

MULTIVARIATE DATA ANALYSIS

Susan Holmes ©

<http://www-stat.stanford.edu/~susan/>

Bio-X and Statistics


IMA Workshop, October, 2013



you do not really understand something unless you can explain it to your grandmother -- Albert Einstein

*you do not really understand something unless you can
explain it to your grandmother -- Albert Einstein*
I am your grandmother

What are multivariate data ?



| | | | | |
|----------|------------|----------------|------------|-----------|
| 26253284 | 876481926 | 6279800699 | 5580033790 | 388230079 |
| 61863161 | 495293951 | 9378430319 | 3913007260 | 278804079 |
| 27775586 | 8743599227 | 4966747796 | 7153682225 | 308494088 |
| 13177994 | 612729431 | 2558470040 | 804761070 | 07688755 |
| 7257988 | 344142417 | 562168094 | 075611814 | 85481009 |
| 68039512 | 021115969 | 1044999945 | 177902019 | 24177365 |
| 40139117 | 437470768 | 064464040 | 231511544 | 830780059 |
| 2674361 | 945072690 | 627978272 | 044051714 | 28845307 |
| 644318 | 600308934 | 23184775 | 32121468 | 9043833 |
| 4778597 | 808605490 | 403390597 | 859285783 | 44871912 |
| 8241892 | 384655258 | SYSTEM FAILURE | 410389084 | 261254009 |
| 8058759 | 712758798 | 323980360 | 320900896 | 9053383 |
| 624352 | 921320472 | 51987824 | 32439148 | 919218 |
| 1294070 | 098858816 | 870652479 | 421749103 | 86883688 |
| 43139097 | 725003637 | 759785853 | 3434484366 | 798899184 |
| 4160853 | 123212592 | 442076488 | 624944695 | 38273787 |
| 480361 | 88563825 | 604073333 | 0244444220 | 648771915 |
| 6832018 | 187235780 | 8933771349 | 779681256 | 42924415 |
| 2178043 | 164382730 | 3222003606 | 2199945594 | 29336466 |

Simplest format: matrices:

If we have measured 10,000 genes on hundreds of patients and all the genes are independent, we can't do better than analyze each gene's behavior by using histograms or box plots, looking at the means, medians, variances and other 'one dimensional statistics'. However if some of the genes are acting together, either that they are positively correlated or that they inhibit each other, we will miss a lot of important information by slicing the data up into those column vectors and studying them separately. Thus important connections between genes are only available to us if we consider the data as a whole. We start by giving a few examples of data that we encounter.

- ▶ Athletes, performances in the decathlon.

| | 100 | long | poid | haut | 400 | 110 | disq | perc | jave | 1500 |
|---|-------|------|-------|------|-------|-------|-------|------|------|------|
| 1 | 11.25 | 7.43 | 15.48 | 2.27 | 48.90 | 15.13 | 49.28 | 4.7 | 61.3 | 15.1 |
| 2 | 10.87 | 7.45 | 14.97 | 1.97 | 47.71 | 14.46 | 44.36 | 5.1 | 61.7 | 15.1 |
| 3 | 11.18 | 7.44 | 14.20 | 1.97 | 48.29 | 14.81 | 43.66 | 5.2 | 64.1 | 15.1 |
| 4 | 10.62 | 7.38 | 15.02 | 2.03 | 49.06 | 14.72 | 44.80 | 4.9 | 64.0 | 15.1 |
| 5 | 11.02 | 7.43 | 12.92 | 1.97 | 47.44 | 14.40 | 41.20 | 5.2 | 57.4 | 15.1 |

- ▶ Clinical measurements (diabetes data).

| | relwt | glufast | glutest | steady | insulin | Group |
|---|-------|---------|---------|--------|---------|-------|
| 1 | 0.81 | 80 | 356 | 124 | 55 | 3 |
| 3 | 0.94 | 105 | 319 | 143 | 105 | 3 |
| 5 | 1.00 | 90 | 323 | 240 | 143 | 3 |
| 7 | 0.91 | 100 | 350 | 221 | 119 | 3 |
| 9 | 0.99 | 97 | 379 | 142 | 98 | 3 |

- ▶ OTU read counts:

| | 469478 | 208196 | 378462 | 265971 | 570812 |
|--------------|--------|--------|--------|--------|--------|
| EKCM1.489478 | 0 | 0 | 2 | 0 | 0 |
| EKCM7.489464 | 0 | 0 | 2 | 0 | 2 |
| EKBM2.489466 | 0 | 0 | 12 | 0 | 0 |

| | | | | | |
|--------------|---|---|----|---|---|
| PTCM3.489508 | 0 | 0 | 14 | 0 | 0 |
| EKCF2.489571 | 0 | 0 | 4 | 0 | 0 |

► RNA-seq, transcriptomic:

| | FBgn0000017 | FBgn0000018 | FBgn0000022 | FBgn000002 |
|------------|-------------|-------------|-------------|------------|
| untreated1 | 4664 | 583 | 0 | 10 |
| untreated2 | 8714 | 761 | 1 | 11 |
| untreated4 | 3150 | 310 | 0 | 3 |
| treated1 | 6205 | 722 | 0 | 10 |
| treated3 | 3334 | 308 | 0 | 5 |

► Mass spec:
Samples × Features.

| mz | 129.9816 | 72.08144 | 151.6255 | 142.0349 | 169.0413 |
|----------|----------|----------|----------|----------|----------|
| KOGCHUM1 | 60515 | 181495 | 0 | 196526 | 25500 |
| WTGCHUM1 | 252579 | 54697 | 412 | 487800 | 48775 |
| WTGCHUM2 | 187859 | 56318 | 46425 | 454226 | 45626 |

Dependencies

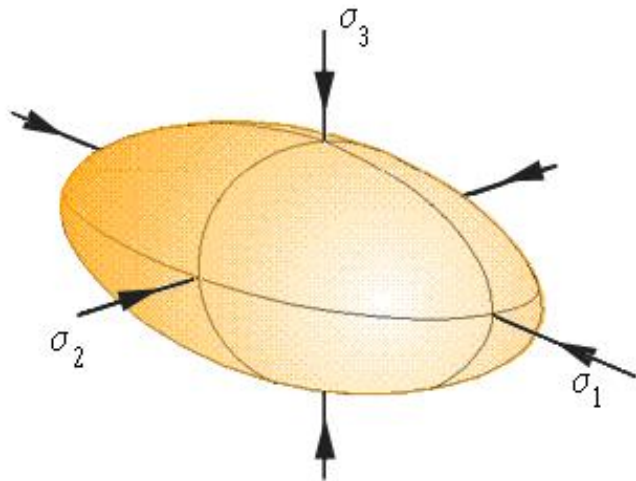
If the data were all independent columns, then the data would have no multivariate structure and we could just do univariate statistics on each variable (column) in turn.

Multivariate statistics means we are interested in how the columns covary.

We can compute covariances to evaluate the dependencies.

If the data were multivariate normal with p variables, all the information would be contained in the $p \times p$ covariance matrix Σ and the mean μ .

