BINF 636: Lecture 10, Oct 29, 2008:

Introduction to Principal Component Analysis and Multidimensional Scaling (Distance Geometry)

Description:

Principal Component Analysis (PCA) can be viewed as a change in coordinate system, chosen so that to the extent possible, most of the variation in the data is captured in the top several coordinates of the data when expressed in terms of the new PCA coordinate system. In cases where enough of the variation is in a few top components of the data, this is a powerful visualization technique. Use of the top several components (enough to capture a majority of the variation in the data) may lead to more efficient data analysis and suppression of the effects of noise. This class will cover what PCA does, how it does it, and when it is advantageous, including visualization examples from tumor subtypes in gene microarray data.

In classical multidimensional scaling (MDS), also called *distance geometry*, one starts with set of distances between the points to be displayed, and attempts to represent these points in a low dimensional space while having the distances between the points approximate, as well as possible, the original distances. The original distances could be the Euclidean distances between the points, or, for example, distances derived from correlations. A standard first step in a distance geometry application is to convert the distances into a matrix of the inner products of the (unknown) position locations. This matrix leads to a set of positions (coordinates of the points) satisfying the distance conditions (assuming the distance data doesn't violate certain geometrical conditions), but in a higher dimensional space. Point locations in 3 coordinates can be obtained by projecting down from the higher dimensional positions in a way that minimizes a certain measure of the error committed in forcing the projected points to have only 3 coordinates, which can be seen to be a PCA projection. The resulting starting positions can be used as the initial values in a nonlinear optimization procedure to search for positions that better satisfy the prescribed distances. The aspects of distance geometry outlined above will also be covered in this class.

Topics to be covered include

- Principal Component Analysis (PCA) as a means of reducing the dimension of a high dimensional dataset and visualizing a high dimensional dataset in 2 or 3 dimensions
- What PCA does - PCA as choice of a new coordinate system capturing as much of the variation in the data as possible in the first several coordinates
- Simple examples and examples from visualization of tumor subtypes in microarray data
- When PCA is effective and when it is not
- PCA in terms of eigenvalues and eigenvectors of the appropriate covariance matrix
- PCA via the Singular Value Decomposition (SVD) of the data matrix
- Viewing an initial step in a distance geometry algorithm as the linear projection of a high dimensional dataset into 3 dimensions that preserves as much of the distance variation as possible, which is a PCA projection. Viewing this step in terms of the singular value decomposition of the matrix of the points in the higher dimensional space.

Alan E. Berger, Ph.D., JHBMC Lowe Family Genomics Core, Johns Hopkins University School of Medicine,
aberger9@jhmi.edu (410) 550-5089

An Introduction to Principal Component Analysis (PCA)

Outline

- Examples of visualization capturing as much of the variation in the data as possible (PCA)
- Illustration of how PCA works – selection of first axis capturing maximal variation
- Precise definition of "variation of the data"
- PCA might not be best for separating subgroups
- Examples of PCA on real data, left out dimensions do matter, comparison with hierarchical agglomerative clustering
- Scree plots for estimating how much variation in the data has been captured
- Matrix singular value decomposition (SVD) gives PCA coord.

Some References for Principal Component Analysis and Singular Value Decomposition (SVD)

[1] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice Hall, 1998, Chapter 8 – Principal Components

[2] I. T. Jolliffe, *Principal Component Analysis* 2nd Edition, Springer, 2002.

[3] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes (FORTRAN VERSION)* Cambridge University Press, 1989, Section 2.9

[4] G. Strang, The Fundamental Theorem of Linear Algebra *American Mathematical Monthly* 100 (1993), pp. 848-855. (explains the singular value decomposition in terms of the fundamental subspaces of a matrix (linear transformation) and its transpose)

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations* 2nd Edition, The Johns Hopkins University Press, 1989, Section 2.5 <Ed. 3, 1996>

[6] B. Carnahan, H. A. Luther and J. O. Wilkes, *Applied Numerical Methods*, John Wiley & Sons, 1969 – see "power method" to get several of the largest eigenvalues and corresponding eigenvectors of a real symmetric positive semidefinite matrix (e.g., a moderate size covariance matrix).
(in general should use a robust algorithm to obtain the SVD)

[7] G. W. Stewart, On the Early History of the Singular Value Decomposition, *SIAM Review* 35 (1993), pp. 551-566.

[8] M. E. Wall, A. Rechtsteiner and L. M. Rocha, Singular Value Decomposition and Principal Component Analysis, in *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky and M. Granzow, eds. pp. 91-109, Kluwer, Norwell, MA (2003)
web page: http://public.lanl.gov/mewall/kluwer2002.html

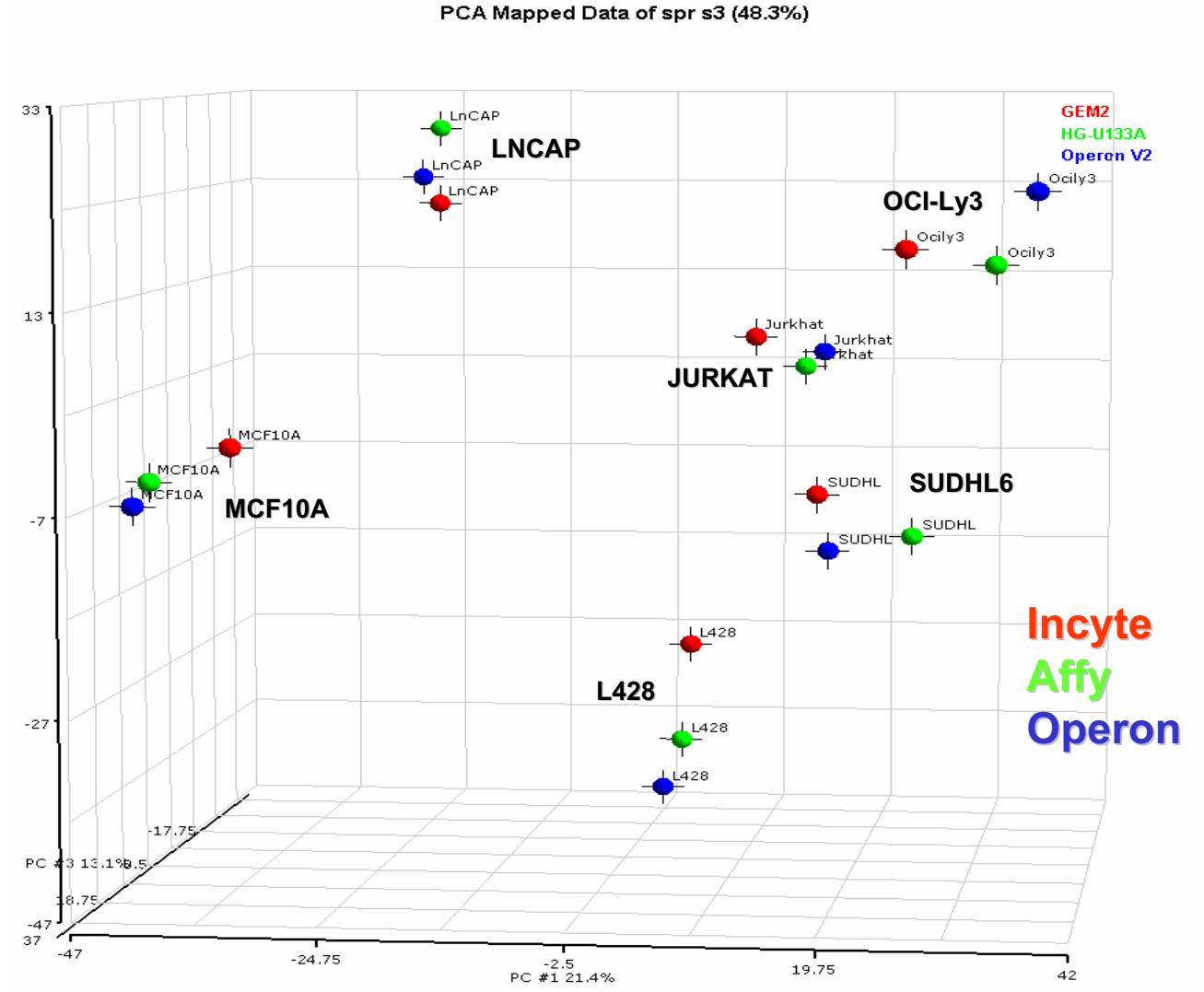


Figure 3
Principal Component Analysis (PCA) of the three microarray platforms and six cell lines using expression of the 3186 genes with signals above background.

may be due to simple errors in gene identification, rather than to the technologies of the platforms. The Incyte library is guaranteed by the manufacturer to be only 90% correct, and an unknown percentage of the Operon and Affymetrix oligonucleotides may have been designed on the basis of incorrect sequences in the public databases. Indeed, we found one oligonucleotide in the Operon set that was apparently designed from an EST sequence that has since been withdrawn from the UniGene database (see RT-PCR studies below). In any case, the concordance is quite high across all platforms with this method of analysis as well as with the others.

Quantitative real-time RT-PCR

In a pilot study with the three platforms, we compared and contrasted gene expression values for only the cell lines MCF10A and LNCaP. RT-PCR data for twelve genes are shown in Figure 6. Most of the values are in reasonable agreement except that there are differences in the *magnitudes* of the expression ratios. As found in other studies, the RT-PCR values are generally higher, probably because ratios are "flattened" with the microarray platforms. Affymetrix ratios are sometimes higher, but that may simply reflect the method of quantitation used in their analysis. The cDNA array ratios are generally lower than those

Sample Section of a Gene Expression Level Matrix

data_set_ALL_AML_train.xls [Read-Only]

1	Gene Description // Tissue Sample # --->	1	2	3	4	5	6	7	8	9
130	Niemann-Pick C disease protein (NPC1) mRNA	54	117	175	235	150	142	221	217	141
131	GB DEF = Angelman Syndrome Gene, E6-AP ubic	180	372	491	528	162	446	527	893	480
132	RET ligand 2 (RETL2) mRNA	373	480	856	506	213	414	772	836	449
133	GB DEF = Delayed rectifier potassium channel (K\	484	485	159	597	155	501	774	982	341
134	GB DEF = Secretory carrier membrane protein (SC	99	21	6	33	110	76	70	76	98
135	Poly(ADP-ribose) glycohydrolase (hPARG) mRNA	65	92	163	139	107	76	68	40	98
136	GB DEF = Importin alpha 6 mRNA	60	62	30	6	16	35	13	56	62
137	Caspase-like apoptosis regulatory protein 2 (clarp)	451	647	842	531	514	487	403	671	492
138	ATF family member ATF6 (ATF6) mRNA	197	119	293	118	200	44	181	206	93
139	Fas-binding protein (DAXX) mRNA, partial cds	946	178	2011	515	478	553	1073	2126	743
140	Arp2/3 protein complex subunit p41-Arc (ARC41) m	1190	866	1408	948	1129	983	1287	566	813
141	Arp2/3 protein complex subunit p20-Arc (ARC20) m	370	1466	1334	552	711	227	530	527	1693
142	GB DEF = RGS3 mRNA, 5' UTR	300	17	330	141	173	156	628	410	451
143	MDM2-like p53-binding protein (MDMX) mRNA	330	43	428	449	122	245	144	178	169
144	Bet1p homolog (hbet1) mRNA	141	21	57	87	244	64	465	374	75
145	Dolichol monophosphate mannose synthase (DPM	260	362	400	361	389	156	283	203	310
146	Phospholipid scramblase mRNA	124	65	102	234	24	16	109	55	90
147	GB DEF = Syntaxin-16C mRNA	339	453	297	403	440	357	317	568	296
148	GB DEF = TEB4 protein mRNA	288	95	93	142	214	64	77	266	14
149	GB DEF = Luman mRNA	1032	635	1175	1079	658	1365	1424	1520	1506

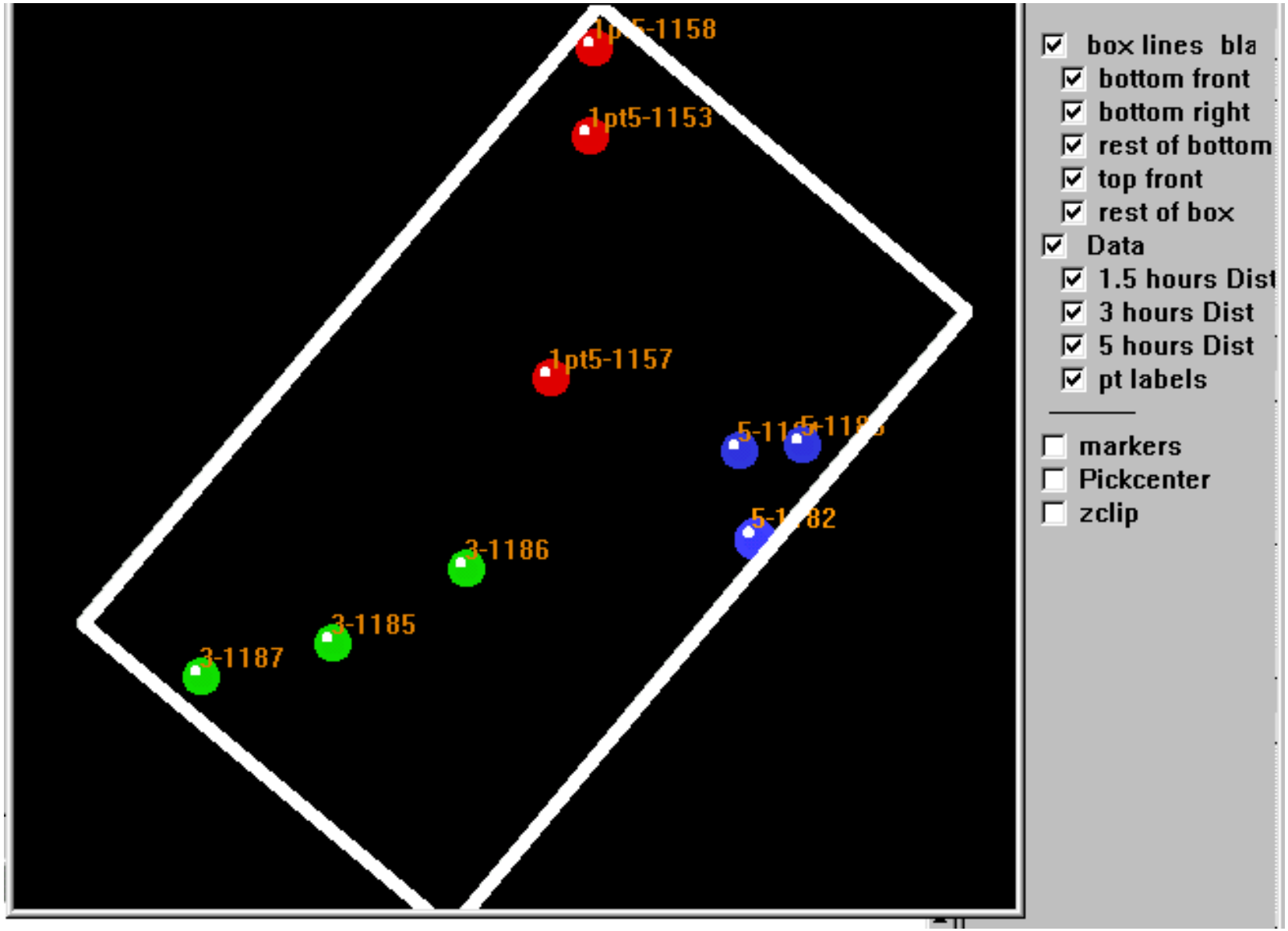
data_set_ALL_AML_train

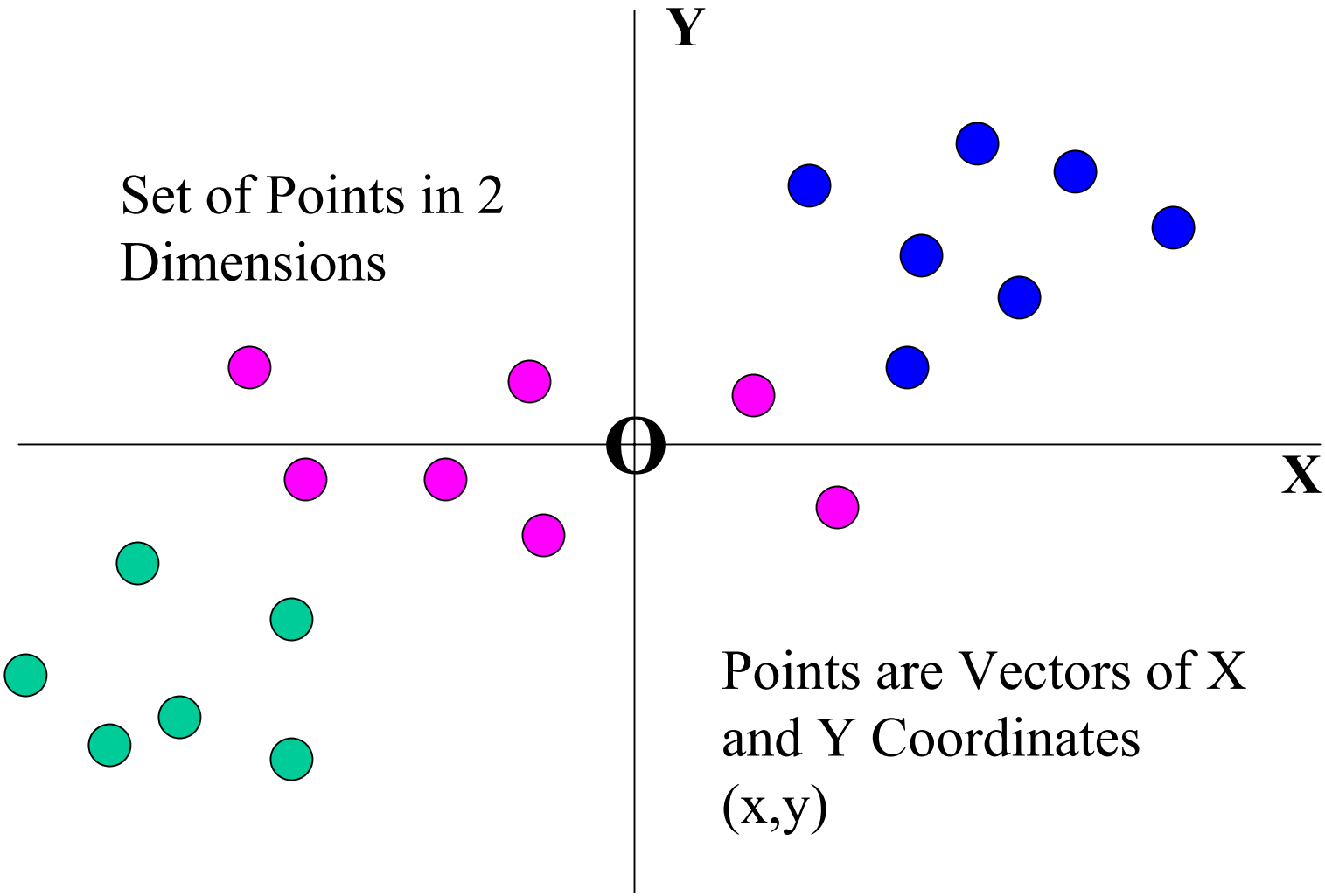
Rows are expression levels of 1 gene, columns are expression profile from 1 microarray chip (here a tissue sample)
(modified from Golub et al. ALL/AML data)

