_{SN} = 1 \tag{21}$$

4 Simulation

My above qualitative argument for the need for an orthogonalization procedure is hopefully persuasive. However, I find it useful to present a very general example using a series of simple simulations. I will consider a three-dimensional world where a single observation x_i is composed of k share attributes, where the sum of k attributes is one. Using a random number generator, I will assign values to the k attributes as well as to the $\frac{k^2}{2} - k$ similarity between attributes. This is equivalent to finding a random point in a random k-space. I will 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 varying values of n and k, with the results displayed in Figures 2 and 1. Here the number of observations are $n = 1, 2, ...120$ [7] which are plotted along the x-axis, and the number of dimensions is $k = 2, 3, ..., 160,$ [8] plotted along the y-axis. The z-axis (vertical) represents the measured distance on k dimensions between a point and the origin, averaged for n observations. Figure 2 ignores the similarity between dimensions and computes Euclidean distance in the normal way. Compare these average values to those in Figure 1 which do take into account the similarity between dimensions and thus compute the true average distances. Figure 3 plots the simple difference between the two surfaces.

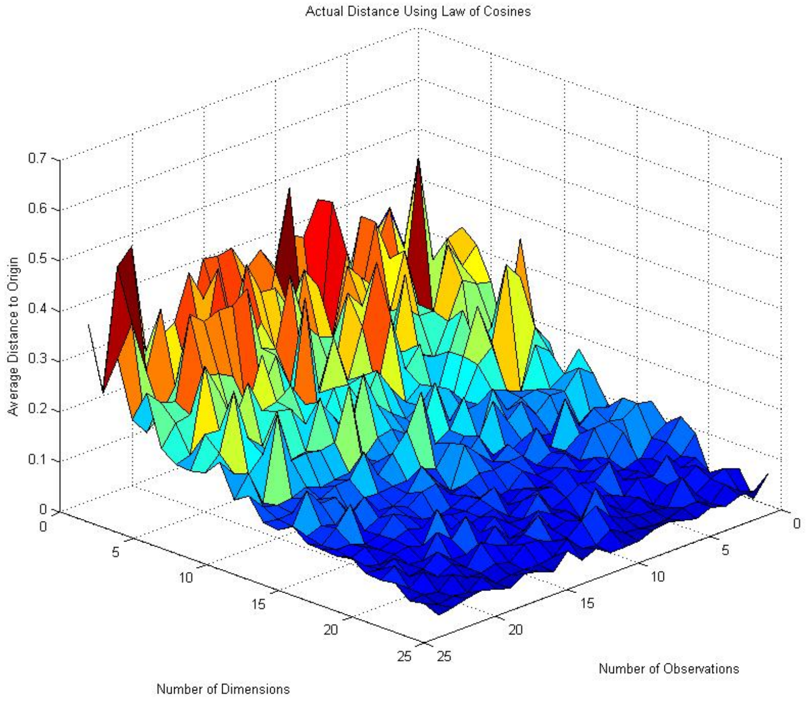
By definition, the distances using the Law of Cosines are correct, it is the Euclidean distances, ignoring the similarity between dimensions, that are incorrect. The difference between the two methods is dramatic. Regular Euclidean distance shows a well-behaved, uniform decrease in distance as the number of dimenions increases and is nearly invariant to the number of observations. On the other hand, the distance metric based on the orthogonalization procedure shows a great deal of variability in its measure, appropriately reflecting the randomness programmed into the simulation. Despite the differences in shape, the correlation between the two measures is 0.8.

This simulation demonstrates that the Euclidean distance metric ignores relevant data in its computation. Furthermore, the values obtained from standard Euclidean distance are almost com-

[7] $n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120$

[8] $k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 160$

Figure 1: Actual Distance Using the Law of Cosines



pletely invariant to the number of observations, and are completely 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.

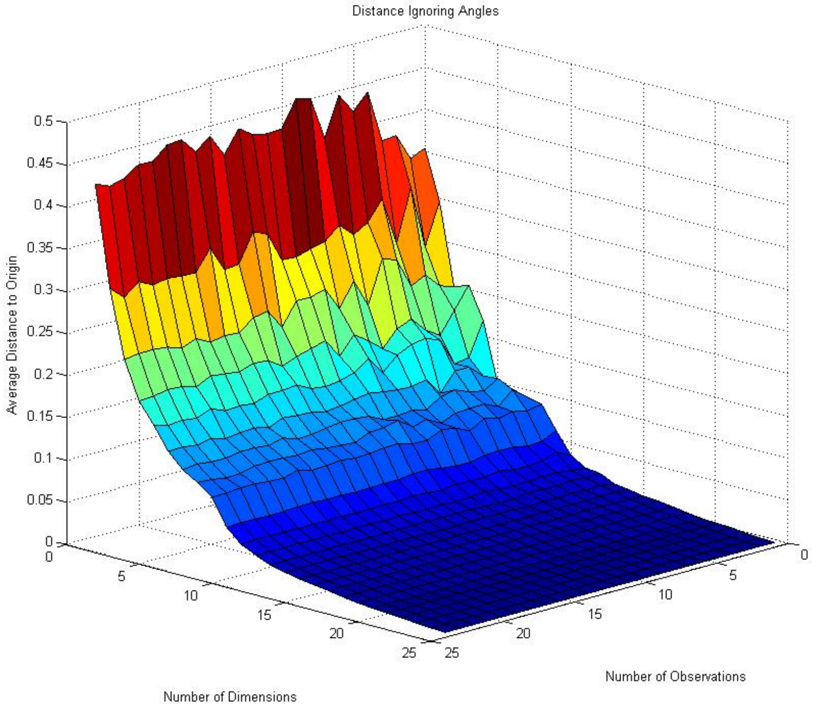
5 Applications

The following section outlines a few examples in the literature where the orthgonalization method can potentially yield great benefit. I plan to academically pursue these topics in the near future. I have drafted or am working on proposals on all of the following topics.

