

Using distances for heterogeneous data analyses.

Susan Holmes

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

Bio-X and Statistics, Stanford University

September 1, 2019

Part I

Heterogeneity

'Homogeneous data are all alike;

all heterogeneous data are heterogeneous

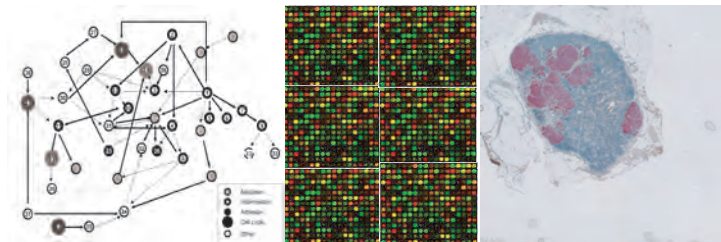
in their own way.'

Heterogeneity of Data

- ▶ Status : response/ explanatory.
- ▶ Hidden (latent)/measured.
- ▶ Types :
 - ▶ Continuous
 - ▶ Binary, categorical
 - ▶ Graphs/ Trees
 - ▶ Images
 - ▶ Maps/ Spatial Information
 - ▶ Rankings
- ▶ Amounts of dependency: independent/time series/spatial.
- ▶ Different technologies used (454, Illumina, PacBio, MassSpec, RNA-seq, Cytof).

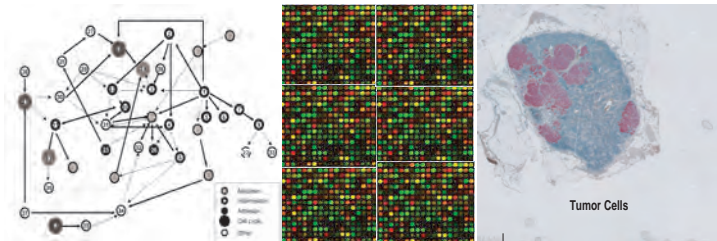
Goals in Modern Biology: Systems Approach

Look at the data/ all the data: data integration



Goals in Modern Biology: Systems Approach

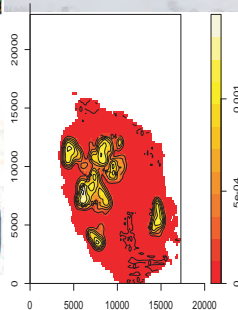
Look at the data/ all the data: data integration



$$\begin{pmatrix} 0110-11000-1 \\ 0110000001 \\ 01-10-10000-1 \\ 0110000101 \\ 01100-11011 \end{pmatrix}$$

$$X_{Blood} = \begin{pmatrix} 0.5 & 1.1 & 1.6 & 1.2 & \dots \\ 0.3 & 1.9 & 2.2 & 1.1 & \dots \\ 1.1 & 0 & 3.2 & 0.4 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 2.7 & 2.3 & 1.2 & 1.1 & \dots \end{pmatrix}$$

$$X_{LN} = \begin{pmatrix} 0.45 & 0.13 & 1.06 & 1.2 & \dots \\ 0.53 & 0.95 & 2.26 & 5.12 & \dots \\ 0.11 & 0 & 3.2 & 1.24 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0.27 & 0.33 & 4.2 & 1.1 & \dots \end{pmatrix}$$



What do statisticians do?

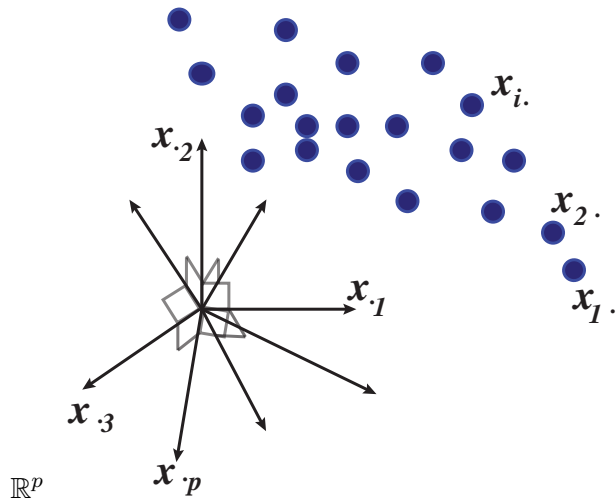
- ▶ Design new experiments to test scientific hypotheses.
- ▶ Visualize and summarize data in ways that account for uncertainties.
- ▶ Look for meaningful differences or structure in high dimensional noisy data.
- ▶ Predict the class of new observations given previously observed ones.
- ▶ Predict the value of a response variable given a whole set of other explanatory variables.
- ▶ Combine different sources of data to understand complex interactions.