Parametric Multivariate Normal



Modern Statistics: Non parametric, multivariate

- ▶ Exploratory Analyses: Hypotheses generating.
 - ▶ Projection Methods (new coordinates)
 - ▶ Principal Component Analysis
 - ▶ Principal Coordinate Analysis–Multidimensional Scaling (PCO,MDS)
 - ▶ Correspondence Analysis
 - ▶ Discriminant Analysis
 - ▶ Tree based methods
 - ▶ Phylogenetic Trees
 - ▶ Clustering Trees
 - ▶ Decision Trees
- ▶ Confirmatory Analyses: Hypothesis verification.
 - ▶ Permutation tests (Monte Carlo).
 - ▶ Bootstrap (Monte Carlo).
 - ▶ Bayesian nonparametrics (Monte Carlo).

Modern Methods: Robust Methods

Variance

Variability of one continuous variable \longrightarrow the variance.

NOT ROBUST, low breakdown.

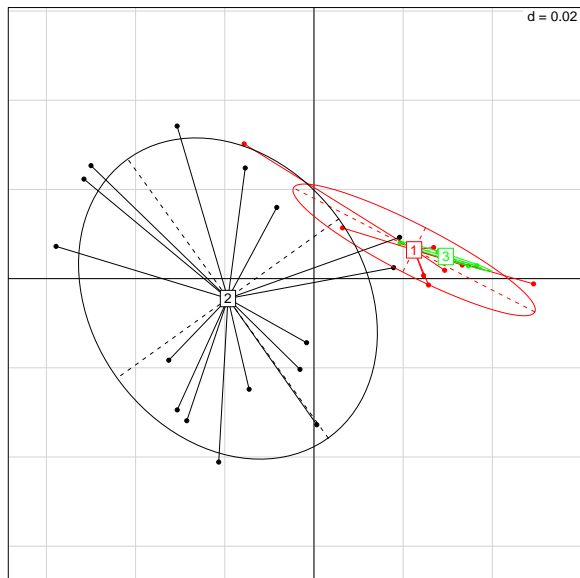
Solution: Take ranks, clumps, logs, or trimming the data.

Part I

EDA: *Exploratory Data Analysis*

Data Checking
Hypothesis Generating

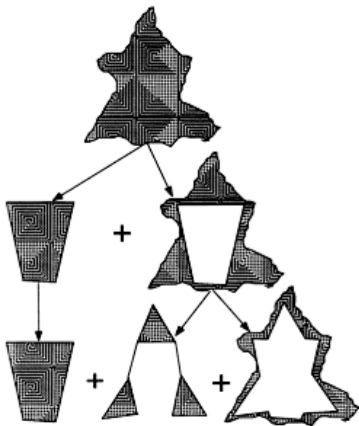
Discovery by Visualization



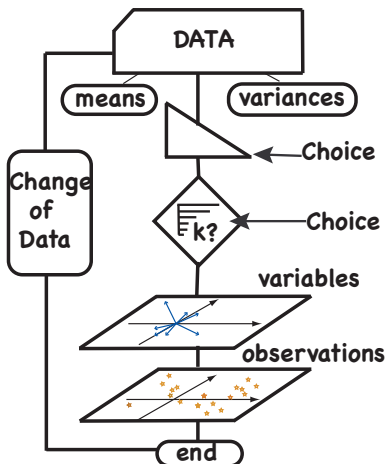
Basic Visualization Tools

- ▶ Boxplots, barplots.
- ▶ Scatterplots of projected data.
- ▶ Scatterplots with binning variable
- ▶ Hierarchical Clustering, heatmaps, Phylogenies.
- ▶ Combination of Phylogenetic Trees and data.

Iterative Structuration (Tukey, 1977)



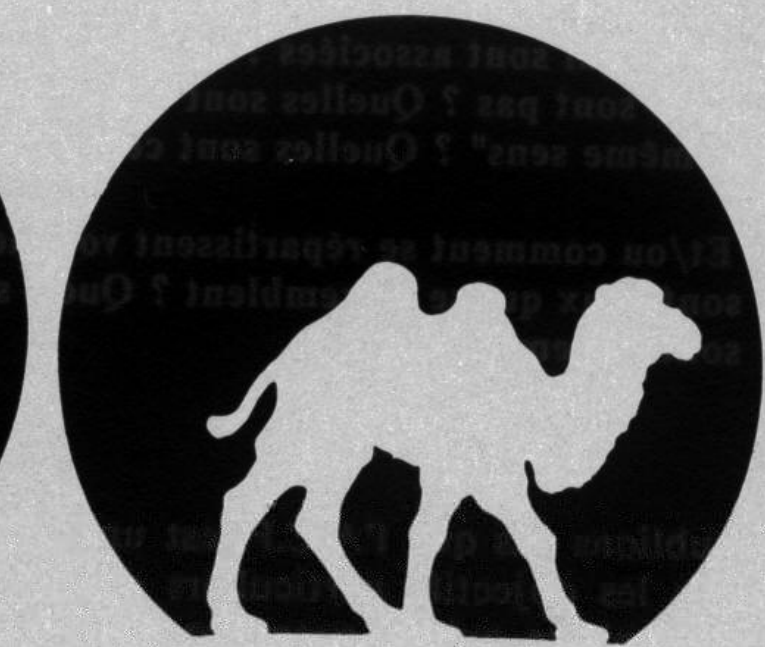
One table methods: PCA, MDS, PCoA, CA,



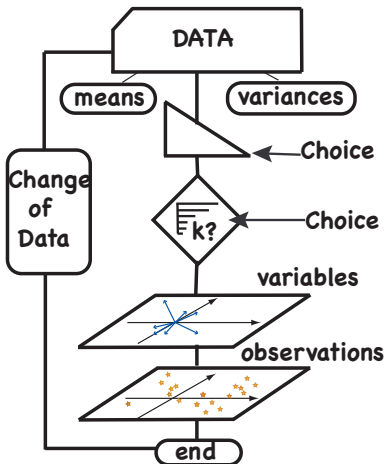
All based on the principle of finding the largest axis of inertia/variability.

New Variables/coordinates from old or distances
Best Projection Directions??





PCA procedure



All based on the principle of finding the largest axis of inertia/variability.

New Variables/coordinates from old or distances
Best Projection Directions because they explain the most variance.

Our first task is often to rescale the data so that all the variables, or columns of the matrix have the same standard deviation, this will put all the variables on the same footing. We also make sure that the means of all columns are zero, this is called centering. After that we will try to simplify the data by doing what we call rank reduction, we'll explain this concept from several different perspectives. A favorite tool for simplifying the data is called Principal Component Analysis (abbreviated PCA).

What is PCA?

PCA is an 'unsupervised learning technique' because it treats all variables as having the same status, there is no particular response variable that we are trying to predict using the other variables as explanatory predictors as in supervised methods. PCA is primarily a visualization technique which produces maps that show the relations between the variables in a useful way.

Useful Facts to Remember

- ▶ Each PC is defined to maximize the variance it explains.
- ▶ The new variables are made to be orthogonal, if the data are multivariate normal they will be independent.
- ▶ Always check the screeplot before deciding how many components to retain (how much signal you have).