5.1 Application: Price Indices

The computation of index numbers suffers from three primary challenges. The first is that the data is in the form of categories, which naturally do not obey the laws of arithmetic. The second is that the weights of categories change over time. These first two challenges are commonly referred to as the "Index Number Problem." The third is that classification and categorical ambiguity

Figure 2: Estimate Using Euclidean Distance

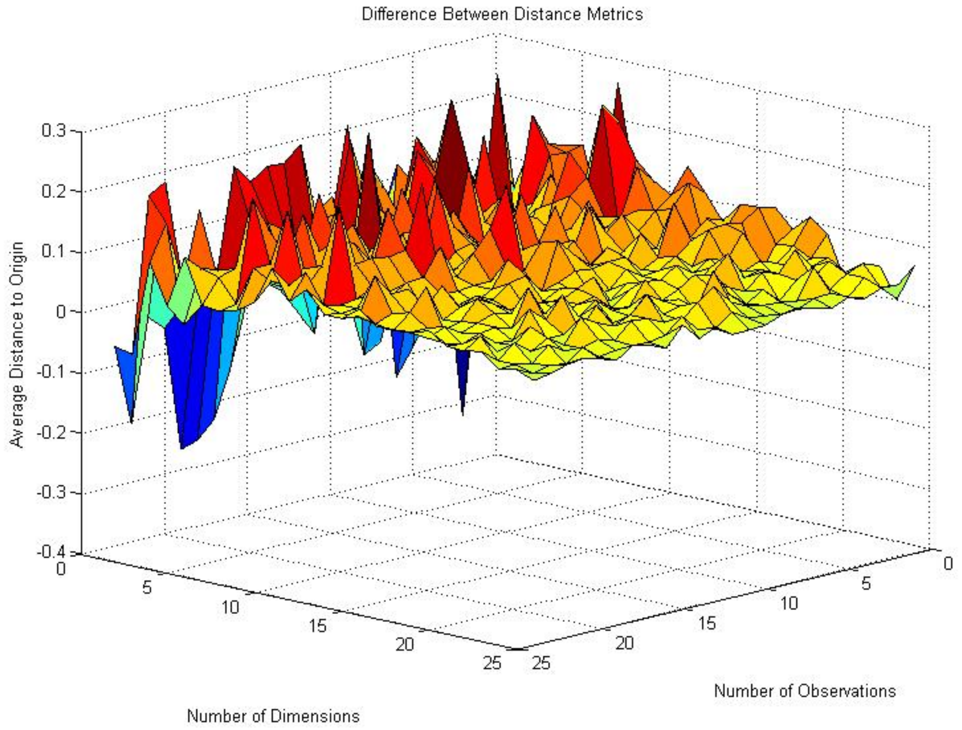


can frequently lead to mis-measured data because categories are arbitrarily chosen. Using the Orthogonalization procedure to produce product category weights, as I've outline above, solves all three problems simultaneously. This process, instead of bypassing the arithmetic problem, uses the categories to its advantage. I solve the weighting problem of index numbers by using time-invariant weights: each orthogonal category carries an equal weight. I also solve the measurement problem by allocating any potentially cross-related categories into appropriate orthogonal categories. Using the orthogonal categories, indexes such Laspreyes, Geometric Mean, Walsh, Paasche, etc. are all equally valid measures of the price index.

5.2 Application: Sociology: The Index of Qualitative Variation

In sociology a common measure of categorical variation is the Index of Qualitative Variation (IQV). The Index is most often used to compare levels of racial diversity across locales or across time. Consider this: in the 18th century, German scientist Johann Friedrich Blumenbach proposed cat-

Figure 3: Difference Between Actual and Estimated



egorizing people as red, yellow, brown, black, and white (Funderburg 2013, 83). The simplicity of this categorization has, quite understandably, been contentiously opposed. The offense is likely not so much in the names used, as it is in the broadness of each category. When being given a label, most individuals would likely want to be recognized as closely as possible to the category in which they self-identify. To this end, a researcher may feel compelled to divide the categories into smaller subcategories. The only problem is that the index monotonically increases with the number of categories. While it's possible that this measure is correct at any aggregation level, the point is that it's not clear which 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 completely different category. However, a multi-racial person is really, by definition, a combination of races, making him or her a combination of racial categories.

The IQV equation is given by:

$$IQV = \frac{1 - \sum(p_k)^2}{n-1} \tag{22}$$

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 with cross-category observations, the bi- or multi-racial observations can be partially grouped into categories, changing the computation in a way that is not immediately clear.

5.3 Application: Development and Political Institutions: The Index of Ethnolinguisitic Fractionalization

Incredibly similar to the IQV is a measure known as the Index of Ethno-linguistic Fractionalization (ELF), which is applied extensively in the literature in the fields of political science and economic development. The equation is given by:

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

Where $k \geq 2$ 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 measure diversity. The problem, as clearly defined by Laitin and Posner (2001) is two-fold. First, a researcher needs to be careful about the level of aggregation used in defining ethnic groups. Second, not all ethnic groups are equally unalike. To this end, Bossert, D'Ambrosio and Ferrara (2005) define a Herfindahl Index that they coin the Generalized Index of Ethno-linguistic Fractionalization Index which accounts for similarties between categories. Indeed they almost define the Law 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 similarity between categories has been recognized before in this literature, but has not had any orthogonalization procedure applied.

5.4 Application: Labor: Location Quotient

A location quotient can refer to a number of economic measures, but the idea and calculation is the same. Consider the location quotient of professional Economists in Boulder, CO. According to the BLS, The percentage of Economists in the workforce in Boulder over the percentage of Economists in the workforce in the USA is 2.28[9]. But aren't Professors of Economics also Economists? In many cases isn't the work of a Statistician, Data Scientiest, or Climate Modeler substantially similar to that of an Economist? In many cases the title of Economist is just that, a title. A collaborator of an economist may be called by some other name, though their work is similar. If you accept the argument that occupations are typically interdisciplinary or potentially haphazardly categorized, then the case for orthogonalization speaks for itself. In fact, the location quotient is exactly the same as the measure of revealed comparative advantage in international trade, which I discuss throughly in the section 5.7.

5.5 Application: Business Analytics

In business, firms benefit from knowing the buying habits of consumers. For example the accuracy and prevalence of the "next best offer" model has simply exploded. This model predicts what a costumer is likely to purchase next. This is based on the buying patterns of previous customers. For example anyone who has had Internet access in the past decade should be familiar with ebay's "My Feed", Amazon's "Recommendations for You", Netflix's "Viewers Also Liked", LinkedIn's "Groups you may like", or really anything with user-level targeted marketing. Every single one of these methods uses a combination of similarity measures[10] and distance metrics to measure the probability that a consumer will buy a specific product, then issues that consumer the sales advertisement for the good which he or she will most likely purchase. This maximizes expected sales, consumer uptake, and customer interaction. The business analytics field would benefit greatly from using corrected distance metrics, since accurate measurements should equate to more sales, at least according to

[9] Data from the Occupational Employment Statistics, accessed online October 7, 2013. The calculation for the location quotient is exactly the same as that for revealed comparative advantage.

[10] Either product similarity based on previous users' probabilities of purchasing or viewing two products, or similarity between the profiles one user and a group of others

Bill Gates (2013, pg 52).