Today's challenge

- ▶ Data are not uniformly distributed from some manifold.
- ▶ Data are not an identically distributed random sample.
- ▶ Data are not independent.
- ▶ Data may be combined from different source types (multiway).



Data can often be seen as points in a state space



Distances in Statistics

- ▶ Euclidean Distances, spatial distances.
- ▶ Weighted Euclidean distances: Mahalanobis distance for discriminant analysis.
- ▶ Chisquare distances for contingency tables and discrete data.
- ▶ Jaccard distances for presence absence is one of 50 distances used in Ecology.
- ▶ Earth Mover's distance **on** trees or graphs.
- ▶ Distances **between** aligned graphs or trees.
- ▶ Biologically meaningful distances (DNA, haplotype, Proteins).

What do statisticians use distances for?

- ▶ Summaries through Fréchet Means and Medians and pseudo variances.
 - ▶ Center of Cloud of Objects T_k (equal weights): Find T_0 that minimizes either $\sum_{k=1}^K d^2(T_0, T_k)$ this is the (L^2) definition of the Fréchet mean object,
 - ▶ or $\sum_{k=1}^K d(T_0, T_k)$ (L^1 or Geometric Median).
 - ▶ Pseudovariance = $\frac{1}{K-1} \sum_{k=1}^K d^2(T_0, T_k) = \hat{s}^2$.

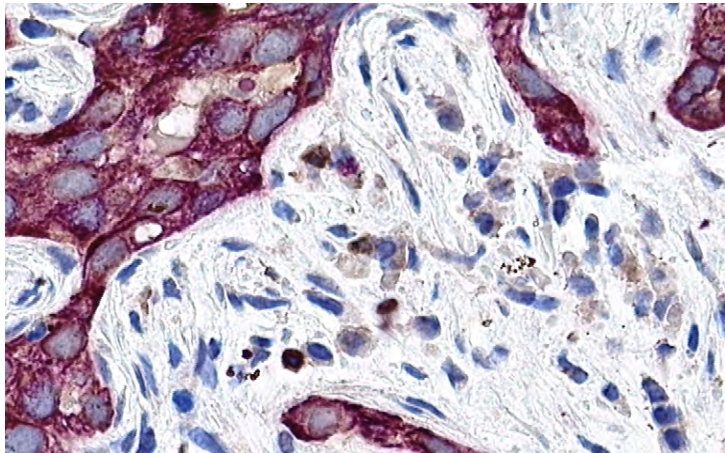
What do statisticians use distances for?

- ▶ Summaries through Fréchet Means and Medians and pseudo variances.
- ▶ Dimension reduction and visualization.
- ▶ Nearest Neighbor Methods.
- ▶ Clustering.
- ▶ Make network edges from close points.
- ▶ Prediction by minimizing weighted residual distances.
- ▶ Cross-products: correlations, autocorrelations.
- ▶ Generalizations of analysis of variance.

Finding the right distance usually solves the statistical problem.