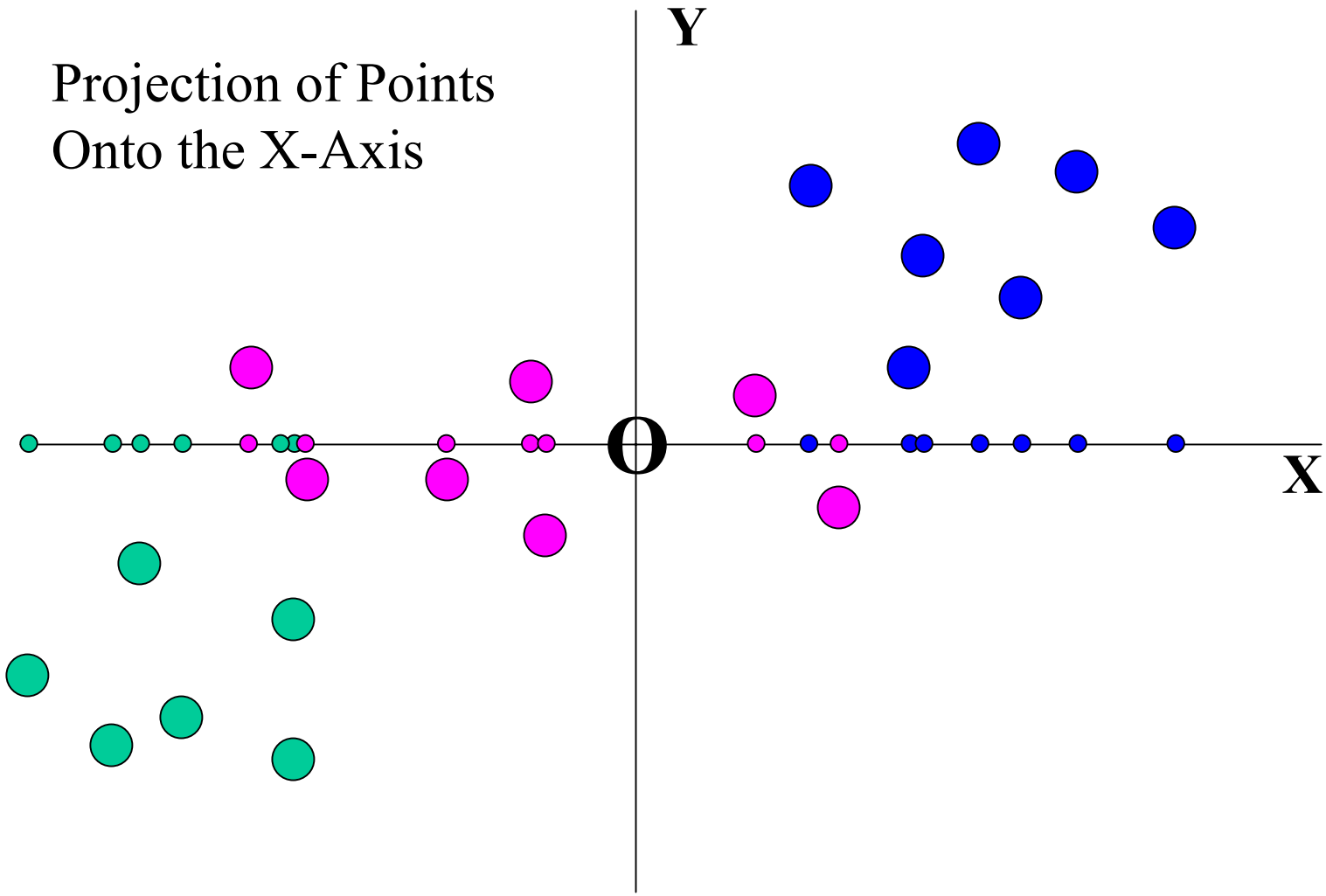
Microarray Expression Data

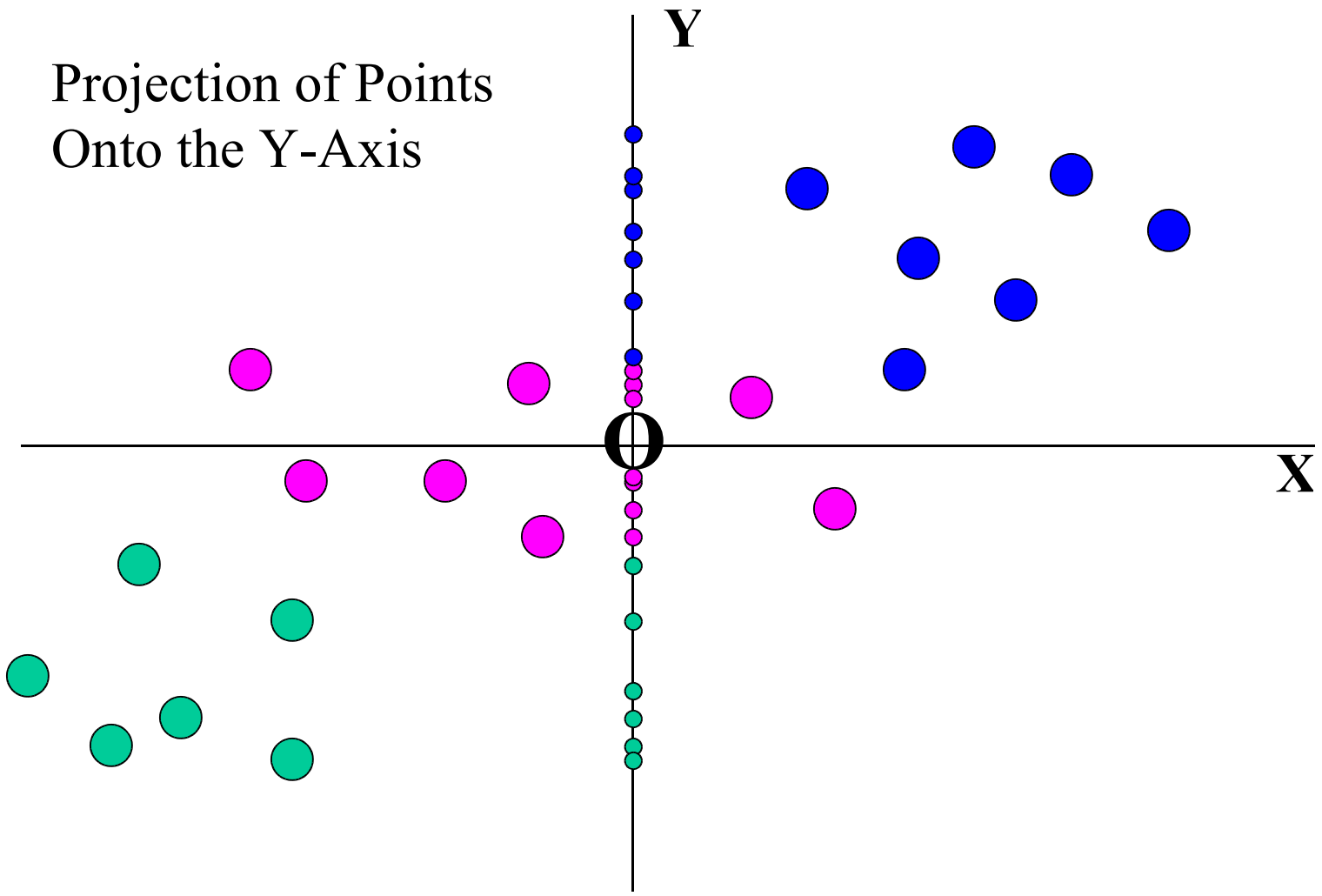
	Sample 1 …	Sample s …	Sample N
gene 1			
gene 2	entry (g,s) of the expression level matrix L contains the expression level L(g,s) of of gene (probe set / spot) g in sample s		
…			
gene g	a *sample* can be, e.g., control cells, treated cells, cells from a particular tissue or disease state		
…			
gene M	for Yale Chip, M = 15,250, for Affymetrix HG-U133a (b) Chip, M = 22,283 (22,645)		

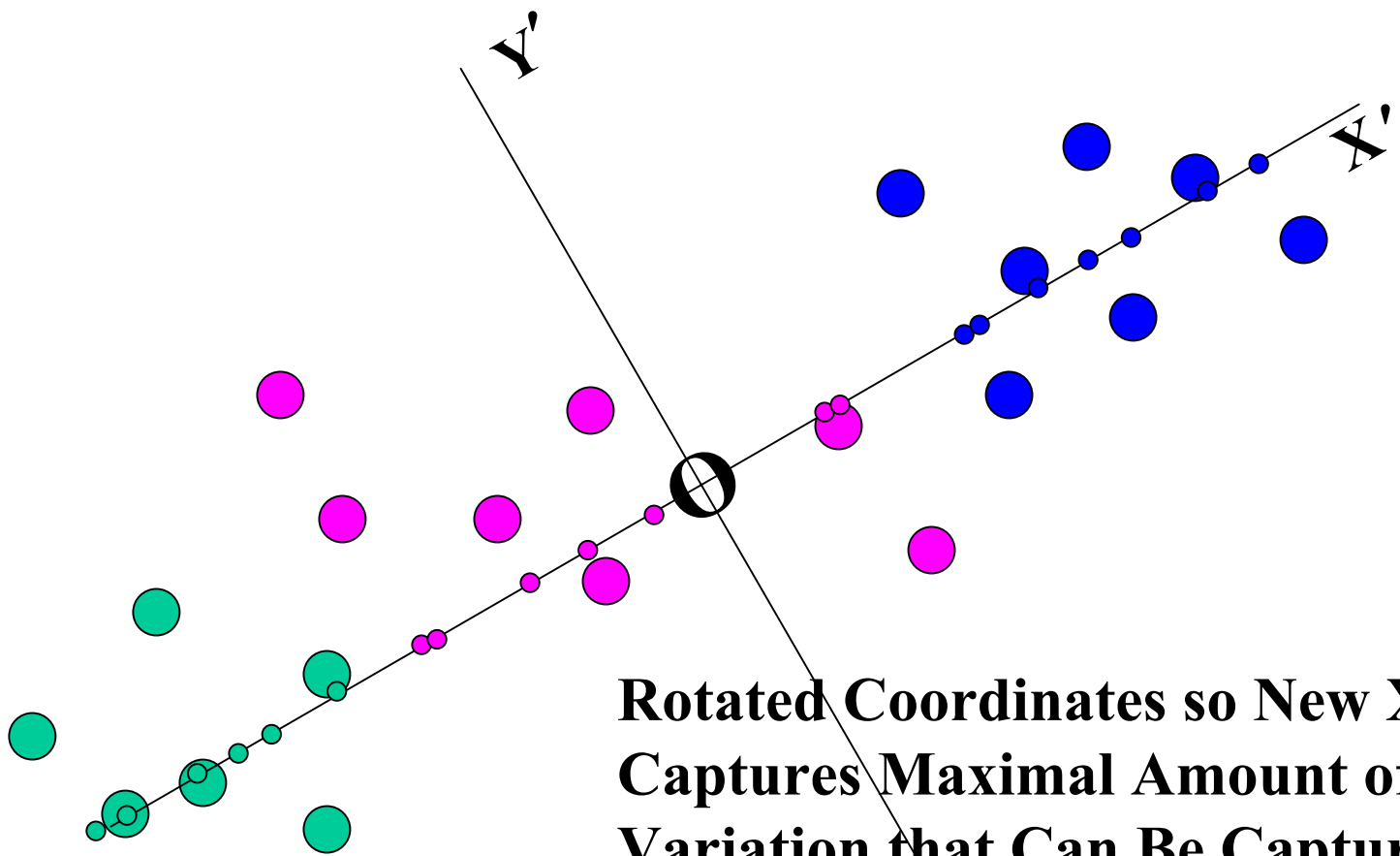
PCA of Three Replicate Chips at Three Times (1.5, 3, 5 hours)




Y
Set of Points in 2 Dimensions
O
X
Points are Vectors of X and Y Coordinates (x,y)

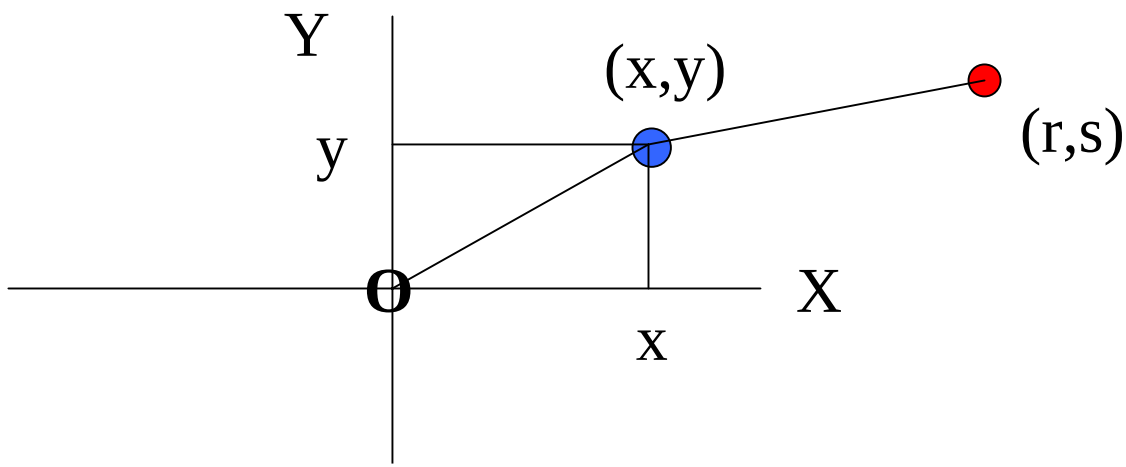

Projection of Points
Onto the X-Axis
Y
O
X


Projection of Points
Onto the Y-Axis
Y
O
X



Rotated Coordinates so New X-Axis Captures Maximal Amount of Variation that Can Be Captured in One Direction

Distance from a point to the origin **0**
Distance between two points



Distance from the point with **coordinates** (x,y) to **O** is sqrt(x*x + y*y)

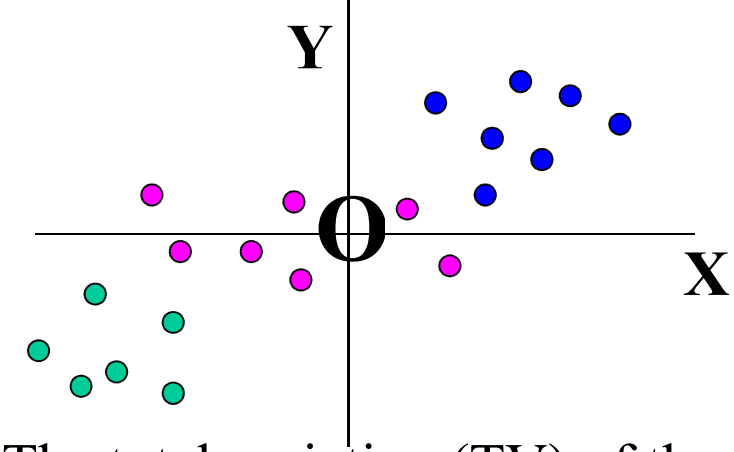
Distance^2 from (X,Y,…,L) to **O** is $X^2 + Y^2 + \cdots + L^2$

Distance^2 between (x,y) and (r,s) is $(x-r)^2 + (y-s)^2$

Total Variation TV = Sum of Distances^2 of Points to Origin =
$TV_x + TV_y + TV_z + \cdots$

TV_x = sum of squares of the X coordinates of all the points, etc.

	point # 1	2	3	…	n	
X-coordinate values →	X_1	X_2	X_3	$\cdots$	X_n	TV_x
Y-coordinate values →	Y_1	Y_2	Y_3	$\cdots$	Y_n	TV_y
Z-coordinate values →	Z_1	Z_2	Z_3	$\cdots$	Z_n	TV_z



Variation of the Data

The total variation (TV) of the data (about the origin **0** – in general will assume the average (center of mass) of the points is at **0**) = the sum of the squares of the distances of the points to **0**. Note that

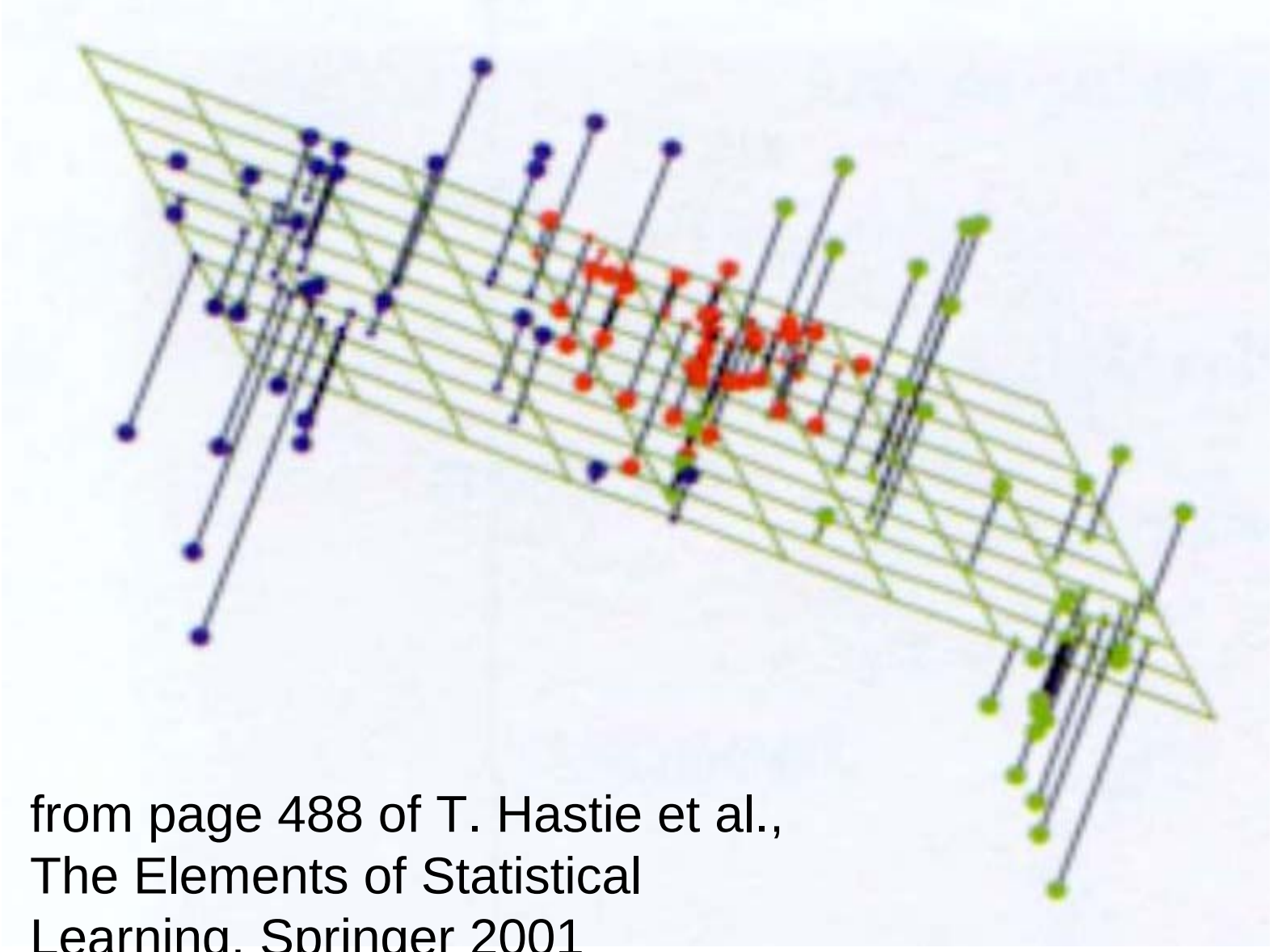
$$TV = TV_x + TV_y + \ldots$$

where TV_x = sum of the squares of the x coordinates of all the points, etc.

PCA picks the new x-axis to maximize TV_x (relative to the new x axis), then over all directions perpendicular to the new x axis PCA picks the new y axis to maximize TV_y (relative to the new y axis), etc.