5.6 Application: Internet Search Indexing

Similar to business analytics, distance metrics are widely used by Internet search algorithms in order to identify the most relevant webpages. The idea behind Google, Bing, Yahoo! and other indexers is to gleen only the necessary data from massive databases that serves to both minimize the amount computational work as well as maximize the relevance of the search results. This work could be sped up by identifying and extracting only the orthogonal dimensions within the dataset. One paper which identifies the importance of hyperspherical coordinates and takes a similar approach was written by Google engineers Nanveet Panda and Arun Qamra, along with Stanford Professor Edward Chang (2006). The paper develops a multi-step matching algorithms for finding nearest-neighbors using hyperspherical coordinates. A similar paper by Wu, Chang and Panda (2005) develops a computationally efficient distance metric for search algorithms. My orthogonalization technique is the same concept: it organizes large amounts of data in order to compute more accurate distance metrics. The difference is that my technique does not have a built-in notion of dimensional similarity, and is general enough to incorporate any class of distance metrics after the data has been orthogonalized.

5.7 Application: Trade: International Trade Indices

This section details a few examples, starting with the lowest-dimensions (easiest) and finishing with higher dimensions. The first example takes place in two dimensions where hyperspherical coordinates are referred to as “polar coordinates”. The second example expands this to a hypersphere of 799 dimensions, taken from an example in International Trade, where the set of all exportable goods is referred to as the “product space”.

5.7.1 International Trade Indices

This section details how every common international trade statistic using micro-trade (import/export categories) data can be viewed as plotting a point in n-space and either finding its absolute distance from zero, or relative distance to another point. These defintions are taken from Ng (2002). Using the same notation as above, denote total exports of country c as X_c, where $X_c = \sum_{i=1}^{n} X_{c,i}$. Define $x_{c,i}$ to be the share of good i in total exports of country c, where $x_{c,i} = \frac{X_{c,i}}{X_c}$, and, consequently, $\sum_{i=1}^{n} x_{c,i} = 1$. Equivalently for a second, but with the subscript d, and for the world, with the subscript w. All measurements except the last two are assumed to be taken in the same time period, so time subscripts are otherwise suppressed.

The Hirschman-Herfindahl Index:

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

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

$$FK_{c,d} = \sum_{i=1}^{n} \min\left(x_{c,i}, x_{d,i}\right) \tag{25}$$

The Grubel-Lloyd Index, Grubel and Lloyd (1971):

$$GL_{c,d} = 1 - \frac{\sum_{i=1}^{n} |X_{c,i} - X_{d,i}|}{\sum_{i=1}^{n} (X_{c,i} - X_{d,i})} \tag{26}$$

Two definitions are common for the Export Diversification Index. The first follows directly from Finger and Kreinin (1979):

$$DX1_c = \sum_{i=1}^{n} \min\left(x_{c,i}, x_{w,i}\right) \tag{27}$$

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

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

The Trade Compatibility Index, Michaely (1996)[11]:

$$TC_{c,d} = 100 - \frac{\sum_{i=1}^{n} |X_{c,i} - M_{d,i}|}{2} \tag{29}$$

[11] Also called the "Trade *Complimentarity* Index" [emphasis added], see Ng (2002).

The Export Specialization Index:

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

Changes in Global Demand for Major Exports:

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

Changes in Global Market Share for Major Exports:

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

And lastly, the Thiel Index of export concentration:

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

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.

5.7.2 Simple Example: 2x2 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 dairy cows 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 relatively more to produce stocks and barrels. By some external measurement process we know the the similarity between guns and butter to be 0.8, or 80 percent of the inputs are alike. Then the similarity matrix, denoted by Φ with individual elements $\phi_{b,b}$, $\phi_{b,g}$, $\phi_{g,b}$, and $\phi_{g,g}$ can be written:

$$\Phi = \begin{bmatrix} \phi_{b,b} & \phi_{b,g} \\ \phi_{g,b} & \phi_{g,g} \end{bmatrix} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

In polar coordinates, 0.8 is the proportion of the angle betwen the two vector dimensions of guns and butter. Note that the bias indicated by this similarity matrix, using equation (1) is 0.8. Now the measure of similarity needs to be converted into either radians or degrees, depending on the software requirements, and so we can redefine the similarity matrix as a matrix of angles:

$$\widehat{\Phi} = \begin{bmatrix} \widehat{\phi_{b,b}} & \widehat{\phi_{b,g}} \\ \widehat{\phi_{g,b}} & \widehat{\phi_{g,g}} \end{bmatrix} = \begin{bmatrix} (1-1)90^\circ & (1-0.8)90^\circ \\ (1-0.8)90^\circ & (1-1)90^\circ \end{bmatrix} = \begin{bmatrix} 0^\circ & 18^\circ \\ 18^\circ & 0^\circ \end{bmatrix}$$

And note that $\cos(0^\circ) = 1$ and $\cos(18^\circ) \approx 0.951$. See Figures 4 and 5 for a graphical representation. Now projecting the $y_{c,b}$ vector onto the $y_{c,1}$ and $y_{c,2}$ axes yields:

$\widehat{y_{c,1,1}} = y_{c,b}\cos(0^\circ) = 0.2(1) = 0.2$

$\widehat{y_{c,2,1}} = y_{c,b}\sin(0^\circ) = 0.2(0) = 0$

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

$\widehat{y_{c,1,2}} = y_{c,g}\cos(18^\circ) = 0.8(0.951) = 0.761$

$\widehat{y_{c,2,2}} = y_{c,g}\sin(18^\circ) = 0.8(0.309) = 0.247$

Then adding together the results of the two projections:

$\widehat{y_{c,1}} = \widehat{y_{c,1,1}} + \widehat{y_{c,1,2}} = 0.2 + 0.761 = 0.961$