A Geometrical Approach

- i. The data are p variables measured on n observations.
- ii. X with n rows (the observations) and p columns (the variables).
- iii. D_n is an $n \times n$ matrix of weights on the "observations", which is most often diagonal.
- iv Symmetric definite positive matrix Q , often

$$Q = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 & \dots \\ 0 & 0 & \frac{1}{\sigma_3^2} & 0 & \dots \\ \dots & \dots & \dots & 0 & \frac{1}{\sigma_p^2} \end{pmatrix}.$$

Euclidean Spaces

These three matrices form the essential "triplet" ($\mathbf{X}, \mathbf{Q}, \mathbf{D}$) defining a multivariate data analysis.

\mathbf{Q} and \mathbf{D} define geometries or inner products in \mathbb{R}^p and \mathbb{R}^n , respectively, through

$$\begin{aligned} \mathbf{x}^\dagger \mathbf{Q} \mathbf{y} &= \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{Q}} & \mathbf{x}, \mathbf{y} &\in \mathbb{R}^p \\ \mathbf{x}^\dagger \mathbf{D} \mathbf{y} &= \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}} & \mathbf{x}, \mathbf{y} &\in \mathbb{R}^n. \end{aligned}$$

An Algebraic Approach

- ▶ Q can be seen as a linear function from \mathbb{R}^p to $\mathbb{R}^{p*} = \mathcal{L}(\mathbb{R}^p)$, the space of scalar linear functions on \mathbb{R}^p .
- ▶ D can be seen as a linear function from \mathbb{R}^n to $\mathbb{R}^{n*} = \mathcal{L}(\mathbb{R}^n)$.

▶

$$\begin{array}{ccc} \mathbb{R}^{p*} & \xrightarrow{x} & \mathbb{R}^n \\ Q \uparrow & & \downarrow D \\ & \mathbb{R}^p & \xleftarrow{x^\dagger} \mathbb{R}^{n*} \\ & \downarrow v & \uparrow w \end{array}$$

An Algebraic Approach

$$\begin{array}{ccc} \mathbb{R}^{p*} & \xrightarrow{\quad X \quad} & \mathbb{R}^n \\ Q \uparrow & & \downarrow D \\ & \mathbb{R}^p & \xleftarrow{\quad X^\dagger \quad} \mathbb{R}^{n*} \\ & & \uparrow W \end{array}$$

Duality diagram

- i. Eigendecomposition of $X^\dagger D X Q = V Q$
- ii. Eigendecomposition of $X Q X^\dagger D = W D$
- iii. Transition Formulae.

Notes

(1) Suppose we have data and inner products defined by Q and D :

$$(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^p \times \mathbb{R}^p \longmapsto \mathbf{x}^\dagger Q \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle_Q \in \mathbb{R}$$

$$(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n \longmapsto \mathbf{x}^\dagger D \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle_D \in \mathbb{R}.$$

$$\|\mathbf{x}\|_Q^2 = \langle \mathbf{x}, \mathbf{x} \rangle_Q = \sum_{j=1}^p q_j (\mathbf{x} \cdot \mathbf{j})^2 \quad \|\mathbf{x}\|_D^2 = \langle \mathbf{x}, \mathbf{x} \rangle_D = \sum_{j=1}^p p_i (\mathbf{x}_i)^2$$

(2) We say an operator O is B -symmetric if $\langle \mathbf{x}, O\mathbf{y} \rangle_B = \langle O\mathbf{x}, \mathbf{y} \rangle_B$, or equivalently $BO = O^\dagger B$.

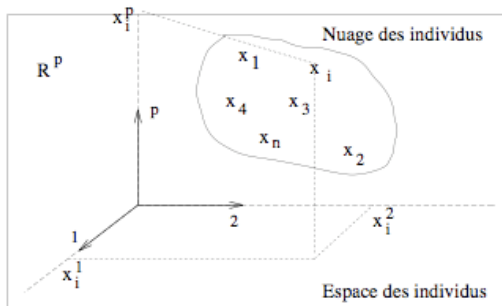
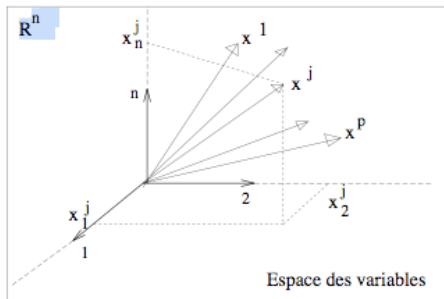
The **duality diagram** is equivalent to $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ such that \mathbf{X} is $n \times p$. Escoufier (1977) defined as $\mathbf{XQX}^\dagger \mathbf{D} = \mathbf{W}\mathbf{D}$ and $\mathbf{X}^\dagger \mathbf{DXQ} = \mathbf{VQ}$ as the characteristic operators of the diagram.

(3) $V = X^{\dagger}DX$ will be the variance-covariance matrix, if X is centered with regards to D ($X'D\mathbf{1}_n = 0$).

Transposable Data

There is an important symmetry between the rows and columns of X in the diagram, and one can imagine situations where the role of observation or variable is not uniquely defined. For instance in microarray studies the genes can be considered either as variables or observations. This makes sense in many contemporary situations which evade the more classical notion of n observations seen as a random sample of a population. It is certainly not the case that the 9,000 species are a random sample of bacteria since these probes try to be an exhaustive set.

Two Dual Geometries



Properties of the Diagram

Rank of the diagram: X, X^\dagger, VQ and WD all have the same rank. For Q and D symmetric matrices, VQ and WD are diagonalisable and have the same eigenvalues.

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_r \geq 0 \geq \dots \geq 0.$$

Eigendecomposition of the diagram: VQ is Q symmetric, thus we can find Z such that

$$VQZ = Z\Lambda, Z^\dagger QZ = \mathcal{I}_p, \text{ where } \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p). \quad (1)$$

Practical Computations

Cholesky decompositions of Q and D , (symmetric and positive definite) $H^\dagger H = Q$ and $K^\dagger K = D$.

Use the singular value decomposition of KXH :

$$KXH = UST^\dagger, \quad \text{with } T^\dagger T = \mathcal{I}_p, U^\dagger U = \mathcal{I}_n, S \text{ diagonal.}$$

Then $Z = (H^{-1})^\dagger T$ satisfies

$$VQZ = Z\Lambda, Z^\dagger QZ = \mathcal{I}_p$$

with $\Lambda = S^2$.

The renormalized columns of Z , $A = SZ$ are called the principal axes and satisfy:

$$A^\dagger QA = \Lambda.$$

Practical Computations

Similarly, we can define $L = K^{-1}U$ that satisfies

$$WDL = L\Lambda, L^\dagger DL = \mathcal{I}_n, \text{ where } \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r, 0, \dots, 0). \quad (2)$$

$C = LS$ is usually called the matrix of principal components. It is normed so that

$$C^\dagger DC = \Lambda.$$

Transition Formulæ:

Of the four matrices Z , A , L and C we only have to compute one, all others are obtained by the transition formulæ provided by the duality property of the diagram:

$$XQZ = LS = C \quad X^\dagger DL = ZS = A$$

General Features

1. Inertia : $\text{Trace}(VQ) = \text{Trace}(WD)$

(inertia in the sense of Huyghens inertia formula for instance).