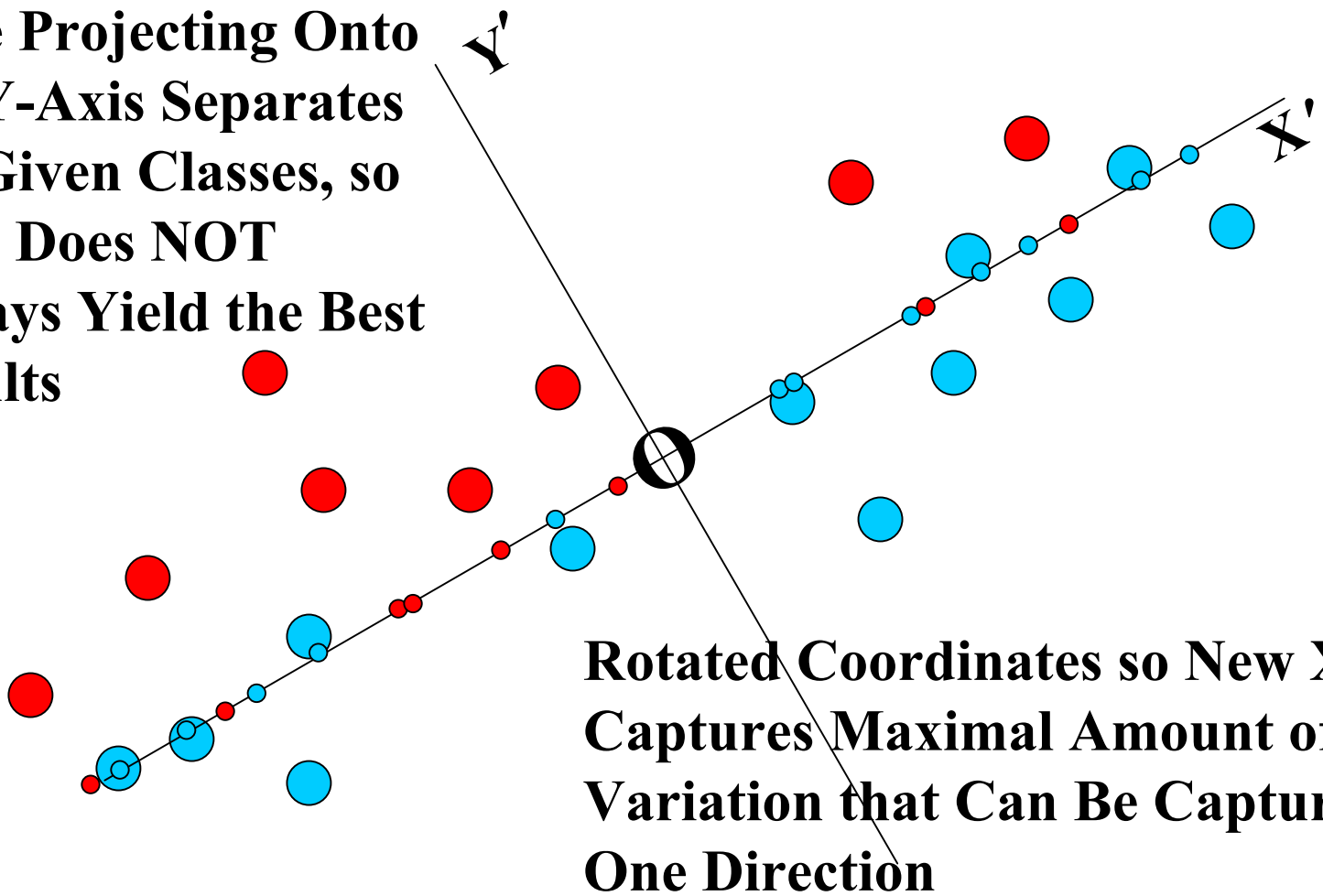
Note while the individual TV_{axis} values change, TV is unchanged.

PCA Projection Down to 2 Dimensions



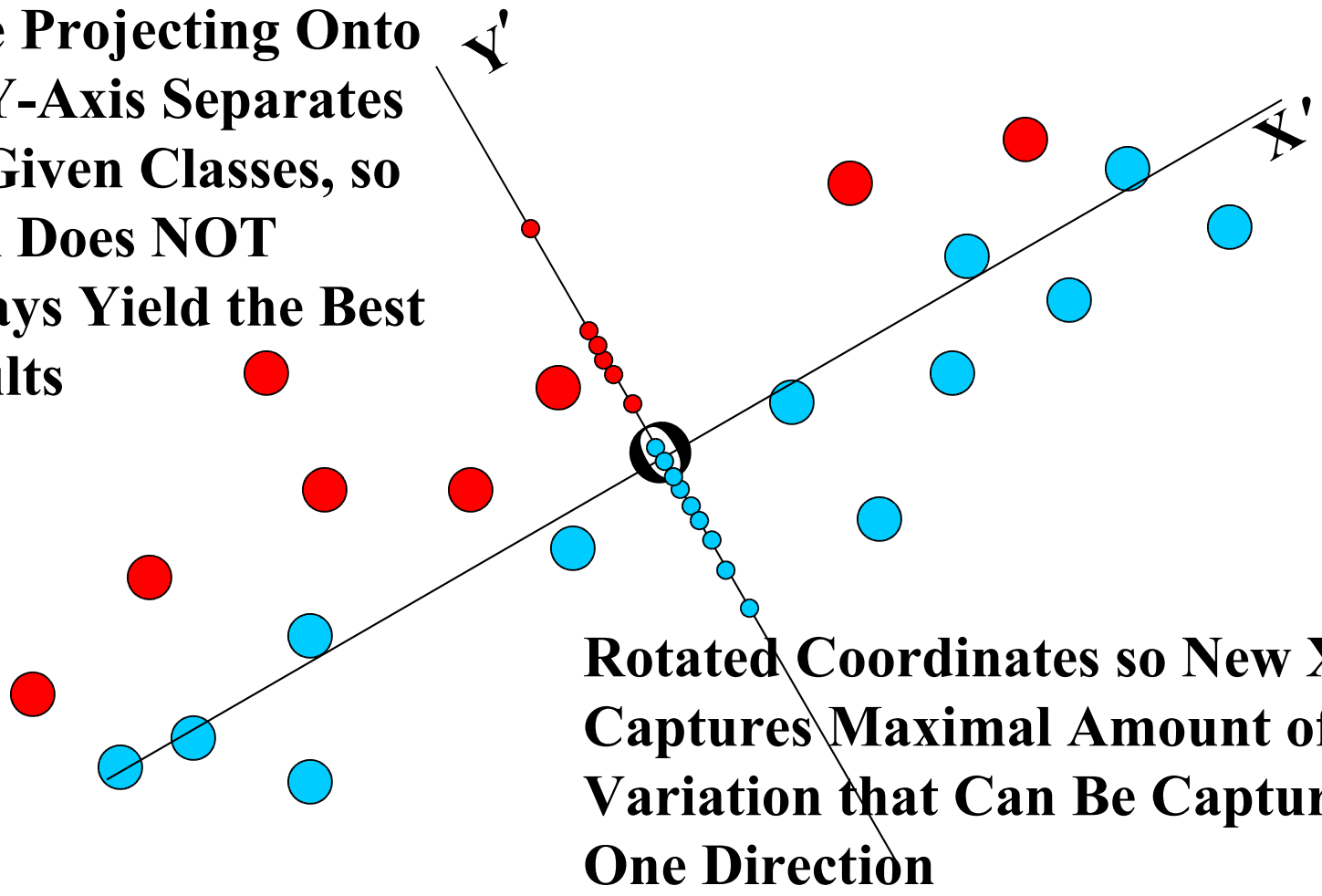
from page 488 of T. Hastie et al., The Elements of Statistical Learning, Springer 2001

Here Projecting Onto the Y-Axis Separates the Given Classes, so PCA Does NOT Always Yield the Best Results



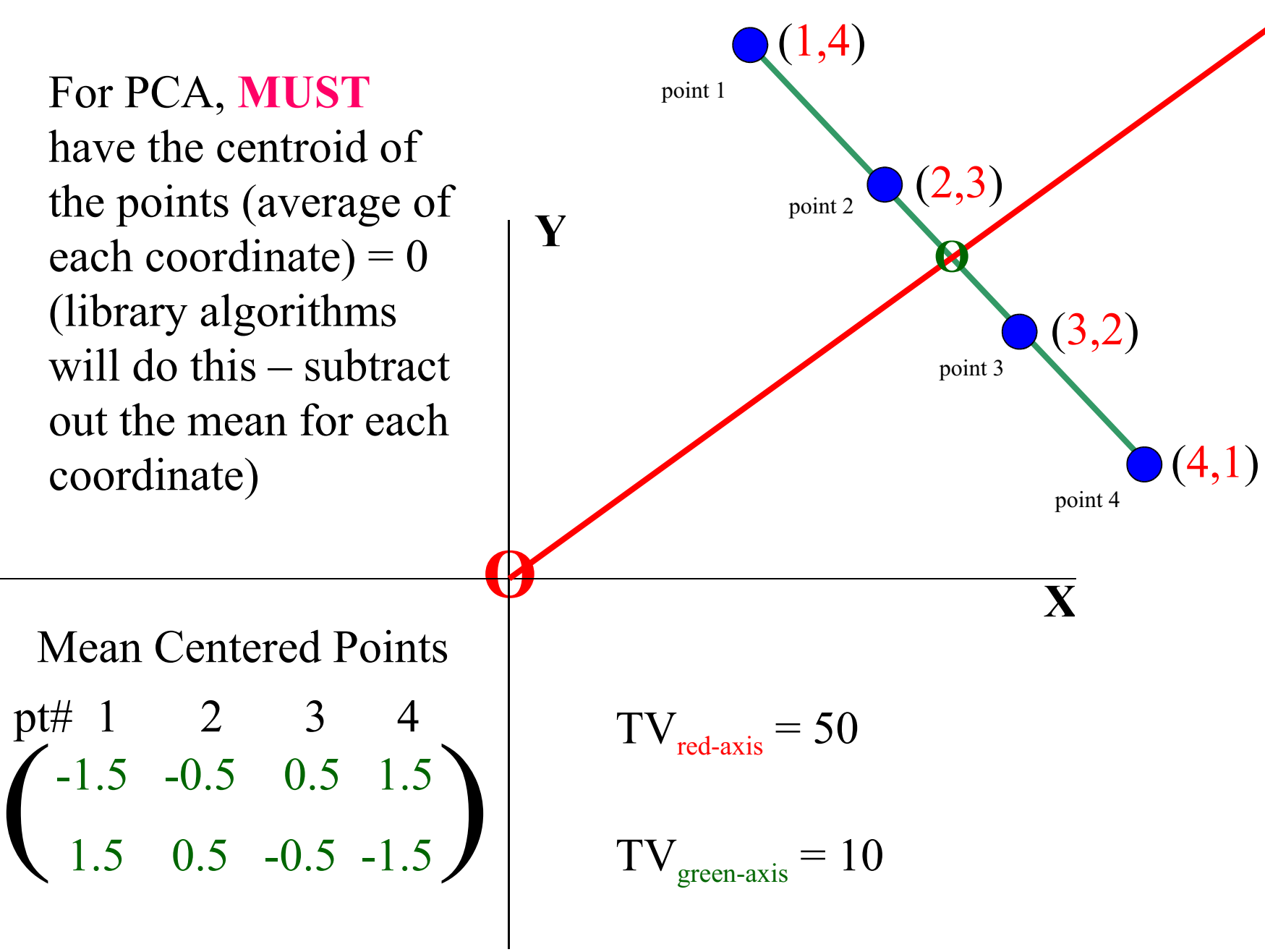
Rotated Coordinates so New X-Axis Captures Maximal Amount of Variation that Can Be Captured in One Direction

Here Projecting Onto the Y-Axis Separates the Given Classes, so PCA Does NOT Always Yield the Best Results



Rotated Coordinates so New X-Axis Captures Maximal Amount of Variation that Can Be Captured in One Direction

For PCA, **MUST** have the centroid of the points (average of each coordinate) = 0 (library algorithms will do this – subtract out the mean for each coordinate)



Mean Centered Points

pt# 1 2 3 4

$$\begin{pmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ 1.5 & 0.5 & -0.5 & -1.5 \end{pmatrix}$$

$TV_{\text{red-axis}} = 50$

$TV_{\text{green-axis}} = 10$

Part of the Anderson Fisher Iris Data Set

5.1	3.5	1.4	0.2	1
4.9	3.0	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5.0	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1
4.6	3.4	1.4	0.3	1
5.0	3.4	1.5	0.2	1
.....				
7.0	3.2	4.7	1.4	2
6.4	3.2	4.5	1.5	2
6.9	3.1	4.9	1.5	2
5.5	2.3	4.0	1.3	2
6.5	2.8	4.6	1.5	2
5.7	2.8	4.5	1.3	2
6.3	3.3	4.7	1.6	2
4.9	2.4	3.3	1.0	2
.....				
6.3	3.3	6.0	2.5	3
5.8	2.7	5.1	1.9	3
7.1	3.0	5.9	2.1	3
6.3	2.9	5.6	1.8	3
6.5	3.0	5.8	2.2	3
7.6	3.0	6.6	2.1	3
4.9	2.5	4.5	1.7	3
7.3	2.9	6.3	1.8	3
.....				

column 5 1 = Setosa
2 = Versicolor
3 = Verginica

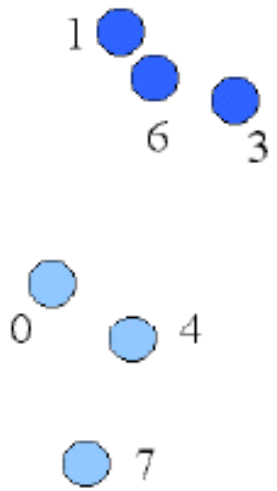
full data set is
50 samples of each iris flower species
(data from R. A. Johnson and D. W. Wichern [1])

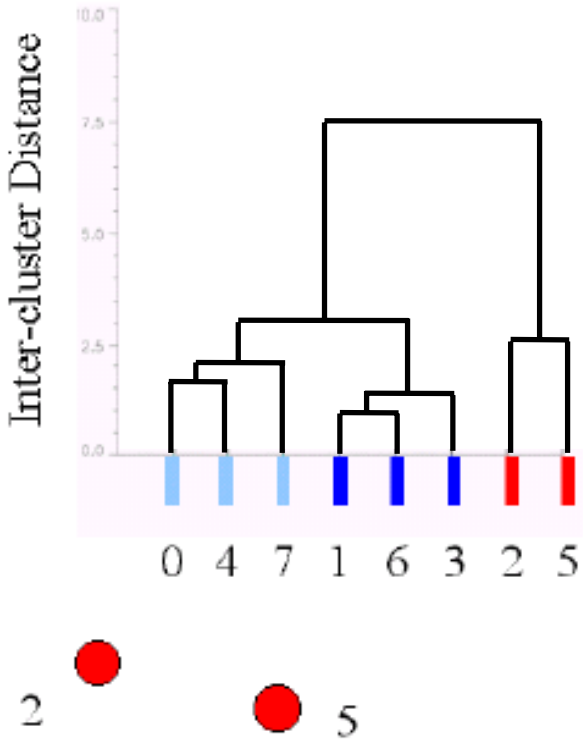
each row is the data from one flower

columns 1, 2, 3, 4 are measured properties
of each flower (sepal length, sepal width,
petal length, petal width)

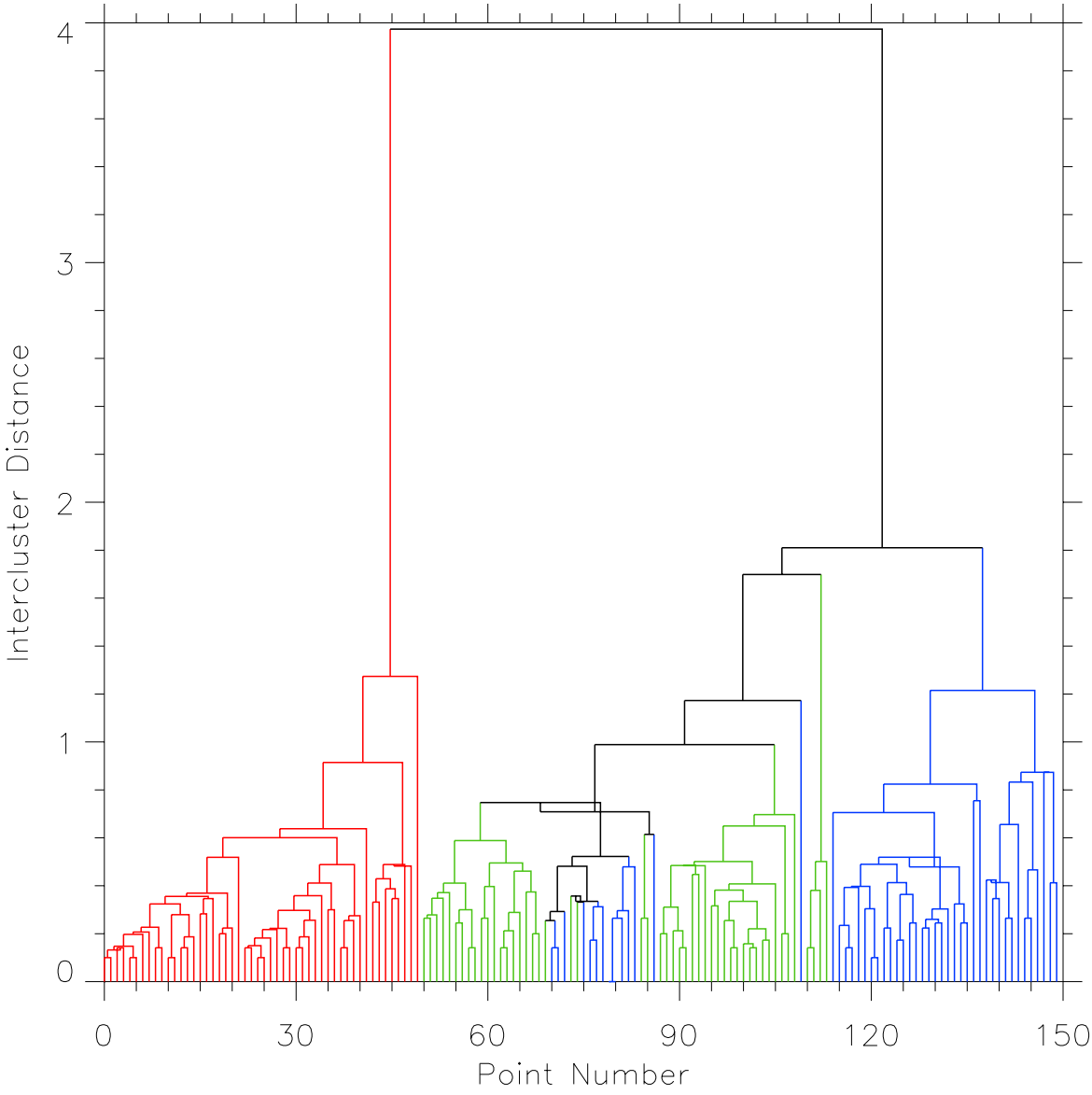
Botany definitions: the *calyx* is the outermost
group of floral parts, usually green; *sepals*
are the individual leaves or parts of the
calyx