Part II

The Geometries of Data



First example: cell segmentation

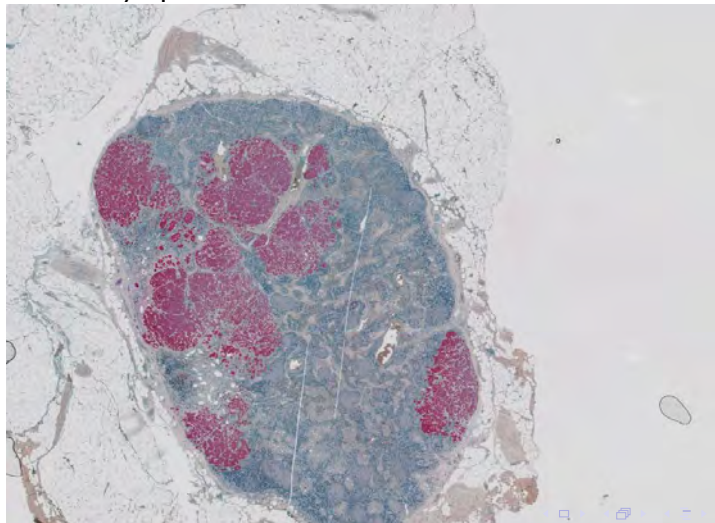
Joint work with Adam Kapelner and PP Lee.

Stained biopsy slides.
levels/wavelengths).

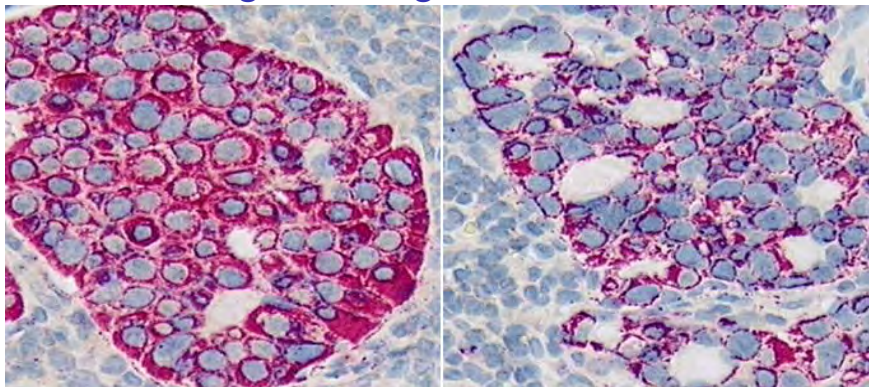
Stained Lymph Node

Multispectral imaging (8

Aim to identify cell.



Problem : Staining is heterogeneous



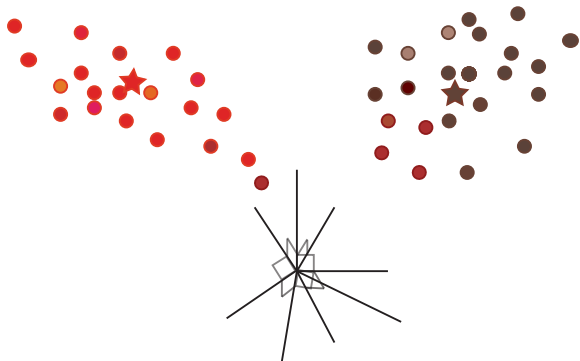
Both images are from the same image set. The stained cells are cancer cells stained with Fast Red red.

Some regions of the tissue stain like the image on the left and other regions stain as the left.

This shows the level of heterogeneity These are two “subclasses” of the same phenotype (the left is named subclass “A,” the right, subclass “B”).

Problem : Staining is heterogeneous

Extreme variability in the image colors/intensity/contrast.
Pixels from a same cell not independent and identically distributed across the different slides or across different cell types.

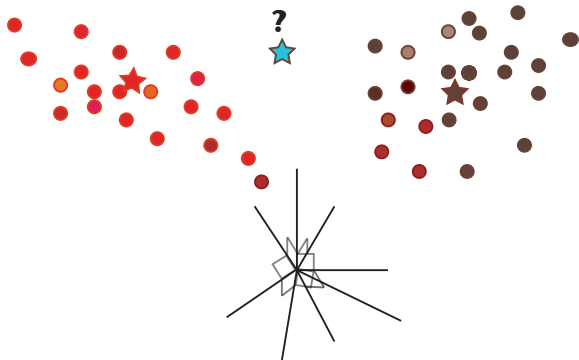


Simple nearest neighbor approach:

- Take 8 dimensional pixels points.
- Assigning the point to the closest neighbor

