Patterns, not genes Characterize Diseases and Optimal Treatments

R.R. Johnson, Professor Emeritus, Computer Science, U/U, 1993
2/22/2010

1. *Overview*

The issue of interest is how many genes need to be considered when trying to classify patients by their genetic patterns in high dimensional genetic and clinical data.
The techniques employed in this analysis are primarily analytic. Vector algebra and Euclidean geometry are the principle techniques but Information Theory as defined by Shannon is sometimes employed.
Traditional techniques for analyzing large data usually start with statistics. All measurements in classical statistics presume that the data is stationary
and that the probability distributions are Normal (Gaussian). If both of these conditions are not satisfied, classical statistics produce results that are not meaningful.
While analytic techniques work independently of the statistical nature of the data. if the data analyzed is not stationary. results from such analyses are also not meaningful.
The first analytical approach used here to define precise patterns and measurements in multidimensional space is vector-fusion. Johnson [1].
Results generated by these studies employ linear algebra, functional analysis, and Euclidean geometry.

The conclusion is that it is patterns of many genes. not one. not 10. but typically thousands of genes that are involved in determining outcomes.
Complicating this situation is the perspective brought to these problems by Francis Collins:
Environmental as well as genetic, proteomic, and clinical data is required to differentiate patients, their maladies, and optimum treatments using only their data.

Described here are several mathematical techniques for recognizing and using patterns.
Objectives:
1. To show that patterns of many if not all values per patient or experiment are required to differentiate patients and classes of patients or experiments.
2. To find people interested in either a science that generates data needing analysis, or in the original pattern identification techniques employed
3. Educate everyone in what can be done now. using this one poster.
4. Illustrate the geometry of 4Dimensional symmetric and asymmetric spheres and tori. The purpose of such models is verification of Vector-fusion.

2. *Analytical techniques that use patterns in multidimensional data to successfully classify specific sets of data into specified classes*

A. Horizontal vector fusion

1). vectorize each column ("dimension") of the n columns of data. assigning each column its own unique phase angle.
In essence this generates what in Excel is called "radial coordinates".
2). compute the vector sum of each of the m rows of this vectorized data.
3). look at. visualize. this vector sum. Plot each of the 2 parts of the vector sum vs row #. and then the scatter-plot of the first dimension against the second dimension.
4). use the vector sums analytically. Horizontal vector-fused results are called ***SBPx,y*** **SBP** stands for **S**ingle-point, **B**roken-line, **P**arallel-coordinates.

B. Vertical, or Transposed, vector fusion

Transpose the matrix of the original data. then repeat the steps of A on this transposed data.
Vertical or transposed horizontal, vector-fused results are called ***SBP-Tx,y***

C. Shannon Information Σ

Shannon Information H = - Σ pi * log pi and Σ where Σ pi = 1 and the sums are over all i And the data is ergodic
Where each patient is characterized by m rows of values. H is computed for each patient and all m of his rows
pi is the probability that each value i (of the m rows) occurs

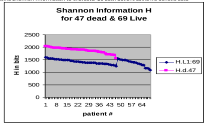
D. Exact Euclidean Models

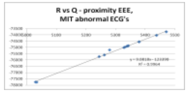
Each dimension here is vectorized and treated as orthogonal to all other dimensions. In Vector-fusion. each dimension is at its own incremental phase or perspective angle to all other dimensions
Projection is then the sum of all orthogonal vectors onto each of 2, 3, 4 or more ortogonal sum dimensions.

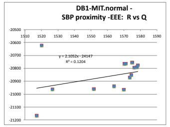
Two charts on the right are scatter plots of 2D Vector-fused sums for each EKG Normal patient in the MIT EKG database
using the separation between the SBP vector fused sums and the Exact Euclidean vectorized values for sum dimensions R and Q
(R is the radi sqrt(EEx^2+EEy^2) and Q=atan(EEy/EEx))
The vector sums are taken longitudinally (in the time direction of the EKG)

Clear distinct region for Normal and for the linear abnormal data using SBP of the EE model
(There are 113,000 rows of longitudinal EKG data in this MIT dataset)
To the issue: are a few or many genes determinative when differentiating classes with this technique.
since class membership here is determined by the slope and intersect of the data for each class,
many gev values would establish class membership more reliably than just the 3 necessary to specify a straight line.
But. 3 points (genes) are all that is required.