Example of Single Linkage Hierarchical Clustering



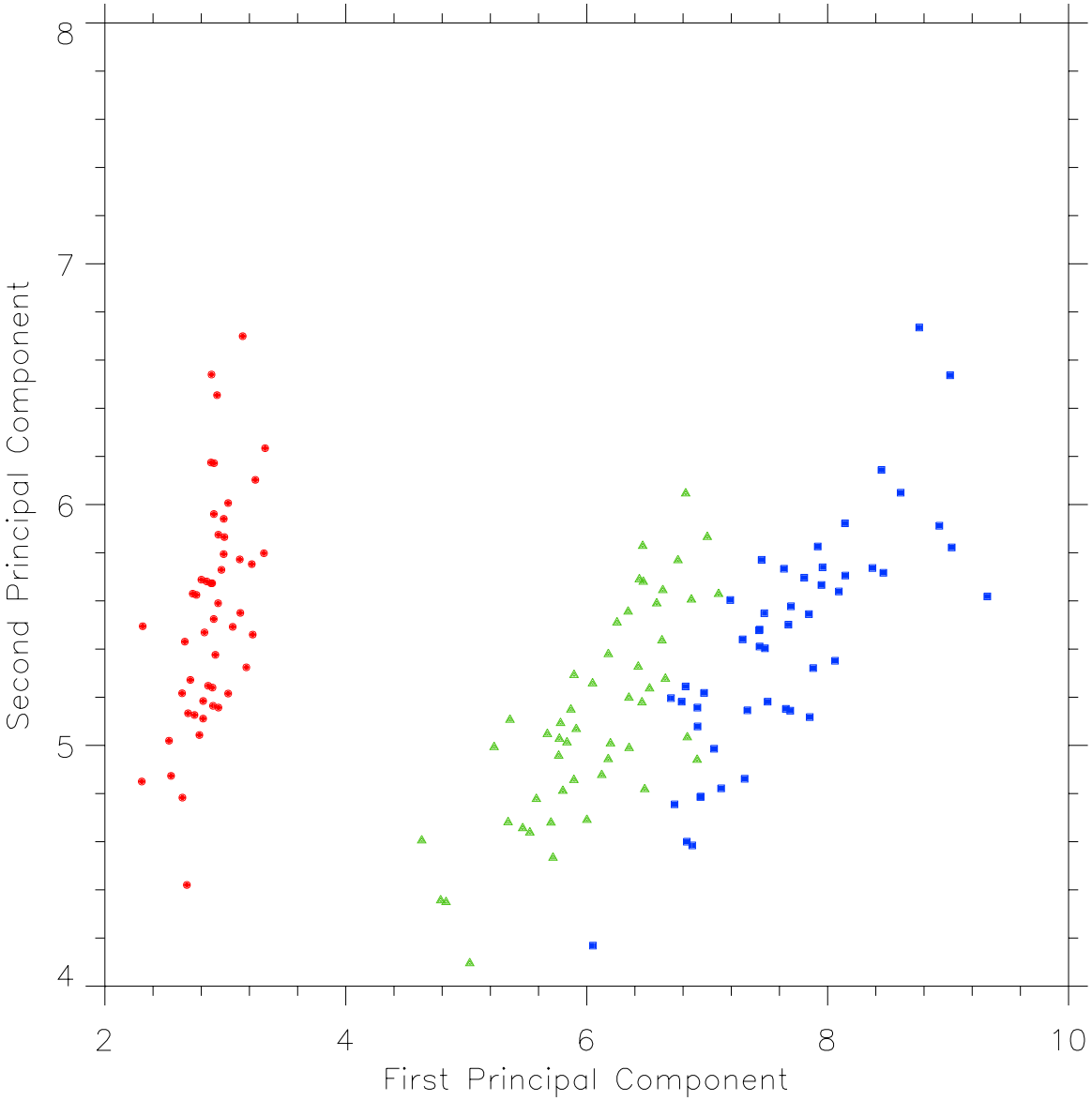


Centroid Hierarchical Clustering of Anderson Iris Data



plotted Wed May 29 12:21:09 2002 cenhieriris.pro--May 29, 2002 cenhiermay29

Principal Component Plot of Anderson Iris Data



cov evalues = 4.228 0.243 0.078 0.024

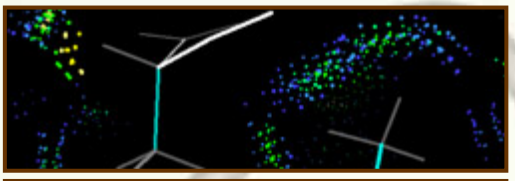
Fri Sep 21 12:36:00 2001 irispc1.pro September 21, 2001 irispc1cov.ps

data from Table 11.5 of Johnson and Wichern 1998

3D Macromolecule Analysis & Kinemage Home Page / at the Richardson Laboratory

Kinemages: 3D interactive graphics

"The Mage Page", kinemage list, KiNG manual, Mage help, ...

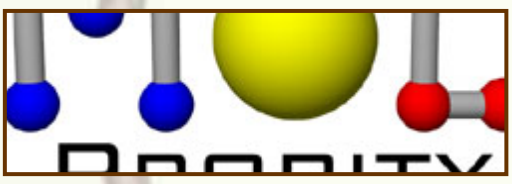


Software: Display; Contacts; & Utilities

Download links, manuals, software descriptions, ...

Databases: protein & nucleic acid compilations

penultimate rotamer library, Top500 proteins with optimized H atoms, RNA backbone rotamers, 100 high resolution proteins with optimized H atoms and fixes, ...

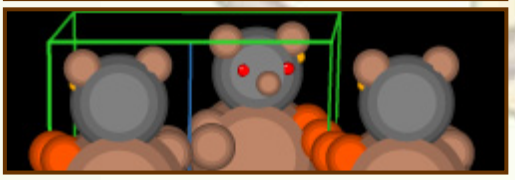


MolProbity: structure validation

Web service for all-atom contact and geometrical analysis of your model

Research: 3D structure of proteins & nucleic acids

All-atom contacts; structure improvement; backbone motions; sidechain rotamers; "RNA Backbone is Rotameric"



Teaching: Course materials

Duke courses: BCH222, BCH258, BCH291, and workshops on Model Quality and software tool use at various places, ...

Gallery: images

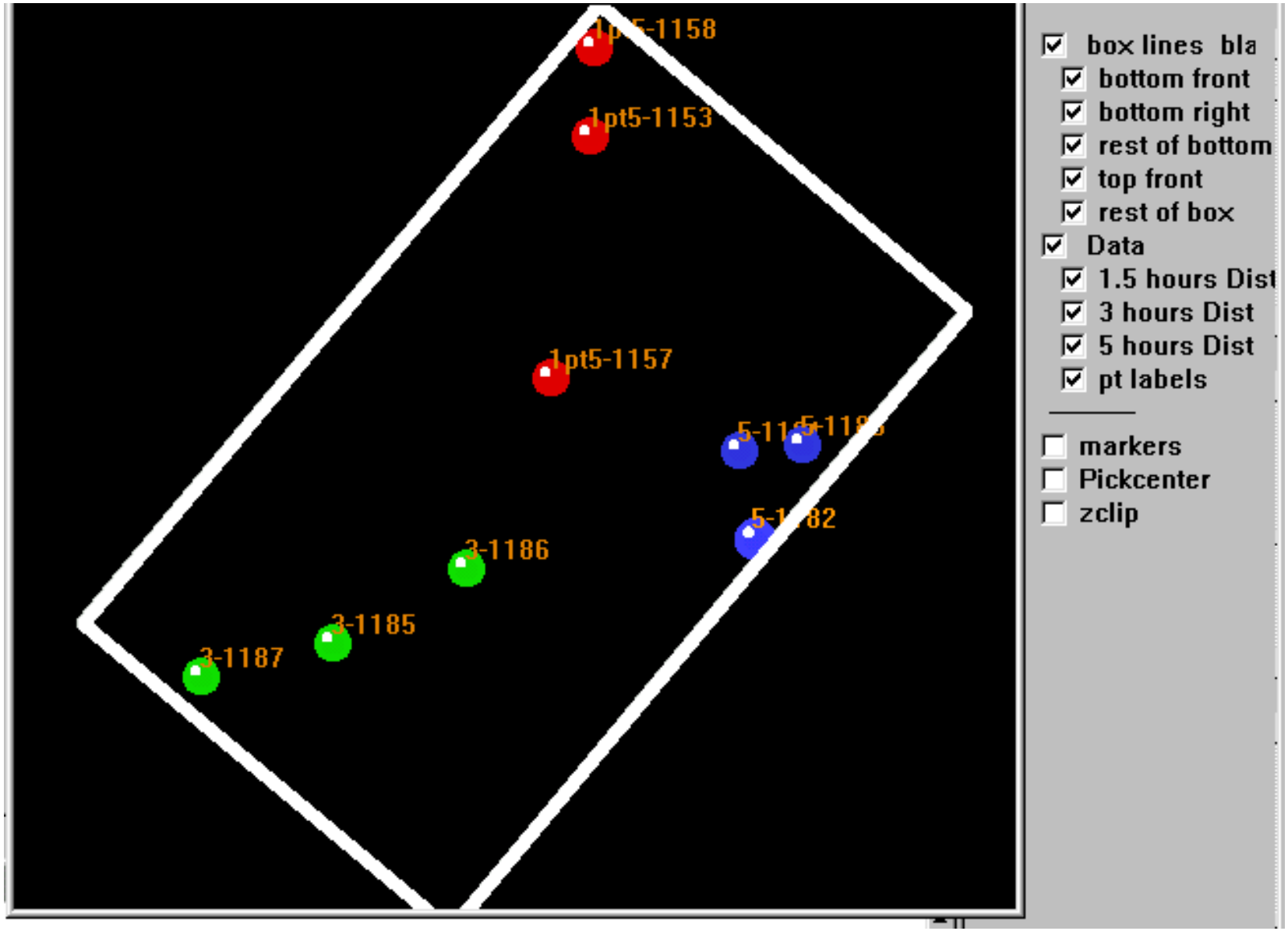
2D images, annotated Anatomy and Taxonomy of Protein Structure (in progress)

About us: Lab Info

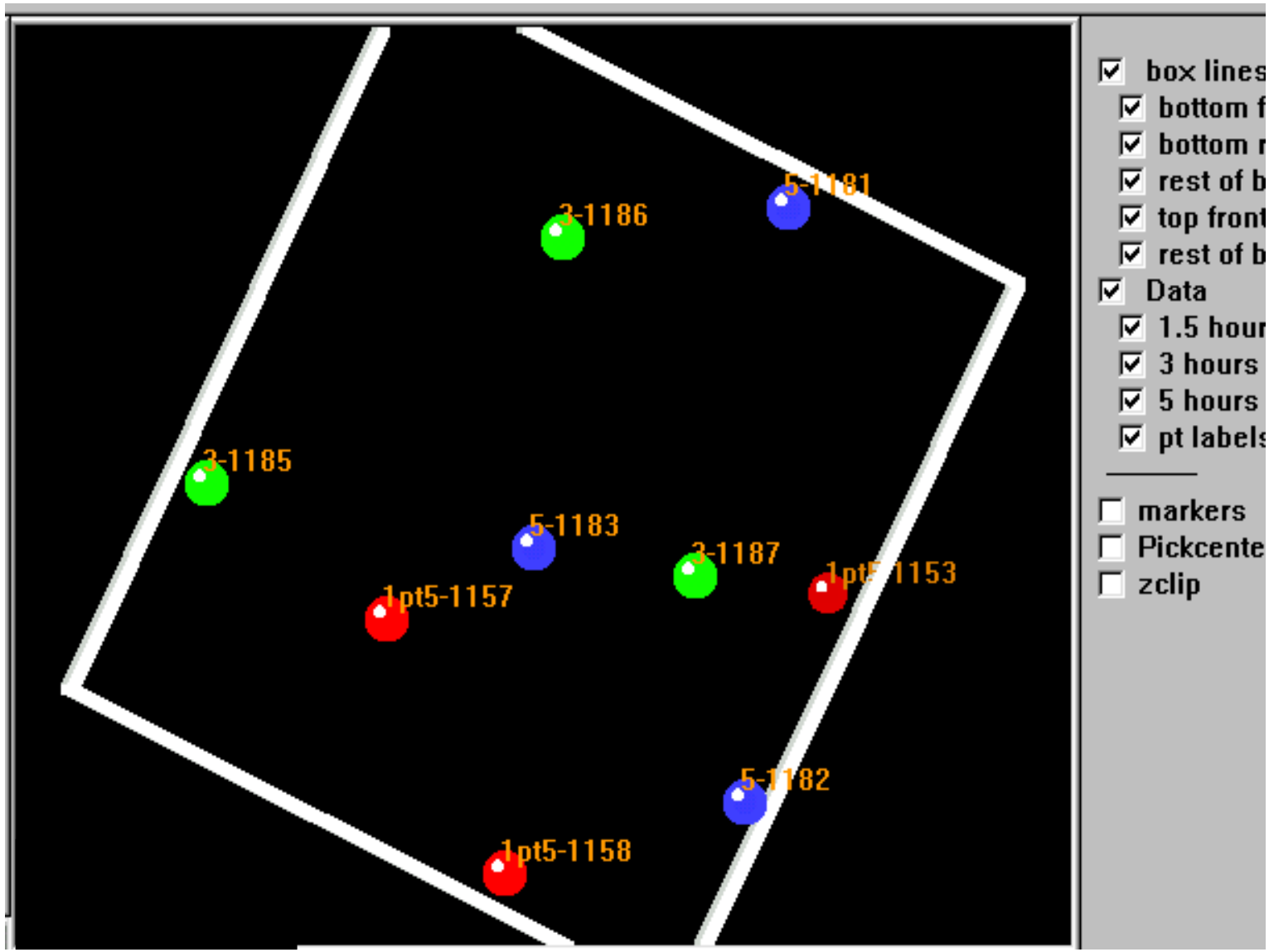
Contacts; travel; publications; SECSG, Biochem. Dept., Biophysics

"This page generated by:"
NIH Grant GM-15000, funding Richardson Lab research for over 34 years; and
NIH Grant GM-61302, funding RLab for over 3 years.

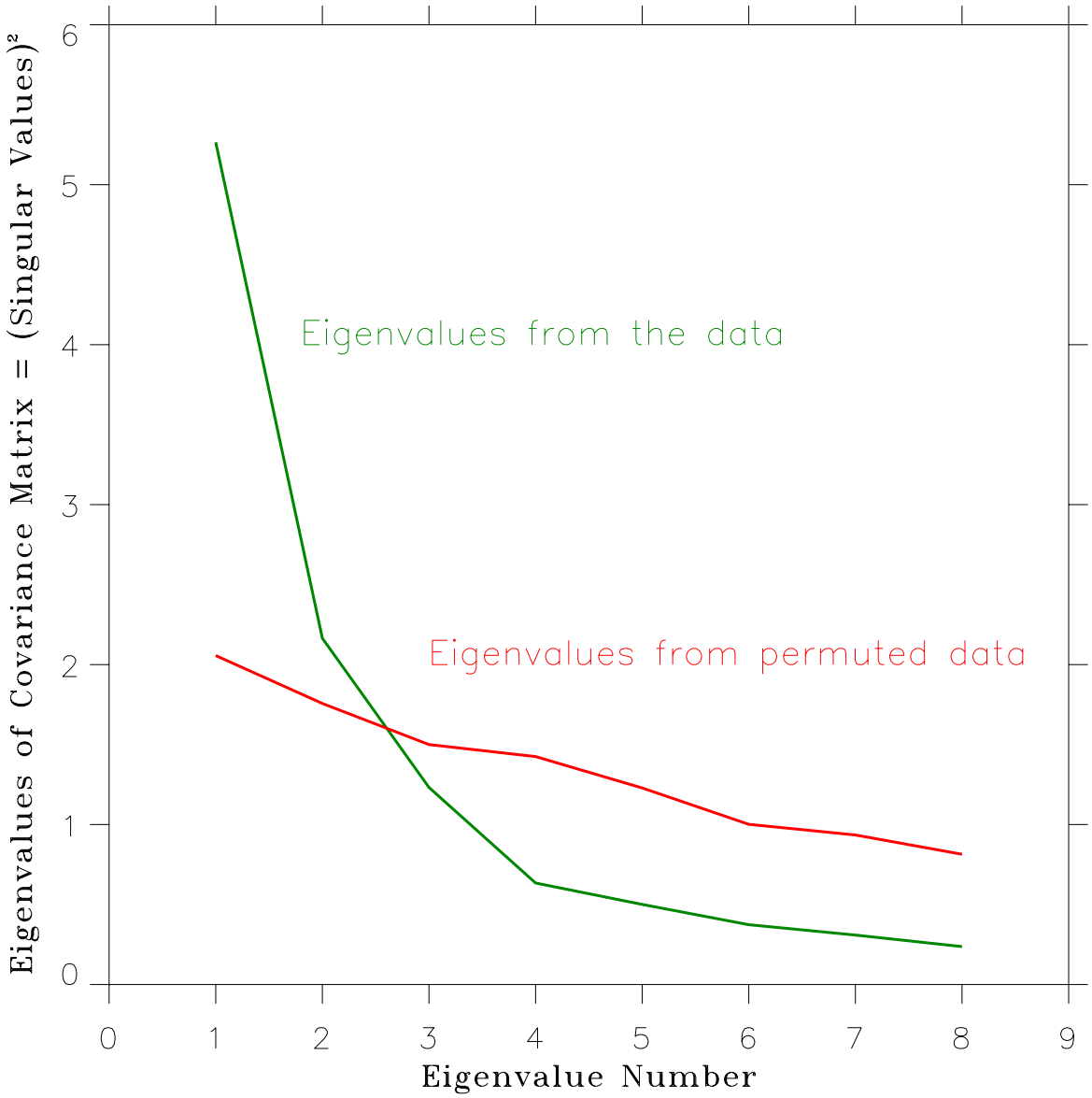
PCA of Three Replicate Chips at Three Times (1.5, 3, 5 hours)



PCA of Permuted Data from Three Replicate Chips at Three Times (1.5, 3, 5 hours)



Scree Plot for Gutting Data and Permutation

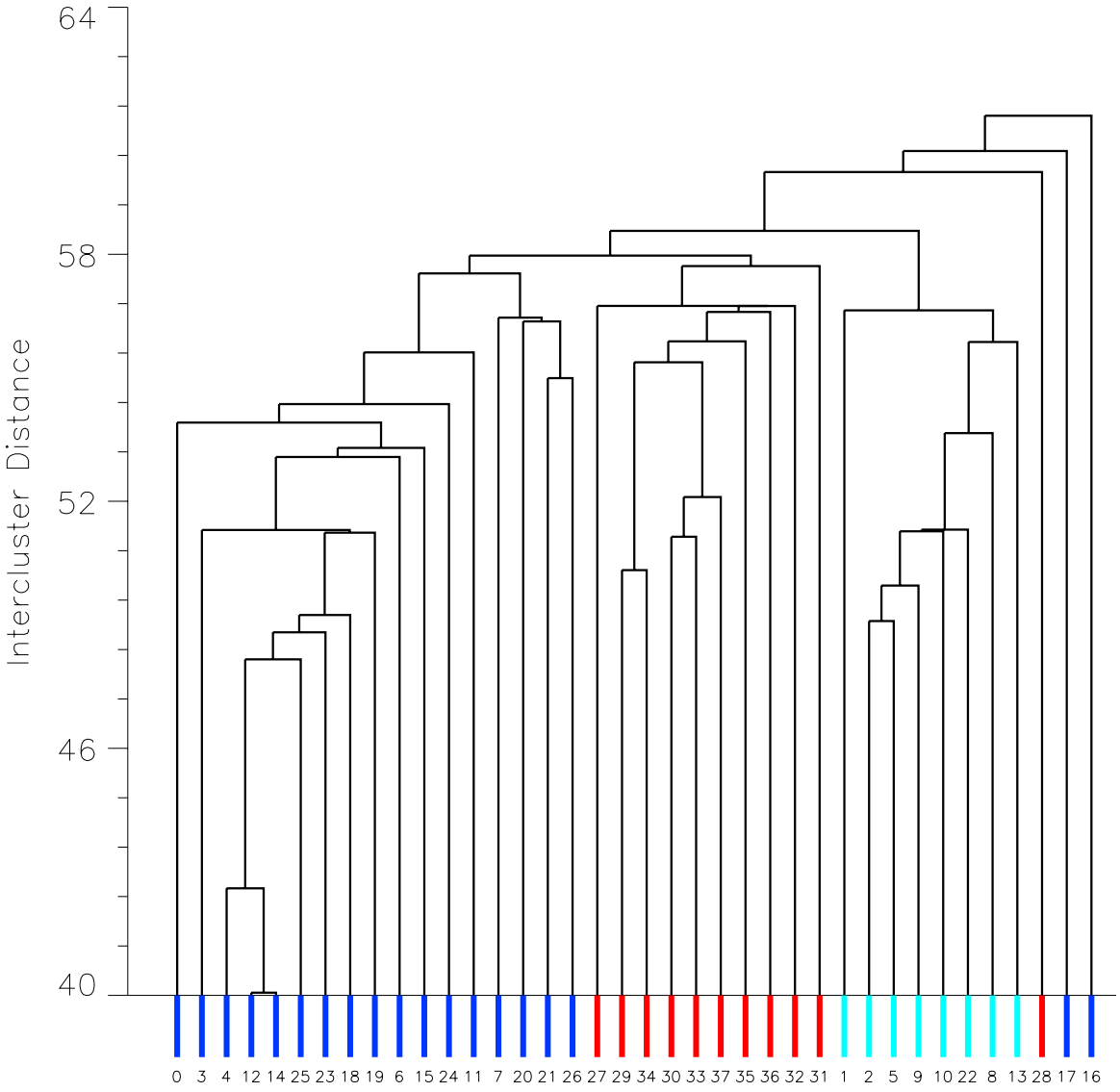


Fri Jun 10 23:37:52 2005 screeplot.pro June 10, 2005

Variances in Principal Coordinate Axes = Eigenvalues of Covariance Matrix

C:\berger\biochemlmuulmay03data\PCAcourse\screeplot.ps

T. Golub et al. ALL/AML Microarray Data Clusters



Single Linkage ℓn(data) keep top 600 var genes

Acute Lymphoblastic Leukemia (ALL): B−cell T−cell

Acute Myeloid Leukemia AML

nnlines =

plotted Tue Oct 08 02:08:08 2002 plothierallaml3n.pro−−Oct 8, 2002

C:\berger\biochem\golubdata\allaml1trainfullploth3n−single.ps

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

The challenge of cancer treatment has been to target specific therapies to pathogenetically distinct tumor types, to maximize efficacy and minimize toxicity. Improvements in cancer classification have thus been central to advances in cancer treatment. Cancer classification has been based primarily on morphological appearance of the tumor, but this has serious limitations. Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. In a few cases, such clinical heterogeneity has been explained by dividing morphologically similar tumors into subtypes with distinct pathogeneses. Key examples include the subdivision of acute leukemias, non-Hodgkin's lymphomas, and childhood "small round blue cell tumors" [tumors with variable response to chemotherapy (*1*) that are now molecularly subclassified into neuroblastomas, rhabdomyosarcoma, Ewing's sarcoma, and other types (*2*)]. For many more tumors, important subclasses are likely to exist but have yet to be defined by molecular markers. For example, prostate cancers of identical grade can have widely variable clinical courses, from indolence over decades to explosive growth causing rapid patient death. Cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtypes. Here we describe such an approach based on global gene expression analysis.