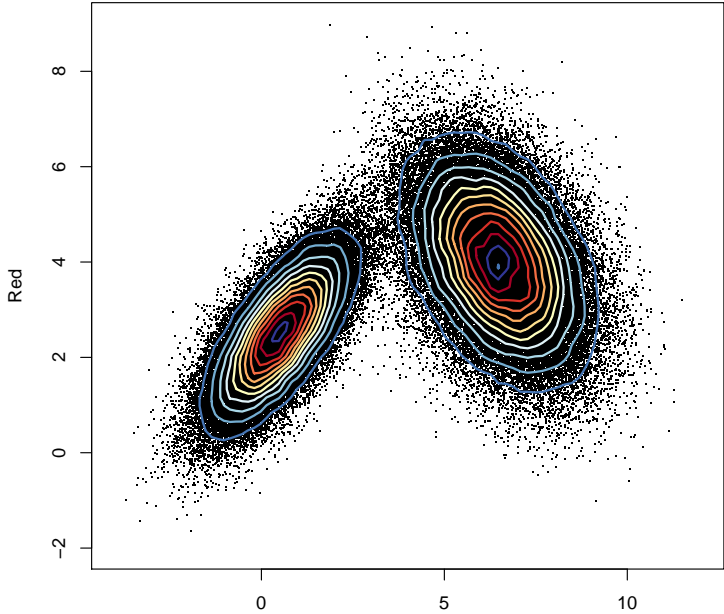
Problem : Staining is heterogeneous

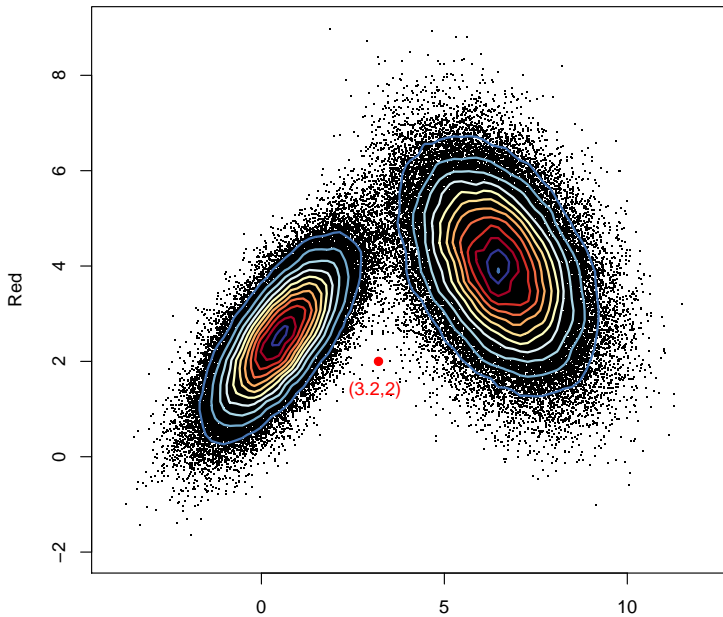
Extreme variability in the image colors/intensity/contrast.
Pixels from a same cell not independent and identically distributed across the different slides or across different cell types.