Huygens, C. (1657),

$$\sum_{i=1}^n p_i d^2(\mathbf{x}_i, \mathbf{a})$$

Inertia with regards to a point \mathbf{a} of a cloud of p_i -weighted points.

PCA with $Q = \mathcal{I}_p$, $D = \frac{1}{n}\mathcal{I}_n$, and the variables are centered, the inertia is the sum of the variances of all the variables.

If the variables are standardized (Q is the diagonal matrix of inverse variances), then the inertia is the number of variables p .
For correspondence analysis the inertia is the Chi-squared statistic.

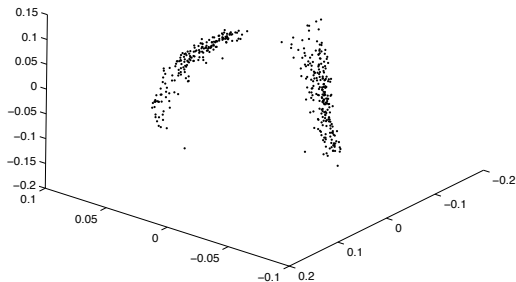
Ordination Methods

Many discrete measurements \longrightarrow Gradients.

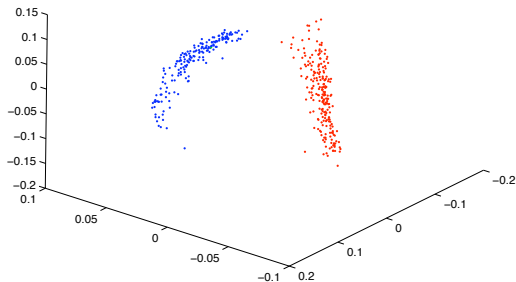
Data from 2005 U.S. House of Representatives roll call votes. We further restricted our analysis to the 401 Representatives that voted on at least 90% of the roll calls (220 Republicans, 180 Democrats and 1 Independent) leading to a 401×669 matrix V of voting data.

The Data

| | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
|----|----|----|----|----|----|----|----|----|----|-----|
| 1 | -1 | -1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 2 | -1 | -1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 |
| 4 | 1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 |
| 5 | 1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 |
| 6 | -1 | -1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 7 | -1 | -1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 |
| 8 | -1 | -1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 9 | 1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 |
| 10 | -1 | -1 | 1 | -1 | 0 | 1 | 1 | 0 | 0 | 0 |



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes.



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. Color has been added to indicate the party affiliation of each representative.

Metric Multidimensional Scaling

Given a distance matrix (or its square) how do we find the points in Euclidean space whose distances are given by this matrix?

Can we always find such a map?

Schoenberg (1935) but also Borschadt 1866.

Think of towns, whose road distances are known for whom we want to reconstruct a map.

Decomposition of Distances

If we started with original data in \mathbb{R}^p that are not centered: Y , apply the centering matrix

$$X = HY, \quad \text{with } H = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right), \text{ and } \mathbf{1}' = (1, 1, 1, \dots, 1)$$

Call $B = XX'$, if $D^{(2)}$ is the matrix of squared distances between rows of X in the euclidean coordinates, we can show that

$$-\frac{1}{2}HD^{(2)}H = B$$

We can go backwards from a matrix D to X by taking the eigendecomposition of B in much the same way that PCA provides the best rank r approximation for data by taking the singular value decomposition of X , or the eigendecomposition of XX' .

$$X^{(r)} = US^{(r)}V' \quad \text{with } S^{(r)} = \begin{pmatrix} s_1 & 0 & 0 & 0 & \dots \\ 0 & s_2 & 0 & 0 & \dots \\ 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & s_r & \dots \\ \dots & \dots & \dots & 0 & 0 \end{pmatrix}$$

Multidimensional Scaling also called PCoA

Simple classical multidimensional scaling.

- ▶ Square D elementwise $D^{(2)} = D_2$.
- ▶ Compute $\frac{-1}{2}HD_2H = B$.
- ▶ Diagonalize B to find the principal coordinates.

Important: What D to use.

Distances, Similarities, Dissimilarities

Distances:

- ▶ Euclidean
- ▶ Chisquare

$$\text{Chisquare}(\text{exp}, \text{obs}) = \sum_j \frac{(\text{exp}_j - \text{obs}_j)^2}{\text{exp}_j}$$

- ▶ Hamming/L1
- ▶ DNA distances (dist.dna in ape)

Similarity Indices:

- ▶ Confusion (cognitive psychology).
- ▶ Matching coefficient

$$\frac{\text{nb of matching attrs}}{\text{nb of attrs}} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{10} + f_{01}}$$

- ▶ Jaccard Similarity Index.

$$\frac{J}{A + B - J} = \frac{a}{a + b + c}$$

See versions in vegan and ade4.

Projection Methods

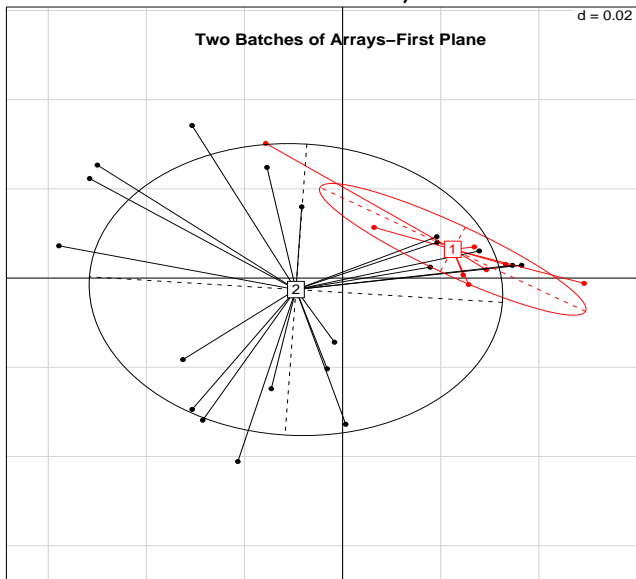
Project Various Factors as class labels onto the first few coordinates (components, factors,...). Projection of supplementary group centers (means) and ellipses of variation (variance) as in the function `s.class` in the `ade4` package Example: Explore batch effects in the laboratory methods used to generate the data. See quality of replicates

| | | | |
|-------------------------------|--|--|---|
| | | | d = 0.5 |
| mMS.43.27.B.11 mMS.43.27.B | | | mSS.72.77.3.1 mSS.72.77.3.1 mSS.72.77.3.1 |
| | | | |
| | | | |
| | | | mLS.73.75.A mLS.73.76.B |
| | | | |