Shannon Information:
gene-expression values for Colon Cancer patients at Moffitt Cancer Institute
47 patients die before 36 months and 69 patients survive beyond 36 months
1 error (H overlaps) or 99% accuracy using Shannon Information to characterize each patient using his genetic data

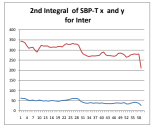






E). Integrals of data prior to SBP x & y, eliminate "noise" or otherwise spurious results.

(First differences of data are used in 2). above)



To the issue: are a few or many genes determinative when differentiating classes with this technique?

In noisy data. this technique employs the magnitude of SBP-T vector sums in both x and y display space
SBP-T is the vector sum down each column of a patient's data.
Small margins differentiate in this chart for the larger or smaller magnitudes in x (red) and y (blue)
of the SBP-T vector sum

In general. it takes many if not most of the genes' gev values to add up precisely within the
small absolute differences identifying whether a patient's SBP-T value places him
in the larger or smaller red (or blue) group.

F). Sorts by gev to determine which largest values of gev occur in patient pairs of kin

The sequence in which these genes appear as the largest varies from each pair compared, but these are almost always the largest gev valued genes in each pair.

		patient:	relative:		
original column #	…………………	6	24		
patient ID ref. gene #:	most expressed	08_5081B	07_5081	most express	ref. gene #:
10710	**16288**	15029.4653	13649.48	5342	10710
16288	**10710**	14071.8888	12721.96	**574**	16288
574	**574**	12787.7416	12472.62	**12529**	574
3623	9339	11174.5831	11044.06	**16288**	3623
12529	**11162**	11136.4277	10978.75	**3623**	12529
2147	**3623**	11126.3452	10975.77	1716	2147
11162	5342	10934.6307	10730.02	9939	11162
4377	2447	10924.7661	10249.38	**10710**	4377
	2147	10154.273	9923.951	4557	
	4557	10039.5752	9752.91	13308	
	12529	9997.2848	8994.3	11162	
	4788	9742.926	8780.324	2147	
	15943	9502.6936	8735.738	4377	
	13871	9292.5437	8618.776	13871	
	8430	8793.6599	8600.828	414	

		patient:	relative:		
		3	20		
ref. gene	most expressed ge	42_566b	41_566	most expresse	ref. gene #:
10710	**574**	15888.5879	16684.44	**3623**	10710
16288	**3623**	15676.817	13499.66	**10710**	16288
574	**16288**	14349.1963	13147.32	**574**	574
3623	**10710**	14331.2977	12454.83	**16288**	3623
12529	**11162**	13419.899	12041.51	8430	12529
2147	**12529**	13176.2429	11871.69	**11162**	2147
11162	4557	12008.1872	11780.29	5342	11162
4377	2147	11610.8373	11441.21	12529	4377
	8430	11526.5944	11191.77	**2147**	
	4788	11193.1262	11036.83	15970	
	2447	10884.9663	10887.88	4788	
	15970	10704.0212	10462.45	14740	
	5342	10662.0481	10362.45	2447	
	5631	10650.1934	10251.26	9339	
	13871	10596.2423	10221.22	5631	

		patient:	relative:			
		11	23			
	most-expres	45_5065b	44_5065	most-expressed		
10710	**10710**	16579.0598	14170.59	**574**	1	10710
16288	**16288**	15020.2432	13252.6	**10710**	2	16288
574	**574**	13988.9654	12943.91	**12529**	3	574
3623	**3623**	13072.8729	12809.12	**3623**	4	3623
12529	**12529**	12379.9058	12784.92	**16288**	5	12529
2147	**2147**	11867.2018	12136.16	**11162**	6	2147
11162	**11162**	11696.4975	11987.18	5342	7	11162
4377	**4377**	11283.3356	10524.77	8430	8	4377
	9339	11241.4518	10245.68	4557	9	
	4788	11137.4838	10011.46	2147	10	
	4557	10822.3243	9915.006	4788	11	
	13871	10666.3596	9711.457	5631	12	
	15943	10401.2045	9602.767	13871	13	
	8430	10314.3324	9523.169	2447	14	
	6772	10124.7123	9519.491	15943	15	