We divided cancer classification into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes, which could reflect current states or future outcomes.

We chose acute leukemias as a test case. Classification of acute leukemias began with the observation of variability in clinical outcome (*3*) and subtle differences in nuclear morphology (*4*). Enzyme-based histochemical analyses were introduced in the 1960s to demonstrate that some leukemias were periodic acid-Schiff positive, whereas others were myeloperoxidase positive (*5*). This provided the first basis for classification of acute leukemias into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL) or from myeloid precursors (acute myeloid leukemia, AML). This classification was further solidified by the development in the 1970s of antibodies recognizing either lymphoid or myeloid cell surface molecules (*6*). Most recently, particular subtypes of acute leukemia have been found to be associated with specific chromosomal translocations—for example, the t(12;21)(p13;q22) translocation occurs in 25% of patients with ALL, whereas the t(8;21)(q22;q22) occurs in 15% of patients with AML (*7*).

Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis. Rather, current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remains imperfect and errors do occur.

Distinguishing ALL from AML is critical for successful treatment; chemotherapy regimens for ALL generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas most AML regimens rely on a backbone of daunorubicin and cytarabine (*8*). Although remissions can be achieved using ALL therapy for AML (and vice versa), cure rates are markedly diminished, and unwarranted toxicities are encountered.

We set out to develop a more systematic approach to cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays (*9*). It has been suggested (*10*) that such microarrays could provide a tool for cancer classification. Microarray studies to date (*11*), however, have primarily been descriptive rather than analytical and have focused on cell culture rather than primary patient material, in which genetic noise might obscure an underlying reproducible expression pattern.

We began with class prediction: How could one use an initial collection of samples belonging to known classes (such as AML and ALL) to create a "class predictor" to classify new, unknown samples? We developed an analytical method and first tested it on distinctions that are easily made at the morphological level, such as distinguishing normal kidney from renal cell carcinoma (*12*). We then turned to the more challenging problem of distinguishing acute leukemias, whose appearance is highly similar.

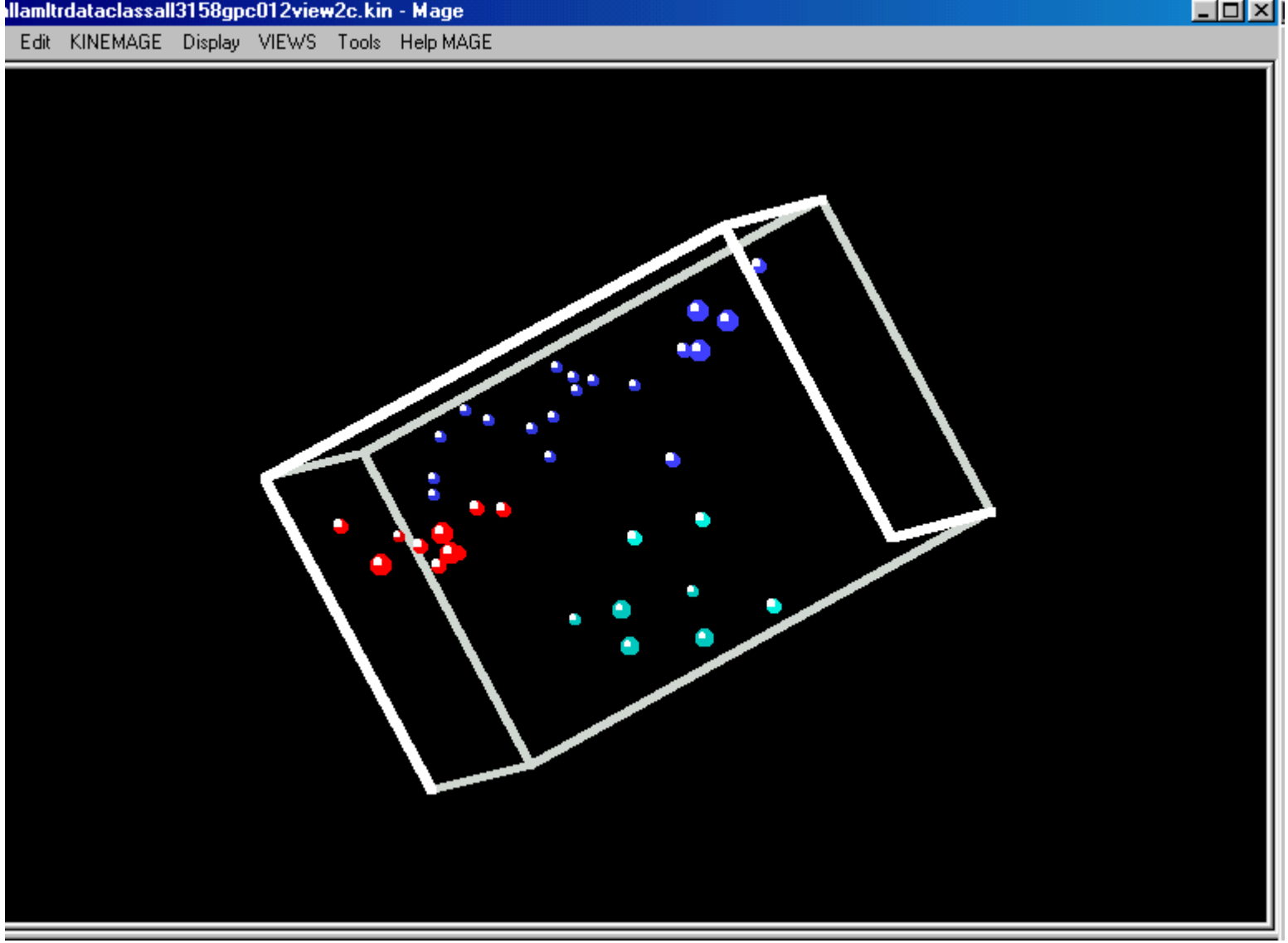
Our initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis (*13*). RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6817 human genes (*14*). For each gene, we obtained a quantitative expression level. Samples were subjected to a priori quality control standards regarding the amount of labeled RNA and the quality of the scanned microarray image (*15*).

The first issue was to explore whether

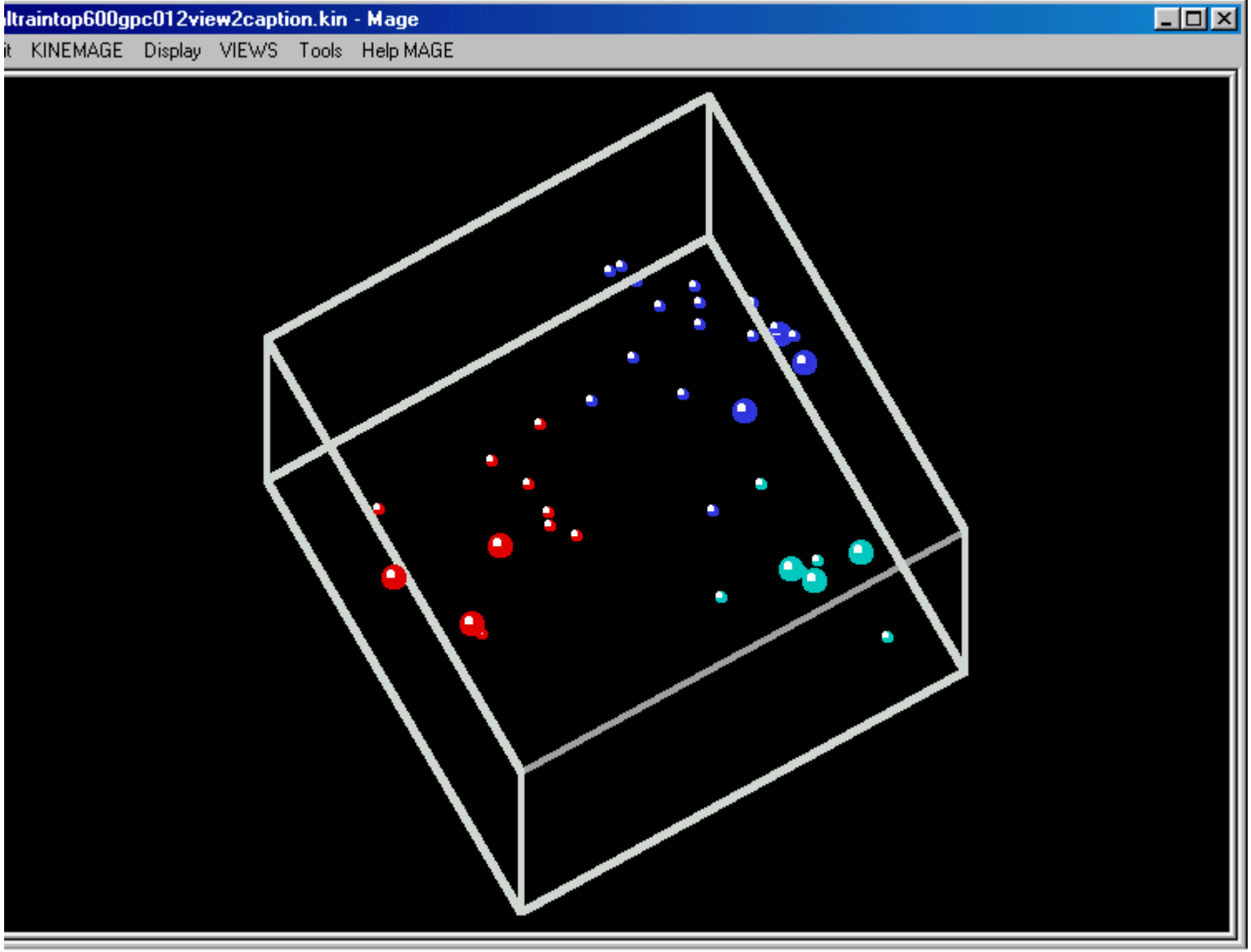
[1]Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139, USA. [2]Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115, USA. [3]St. Jude Children's Research Hospital, Memphis, TN 38105, USA. [4]Comprehensive Cancer Center and Cancer and Leukemia Group B, Ohio State University, Columbus, OH 43210, USA. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: golub@genome.wi.mit.edu; lander@genome.wi.mit.edu.

†These authors contributed equally to this work.

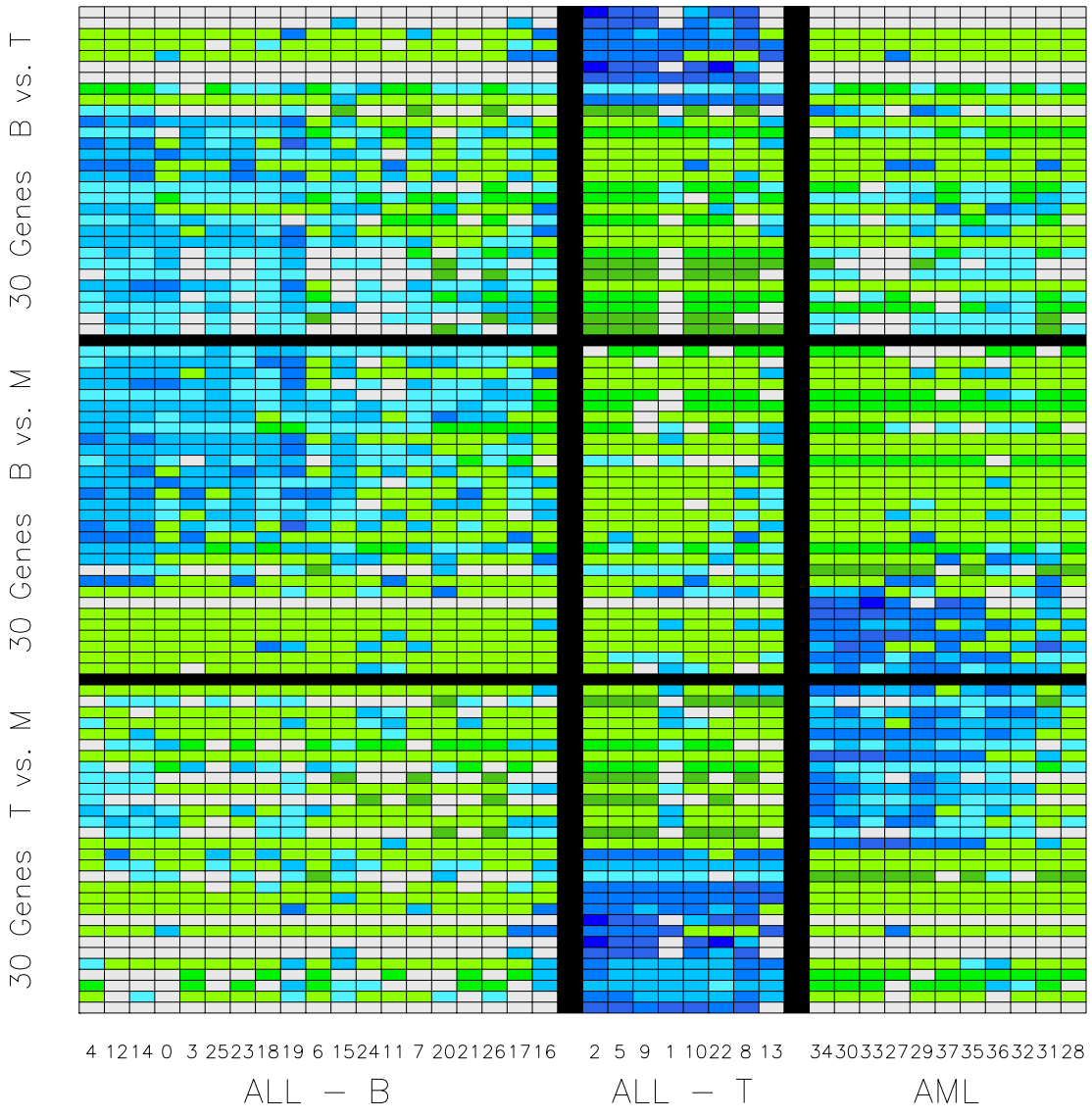


PCA of Golub training data, using all 3158 genes that passed filter

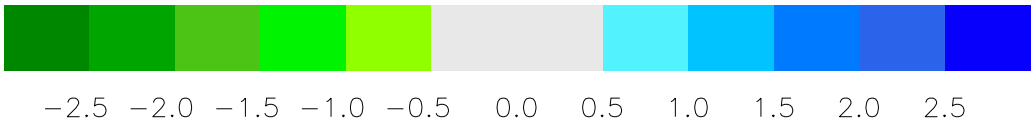


PCA of Golub training data, using 600 top variance genes

Expression Levels for Genes Classifying Tumor Type



Color scale for z ranges of expression for each gene (σ units)



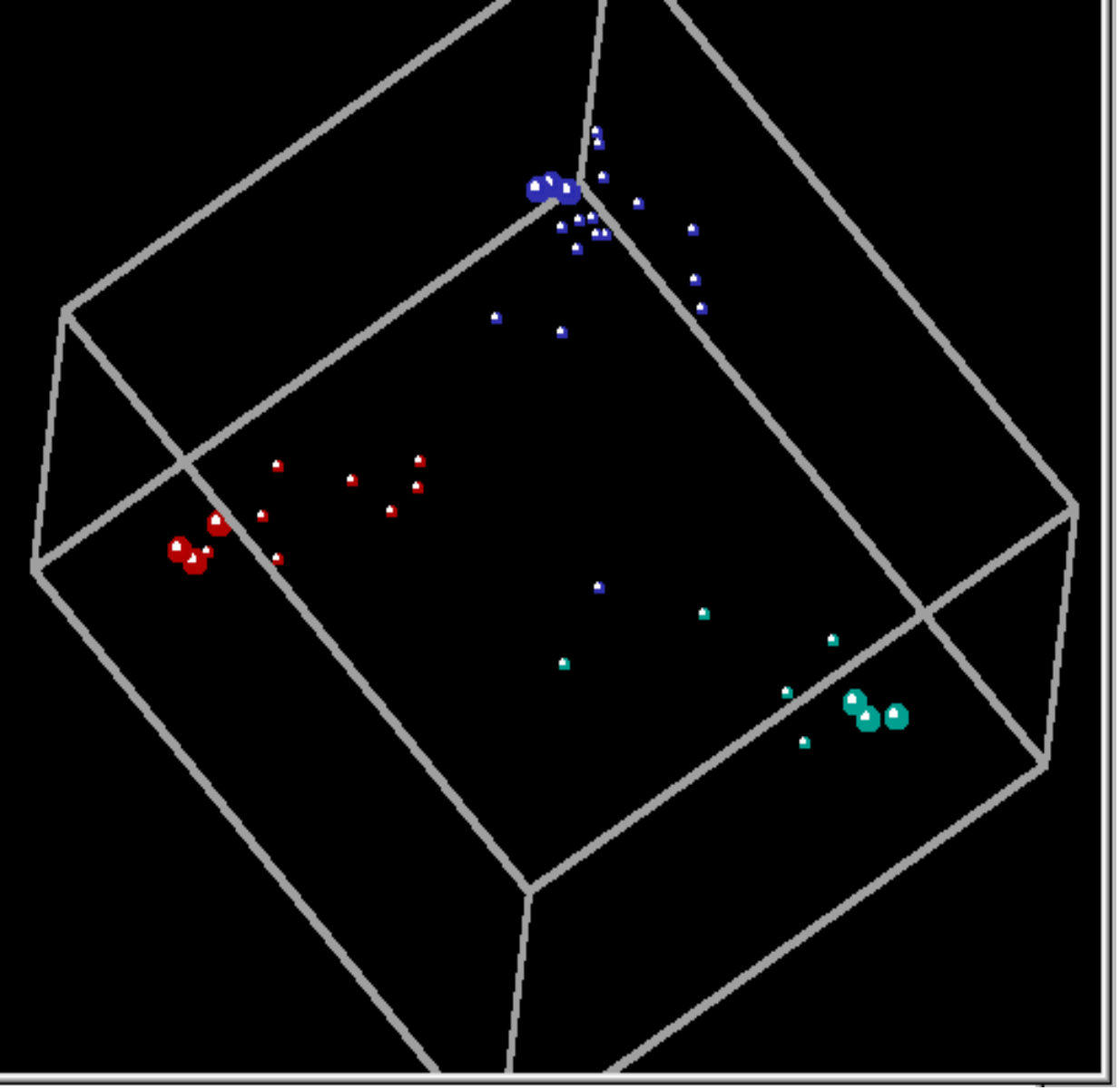
T. R. Golub et al. Acute Leukemia Data (Science, V286, 15 Oct 1999)