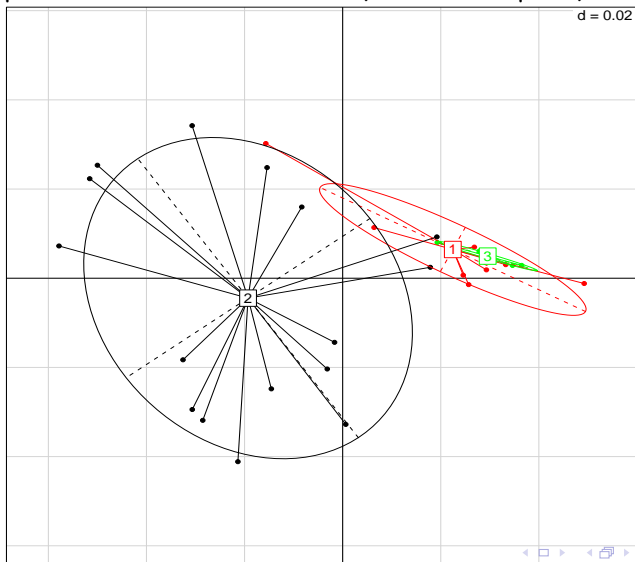
Projection of a categorical variable on a PCA

In the case of PCA: The ellipses are computed using the means, variances and covariance of each group of points on both axes, and are drawn with these parameters: the center of the ellipse is centered on the means, its width and height are given by the variances, and the covariance sets the slope of the main axis of the ellipse.

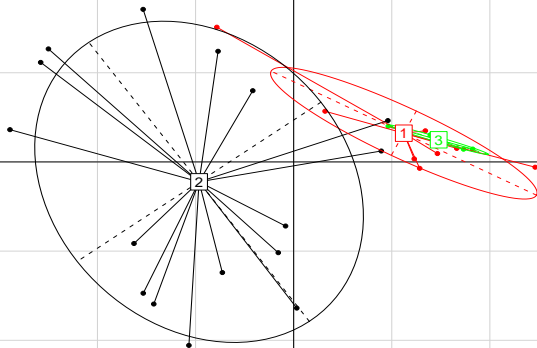
First two batches (in black and red)
(although both balanced with regards to IBS and healthy rats)
were extremely different in variability and overall multivariate
location. Batches were done different days with different sets of



A third batch was generated with the same arrays as batch 2 but the same experimental protocol as batch 1. The third group faithfully overlaps with batch 1 thus showing that the batch effect was not due to a difference in arrays but to the experimental protocol. This shows the utility of PCA in quality control.



$d = 0.02$



Chisquare: Aims and Relevant Data

What is a contingency table?

An example (thanks to the xkcd blog):

| | black | blue | green | grey | orange | purple | white |
|-----------|--------|-------|-------|-------|--------|--------|-------|
| quiet | 27700 | 21500 | 21400 | 8750 | 12200 | 8210 | 25100 |
| angry | 29700 | 15300 | 17400 | 7520 | 10400 | 7100 | 17300 |
| clever | 16500 | 12700 | 13200 | 4950 | 6930 | 4160 | 14200 |
| depressed | 14800 | 9570 | 9830 | 1470 | 3300 | 1020 | 12700 |
| happy | 193000 | 83100 | 87300 | 19200 | 42200 | 26100 | 91500 |
| lively | 18400 | 12500 | 13500 | 6590 | 6210 | 4880 | 14800 |
| perplexed | 1100 | 713 | 801 | 189 | 233 | 152 | 1090 |
| virtuous | 1790 | 802 | 1020 | 200 | 247 | 173 | 1650 |

Correspondence analysis (CA, also called homogeneity analysis and reciprocal averaging), can be used to analyse several types of multivariate data. All involve some categorical variables. Here are some examples of the type of data that can be decomposed using this method:

- ▶ Contingency Tables (cross between two categorical variables)
- ▶ Multiple Contingency Tables (cross between several categorical variables).
- ▶ Binary tables obtained by cutting continuous variables into classes and then recoding both these variables and any extra categorical variables into 0/1 tables, 1 indicating presence in that class. So for instance a continuous variable cut into three classes will provide three new binary variables of which only one can take the value 1 for any given observation.

For a complete treatment, see the paper: Multivariate Data Analysis: the French Way[6] (see my homepage for papers).

To first approximation, correspondence analysis can be understood as an extension of principal components analysis (PCA) where the variance in PCA is replaced by an inertia proportional to the χ^2 distance of the table from independence. CA decomposes this measure of departure from independence along axes that are orthogonal according to the χ^2 inner product.

Plato's works

In statistics the most commonplace use of Correspondence Analysis is in ordination or seriation, that is , the search for a hidden gradient in contingency tables. As an example we take data analysed by Cox and Brandwood who wanted to seriate Plato's works using the proportion of sentence endings in a given book, with a given stress pattern. We propose the use of correspondence analysis on the table of frequencies of sentence endings.

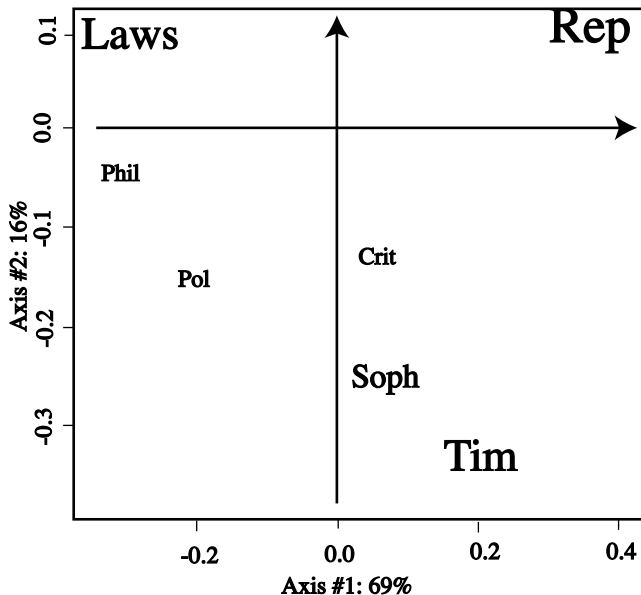
The first 10 profiles (as percentages) look as follows:

| | Rep | Laws | Crit | Phil | Pol | Soph | Tim |
|-------------------------------------|-----|------|------|------|-----|------|-----|
| UUUUU | 1.1 | 2.4 | 3.3 | 2.5 | 1.7 | 2.8 | 2.4 |
| -UUUU | 1.6 | 3.8 | 2.0 | 2.8 | 2.5 | 3.6 | 3.9 |
| U-UUU | 1.7 | 1.9 | 2.0 | 2.1 | 3.1 | 3.4 | 6.0 |
| UU-UU | 1.9 | 2.6 | 1.3 | 2.6 | 2.6 | 2.6 | 1.8 |
| UUU-U | 2.1 | 3.0 | 6.7 | 4.0 | 3.3 | 2.4 | 3.4 |
| UUUU- | 2.0 | 3.8 | 4.0 | 4.8 | 2.9 | 2.5 | 3.5 |
| --UUU | 2.1 | 2.7 | 3.3 | 4.3 | 3.3 | 3.3 | 3.4 |
| -U-UU | 2.2 | 1.8 | 2.0 | 1.5 | 2.3 | 4.0 | 3.4 |
| -UU-U | 2.8 | 0.6 | 1.3 | 0.7 | 0.4 | 2.1 | 1.7 |
| -UUU- | 4.6 | 8.8 | 6.0 | 6.5 | 4.0 | 2.3 | 3.3 |
|etc (there are 32 rows in all) | | | | | | | |