Figure 4: Unadjusted Shares

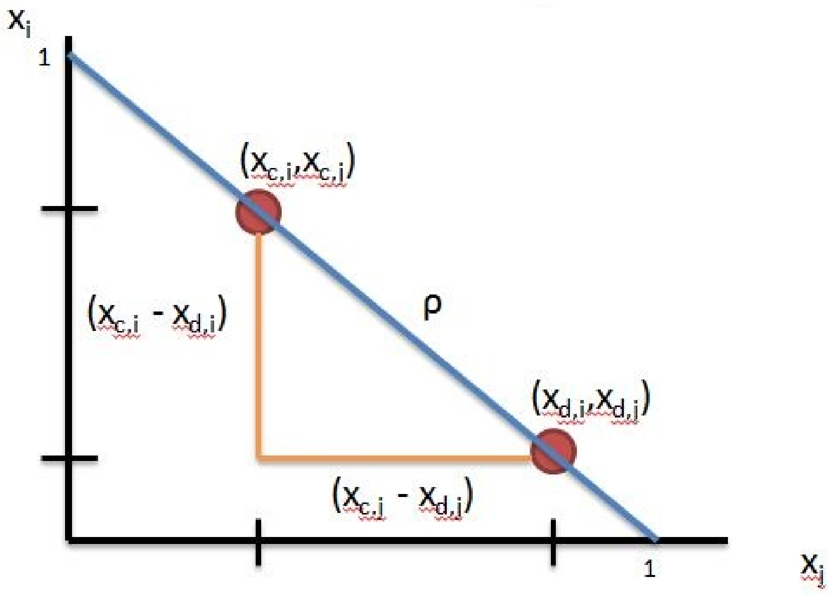
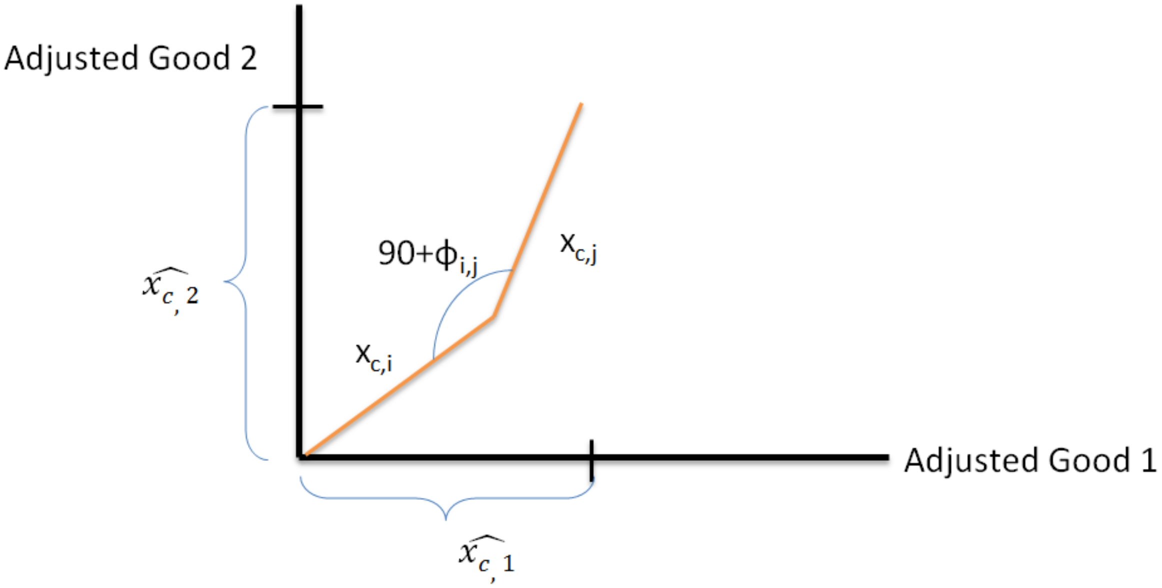


Figure 5: Projection onto Principal Axes



$$\widehat{y_{c,2}} = \widehat{y_{c,2,1}} + \widehat{y_{c,2,2}} = 0 + 0.247 = 0.247$$

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

$$\widehat{y_{c,1}} + \widehat{y_{c,2}} = 0.961 + 0.247 = 1.208$$

And then:

$\widehat{y^a_{c,1}} = \frac{0.961}{1.208} = 0.796$

$\widehat{y^a_{c,2}} = \frac{0.247}{1.208} = 0.204$

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

$\widehat{y_{d,1,1}} = y_{d,b}\cos(0°) = 0.7(1) = 0.7$

$\widehat{y_{d,2,1}} = y_{d,b}\sin(0°) = 0.7(0) = 0$

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

$\widehat{y_{d,1,2}} = y_{d,g}\cos(18°) = 0.3(0.951) = 0.285$

$\widehat{y_{d,2,2}} = y_{d,g}\sin(18°) = 0.3(0.309) = 0.093$

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

$\widehat{y_{d,1}} = \widehat{y_{d,1,1}} + \widehat{y_{d,1,2}} = 0.7 + 0.285 = 0.985$

$\widehat{y_{d,2}} = \widehat{y_{d,2,1}} + \widehat{y_{d,2,2}} = 0 + 0.093 = 0.093$

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

$\widehat{y_{d,1}} + \widehat{y_{d,2}} = 0.985 + 0.093 = 1.078$

And then:

$\widehat{y^a_{d,1}} = \frac{0.985}{1.078} = 0.914$

$\widehat{y^a_{d,2}} = \frac{0.093}{1.078} = 0.086$

Where the final orthogonalized values can be rewritten in vector form as:

$$\widehat{y^a_c} = \begin{bmatrix} 0.796 \\ 0.204 \end{bmatrix}, \quad \widehat{y^a_d} = \begin{bmatrix} 0.914 \\ 0.086 \end{bmatrix}.$$

One must be careful to realize that after adjusting the vectors, the magnitudes no longer have the original interpretation. In other words, the components of the vectors no longer represent shares of guns and butter. Each component of each new vector is now a non-linear composite of the other components.

If one wanted to compare the structure of these economies, one could now find the normalized Euclidean distance between the two:

$$dist^a_{c,d} = \frac{1}{\sqrt{2}}\sqrt{(0.795 - 0.914)^2 + (0.204 - 0.086)^2} = \frac{0.168}{\sqrt{2}} = 0.119, \qquad (34)$$

And notice that for the unadjusted vectors, the normalized Euclidean distance would 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 Law of Cosines distance metric in Knippenberg (2012) finds the Euclidean distance between the unadjusted vectors which is equivalent to the distance between the adjusted but non-normalized vectors, see Appendix B for the proof.

The reader can hopefully see that when similarity is zero, then $\cos(0) = 1$, allowing the orthogonalization process to return the original vectors of guns and butter. Now, to take the analysis a step further, assume that the export share vector of each country is in exactly the same proportion as their production vectors:

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

To abstract from any confounding effects, assume that each country has equal economic output, that these are the only two countries in the world, and that each exports goods equal to 1 normalized unit of value. Abstracting away from any theory on why 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 by $GL^{un}_{c,d} = 0.5$. Contrast this to using the adjusted vectors: $GL^a_{c,d} = 0.833$. Is this difference economically substantial? This is unclear from our arbitrary example, but if any of the values were different, this index could potentially change substantially. Note that the orthogonalization producedure makes the absolute difference between the two vectors equal. For the first good: $|0.705 - 0.872| = 0.167$ and for the second good: $|0.295 - 0.128| = 0.167$. This implies a future application of this procedure: that this procedure should produce a continuum of theoretical goods which, which arranged in a line, should have properties which mimic those of a continuum of goods assumption, as in Dornbusch, Fischer, and Samuelson (1977), Krugman (1979, 1980), Helpman and Krugman (1984), Melitz (2003), and so forth. Again, this is another area for future research.