ALL – B is B–cell acute lymphoblastic leukemia

ALL – T is T–cell acute lymphoblastic leukemia

AML is acute myeloid leukemia

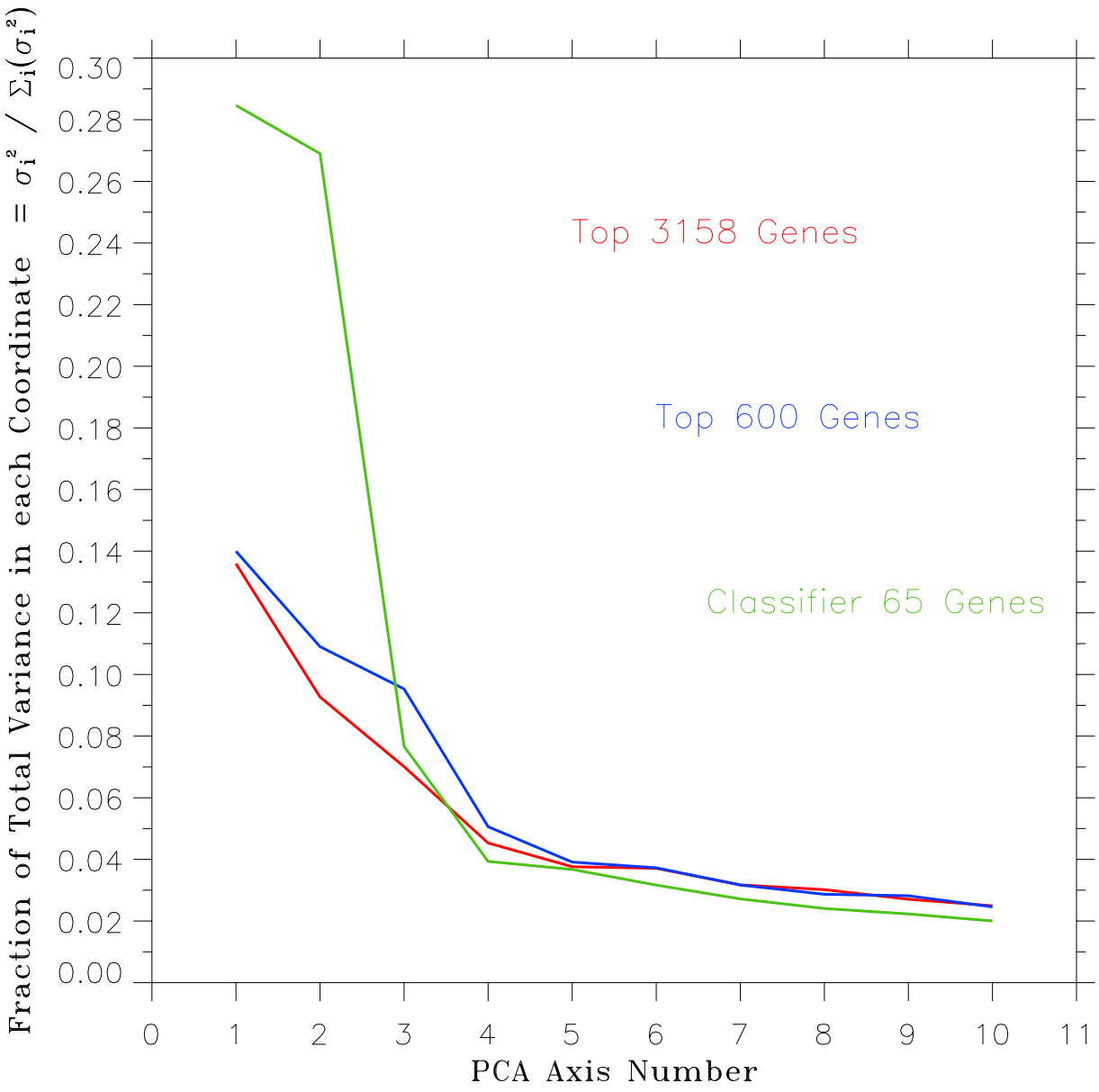
C:\berger\biochem\golubdata\allaml3x30genes3ps.ps



Top 3 principal components projection of Golub et al. ALL/AML training data, using the top genes discriminating the 3 "close neighbors" points in each of the 3 classes pair-wise from each other (30 genes for each of the 3 pairs yielding 65 distinct genes)

Blue = ALL-B cell cyan = ALL-T cell red = AML
Larger spheres = points chosen from "close neighbors criterion" (3 for each class, used to define classifier)

Scree Plot for Golub Leukemia Data



Tue Nov 21 14:54:36 2006 screeplotGolubData2.pro Nov 21, 2006

Variances in Principal Coordinate Axes

C:\berger\biochemlmuulmay03data\PCAcourse\screeplotGolubData2all3.ps

Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 6745–6750, June 1999
Cell Biology

Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays

U. Alon*†, N. Barkai*†, D. A. Notterman*, K. Gish‡, S. Ybarra‡, D. Mack‡, and A. J. Levine*§

Departments of *Molecular Biology and †Physics, Princeton University, Princeton, NJ 08540; and ‡EOS Biotechnology, 225A Gateway Boulevard, South San Francisco, CA 94080

Contributed by A. J. Levine, April 13, 1999

ABSTRACT **Oligonucleotide arrays can provide a broad picture of the state of the cell, by monitoring the expression level of thousands of genes at the same time. It is of interest to develop techniques for extracting useful information from the resulting data sets. Here we report the application of a two-way clustering method for analyzing a data set consisting of the expression patterns of different cell types. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. An efficient two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Coregulated families of genes clustered together, as demonstrated for the ribosomal proteins. Clustering also separated cancerous from noncancerous tissue and cell lines from *in vivo* tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on gene expression.**

Recently introduced experimental techniques based on oligonucleotide or cDNA arrays now allow the expression level of thousands of genes to be monitored in parallel (1–9). To use the full potential of such experiments, it is important to develop the ability to process and extract useful information from large gene expression data sets. Elegant methods recently have been applied to analyze gene expression data sets that are comprised of a time course of expression levels. Examples of such time-course experiments include following a developmental process or changes as the cell undergoes a perturbation such as a shift in growth conditions. The analysis methods were based on clustering of genes according to similarity in their temporal expression (5, 6, 9–11). Such clustering has been demonstrated to identify functionally related families of genes, both in yeast and human cell lines (5, 6, 9, 11). Other methods have been proposed for analyzing time-course gene expression data, attempting to model underlying genetic circuits (12, 13).

Here we report the application of methods for analyzing data sets comprised of snapshots of the expression pattern of different cell types, rather than detailed time-course data. The data set used is composed of 40 colon tumor samples and 22 normal colon tissue samples, analyzed with an Affymetrix oligonucleotide array (8) complementary to more than 6,500 human genes and expressed sequence tags (ESTs) (14). We focus here on generally applicable analysis methods; a more detailed discussion of the cancer-specific biology associated with this study will be presented elsewhere (D.A.N. and A.J.L., unpublished work). The correlation in expression levels across different tissue samples is demonstrated to help identify genes that regulate each other or have similar cellular function. To detect large groups of related genes and tissues we applied two-way clustering, an effective technique for detecting patterns in data sets (see e.g., refs. 15 and 16). The main result is that an efficient clustering algorithm revealed broad, coherent patterns of genes whose expression is correlated, suggesting a high degree of organization underlying gene expression in these tissues. It is demonstrated, for the case of ribosomal proteins, that clustering can classify genes into coregulated families. It is further demonstrated that tissue types (e.g., cancerous and noncancerous samples) can be separated on the basis of subtle distributed patterns of genes, which individually vary only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on their gene expression similarity.

MATERIALS AND METHODS

Tissues and Hybridization to Affymetrix Oligonucleotide Arrays. Colon adenocarcinoma specimens (snap-frozen in liquid nitrogen within 20 min of removal) were collected from patients (D.A.N. and A.J.L., unpublished work). From some of these patients, paired normal colon tissue also was obtained. Cell lines used (EB and EB-1) have been described (17). RNA was extracted and hybridized to the array as described (1, 8).

Treatment of Raw Data from Affymetrix Oligonucleotide Arrays. The Affymetrix Hum6000 array contains about 65,000 features, each containing $\approx 10^7$ strands of a DNA 25-mer oligonucleotide (8). Sequences from about 3,200 full-length human cDNAs and 3,400 ESTs that have some similarity to other eukaryotic genes are represented on a set of four chips. In the following, we refer to either a full-length gene or an EST that is represented on the chip as EST. Each EST is represented on the array by about 20 feature pairs. Each feature contains a 25-bp sequence, which is either a perfect match (PM) to the EST, or a single central-base mismatch (MM). The hybridization signal fluctuates between different features that represent different 25-mer oligonucleotide segments of the same EST. This fluctuation presumably reflects the variation in hybridization kinetics of different sequences, as well as the presence of nonspecific hybridization by background RNAs. Some of the features display a hybridization signal that is many times stronger than their neighbors ($\approx$4% of the intensities are $>$3 SD away from the mean for their EST). These outliers appear with roughly equal incidence in PM or MM features. If not filtered out, outliers contribute significantly to the reading of the average intensity of the gene. Because most features

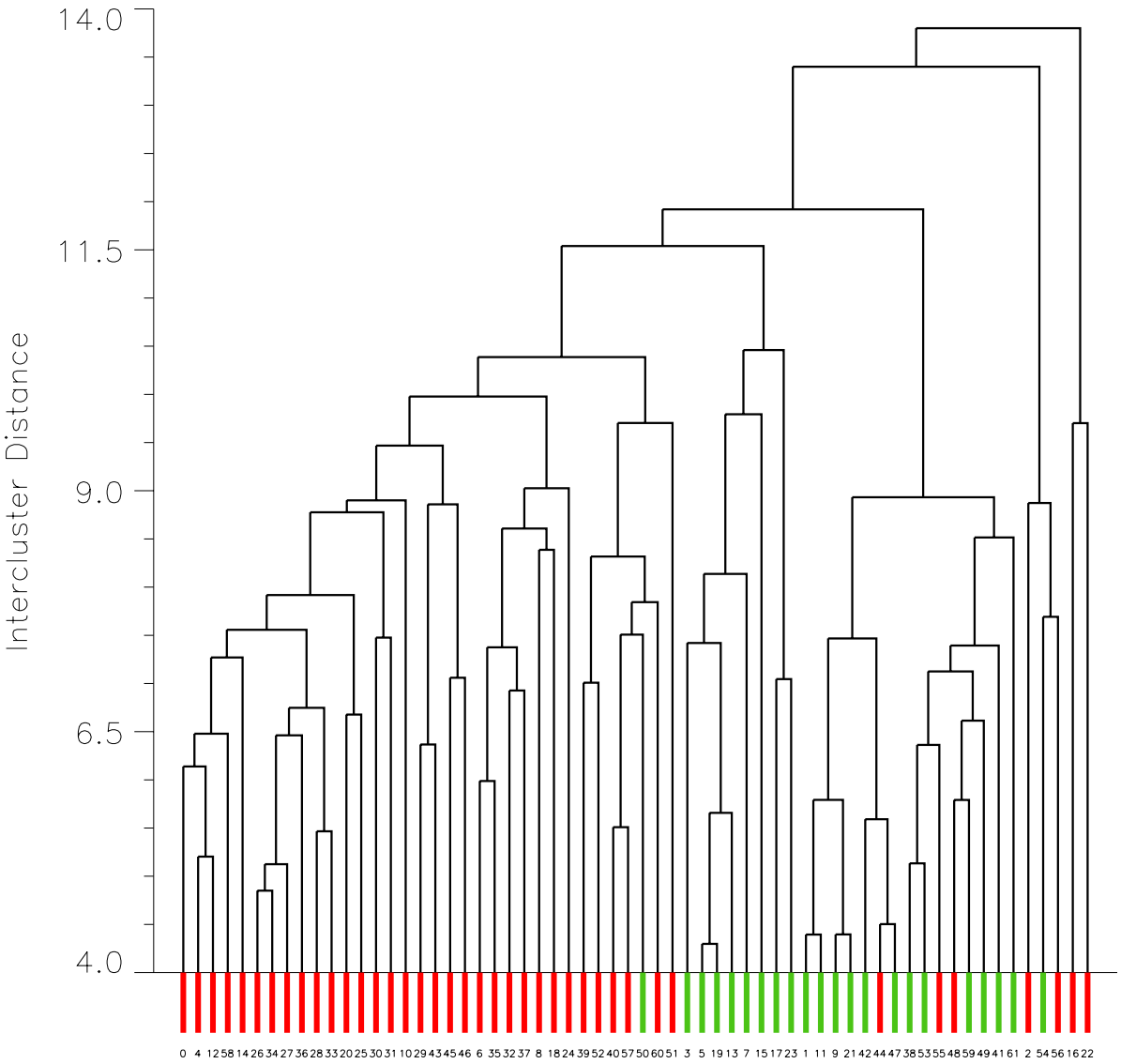
The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviation: EST, expressed sequence tag.

§To whom reprint requests should be sent at present address: President's Office, Rockefeller University, 1230 York Avenue, New York, NY 10021. e-mail: ajlevine@rockvax.rockefeller.edu.

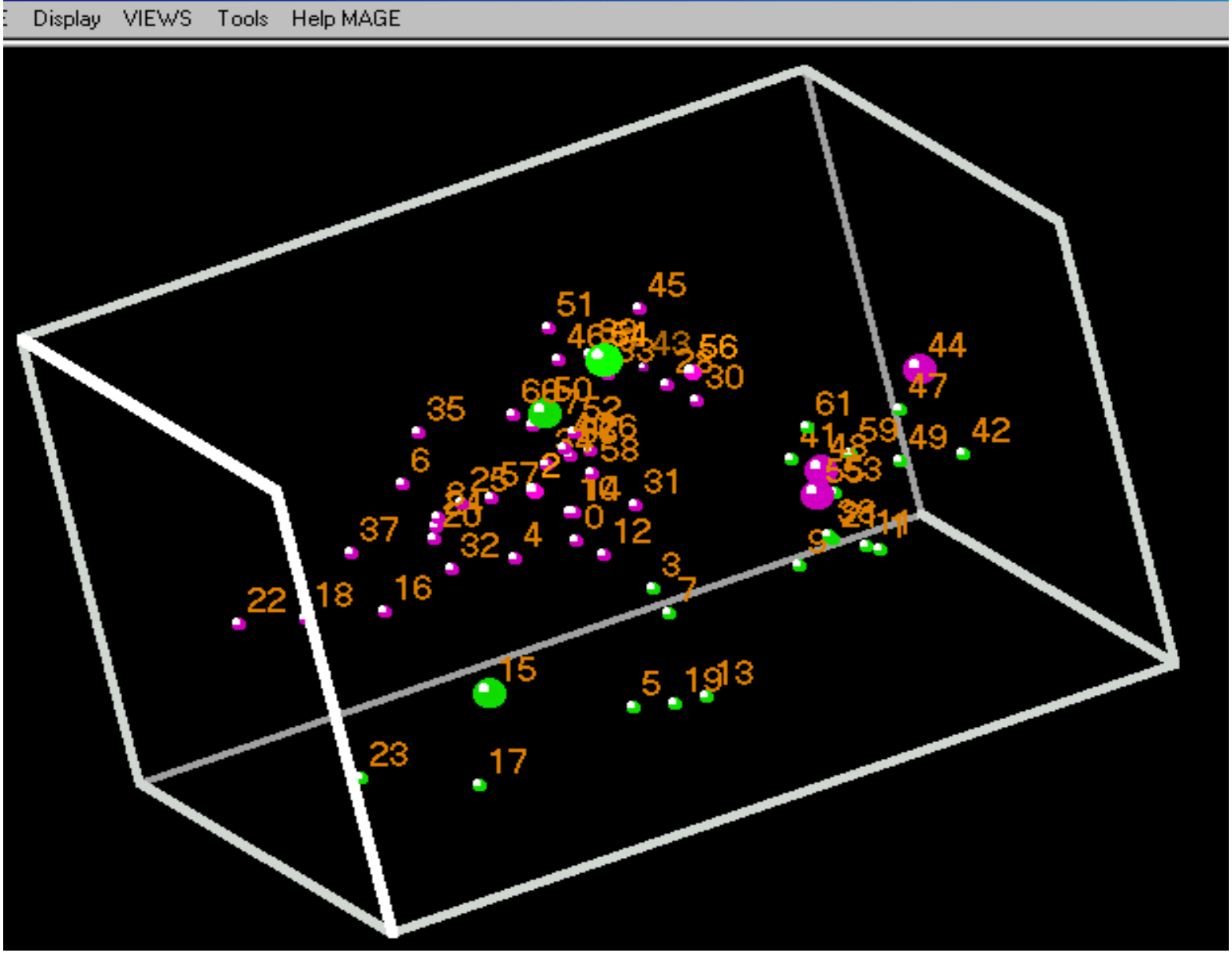
U. Alon et al. Colon Microarray Data Clusters



Average Linkage ln(data) keep top 50 var genes

Colon Tissue Samples: Normal Cancer

plotted Thu Jun 09 22:34:28 2005 plothieralon3nnotrue.pro--July 11, 2003
C:\berger\biochem\alondata\alonallploth3nnotrue-average-nk50PCAclass.ps



Leave-One-Out Crossvalidation for Alon et al. Data:
Misclassified points (3 tumor ● , 3 normal ●) are larger size

Summary for PCA

- PCA projects into orthogonal coordinates that capture as much variation as possible in the top coordinates
- PCA is not necessarily the best way to discover clusters in the data
- PCA works best if most of the variation in the data occurs in the coordinates being kept
- Need to first translate so mean of each component of the data is 0 unless software does so

Non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999, 2001) for Clustering Data

Expresses data (column vectors) in terms of a reduced number of basis vectors. In contrast with SVD (PCA) here the entries in the basis vectors are ≥ 0, and the basis vector coefficients are ≥ 0 and "sparse."

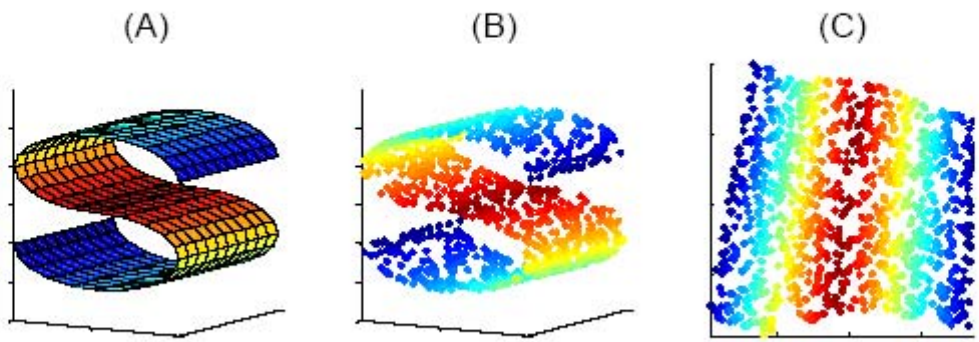
Is the natural distance the distance in R^3 or the distance within the curved surface?