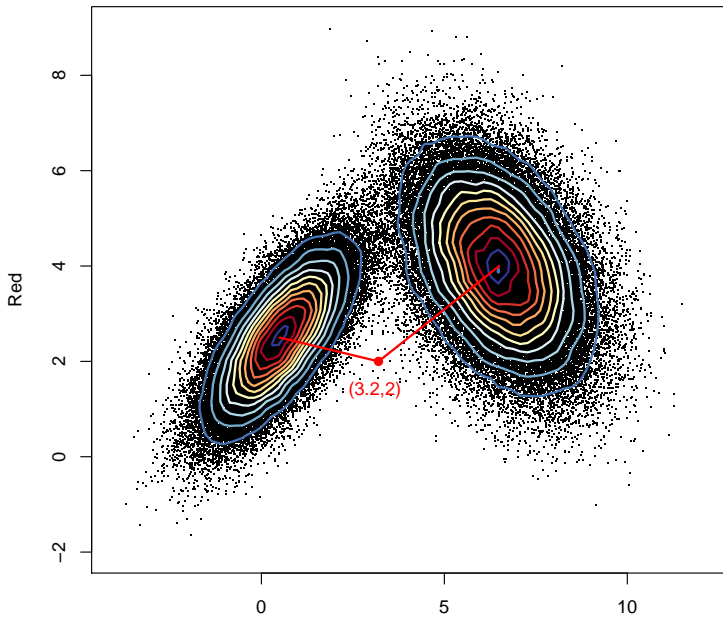


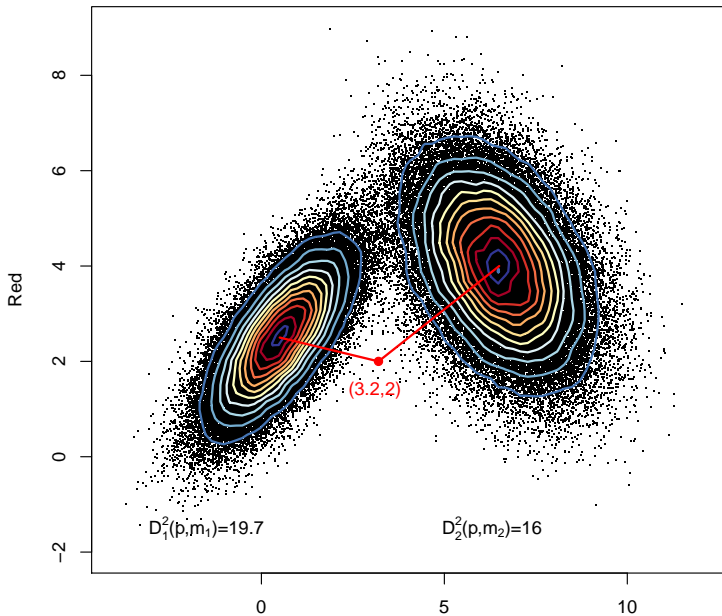
Simple nearest neighbor approach:

- Take 8 dimensional pixels points.
- Assigning the point to the closest neighbor









Multivariate Normal Data

Mahalanobis Transformation.

Several different clusters with different variance-covariance matrices and different means.

(μ_1, Σ_1) (μ_2, Σ_2)

$$D_1^2(x, \mu_1) = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$$

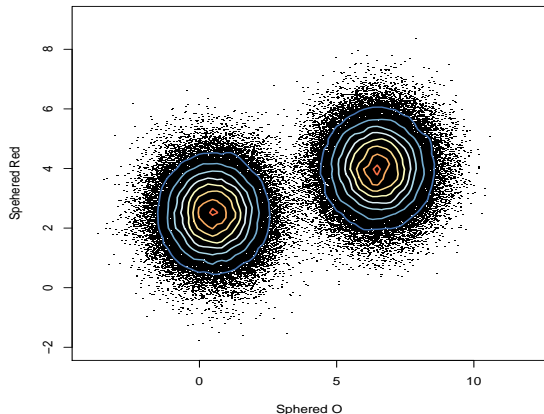
$$D_2^2(x, \mu_2) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$

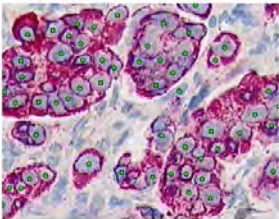
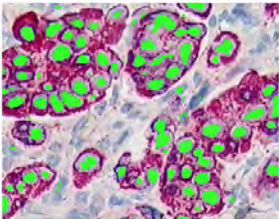
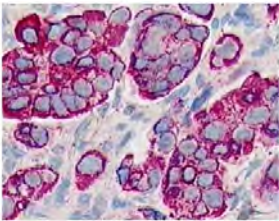
Corresponding Data Transformation


$$H = I - 1D_n1^T, \quad S = X'HD_nHX$$

$$z_{i.} = S^{-\frac{1}{2}}(x_{i.} - \bar{x})$$

This is sometimes called 'data sphering'.