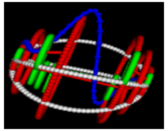
G). Gene Patterns in an Identical twin & between a Parent and an offspring Reference: file parent.bro.sis.match.xlsb

These are known parental or sibling relationships - i(dataset: Williams.Normal from Bob Weiss 10/19/09
At least 8 genes are the largest gev valued genes among the set of parents/children and Identical twin available data:

ref. gene # 10710
16288
574
3623
12529
2147
11162
4377

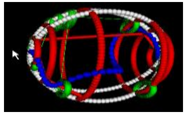
Multi-dimensional Geometry - Visualization Examples

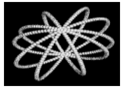
4D asymmetric sphere



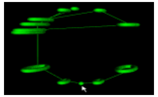
blue - Helix on 4D sphere

White - longitudes
Green - one set of latitudes
Red - 2nd set of latitudes





longitudes on 4D asymm.sphere



first set of latitudes
on 4D sphere

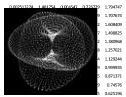
Note there are
two poles on this
4D sphere

4D asymmetric torus:

(1+SIN(200*K2))
(1+COS(200*K2))*SIN(K1)
(1+COS(200*K2))*COS(K1)*SIN(50*K1)
(1+COS(200*K2))*COS(K1)*COS(50*K1)

4Dtorus.wrl (asymmetric) **helix on 4D asym. torus**

α	x	y	z	w
0	0	0	0	1.992115
0.000628319	1.125333	0.001237	0.061835	1.967611
0.001256637	1.24869	0.002425	0.121172	1.925967
0.001884956	1.368125	0.003537	0.176576	1.867976
0.002513274	1.481754	0.004547	0.226729	1.794747
			9	1.707674
			2	1.608409
			7	1.498825
			2	1.380968
			8	1.257021
			4	1.129244
			9	0.999935
			5	0.871371
			9	0.74576
			5	0.625196
0.009424778	1.951057	0.005412	0.260679	0.511612
0.010053096	1.904827	0.004666	0.223606	0.406738
0.010681415	1.844328	0.003873	0.184556	0.312067
0.011309734	1.770513	0.003065	0.145217	0.228825
0.011938052	1.684547	0.00228	0.107341	0.157947
0.012566371	1.587785	0.001554	0.072699	0.100062
0.013194689	1.481754	0.000927	0.043037	0.055483
0.013823008	1.368125	0.000434	0.020024	0.024205
0.014451326	1.24869	0.000114	0.005214	0.005914
0.015079645	1.125333	0	0	0
0.015707963	1	0.031415	1.414039	1.414039
	1			



4DsymTorus1.wrl:

x = SIN(A4)*(3+SIN(100*A4))
y = COS(A4)*(3+SIN(100*A4))
z = SIN(50*A4)*(3+COS(100*A4))
w = COS(50*A4)*(3+COS(100*A4))

Equations:

x = SIN(AF5)*(5+SIN(100*AF5))
y = COS(AF5)*(5+SIN(100*AF5))
z = SIN(50*AF5)*(5+COS(100*AF5))
w = COS(50*AF5)*(5+COS(100*AF5))

The matrix of data above is typical of data of any kind that can be visualized by Vector-fusion.
The name of each column of data in this case of a 4D torus. is:
α is the basic radial coordinate (R = 1 in this example)
x, y, z, w are the names of each of the 4 dimensions of the torus
This is a typical parametric equation expression for any multi-diimensional geometry.
Any number of parameters can be specified. and any multi-dimensional geometry that can be defined by parametric equations can be generated and seen.
One preferred locus on any surface is a helix generated as increments in the basic radial coordinate α.

A geometry of particular interest is the single surface geometry of the Klein Bottle.
Using the parametric equations for a symmetric Klein Bottle in 4 dimensions,
a generalized Klein Bottle has been defined for any dimensionality that is a power of 2.