Second, consider the non-normalized Hirschman-Herfindahl Index of export concentration (Hirschman

1945, 1964):

$$H_c = \sum_{i=1}^{2} \left(\frac{x_{c,i}}{X_c} \right)^2 \tag{36}$$

Where $x_{c,i}/X_c$ is the share of good i in the export bundle of country 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 between these two measures is subjective, but what is interesting is that the ordering has reversed. Where in the unadjusted index, country c was the more concentrated country, in the adjusted index, the more concentrated country is now d.

5.7.3 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 by Hidalgo, et al. (2007), who use export shares to find a measure of similarity between export product categories and then map them using a network analysis approach. I take their analysis a step further by using the similarity measures to adjust the original country export vectors, and I show that the measurements, while clearly correlated, are very different. Because of the computational intensity of the orthogonalization producedure [12] , I have only produced estimates on the export similarity measure. A detailed treatment of the consequences of changing the export similarity measure can be found in Knippenberg (2012), where I insert the new export similarity measure into a gravity equation of international trade and find very different results from previous studies.

Export similarity was first conceived by 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 $FK_{c,d}$, which is derived in Sun and Ng (2000), and is given in equation (19). The measure has been used in hundreds of academic papers on international trade.

[12] Computing this variable for 47,653 observations took approximately four weeks on a desktop computer with a quad-core 3.3Ghz processor.

The steps taken to arrive at these export similarity indices are as follows. First I downloaded the export data from Feenstra's website. The data is 4-digit SITC trade data with 799 categories. I am using 5-year intervals from 1970 to 2000 for 133 countries. Second, I transform export values into export shares. Third, I follow Hidalgo, et al. (2007) to calculate similarity between export categories. Fourth, using this similarity matrix and export shares, I apply the orthogonalization procedure to obtain the adjusted export vectors. Lastly, I apply a Euclidean Distance algorithm to the adjusted export vectors to calculate export similarity between countries c and d at time t. Strictly for the sake of comparison, I then use the original shares data to calculate export similarity in the way that Finger and Kreinin suggest, as given in equation (25).

Now look below at Figure 1 which compares histograms of the adjusted (left) and unadjusted (right) export similarity values. For both histograms, 0 represents two countries having no exports in common and 1 represents two countries having exactly the same proportion of exports. The adjusted measure is approximately normally distributed, with a mean of 0.55. In contrast, the Finger and Kreinin measure is approximately exponentially distributed with a mean of 0.1066 [13]. The point of this figure is to show that a simple distance metric can have extremely different values based on its underlying assumptions, even though they claim to be computing equivalent measures.

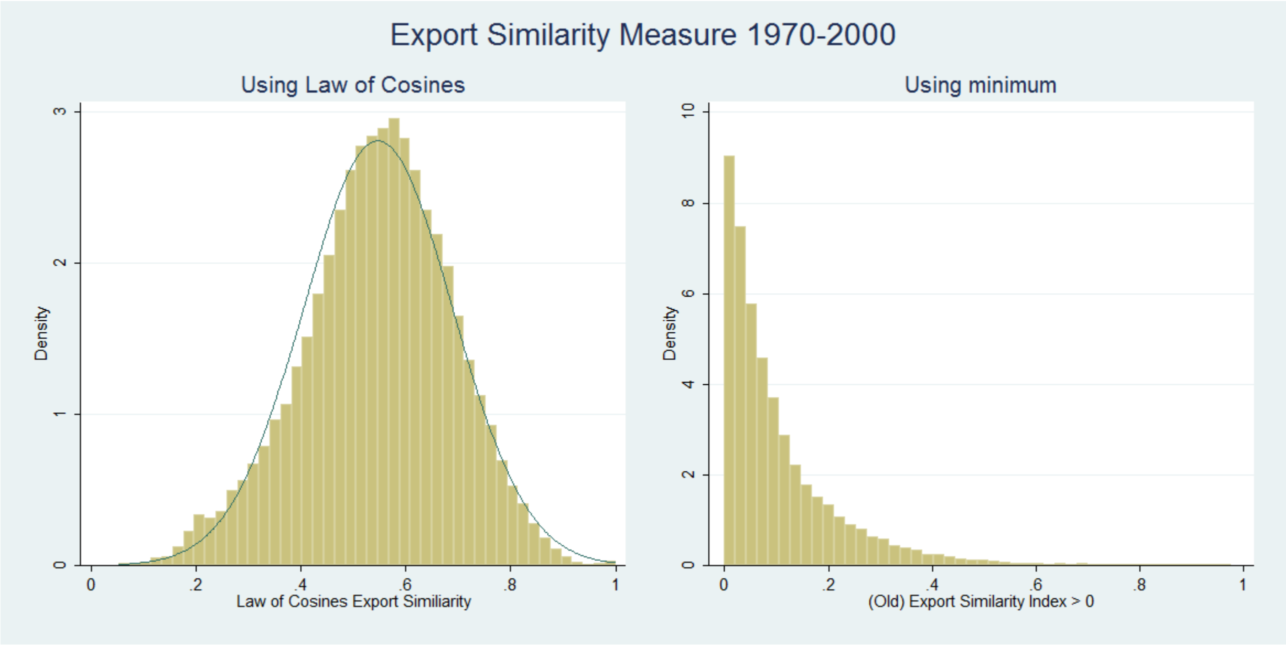
Next, Figure 2 shows the same data, but combined in a scatterplot. Each point represents the value of the adjusted export similarity measurement (y-axis) with the unadjusted export similarity measurement (x-axis). Again, observations are for each country pair (c, d) at time t. Notice that all points lie on or above the 45° line; where the 45° line represents no difference between the export similarity measures. This is because the adjusted export similarity measure accounts for both the typical categorical similarities as well as cross-category product similarities, something which is not measured by the traditional metric.

5.8 Other Applications

The orthogonalization producedure described herein applies to situations where aggregates of arbitrarily chosen categories are used to measure differences between structures. It is not applicable

[13] The lambda parameter of the exponential distribution is $1/E(X) = 9.381$.