Color: 

rMin: 1

rMax: 9

Pix/Cent:


Name: tic cells 26 


Color: 

rMin: 1

rMax: 9

Pix/Cent:

Name: Tumor 111 

Color: 

rMin: 1


rMax: 10

Settings: Visualize:

See points:

See Typel Errors (1)

Image Colors:

Image options: 

T_cells identified pixels visibility

T_cells centroids visibility

Dendritic cells identified pixels visibility

Dendritic cells centroids visibility

Tumor identified pixels visibility


Tumor centroids visibility


other_cells identified pixels visibility


other_cells centroids visibility


Mark / Delete Training Point: Location: (962,319)


Boost on classified images:


0-7-0-0-0 stage 0232 


185-18-7-31-12 stage 0093 


258-14-6-35-11 stage 0147 


112-10-5-11-9 stage 0181 

165-12-11-10-10 stage 0096 

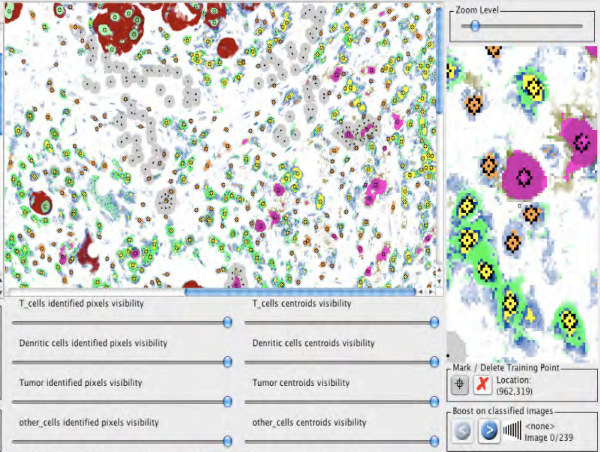
0-0-0-2-0 stage 0278 

237-4-3-20-12 stage 0318 

0-0-4-0-0 stage 0255 

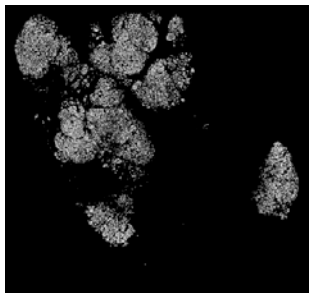
0-0-0-2-0 stage 0126 

Zoom Level:



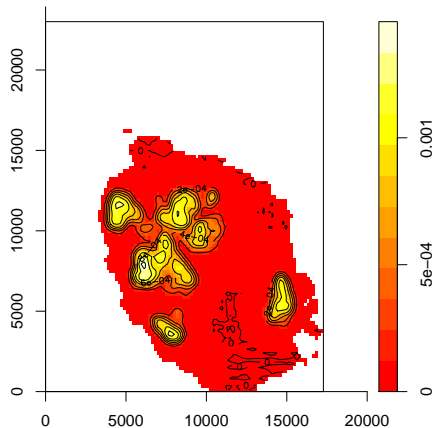
Output Data

Tumor



Number of Tumor cells: 27,822

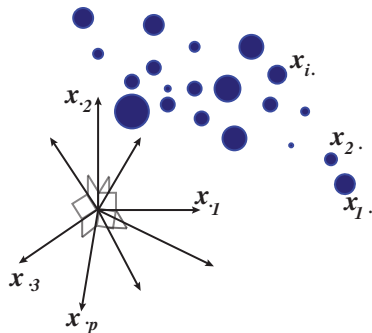
Tumor Cells



We can add information through choice of distances

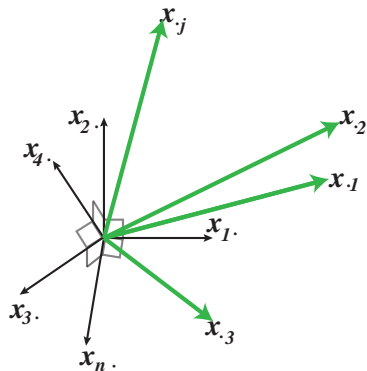
Sample data can often be seen as points in a state space.