The eigenvalue decomposition (called the scree plot) of the chisquare distance matrix shows that two axes out of a possible 6 (the matrix is of rank 6) will provide a summary of 85% of the departure from independence, this suggests that a planar representation will provide a good visual summary of the data.

| | Eigenvalue | inertia % | cumulative % |
|---|------------|-----------|--------------|
| 1 | 0.09170 | 68.96 | 68.96 |
| 2 | 0.02120 | 15.94 | 84.90 |

| | | | |
|---|---------|------|--------|
| 3 | 0.00911 | 6.86 | 91.76 |
| 4 | 0.00603 | 4.53 | 96.29 |
| 5 | 0.00276 | 2.07 | 98.36 |
| 6 | 0.00217 | 1.64 | 100.00 |



Correspondence Analysis of Plato's Works

We can see from the plot that there is a seriation that as in

most cases follows a parabola, horseshoe or arch from Laws on one extreme being the latest work and Republica being the earliest among those studied.

We will also consider 'multiple contingency tables' where more than two categorical variables are compared.

, , Sex = Male

Eye

| Hair | Brown | Blue | Hazel | Green |
|-------|-------|------|-------|-------|
| Black | 32 | 11 | 10 | 3 |
| Brown | 53 | 50 | 25 | 15 |
| Red | 10 | 10 | 7 | 7 |
| Blond | 3 | 30 | 5 | 8 |

, , Sex = Female

Eye

| Hair | Brown | Blue | Hazel | Green |
|-------|-------|------|-------|-------|
| Black | 36 | 9 | 5 | 2 |
| Brown | 66 | 34 | 29 | 14 |
| Red | 16 | 7 | 7 | 7 |
| Blond | 4 | 64 | 5 | 8 |

```
> HairColor=HairEyeColor[, ,2]  
> chisq.test(HairColor)
```

Pearson's Chi-squared test

```
data:  HairColor  
X-squared = 106.6637, df = 9, p-value < 2.2e-16
```

Warning message:

In chisq.test(HairColor) : Chi-squared approximation m

The data do not actually come in the form of a table usually

```
> HairColor
      Eye
Hair   Brown Blue Hazel Green
Black   36    9     5     2
Brown   66   34    29    14
Red     16    7     7     7
Blond    4   64     5     8
> sum(HairColor)
[1] 313
```

but as categorical variables, for instance:

| Id | Hair_color | Eye_color | Sex |
|-------|------------|-----------|-----|
| 1 | Brown | Brown | F |
| 2 | Blonde | Blue | M |
| 3 | Black | Hazel | F |
| | | .. | |
| 313 | Red | Brown | M |

We cross tabulate before we start the analysis.

```
> res.coa=dudi.coa(HairColor)
> s.label(res.coa$c1,boxes=F)
> s.label(res.coa$li,add.plot=TRUE)
```



Independence

If we are comparing two categorical variables, (hair color, eye color), (color, emotion), the simplest possible model is that of independence in which case the counts in the table would obey approximately the margin products identity for a $I \times J$ contingency table with a total sample size of $n = \sum_{i=1}^I \sum_{j=1}^J n_{ij} = n \dots$

$$n_{ij} \doteq \frac{n_{i.}}{n} \frac{n_{.j}}{n} n$$

can also be written: $\mathbf{N} \doteq \mathbf{c} \mathbf{r}' \mathbf{n}$, where

$$\mathbf{c} = \frac{1}{n} \mathbf{N} \mathbf{1}_m \quad \text{and} \quad \mathbf{r}' = \frac{1}{n} \mathbf{N}' \mathbf{1}_p$$

The departure from independence is measured by the χ^2 statistic

$$\chi^2 = \sum_{i,j} \left[\frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}} \right]$$

Verification in R:

```
> F=HairColor/sum(HairColor)
> r=apply(F,1,sum)
> c=apply(F,2,sum)
> E=outer(r,c)
> E
```

| | Brown | Blue | Hazel | Green |
|-------|------------|------------|------------|------------|
| Black | 0.06475518 | 0.06050894 | 0.02441589 | 0.01645418 |
| Brown | 0.17807674 | 0.16639958 | 0.06714369 | 0.04524901 |
| Red | 0.04607580 | 0.04305444 | 0.01737284 | 0.01170779 |
| Blond | 0.10086864 | 0.09425430 | 0.03803244 | 0.02563056 |

```
> sum((F-E)^2/E)
[1] 0.3407787
> sum((F-E)^2/E)*313
[1] 106.6637
> sum(E)
[1] 1
```

Matrix decomposition and χ^2 distances

To compute the distance between profiles, each column is reweighted by the inverse of its sum, this gives the χ^2 distance between row profiles.

$$\begin{aligned}\chi^2 &= n \operatorname{trace} ((\mathbf{F} - \mathbf{rc}')' \mathbf{D}_r^{-1} (\mathbf{F} - \mathbf{rc}') \mathbf{D}_c^{-1}) \\ &= \operatorname{trace} (\mathbf{A}' \mathbf{A}) \text{ where } \mathbf{A} = \mathbf{D}_{\sqrt{r}}^{-1} (\mathbf{F} - \mathbf{rc}') \mathbf{D}_{\sqrt{c}}^{-1}\end{aligned}$$

The latter decomposition shows a justification for choosing the matrix \mathbf{A} as a natural square root. $\mathbf{W} = \mathbf{A}' \mathbf{A}$ is in a sense the characteristic matrix-operator of the analysis, in the same way the covariance or correlation matrices are those of principal components analysis.

Correspondence analysis decomposes the matrix \mathbf{W} : its eigenvectors give the axes that account for the largest part of the departure from independence, just as principal components provides the axes accounting for the largest variability. Computationally this is achieved by a generalized singular value decomposition

$$\mathbf{D}_r^{-1} \mathbf{F} \mathbf{D}_c^{-1} - \mathbf{1}_m' \mathbf{1}_p = \mathbf{U} \mathbf{S} \mathbf{V}',$$

with $\mathbf{V}' \mathbf{D}_c \mathbf{V} = \mathbf{I}_p, \mathbf{U}' \mathbf{D}_r \mathbf{U} = \mathbf{I}_m$

equivalent to the eigendecomposition $\mathbf{W} = \mathbf{A}' \mathbf{A} = \mathbf{V}' \mathbf{S}^2 \mathbf{V}$ or the singular value decomposition

$$\mathbf{D}_r^{-\frac{1}{2}} \mathbf{F} \mathbf{D}_c^{-\frac{1}{2}} - \sqrt{\mathbf{r}} \sqrt{\mathbf{c}}' = (\mathbf{D}_r^{\frac{1}{2}} \mathbf{U}) \mathbf{S} (\mathbf{D}_c^{\frac{1}{2}} \mathbf{V})',$$

where $(\mathbf{D}_c^{\frac{1}{2}} \mathbf{V})' (\mathbf{D}_c^{\frac{1}{2}} \mathbf{V}) = \mathbf{I}_p$, and $(\mathbf{D}_r^{\frac{1}{2}} \mathbf{U})' (\mathbf{D}_r^{\frac{1}{2}} \mathbf{U}) = \mathbf{I}_p$.