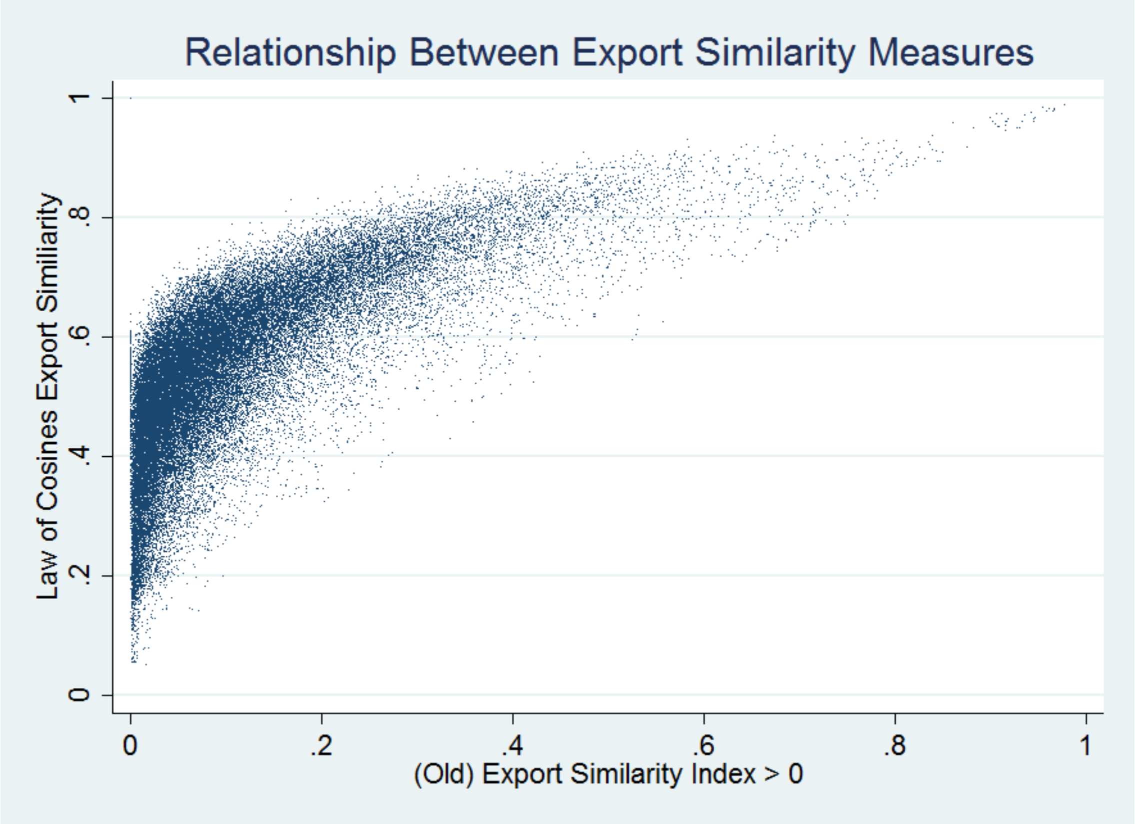
Figure 6: Histograms of Export Similarity Measures



to situations in which similarity between variates or correlation is already accounted for, such as regression analysis or principal components analysis. Given the nature of international trade shares data, this orthogonalizaiton procedure is clearly applicable. Furthermore, New Trade Theory models assume an equal marginal rate of substitution between varieties of a good. However, if two varieties are more similar than either are to any third, then equal marginal rates of substitution cannot mathematically hold. After applying this orthogonalization procedure, the marginal rates of substitution between the adjusted goods should be equal because the variables are orthogonal to one another. This would make the data consistent with the theory, and is a promising area for future research.

This procedure works only when a bivariate notion of “similarity” or “distance” is computable, as these similarity measures directly feed into the equation. This procedure is not applicable where similarity is not defined or calculable. Finding a way to calculate this similarity in many different contexts is an area for future research where notions of covariance, correlation, may be very important. Furthermore, a simple lack of a way to calculate similarity doesn’t make the previous distance metrics any more valid - they are still computed using the incorrect coordinate system.

Figure 7: Scatterplot of Original and Adjusted Export Similarity Measures



This procedure does not apply in regression because regression already adjusts for covariance between variables. All types of regression procedures accounts for any similarity between dimensions. Also, this procedure does not apply to computed indices which have mutually-exclusive categories which clearly never overlap. For example, the Herfindahl Index of industry concentration is a sum of the squared market shares of firms (see for example Hirschman (1964), among many others), with the keyword being *firms*. This is different from the international trade definition, which defined on *categories*, not *firms*. The orthogonalization procedure would not work here because firms are independent entities, and a similarity matrix would already be the identity matrix, meaning no adjustment is necessary.

6 Conclusion

I like the following quote from a linear algebra textbook: “Physical Laws must be independent of any particular coordinate system used in describing them mathematically, if they are to be valid” Spiegel (1959 pg 166). It reminds me that just because you can measure something doesn’t mean that what you have measured must necessarily obey the laws of your theory: sometimes a researcher has to manipulate data to make sense of it. In the case of shares data, often it is in the wrong coordinate system and must be converted to the proper system before familiar measures can be applied, like measures of distance in the rectangular coordinate system. I have argued throughout this paper that arbitrary classifications are not automatically defined by the rectangular coordinate system. However the rectangular coordinate system is the only requirement for applying familiar statistical distance metrics. In other words, the principle axes of the coordinate system are rarely the same as the axes of the data, so distance metrics cannot be immediately 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 would include the effect of export similarity or diversification on bilateral trade (Knippenberg 2012), or likewise the effect of the Grubel-Lloyd 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 underlying good. I have written this paper in an attempt to stay as general as possible about its applications: the extensive examples in international trade are merely a consequence of my own experience. The concepts described herein have wide applicability in all areas of empirical research and I look forward to conducting these applications in the near future.

References

[1] Axler, Sheldon Jay. (1997). *Linear Algebra Done Right* Springer Publishing, New York, NY: 251 pgs. ISBN 0387982582.

[2] Borg, Ingwer and Patrick J.F. Groenen. (1997). *Modern Multidimensional Scaling: Theory and Applications.* Springer Publishing: New York, NY.

[3] Dauxois, Jacques and Guy Martial Nkiet. (2002). "Measures of Association for Hilbertian Subspaces and Some Applications" *Journal of Multivariate Analysis* 82: 263-298.

[4] Ding, Yiren. (2008). "The law of cosines for an n-dimensional simplex" . *International Journal of Mathematical Education in Science and Technology* 39:3, 407-410.

[5] Eaton, Jonathan and Samuel Kortum. (2002). "Technology, Geography, and Trade" *Econometrica* 70(5): 1741-1779.

[6] Funderburg, Lise. (2013). "The Changing Face of America", *National Geographic*, October 2013.

[7] Gates, Bill. (2013). *Time* September 30, 2013: 52.

[8] Gentle, James. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics.* Springer Publishing: New York, NY.

[9] Gower, John C. (1966). "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis", *Biometrika* 53(3/4): 325-338.

[10] Gower, John C. (1967). "Multivariate Analysis and Multidimensional Geometry", *Journal of the Royal Statistical Society. Series D (The Statistician)* 17(1): 13-28.

[11] Grubel, Herbert, and P. J. Lloyd. (1975). *Intra-industry trade: The theory and measurement of international trade in differentiated products*, Macmillan, London, U.K.