\mathbb{R}^p



Variables are 'vectors' in data point space

\mathbb{R}^n



$$x^t Q y = \langle x, y \rangle_Q$$

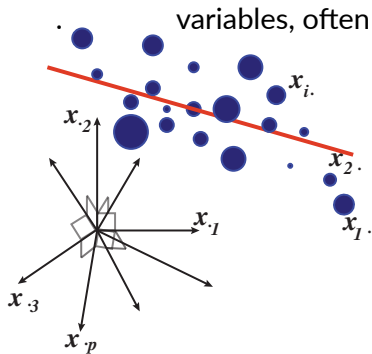
Duality : Transposable data.

$$x^t D y = \langle x, y \rangle_D$$

Data Analysis: 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 is an $n \times n$ matrix of weights on the “observations”, which is most often diagonal but not always.
- iv Symmetric definite positive matrix Q , weights on

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



Euclidean Space and dimension reduction

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

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

$$\begin{aligned}x^t Q y &= \langle x, y \rangle_Q & x, y &\in \mathbb{R}^p \\x^t D y &= \langle x, y \rangle_D & x, y &\in \mathbb{R}^n.\end{aligned}$$

This can be extended to more inner products giving what is known as **Kernel** methods.

Principal Component Analysis: Dimension Reduction

PCA seeks to replace the original (centered) matrix X by a matrix of lower rank, this can be solved using the singular value decomposition of X :

$$X = USV', \text{ with } U'DU = I_n \text{ and } V'QV = I_p \text{ and } S \text{ diagonal}$$

$$XX' = US^2U', \text{ with } U'DU = I_n \text{ and } S^2 = \Lambda$$

PCA is a linear nonparametric multivariate method for dimension reduction. D and Q are the relevant metrics on the dual row and column spaces of n samples and p variables.

A Commutative Diagram Approach

Caillez and Pages, 1976. Escoufier, 1977.

Statisticians search for approximations with certain properties, for the case of PCA for instance, we rephrase the problem as follows:

- ▶ 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 $\mathbb{R}^{n^*} = \mathcal{L}(\mathbb{R}^n)$.
- ▶

$$\begin{array}{ccccc} \mathbb{R}^{p^*} & \xrightarrow{\quad X \quad} & \mathbb{R}^n & & \\ & & & & \\ V = X^t D X & Q \uparrow & \downarrow V & D \downarrow & \uparrow W & W = X Q X^t \\ & & \mathbb{R}^p & \xleftarrow{\quad X^t \quad} & \mathbb{R}^{n^*} & \end{array}$$

This duality gives 'transposable' data.

Properties of the Diagram

Rank of the diagram:

X, X^t, 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

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

Modern extensions to this approach include Kernel methods in Machine Learning.

Comparing Two Diagrams: the RV coefficient

Many problems can be rephrased in terms of comparison of two “duality diagrams” or put more simply, two characterizing operators, built from two “triplets”, usually with one of the triplets being a response or having constraints imposed on it. Most often what is done is to compare two such diagrams, and try to get one to match the other in some optimal way. ($O = WD$)

To compare two symmetric operators, there is either a vector covariance as inner product

$covV(O_1, O_2) = Tr(O_1^t O_2) = \langle O_1, O_2 \rangle$ or a vector correlation (Escoufier, 1977)

$$RV(O_1, O_2) = \frac{Tr(O_1^t O_2)}{\sqrt{Tr(O_1^t O_1) tr(O_2^t O_2)}}.$$

If we were to compare the two triplets $(X_{n \times 1}, 1, \frac{1}{n} I_n)$ and $(Y_{n \times 1}, 1, \frac{1}{n} I_n)$ we would have $RV = \rho^2$.

PCA: Approximating one diagram by another

PCA can be seen as finding the matrix Y which maximizes the *RV* coefficient between characterizing operators, that is, between $(X_{n \times p}, Q, D)$ and $(Y_{n \times q}, I, D)$, under the constraint that Y be of rank $q < p$.

$$RV(XQX^tD, YY^tD) = \frac{Tr(XQX^tDYY^tD)}{\sqrt{Tr(XQX^tD)^2 Tr(YY^tD)^2}}.$$

This maximum is attained where Y is chosen as the first q eigenvectors of XQX^tD normed so that $Y^tDY = \Lambda_q$. The maximum RV is

$$RV_{max} = \frac{\sum_{i=1}^q \lambda_i^2}{\sum_{i=1}^p \lambda_i^2}.$$

Of course, classical PCA has $D = \frac{1}{n}\mathcal{I}$, $Q = \mathcal{I}$, but the extra flexibility is often useful. We define the distance between triplets (X, Q, D) and (Z, Q, M) where Z is also $n \times p$, as the distance deduced from the RV inner product between operators XQX^tD and ZMZ^tD .