Matrix Diagonalised and Diagram

```
> res.eigen=eigen(t(X)%*%diag(r)%*%X)%*%diag(c))
> res.eigen
$values
[1] 1.000000000 0.302459246 0.032631660 0.005687796

$vectors
      [,1]      [,2]      [,3]      [,4]
[1,]  0.5  0.58634594 -0.2752916  0.08027417
[2,]  0.5 -0.74350003 -0.1444609 -0.05137694
[3,]  0.5  0.31830243  0.5814691 -0.61374293
[4,]  0.5 -0.04571334  0.7518240  0.78373215

> res.coa$eig
[1] 0.302459246 0.032631660 0.005687796

> t(res.coa$co[,1])%*%diag(c)%*%res.coa$co[,1]
      [,1]
[1,] 0.3024592
```

Matrix Diagonalized and Diagram

```
> res.eigen=eigen(t(X)%*%diag(r)%*%X)%*%diag(c))
> res.eigen
$values
[1] 1.000000000 0.302459246 0.032631660 0.005687796

$vectors
      [,1]      [,2]      [,3]      [,4]
[1,]  0.5  0.58634594 -0.2752916  0.08027417
[2,]  0.5 -0.74350003 -0.1444609 -0.05137694
[3,]  0.5  0.31830243  0.5814691 -0.61374293
[4,]  0.5 -0.04571334  0.7518240  0.78373215

> res.coa$co
      Comp1      Comp2
Brown  0.54472970  0.13159215
Blue   -0.69072969  0.06905379
Hazel   0.29571073 -0.27794807
Green  -0.04246881 -0.35937943

> res.coa$co/res.eigen$vectors[,2:3]
```

Comp1

Comp2

Part II

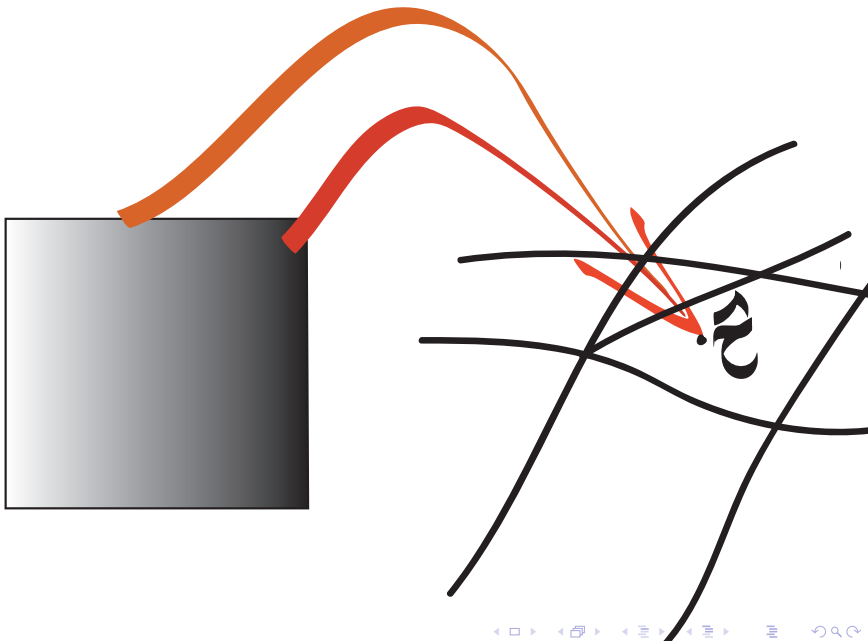
Inference: *Confirmatory Data Analysis*

Stability (Internal and External)
Hypothesis Testing
Evaluation

- ▶ The statistical paradigm.
 - ▶ Estimates and confidence.
 - ▶ Statistical approaches to variability.
 - ▶ Robustness
 - ▶ Not always just multivariate vectors.
- ▶ Building a space for any object with a natural distance.
- ▶ The Bootstrap.
- ▶ Special cases: multivariate statistics and Treespace

Estimate $\hat{\tau}$ computed from the data:

Data



- ▶ **Binary**

```
Lemur_cat    0000000000000001010100000
Tarsius_s    100000100000000010000000
Saimiri_s     100000100000001010000000
Macaca_sy     000000000000000010000000
Macaca_fa     100000100000000010000000
```

- ▶ **Aligned**

DNA Data for 12 species of primates

Mitochondria, 898 characters on 12 species,(Hayasaka, K., T. Go

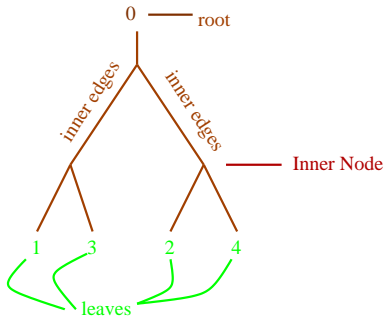
```

12  60
Lemur_cat  AAGCTTCATA GGAGCAACCA TTCTAATAAT CGCACATGGC CTTACATCAT CCATATTATT
Tarsius_s  AAGTTTCATT GGAGCCACCA CTCTTATAAT TGCCCATGGC CTCACCTCCT CCCTATTATT
Saimiri_s  AAGCTTCACC GCGCAATGA TCCTAATAAT CGCTCACGGG TTTACTTCGT CTATGCTATT
Macaca_sy  AAGCTTCTCC GGTGCAACTA TCCTTATAGT TGCCCATGGA CTCACCTCTT CCATATACTT
Macaca_fa  AAGCTTCTCC GCGCAACCA CCCTTATAAT CGCCCACGGG CTCACCTCTT CCATGTATTT
Macaca_mu  AAGCTTTTCT GCGCAACCA TCCTCATGAT TGCTCACGGA CTCACCTCTT CCATATATTT

```

These data sets usually come with their own metrics.

The parameter is a semi-labeled binary Tree



Statistical Paradigms

Classical Frequentist

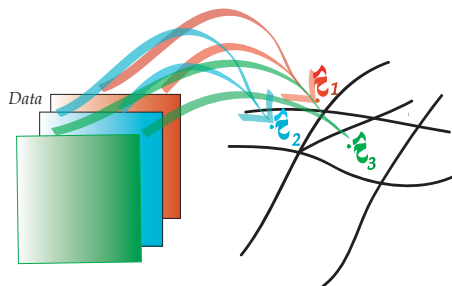
- estimate the parameter,
(either in a parametric (ML) way,
semiparametric (Distance based methods),
or nonparametric way (Parsimony))
- find the sampling Distribution of the estimator.