[12] Hausmann, Ricardo and Bailey Klinger. (2006). "Structural Transformation and Patterns of Comparative Advantage in the Product Space" *The Center for International Development Working Paper* No. 128, August 2006.

[13] Hausmann, Ricardo and Bailey Klinger. (2007). "The Structure of the Product Space and the Evolution of Comparative Advantage" Working Paper *The Center for International Development Working Paper* No. 146, April 2007.

[14] Hidalgo, César, Bailey Klinger, A.-L. Barabási, and Ricardo Hausmann. (2007). "The Product Space Conditions the Development of Nations" . *Science* 317: (27 July 2007), 482-487.

[15] Hirschman, Albert. (1945). *National Power and the Structure of Foreign Trade* Berkeley: University of California Press.

[16] Hirschman, Albert. (1964). "The Paternity of an Index." *American Economic Review* 54(5): 761.

[17] Knippenberg, Ross. (2012). "Is There Any Space for Comparative Advantage in the Gravity Model?" *Working Paper*, March 2012.

[18] Lin, Chii-Dong. (1995) “Hyperspherical Coordinate Approach to Atomic and Other Coulombic Three-Body Systems.” *Physics Reports* 257: 1-83.

[19] Malaney, Pia. (1996) “The Index Number Problem: A Differential Geometric Approach.” *Ph.D. Thesis*, Harvard University, Department of Economics, December 1996.

[20] Marsaglia, George. (1972). “Picking a Point from the Surface of a Sphere”. *The Annals of Mathematical Statistics* 43(2): 645-646.

[21] Michaely, Michael. (1996). “Trade Preferential Agreements in Latin America: An Ex-Ante Assessment ” World Bank Policy Research Paper 1583, March 2006.

[22] Mikic, Mia. (2005). “Commonly Used Trade Indicators: A Note” ARTNeT Capacity Building Workshop on Trade Research.

[23] Munkres, James. (1991). *Analysis on Manifolds* Westview Press, 1991, Chapter six.

[24] Ng, Francis. (2002). “Appendix B: Trade Indicators and Indices” in *Development, Trade and the WTO: A Handbook* Worldbank Bank Publications, Washington, D.C, pg 585-588.

[25] Panda, Navneet, Edward Y. Chang and Arun Qamra. (2006).“Hypersphere Indexer”. Lecture Notes in Computer Science, Volume 4080, 2006.

[26] Spiegel, Murray. (1959). *Vector Analysis With an Introduction to Tensor Analysis* McGraw-Hill, New York, New York. ISBN 07-060228-X.

[27] Sun, Guang-Zhen and Yew Kwang Ng. (2000). “The measurement of structural differences between economies: An axiomatic characteization.” *Economic Theory* 16: 313-321.

A Appendix: Proof to Equivalence of Finger-Kreinin and Sun-Ng Distance Measures

This section provides a proof that the export similarity measures from Finger and Kreinin (1979) (FK) and Sun and Ng (2000) (SN) are perfectly negatively correlated. Because of the minimum function in FK and the absolute function in SN, this proof is not conducive to deduction, but an inductive argument is easier to show. Define FK and SN according to their authors:

$$FK = \sum_{i=1}^{n} \min(\frac{x_{c,i}}{X_c}, \frac{x_{d,i}}{X_d}), \tag{37}$$

and:

$$SN = \sum_{i=1}^{n} \frac{|x_{c,i} - x_{d,i}|}{2} \tag{38}$$

Proposition:

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,

$$\sum_{i=1}^{n} x_{c,i} = 1, \tag{39}$$

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:

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

Proof:

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,

$$SN = \frac{x_{c,1} - x_{d,1}}{2} + \frac{x_{c,2} - x_{d,2}}{2} \tag{42}$$

By assumption, $x_{c,1} - x_{d,1} = 0$ and since $x_{c,2} = x_{d,2}$, then $x_{c,2} - x_{d,2} = 0$, making $SN = 0$. Thus, trivially,

$$SN = 1 - FK. \tag{43}$$

A.2 Case 1.2

Let $n = 2$ and $x_{c,1} > x_{d,1}$. Then the definition of shares data, (39), implies that $x_{c,2} < x_{d,2}$. So by the definition of FK:

$$FK = x_{c,2} + x_{d,1}. \tag{44}$$

Rewrite the shares definition (39) for country c as:

$$x_{c,1} = 1 - x_{c,2}. \tag{45}$$

Likewise, rewrite the shares definition (39) for country d as:

$$x_{d,2} = 1 - x_{d,1}. \tag{46}$$

The SN index (38) can be written as:

$$SN = \frac{x_{c,1} - x_{d,1}}{2} + \frac{x_{d,2} - x_{c,2}}{2} \tag{47}$$

Now directly plug-in (45) and (46):

$$SN = \frac{1}{2}[(1 - x_{c,2} - x_{c,2}) + (1 - x_{d,1} - xd, 1)] \tag{48}$$

Simplifying:

$$SN = \frac{1}{2}[(1 - 2x_{c,2}) + (1 - 2x_{d,1})] \tag{49}$$

and:

$$SN = \frac{1}{2} - x_{c,2} + \frac{1}{2} - x_{d,1}, \tag{50}$$

and:

$$SN = 1 - (x_{c,2} + x_{d,1}), \tag{51}$$

and notice that the two terms in the parentheses are exactly equal to FK (44), so subbing into the equation gives:

$$SN = 1 - FK, \tag{52}$$

A.3 Case 2.1

Let $n \geq 2$ and $x_{c,i} = x_{d,i}$ for $i = 1, ..., n$. Then, trivially,

$$FK = \sum_{i=1}^{n} x_{c,i} = 1 \tag{53}$$

And, similarly,

$$SN = \sum_{i=1}^{n} \frac{1}{2}[x_{c,i} - x_{d,i}], \tag{54}$$

and since $x_{c,i} = x_{d,i}$ $\forall i = 1, ..., n$ by assumption, $SN = 0$. Thus, trivially:

$$SN = 1 - FK \tag{55}$$

A.4 Case 2.2

Let $n \geq 2$ and $x_{c,i} > x_{d,i}$ for $i = 1, ..., j$. Let $x_{c,i} > x_{d,i}$ for $i = 1, ..., k$. Let $x_{c,i} > x_{d,i}$ for $i = 1, ..., l$. 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 $x_{d,i}, i = 1, ..., l$. 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:

$$\sum_{i=1}^{j} x_{c,i} + \sum_{i=1}^{l} x_{c,i} = 1 - \sum_{i=1}^{k} x_{c,i}, \tag{61}$$

and likewise from rearranging (58):

$$\sum_{i=1}^{k} x_{d,i} = 1 - \sum_{i=1}^{j} x_{d,i} - \sum_{i=1}^{l} x_{d,i}. \tag{62}$$