High dimensional Klein Bottles are used as single surface reference points for proximity comparisons
65,532 dimensions (for Klein Bottle proximities to comparable high dimensional micro-array data).

Question: In topology of 4D. symmetric spheres do not have any poles common to all radial angle components
while asymmetric 4D spheres have common poles. The ordering of points on those 2 kinds of spheres
are different, and there are points on one kind of sphere that are not on the other.
Are their shapes different in some way? **The latitudes on each type also differ.**
Are both types of 4D spheres topologically the same?

Note: As illustrated above, the 4D symmetrical torus has 2 apparent "holes",
and the 4D asymmetrical torus has only one "hole" -
as seen projected into 3D display space.

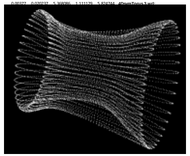
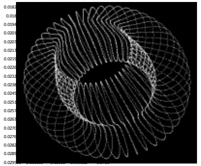
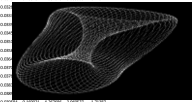
Are both types of 4D tori typologically the same?

The Purpose of Studies of nD geometries is to test the pattern identification methodologies being used for analyzing patterns in high dimensional data since there are few trusted datasets with known patterns. Seeing that the patterns generated by those data mining techniques are plausible for geometries builds confidence that those techniques are plausible for biological. genetic. and other kinds of unknown data.

4D Symmetric Torus. 3.wrl

α	x	y	z	w
0	0	5	0	6
0.000628	0.003181	5.06279	0.188403	5.995067
0.001257	0.006441	5.125329	0.376248	5.980291
0.001885	0.009778	5.187372	0.562983	5.955738
0.002513	0.013191	5.248673	0.748062	5.921519
0.003142	0.016679	5.308991	0.93095	5.877789
0.00377	0.020237	5.368086	1.111129	5.824744

4DsymTorus3.wrl:



0.01508	0.090445	5.997345	3.465719	3.690615
0.015708	0.094244	5.99926	3.535534	3.535534
0.016336	0.097981	5.997226	3.599071	3.379752
0.016965	0.101649	5.991252	3.656542	3.223675
0.017593	0.10524	5.981361	3.708186	3.067679
0.0182				
0.018				
0.0194				
0.0201				
0.0207				
0.0213				
0.0219				
0.0226				
0.0232				
0.0238				
0.0245				
0.0251				
0.0257				
0.0263				
0.0270				
0.0276				
0.0282				
0.0289				
0.0295				
0.030159	0.154553	5.123002	3.999977	0.251657
0.030788	0.155847	5.060391	3.999999	0.125705
0.031416	0.157054	4.997533	4	1.13E-15
0.032044	0.158182	4.934675	3.999999	-0.12571
0.0326				
0.0333				
0.0339				
0.0345				
0.0351				
0.0358				
0.0364				
0.0370				
0.0376				
0.0383				
0.0389				
0.039584	0.169021	4.267686	3.960527	-1.71387
0.040212	0.170032	4.226068	3.947377	-1.85749
0.040841	0.171115	4.187488	3.931312	-2.0031
0.041469	0.172282	4.152099	3.911985	-2.15063
0.042097	0.173545	4.12004	3.889045	-2.29997
0.042726	0.174916	4.091436	3.862142	-2.45099
0.043354	0.176405	4.066399	3.830934	-2.6035
0.043982	0.178024	4.045028	3.795085	-2.75729
0.044611	0.179784	4.027406	3.754272	-2.91211
0.045239	0.181695	4.013602	3.708186	-3.06768
0.045867	0.183766	4.00367	3.656542	-3.22367
0.046496	0.186007	3.997648	3.599071	-3.37975
0.047124	0.188426	3.995559	3.535534	-3.53553
0.047752	0.19103	3.997411	3.465719	-3.69062
0.048381	0.193828	4.003196	3.389444	-3.84457
0.049009	0.196825	4.012889	3.306561	-3.99695
0.049637	0.200026	4.026451	3.216959	-4.14728
0.050265	0.203436	4.043829	3.120562	-4.29508