Journal of Machine Learning Research 4 (2003) 119-155 Submitted 6/02; Published 6/03

Think Globally, Fit Locally:
Unsupervised Learning of Low Dimensional Manifolds

Lawrence K. Saul LSAUL@CIS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
200 South 33rd Street
557 Moore School - GRW
Philadelphia, PA 19104-6389, USA

Sam T. Roweis ROWEIS@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road
Pratt Building 283
Toronto, Ontario M5S 3G4, CANADA

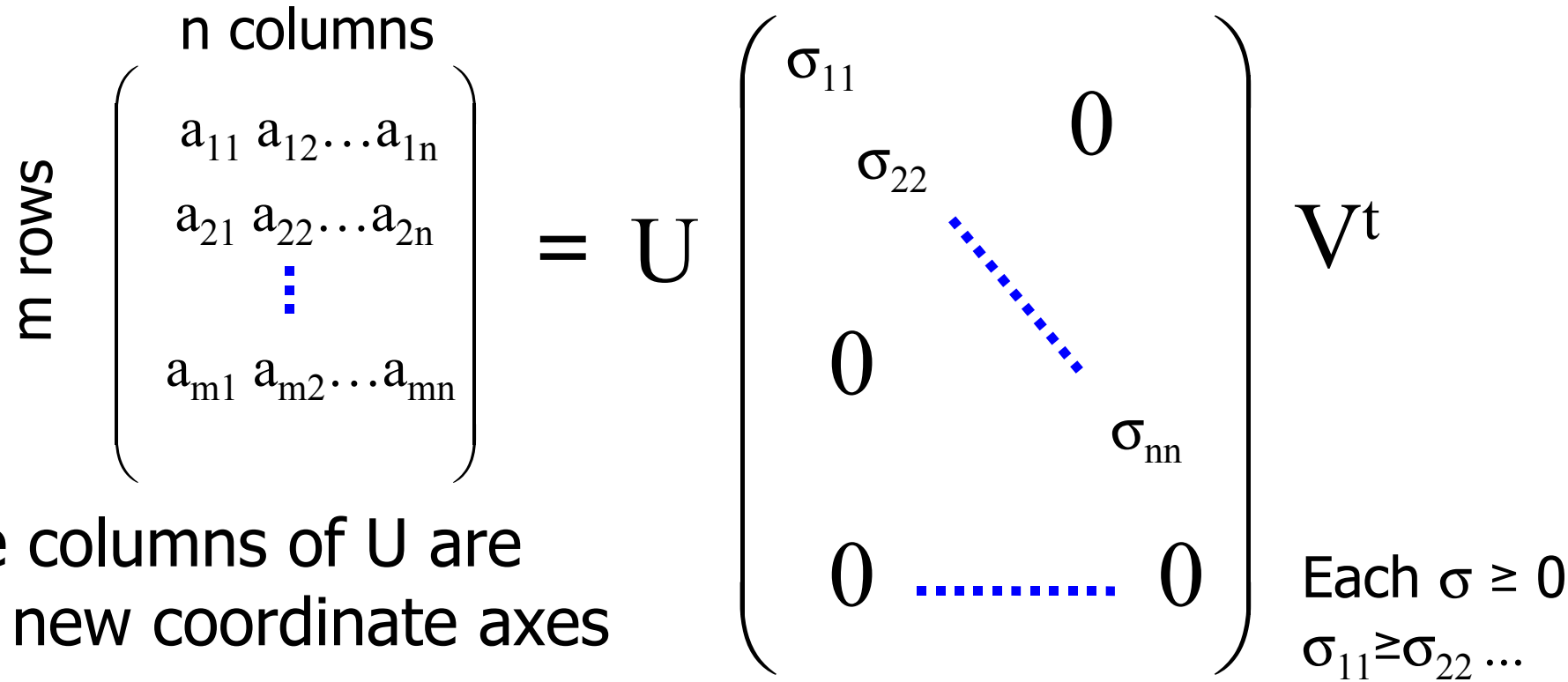


From Figure 1, page 121

Singular Value Decomposition (SVD)

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^t_{n \times n}$$

Here draw schematic for case m>n corr. to data points = cols of A and more rows than columns (e.g., microarray data – viewing samples)



The columns of U are the new coordinate axes

Each $\sigma \geq 0$

$\sigma_{11} \geq \sigma_{22} \ldots$

Matrix Singular Value Decomposition (SVD)

Outline

- The SVD of a Matrix
- PCA in terms of SVD
- PCA, SVD, eigenvalues & eigenvectors of the inner product matrix of the rows of A or of the covariance matrix of the rows of A (same matrices except for a factor of (n-1))
- Statistical Viewpoint: PCA coordinates are uncorrelated

Vectors with n entries

$V = (v_1, v_2, \ldots, v_n)$, $W = (w_1, w_2, \ldots, w_n)$ are called n-vectors, and are said to be points in $\mathbb{R}^n$

the difference (V-W) between V and W
(the vector from W to V) is
$V - W = (v_1 - w_1, \ldots, v_n - w_n)$

the square of the Euclidean distance between V and W =
$(V-W) \bullet (V-W) = (v_1 - w_1)^2 + \ldots + (v_n - w_n)^2$

in general, $V \bullet W = v_1 w_1 + \ldots + v_n w_n$

Eigenvalues and Eigenvectors of a Matrix & Transpose Notation

if $A\,v = \lambda\, v$

for n by n matrix A, n-vector v, number λ:
then λ is an eigenvalue of A with corresponding eigenvector v

$$\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{t} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \quad \text{and} \quad (1,2,3)^{t} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

Orthogonality

Vectors **v** and **w** are orthogonal if

$$\mathbf{V} \bullet \mathbf{W} = 0$$

An n×n matrix A **is** orthogonal if any two distinct rows <columns> of A are orthogonal and each row <column> has length 1 (in which case $A^{-1} = A^{t}$)

e.g.,

$$\begin{pmatrix} \sqrt{3}/2 & -1/2 \\ 1/2 & \sqrt{3}/2 \end{pmatrix}$$

Picking PCA Directions w to Maximize Variation Along w

Consider the points to be the columns of the m by n matrix A

Want to pick the m-vector w of length 1 to maximize the sum of the squares of the entries in the row vector $w^t A$, i.e., want to maximize

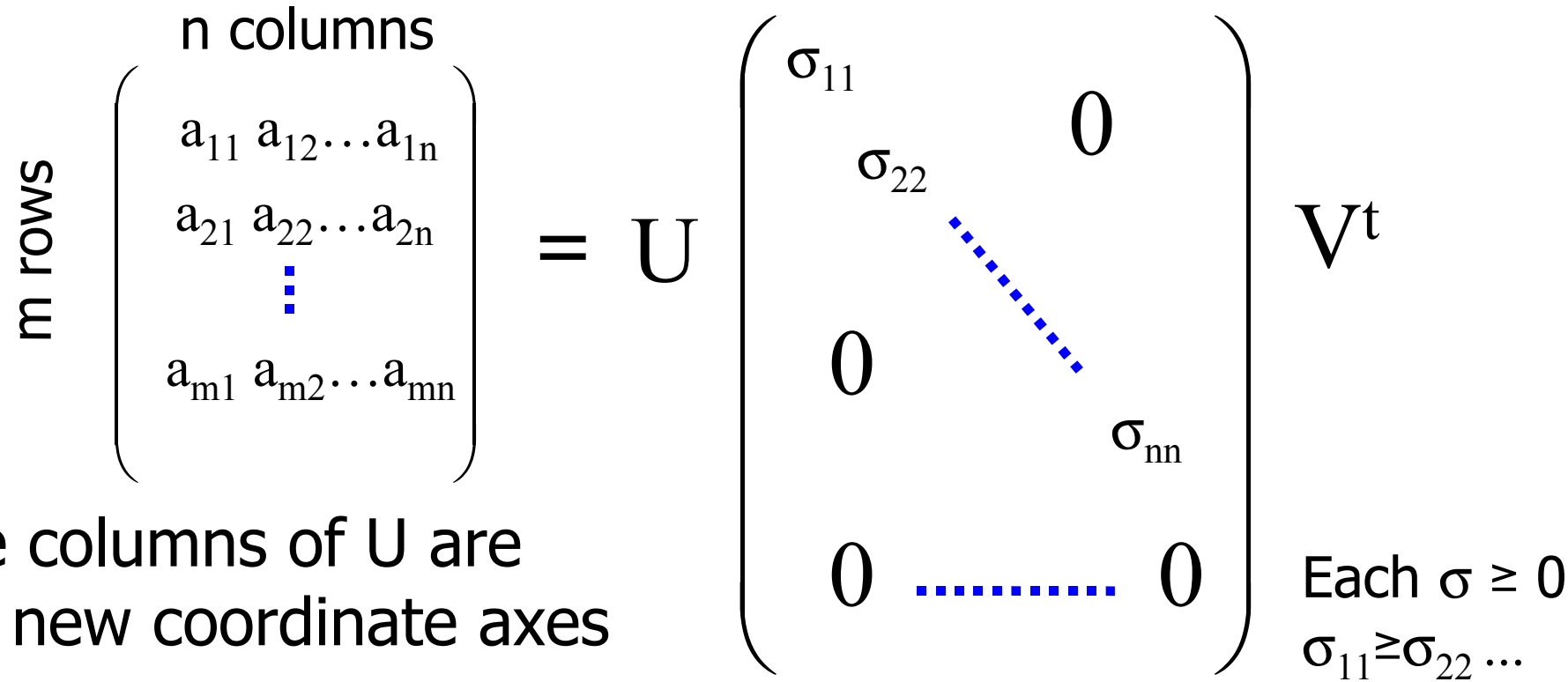
(Q) $\qquad\qquad w^t (A\ A^t)\ w$

This means w is the eigenvector corresponding to the largest eigenvalue of the covariance matrix of the rows of A; successive principal component directions maximize (Q) subject to the current w being perpendicular to all the previous ones.

Singular Value Decomposition (SVD)

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^t_{n \times n}$$

Here draw schematic for case m>n corr. to data points = cols of A and more rows than columns (e.g., microarray data – viewing samples)



The columns of U are the new coordinate axes

Each $\sigma \geq 0$

$\sigma_{11} \geq \sigma_{22} \ldots$

Singular Value Decomposition (SVD)

Decompose $m \times n$ matrix A as the product

$$A = U \; \Sigma \; V^{\mathrm{T}}$$

where

- **Columns (& rows) of U $(m \times m)$ are orthonormal**
- **Rows (& cols) of V^{T} $(n \times n)$ are orthonormal**
- **Σ is an $m \times n$ diagonal matrix**
  Σ = $m \times n$ diag$(\sigma_1 \geq \ldots \geq \sigma_{\mathrm{r}} > 0, \;\; \sigma_{\mathrm{r}+1} = \ldots = \sigma_{\min(m,n)} = 0)$
  r = rank(A) is the # of indep. rows / cols in A
- **Use the k "most significant" components to do k-dimensional Principal Components Analysis (PCA) – project the data (the n cols of A) into the linear subspace spanned by the first k cols of U: $A_k \equiv U_k^{\mathrm{t}} A$ where U_k is columns 1 through k of U.**
- **Note $AA^{\mathrm{T}} = U\Sigma\Sigma^{\mathrm{T}}U^{\mathrm{T}}, \;\; A^{\mathrm{T}}A = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}$**
- **The range of A is spanned by cols 1,…, r of U**
  The null space of A is spanned by cols r+1,…, n of V

Singular Value Decomposition (SVD)

Decompose $m \times n$ matrix A as the product

$$A = U\ \Sigma\ V^{\mathrm{T}}$$

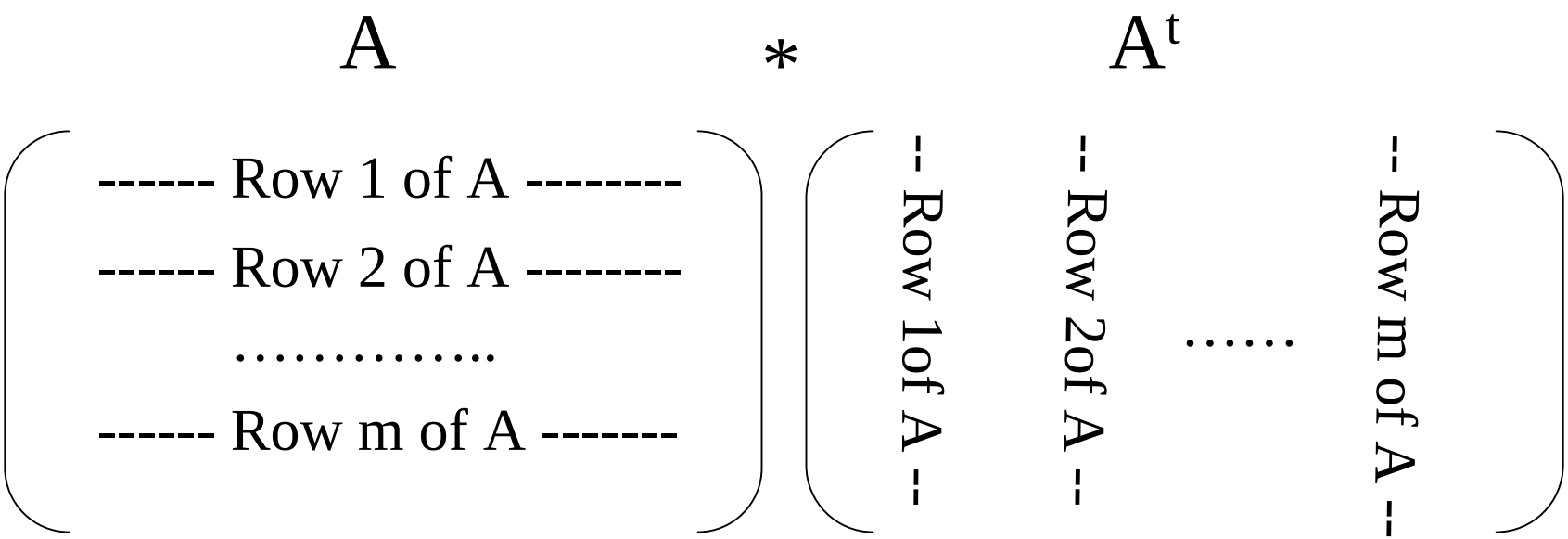
- **Columns of U $(m \times m)$ are orthonormal**
  Rows of V^{T} $(n \times n)$ are orthonormal

- **Σ is an $m \times n$ diagonal matrix**
  $\Sigma = m \times n$ diag$(\sigma_1 \geq \ldots \geq \sigma_{\mathrm{r}} > 0,\ \sigma_{\mathrm{r}+1} = \ldots = \sigma_{\min(m,n)} = 0)$
  r = rank(A) is the # of indep. rows / cols in A

- **Note $AA^{\mathrm{T}} = U\Sigma\Sigma^{\mathrm{T}}U^{\mathrm{T}},\ A^{\mathrm{T}}A = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}$**

- **AA^{T} gives a constant, namely, $n-1$, times the covariance matrix for the rows of A (assuming each row mean = 0).**
  The eigenvectors of AA^{T} $<A^{\mathrm{T}}A>$ (the cols of U $<V>$) are the principal components basis for the columns of A $<A^{\mathrm{T}}>$.
  The eigenvalues of AA^{T} are $n-1$ times the variances V_i of the data along the principal component axes (the cols of U). ($V_i = 0$ for $i > r$, $V_i = TV_i/(n-1)$.)
  The data in principal component coordinates is $Y = U^{\mathrm{T}}A$, and the covariance matrix $YY^{\mathrm{T}}/(n-1)$ is diag$(V_i)_{m \times m}$: the rows of Y are uncorrelated.
  Statistical viewpoint: one is taking linear combinations of random variables (rows of A).

Matrix of Inner Products of the Rows of A_{mxn}



The *covariance* C_{ij} of A_i = row i of A and A_j = row j of A is: $(1/(n-1)) * (A_i - \mu(A_i)) * (A_j - \mu(A_j)) = \Sigma_{k=1}^{n} (A_{ik} - \mu(A_i)) \bullet (A_{jk} - \mu(A_j)) / (n-1)$

Here $\mu(A_i)$ and $\mu(A_j)$ are assumed to have already been arranged to be 0; so C = cov. matrix of rows of A is just $A A^t / (n-1)$. Thus e. vectors of C equal the e.vectors of $A A^t$ & e. values of C = (e.values of $A A^t$) / (n-1).

Distance Geometry
Outline

- Here motivated by: distance constraints → molecular conformations
- Distance constraints $\{m \le d_{ij} \le M\}$ → random sample sets of $\{d_{ij}\}$
- for each given set of $\{d_{ij}\}$ → metric (inner product matrix) $G = A^t A$
- Eigenvectors and eigenvalues of G give construction of A
- This construction is a principal component representation of points
- Duality of this construction, PCA in terms of the SVD
- Simple Example
- Need to follow this DG/PCA linear projection into 3-D by nonlinear optimization to best satisfy the distance constraints when have experimental errors / incomplete info. / inexact choices of the $\{d_{ij}\}$

Some References for Classical Distance Geometry and Restrained Molecular Dynamics

[1] T. F. Havel, I. D. Kuntz and G. M. Crippen, The Theory and Practice of Distance Geometry, *Bull. Math. Biol.* 45 (1983), pp. 665-720. (geometry)

[2] G. M. Crippen and T. F. Havel, Stable Calculation of Coordinates from Distance Information, *Acta Cryst.* A34 (1978), pp. 282-284. D → metric matrix

[3] T. F. Havel, An Evaluation of Computational Strategies for Use in the Determination of Protein Structure from Distance Constraints Obtained by Nuclear Magnetic Resonance, *Prog. Biophys. Molec. Biol.* 56 (1991), pp. 43-78. (practical alg.)

[4] G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation*, Research Studies Press - Chemometrics Series Vol. 15, 1988.

[5] I. T. Jolliffe, *Principal Component Analysis* 2nd Edition, Springer, 2002 – cf. Section 5.2.

[6] http://www.statsoft.com/textbook/stmulsca.html (web site explaining multidimensional scaling)

[7] J. M. Blaney, G. M. Crippen, A. Dearing, J. S. Dixon and D. C. Spellmeyer, DGEOM 95: Distance Geometry, QCPE Program Number 590, Quantum Chemistry Program Exchange, Indiana University Department of Chemistry, 1995.

[8] D. Bassolino-Klimas, R. Tejero, S. R. Krystek, W. J. Metzler, G. T. Montelione and R. E. Bruccoleri, Simulated Annealing with Restrained Molecular Dynamics Using a Flexible Restraint Potential: Theory and Evaluation with Simulated NMR Constraints, *Protein Science* 5 (1996), pp. 593-603.

[9] R. Tejero, D. Bassolino-Klimas, R. E. Bruccoleri, and G. T. Montelione, Simulated Annealing with Restrained Molecular Dynamics Using CONGEN: Energy Refinement of the NMR Solution Structures of Epidermal and Type-alpha Transforming Growth Factors, *Protein Science* 5 (1996), pp. 578-592.

A Couple References for Molecular Modeling

[1] A. R. Leach, *Molecular Modelling: Principles and Applications*, Second Edition, Prentice Hall, 2001.

[2] T. Schlick, *Molecular Modeling and Simulation*, Springer, 2002.

Given a Distance Matrix for n points (that satisfy certain geometrical conditions) there exists n points that are vectors with (n-1) components that satisfy the distance conditions

2 points fit in a line (1-D), 3 points form a plane triangle (2-D), 4 points form a tetrahedron (3-D)…

For molecular structure (or general visualization (MDS)) want to project back down into 3-D while minimizing loss of information / extent of discrepancy with the constraints – sounds like PCA except we don't yet have the points

Distance Geometry Construction

Given distances between n points, use “fundamental equality” to construct *metric matrix* = inner product of point vectors (even though don’t have the points!!)

Having obtained this inner product matrix G, find its eigenvalues and eigenvectors and use them to directly construct n-dimensional points satisfying the distances.

This construction naturally is in the PC coordinates for the points (so using the first 3 coord. gives the PCA projection).