Plugging these two expressions into (60) yields:

$$SN = \frac{1}{2}\left[1 - \sum_{i=1}^{k} x_{c,i} - \sum_{i=1}^{k} x_{c,i} + 1 - \sum_{i=1}^{j} x_{d,i} - \sum_{i=1}^{j} x_{d,i} - \sum_{i=1}^{l} x_{d,i} - \sum_{i=1}^{l} x_{d,i}\right]. \tag{63}$$

Grouping like terms gives:

$$SN = \frac{1}{2}\left[2 - 2\sum_{i=1}^{k} x_{c,i} - 2\sum_{i=1}^{j} x_{d,i} - 2\sum_{i=1}^{l} x_{d,i}\right]. \tag{64}$$

Simplifying:

$$SN = 1 - \left[\sum_{i=1}^{k} x_{c,i} - \sum_{i=1}^{j} x_{d,i} - \sum_{i=1}^{l} x_{d,i} \right]. \tag{65}$$

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

$$SN = 1 - FK. \tag{66}$$

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

B Appendix: Proof of Equivalence Between Orthogonalization and the n-Dimesional Law of Cosines

Proposition:

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

I will only prove this equivalence for the two-good case. The notation needed to prove the proposition in higher dimensions is incredibly cumbersome, but the concepts needed are simply algebraic. One can easily see how to extend the proof to higher dimensions.

Proof: (2-dimensions) According to Knippenberg (2012), the distance measure using the n-dimensional Law of Cosines is:

$$dist_{c,d} = \frac{1}{\sqrt{2}}\sqrt{(x_{c,1} - x_{d,1})^2 + (x_{c,2} - x_{d,2})^2 - 2(x_{c,1} - x_{d,1})(x_{c,2} - x_{d,2})cos(90 - \phi)} \quad (67)$$

According to the orthogonalization procedure described in the text, the adjusted vectors can be written as:
$\widehat{x_{c,1}} = x_{c,1} + x_{c,2}\cos(90 - \phi)$,
$\widehat{x_{c,2}} = x_{c,2}\sin(90 - \phi)$,
$\widehat{x_{d,1}} = x_{d,1} + x_{d,2}\cos(90 - \phi)$,
$\widehat{x_{d,2}} = x_{d,2}\sin(90 - \phi)$.
Note that: $\cos(90 - \phi) = \sin(\phi)$,
and $\sin(90 - \phi) = \cos(\phi)$.
Sub these into the usual definition of Euclidean distance, the norm distance in a Hilbert subspace, normalized to the unit inverval:

$$2dist^2_{c,d} = [(x_{c,1} + x_{c,2}\cos(90 - \phi)) - (x_{d,1} + x_{d,2}\cos(90 - \phi))]^2 \quad (68)$$

$$+ [(x_{c,2}\sin(90 - \phi)) - (x_{d,2}\sin(90 - \phi))]^2 \quad (69)$$

Rearrange, grouping like terms:

$$dist_{c,d} = \frac{1}{\sqrt{2}}\sqrt{(x_{c,1} - x_{d,1} + (x_{c,2} - x_{d,2})\cos(\phi))^2 + ((x_{c,2} - x_{d,2})\sin(\phi))^2} \quad (70)$$

Multiply out the squared terms, group like terms and simplify:

$$dist_{c,d} = \frac{1}{\sqrt{2}}\sqrt{(x_{c,1} - x_{d,1})^2 + (x_{c,2} - x_{d,2})^2(\cos^2\phi + sin^2\phi) - 2(x_{c,1} - x_{d,1})(x_{c,2} - x_{d,2})cos(\phi)} \quad (71)$$

By definition, $\sin^2\phi + \cos^2\phi = 1$, which yields:

$$dist_{c,d} = \frac{1}{\sqrt{2}}\sqrt{(x_{c,1} - x_{d,1})^2 + (x_{c,2} - x_{d,2})^2 - 2(x_{c,1} - x_{d,1})(x_{c,2} - x_{d,2})cos(90 - \phi)} \quad (72)$$

Which exactly equals the distance measure obtained from the Law of Cosines. □

C Appendix: The Almost Orthogonal Property

A potentially damaging Proposition is the Almost Orthogonal Property, found for example in Gentle (2007, pg 38). It states that the angle between a vector and an axis increases to 90° as the number of dimensions increases. The property is given below and is followed by an explanation as to why it does not apply to this study.

Proposition:

For an arbitrary vector x_c, the angle between the vector and any given axis approaches 90 degrees as the number of dimensions increases. For purely illustrative purposes, consider the property in lower dimensions: Let

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

and define each unit axis as

$$y_c = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad y_d = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Define the angle between any two vectors x and y as: $\text{angle}(x, y) = \cos^{-1}\left(\frac{<x,y>}{||x||||y||}\right)$. Then $\text{angle}(x, y_c) = \text{angle}(x, y_d) = \cos^{-1}\left(\frac{1}{\sqrt{1^2+1^2}\sqrt{1^2}}\right) = 45°$. For three dimensions, the angle is about 54.74°.

Proof: (for n-dimensions) Let

$$x = \begin{bmatrix} 1 \\ 1 \\ ... \\ 1 \end{bmatrix},$$

and define each axis as

$$y_i = \begin{bmatrix} 0 \\ ... \\ 1 \\ ... \\ 0 \end{bmatrix}, \quad \forall i = 1, ...n$$

where y_i is a vector of zeroes except in the i-th entry.

Then the angle between x and any y_i is:
$\text{angle}(x, y_i) = cos^{-1}\left(\frac{1}{\sqrt{n}\sqrt{1^2}}\right)$.
and taking the limit as n approaches infinity:

$$\begin{aligned} \lim_{n\to\infty} angle(x, y_i) &= \cos^{-1}\left(\frac{1}{\sqrt{\infty}}\right) \\ &= \cos^{-1}\left(\frac{1}{\infty}\right) \\ &= \cos^{-1}(0) \\ &= 90°, \forall i \end{aligned} \tag{73}$$

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 can be interpreted as unit axes having multiple entries with length perhaps summing to more or less than unity.

Secondly, and probably more importantly, the distribution of x_i (export shares) is neither uniform nor random. Most countries in the world do not export or import every possible good, reducing the number of dimensions in most cases. Also, the patterns of similar goods that countries do export is not random nor evenly divided between categories. In other words, in an orthogonal coordinate system, all axes are equally different from one another. But in a non-orthogonal coordinate system, such as one that naturally arises with categorical dimensions, the

Because of these two reasons, the Almost Orthogonal property does not apply to the Product Space of International Trade, nor to any part of economics made up of industry classifications, though I cannot necessarily speak for disciplines other than economics.