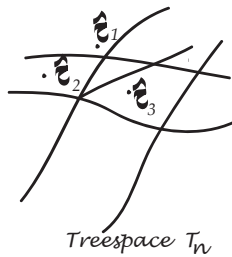
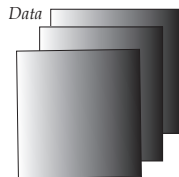
Bayesian

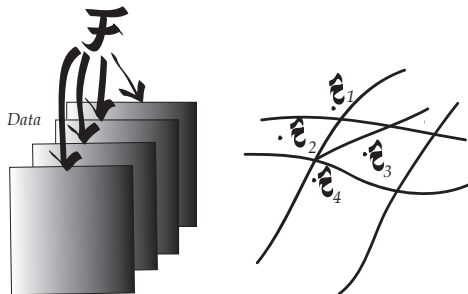
- Specify a Prior Distribution
- Update the prior using the Data
- Compute the Posterior Distribution

Difficulties arise as the estimators lie in a non Euclidean space.

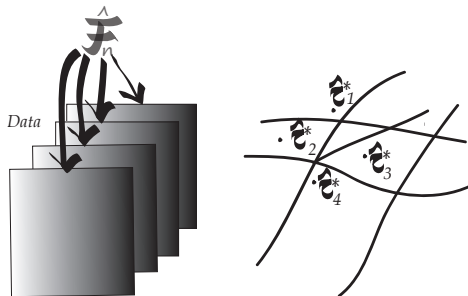
Sampling Distribution for Trees



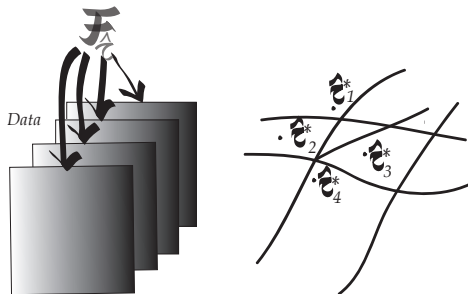




True Sampling Distribution



*Bootstrap Sampling Distribution
(non parametric)*



Bootstrap Sampling Distribution
(parametric)

How do we define distributions on Treespace

- ▶ Not the uniform distribution.
- ▶ By inspiration from ranked Data: Mallows's model (1957)

$$P(\tau_i) = K e^{-\lambda d(\tau_i, \tau_0)}$$

- ▶ Uses a central tree τ_0 .
 - ▶ Uses a distance d in treespace.
 - ▶ But very symmetrical, maybe need a mixture model.
- ▶ Other distributions (see Aldous, 2001), one might want to include information about the estimation method used as this influences the shape of the tree.

Classical Statistical Summaries

- ▶ Expectation (center of the distribution).
Open question: What distribution is the consensus such a center of?
- ▶ Median (multivariate median (Tukey, 1972)).
- ▶ Variance (second moment $E_{p_n} d^2(\hat{\tau}, \tau)$).
- ▶ Presence/Absence of a clade.
- ▶ Summaries of Multivariate Variabilities. (PCA, MDS, ...).

Frequentist Confidence Regions

$$P(\tau \in \mathcal{R}_\alpha) = 1 - \alpha$$

We will use the nonparametric approach of Tukey who proposed peeling convex hulls to construct successive 'deeper' confidence regions. But we need a geometrical space to build these regions in.

The Bootstrap

- ▶ What is the best estimate?(unbiased, most efficient)
- ▶ How sure are we of the estimate? Confidence regions.

Bootstrap support for Phylogenies Taking as observations the columns of the matrix X of aligned sequences, the rows representing the species.

The sampling distribution of the estimated tree is estimated by resampling with replacement among the characters or columns of the data.

This provides a large set of plausible alternative data sets, each be used in the same way as the original data to give a separate tree (see [5] for a review).

Parametric Bootstrapping for Microarray Clusters

Bayesian posterior distributions for phylogenetic trees ▶

Prior distributions on the DNA mutation rates that occur during the evolutionary process and a uniform distribution on the original tree.

- ▶ Use of MCMC to generate instances of the posterior distribution.
- ▶ Implementations MrBayes MrBayes and Beast provide a sample of trees from the posterior distribution.
- ▶ The posterior distribution provides an estimate of variability.

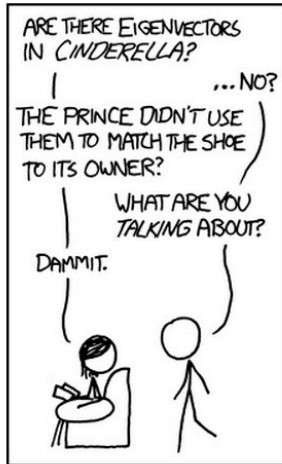
Bayesian methods in hierarchical clustering Heller[9]
provide a Bayesian nonparametric method
for generating posterior distributions of
hierarchical clustering trees.

Summary of some useful vocabulary

- ▶ Parameters are fixed and unknown (not to everyone?)
- ▶ Use estimates computed from samples for these unknowns.
- ▶ Linear methods with differing metrics.
- ▶ Confidence regions depends on paradigm, all use probability, many use simulation.
- ▶ Robustness, nonparametric (doesn't mean no parameters).
- ▶ Distributions in multivariate space are rarely parametric (never uniform)
- ▶ Reproducible Research: how to solve the problem that there are an exponential number of choices of metrics, thresholds, models, methods?
Kepp a diary, show your work: reproducible research, see an example: [?].

Extensions

- ▶ Transformations that allow for nonlinearity (Kernel PCA, etc).
- ▶ Transformation of variables to make them comparable (sphering).
- ▶ Transformation of variables to make them robust (ranks).



Papers available here:

<http://www-stat.stanford.edu/~susan/>



L. Billera, S. Holmes, and K. Vogtmann.

The geometry of tree space.

Adv. Appl. Maths, 771--801, 2001.



J. Chakerian and S. Holmes.

Computational methods for evaluating phylogenetic trees, 2010.

arXiv.



P. Diaconis, S. Goel, and S. Holmes.

Horseshoes in multidimensional scaling and kernel methods.

Annals of Applied Statistics, 2007.



B. Efron.

Bootstrap methods: Another look at the jackknife.

The Annals of Statistics, 7:1--26, 1979.



S. Holmes.

Bootstrapping phylogenetic trees: theory and methods.

Statist. Sci., 18(2):241--255, 2003.

Silver anniversary of the bootstrap.



Susan Holmes.

Multivariate Analysis: The French way, volume 56 of
IMS Lecture Notes--Monograph Series.

IMS, Beachwood, OH, 2006.



R. Ihaka and R. Gentleman.

R: A language for data analysis and graphics.

Journal of Computational and Graphical Statistics,
5(3):299--314, 1996.



K. Mardia, J. Kent, and J. Bibby.

Multivariate Analysis.

Academic Press, NY., 1979.



R Savage, K Heller, Y Xu, and Z. Ghahramani.

R/BHC: fast Bayesian hierarchical clustering for
microarray data.

BMC, Jan 2009.



I.J. Schoenberg.

Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de Hilbert.

The Annals of Mathematics, 36(3):724--732, July 1935.