If the distances are "consistent," the algorithm below produces a set of points P_k (n-vectors) that
(1) satisfy the distances $\text{dist}(P_i, P_j) = d_{ij}$
(2) have centroid = **0** (average = **0**)

(3) have coordinates that are the principal component coordinates, so just using the first 3 coordinates gives the best linear projection into 3-D, i.e., the projection that maximizes the sum over the projected points of $(\text{dist}(\mathbf{0}, P_k))^2$ or equivalently that maximizes the sum over i and j of the inter-point distances $(\text{dist}(P_i, P_j))^2$

METRICS SERIES

Editor: **Dr. D. Bawden**
Central Research, Sandwich, Kent, England

- Distance Geometry and Conformational Calculations*
  G.M. Crippen
- Clustering of Large Data Sets*
  Jure Zupan
- Multivariate Data Analysis in Industrial Practice
  Paul J. Lewi

Correlation Analysis of Organic Reactivity: With particular reference to multiple regression*
John Shorter

Information Theoretic Indices for Characterization of Chemical Structures
Danail Bonchev

Logical and Combinatorial Algorithms for Drug Design
V.E. Golender *and* **A.B. Rozenblit**

Minimum Steric Difference
The MTD method for QSAR studies
Z. Simon, A. Chiriac, S. Holban, D. Ciubotaru *and* **G.I. Mihalas**

Analytical Measurement and Information
Advances in the information theoretic approach to chemical analyses
K. Eckschlager *and* **V. Štěpánek**

Molecular Connectivity in Structure-Activity Analysis
Lemont B. Kier *and* **Lowell H. Hall**

Potential Pattern Recognition in Chemical and Medical Decision Making
D. Coomans *and* **I. Broeckaert**

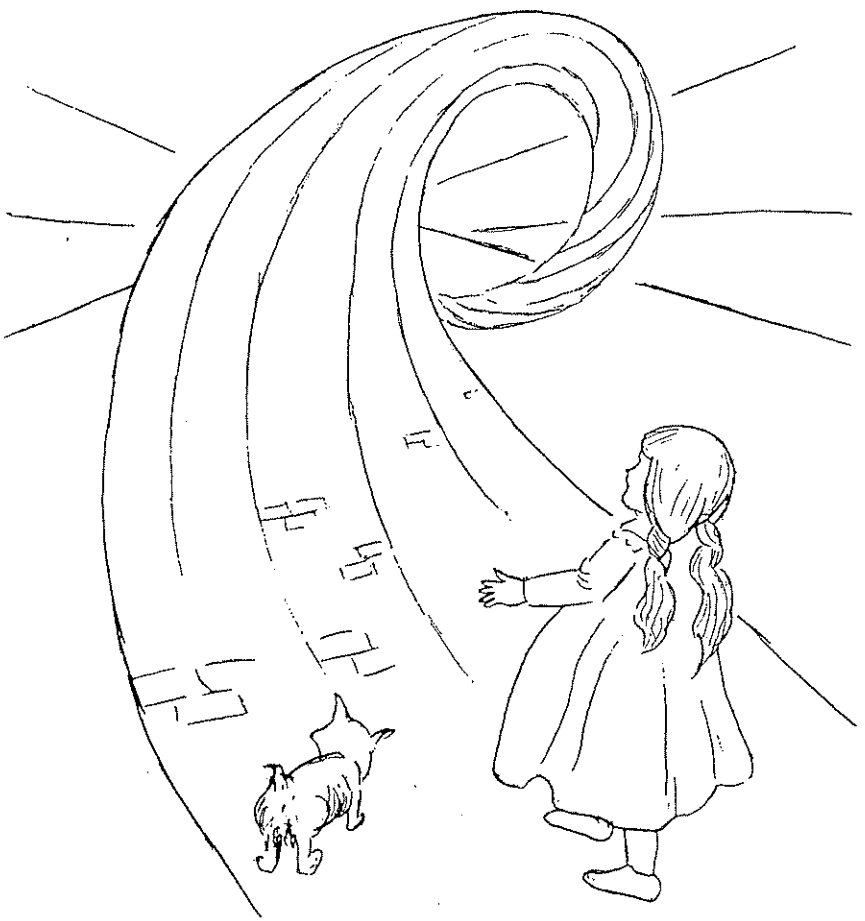
Chemical Pattern Recognition
O. Strouf

Similarity and Clustering in Chemical Information Systems
Peter Willett

Multivariate Chemometrics in QSAR: A Dialogue
Peter P. Mager

Application of Pattern Recognition to Catalytic Research
I.I. Ioffe

Distance Geometry and Molecular Conformation
G.M. Crippen *and* **T.F. Havel**



"My goodness, Toto, I don't think we're in $\mathbb{R}^n$ anymore!"

EMBEDDING

following G. M. Crippen & T. F. Havel,
Stable Calculation of Coordinates from Distance Information,
Acta Cryst. A**34** (1978), pp. 282–284.

Stable algorithm to obtain initial atom positions $\{P^0_{\mathrm{k}}\}$ approximately satisfying given distance constraints:

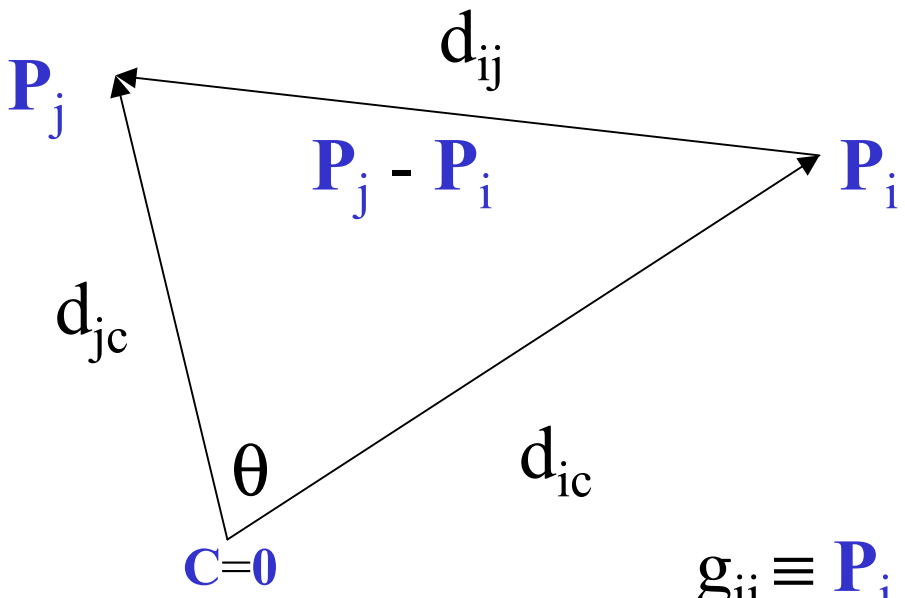
1. $d^2_{\mathrm{i\,c}} \equiv ||P_{\mathrm{i}} - \text{average } A \text{ of } \{P_{\mathrm{k}}\}||^2 = \mathrm{fcn}(d_{\mathrm{j\,k}})$

$$d^2_{\mathrm{i\,c}} = \frac{1}{N}\sum_{\mathrm{j}=1}^{\mathrm{N}} d^2_{\mathrm{i\,j}} - \frac{1}{N^2}\sum_{\mathrm{j}=2}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{j\text{-}1}} d^2_{\mathrm{j\,k}}.$$

2. Calculate $N \times N$ metric matrix $\mathbb{G} = g_{\mathrm{i\,j}} = (P_{\mathrm{i}} - A) \cdot (P_{\mathrm{j}} - A)$ using the law of cosines:

$$g_{\mathrm{i\,j}} = \frac{1}{2}(d^2_{\mathrm{i\,c}} + d^2_{\mathrm{j\,c}} - d^2_{\mathrm{i\,j}}).$$

Trig. – Law of Cosines



$$g_{ij} \equiv \mathbf{P}_i \bullet \mathbf{P}_j = d_{ic}\, d_{jc} \cos(\theta)$$

$$(\mathbf{P}_j - \mathbf{P}_i) \bullet (\mathbf{P}_j - \mathbf{P}_i) = d_{ij}^2$$

$$= d_{ic}^2 + d_{jc}^2 - 2g_{ij}$$

so $$g_{ij} = (d_{ic}^2 + d_{jc}^2 - d_{ij}^2) / 2$$

The fundamental equality (1) in the "Embedding Page" says that if there are points satisfying the distance constraints, their inner product (dot product) metric matrix g_{ij} can be explicitly expressed completely in terms of the inter-point distances.

Conversely given an inner product matrix g_{ij} from a set of points (centroid at origin) one can directly write down the inter-point distances

If one sums the fundamental embedding identity (1) over i=1,…,n, (when the centroid of the points is **0**) one finds that:

sum over i of $(\text{dist}(\mathbf{0}, P_i))^2$ =
sum over i and j of $(\text{dist}(P_i, P_j))^2 / (2n)$

the "variation" of the points about the origin = the sum of the squares of the inter-point distances / (2n) (note this counts d_{ij} **&** d_{ji})

ISOMORPHISMS BETWEEN the SETS of DISTANCE and METRIC MATRICES

$$\left\{ d^2_{\mathrm{i\,j}} : \; d^2_{\mathrm{i\,j}} = d^2_{\mathrm{j\,i}},\; d^2_{\mathrm{i\,i}} = 0 \right\} \underset{\mathrm{W}}{\overset{\mathrm{T}}{\rightleftharpoons}} \left\{ g_{\mathrm{i\,j}} : \; g_{\mathrm{i\,j}} = g_{\mathrm{j\,i}},\; \sum_{\mathrm{i}} g_{\mathrm{i\,j}} = 0 \right\}$$

$$\mathrm{T}\left\{ d^2_{\mathrm{r\,s}} \right\}_{\mathrm{i\,j}} = \frac{1}{2}\left(d^2_{\mathrm{i\,c}} + d^2_{\mathrm{j\,c}} - d^2_{\mathrm{i\,j}} \right)$$

$$\mathrm{W}\left\{ g_{\mathrm{r\,s}} \right\}_{\mathrm{i\,j}} = g_{\mathrm{i\,i}} + g_{\mathrm{j\,j}} - 2g_{\mathrm{i\,j}}$$

$$d^2_{\mathrm{i\,c}} = \frac{1}{N}\sum_{\mathrm{j}=1}^{\mathrm{N}} d^2_{\mathrm{i\,j}} - \frac{1}{N^2}\sum_{\mathrm{j}=2}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{j\text{-}1}} d^2_{\mathrm{j\,k}}.$$

$\mathbb{G}$ Positive Semi-Definite is a Necessary and Sufficient Condition for Existence of a Corresponding Set of Points in $\mathbb{R}^{\mathrm{n}}$

Let the eigenvalues of $\mathbb{G}$ be $\lambda_1,\ \lambda_2,\ \lambda_3 \ldots$, and let the corresponding eigenvectors (here row vectors) be $W_1,\ W_2,\ W_3 \ldots$ with the order $\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots$

Then, when $\mathbb{G}$ is Positive Semi-Definite, define the points (column vectors) $\{P_{\mathrm{k}}\}$ by

$$(P_1 \quad P_2 \quad \ldots \quad P_{\mathrm{N}})_{\mathrm{N\,x\,N}} = \begin{pmatrix} \lambda_1^{1/2} W_1 \\ \lambda_2^{1/2} W_2 \\ \vdots \\ \lambda_{\mathrm{N}}^{1/2} W_{\mathrm{N}} \end{pmatrix}_{\mathrm{N\,x\,N}}$$

In general the top few eigenvalues will be positive even when the geometric conditions on the distances for exact embedding are not satisfied.

Defining the Points from the Eigenvectors and Eigenvalues

$P_1\ P_2\ P_3\ P_4\ P_5\ P_6\ P_7\ P_8\ P_9\ \ldots\ldots\ldots\ P_n$

row 1	$(\lambda_1)^{1/2}$	X X X X X X X X X X X X X
		← eigenvectors of G →
row 2	$(\lambda_2)^{1/2}$	X X X X X X X X X X X X X
row 3	$(\lambda_3)^{1/2}$	X X X X X X X X X X X X X
		If data were "perfect," rows beyond 3 would be 0
row n-1	$(\lambda_{n-1})^{1/2}$	X X X X X X X X X X X X X
row n must be 0	0	u u u u u u u u u u u u u

$$u = 1/\sqrt{n}$$

Distance Geometry / PCA duality

If have distances from n points, use “vector formula” to construct nxn *metric matrix* G = inner product of point vectors (even though don’t have the points!!)

Having obtained this inner product matrix G, find its eigenvalues and eigenvectors and use them to directly construct n-dimensional points satisfying the distances.

This construction naturally is in the PC coordinates for the points (so using the first 3 coord. gives the PCA projection).

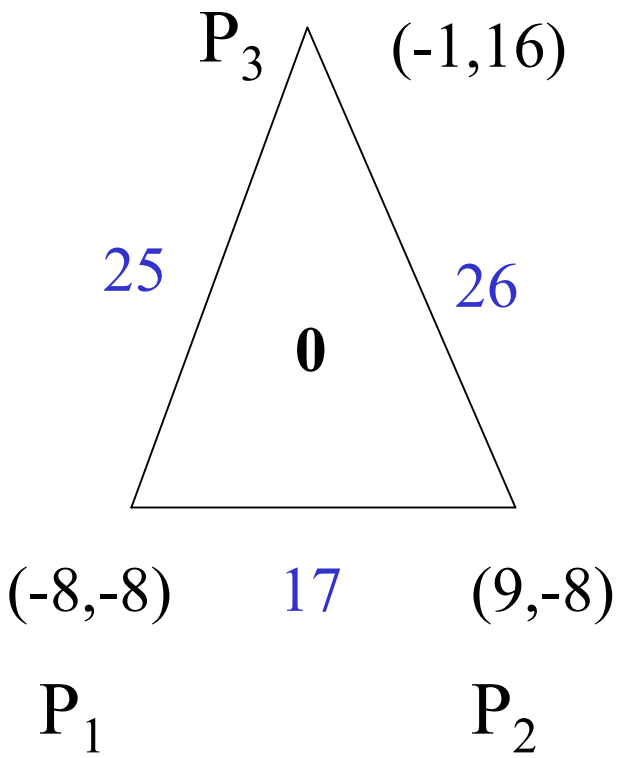
Distance Geometry / PCA duality II

If have n points, centroid=0, as columns of matrix A, then their metric matrix $G = A^t * A$; use SVD of A – $A = U \Sigma V^t$ to write $G = V \Sigma^t U^t U \Sigma V^t$ so

$$G = V \Lambda V^t = (V \Lambda^{1/2}) (\Lambda^{1/2} V^t) = P^t P$$

The line above is the distance geometry construction that only required knowing the distances between the points. Since G = the inner product matrix from P, and G came from the distance matrix, the isomorphism between Dij's and Gij's guarantees the points P satisfy the inter-point distances. If have A; U^t A directly gives the principal component coordinates of the points. (note the non-zero entries of $\Lambda^{1/2}$ and Σ are the same).

Distance Geometry – Simple Example

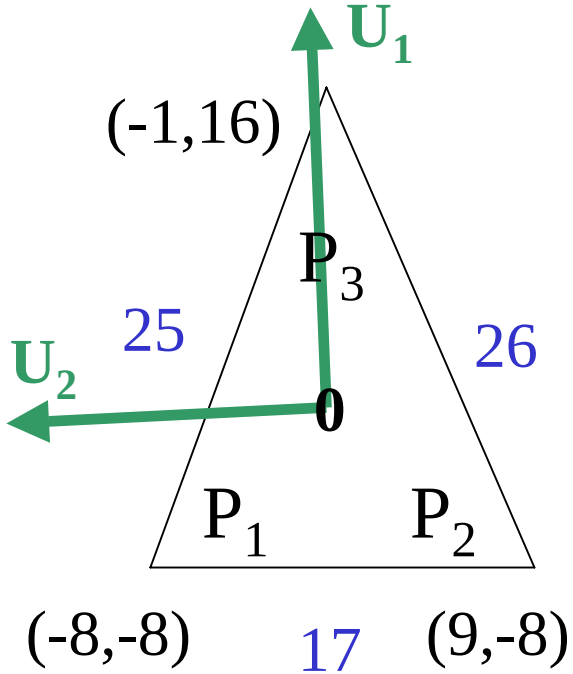


metric matrix G

$$\begin{pmatrix} 128 & -8 & -120 \\ -8 & 145 & -137 \\ -120 & -137 & 257 \end{pmatrix}$$

eigenvalues = 386.39605, 143.60395 sum = 530 = sum of squares of distances of the 3 points to the origin

Distance Geometry – Simple Example



points from e-values & e-vectors of G

$$\begin{pmatrix} (\lambda_1)^{1/2} & \mathbf{P}_1 & \mathbf{P}_2 & \mathbf{P}_3 \\ 19.657 & (-0.36454 & -0.45045 & 0.81499) \\ & \leftarrow \text{first eigenvector} \rightarrow & & \\ (\lambda_2)^{1/2} & & & \\ 11.9835 & (0.73060 & -0.68100 & -0.04960) \\ & \leftarrow \text{second eigenvector} \rightarrow & & \\ & 0 & 0 & 0 \end{pmatrix}$$

1-D projection = (-7.1657, -8.8545, 16.0202)

"Best" 1-D (min $\Sigma_{ij}(D_{ij}-d_{ij})^2$) is ($p_2$ = -14.333, p_1=-2.667, p_3 = 17)
error is $3*(5.333)^2$ so need nonlinear optimization after projection

sum of squares of inter-point distances (*double count*) =
$2(289 + 676 + 625) = 3180 = 2*(n=3)*(\Sigma(\text{dist of pt to } \mathbf{0})^2 = 530)$

Singular Value Decomposition on the array A of the 3 points
from the distance geometry example - each point is a column of A

A =

-8.0000000	9.0000000	-1.0000000
-8.0000000	-8.0000000	16.000000

Note the sum of the entries in each row of A has
already been arranged to be 0

the singular value decomposition of A:
A = U Σ Vt

(here Vt means the transpose of V)
t after any matrix name means take the transpose of the matrix:
the rows of the transposed matrix = the columns of the original matrix

U =

-0.099341496	-0.99505338
0.99505340	-0.099341644

The columns of U are the PCA axes for the columns of A.
These are the eigenvectors of C_r = A At;
C_r = (n-1)*covariance matrix for rows of A.

Σ =

19.656959	0.00000000	0.0
0.00000000	11.983487	0.0

The variation (sum of squares of coordinate values) for each
nontrivial PCA coordinate = (diagonal of Σ)^2 = eigenvalues of the
metric matrix G = At A. Since centroid of points = 0, at least one
eigenvalue of G is always 0

Vt =

-0.36453728	-0.45045123	0.81498851	The rows of Vt are the
0.73060204	-0.68099945	-0.049602588	eigenvectors of the
0.57735026	0.57735026	0.57735026	metric matrix G = At A

verify A = U Σ Vt
U Σ Vt =

-8.0000006	8.9999992	-0.99999859
-8.0000006	-7.9999996	16.000000

write the points (columns of A) in terms of the PCA coordinates
(the columns of U)

Ut A = Σ Vt =

-7.1656945	-8.8545014	16.020196
8.7551598	-8.1607478	-0.59441195

These are the coordinates of the 3 points
(columns of A) in terms of the PCA
coordinate system (the 2 columns of U). Note these 2 rows are
uncorrelated (their inner product is 0). Distances are
preserved by the full (all coordinate) PCA representation.

note the sums of the columns of A are 0
(the centroid is at the origin) so
(n-1) x (the covariance matrix of the rows of A) = A At =

146.00000	-24.000000
-24.000000	384.00000

the Metric Matrix = the array of inner products of
the points (the columns of A) = At A =

128.00000	-8.0000000	-120.00000
-8.0000000	145.00000	-137.00000
-120.00000	-137.00000	257.00000

Ut A At U = ΣΣt =

386.39605	-7.8708505e-005
-7.8708505e-005	143.60395

so A At U = U ΣΣt (the eigenvectors of A At are the columns of U)

and

Vt At A V = Σt Σ =

386.39605	-9.6387622e-005	0.0
-9.6387622e-005	143.60395	0.0
1.6403545e-014	-4.1017537e-015	0.0

so At A V = V Σt Σ

(the eigenvectors of At A are the columns of V = the rows of Vt)