Discriminant Analysis as a duality diagram

Case of a categorical response variable (group labels).

Let A be the $g \times p$ matrix of group means in each of the p variables. This satisfies

$$Y^t D X = \Delta_Y A \quad \text{where } \Delta_Y = Y^t D Y = \text{diag}(w_1, w_2, \dots, w_g),$$

and $w_k = \sum_{i:y_{ik}=1} d_i$, the w_k 's are the group weights, as they are the sums of the weights as defined by D for all the elements in that group.

Call T the matrix $T = X^t D X$, in the standard case with all diagonal elements of D equal to $\frac{1}{n}$ this is just the standard variance-covariance, otherwise it is a generalization thereof.

The generalized between group variance-covariance is $B = A^t \Delta_Y A$ and call the between group variance covariance the matrix $W = (X - Y A)^t D (X - Y A)$.

A generalized Huyghens' formula:

$$T = B + W$$

Proof: Expanding W gives

$$\begin{aligned} W &= X^t DX - X^t DYA - A^t Y^t DX + A^t Y^t DYA \\ &= T - A' \Delta_Y A - A' \Delta_Y A + A' \Delta_Y A = T - B \end{aligned}$$



Duality Diagram for LDA

The duality diagram for linear discriminant analysis is

$$\begin{array}{ccc} \mathbb{R}^{p^*} & \xrightarrow{\quad A \quad} & \mathbb{R}^g \\ T^{-1} \uparrow & & \downarrow \Delta_Y \\ \mathbb{R}^p & \xleftarrow{\quad A^t \quad} & \mathbb{R}^{g^*} \\ & & \uparrow AT^{-1}A^t \end{array}$$

This corresponds to the triple (A, T^{-1}, Δ_Y) , because

$$(X^t D Y) \Delta_Y^{-1} (Y^t D X) = A^t \Delta_Y A$$

and gives equivalent results to the triple $(Y^t D X, T^{-1}, \Delta_Y^{-1})$.

The discriminating variables are the eigenvectors of the operator

$$A^t \Delta_Y A T^{-1}.$$

Part III

Combine and Compare Trees,
Graphs and Contingent Count Data
for the Human Microbiome

Layers of Data in the Microbiome

Joshua Lederberg: 'the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease'

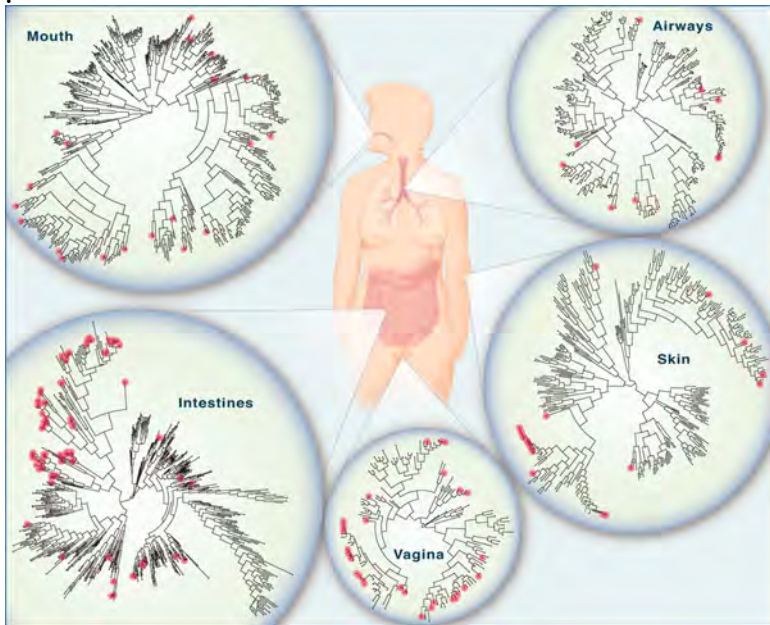
Microbiome Complete collection of genes contained in the genomes of microbes living in a given environment.

Numbers Humans shelter 100 trillion microbes (10^{14}), (we are made of 10×10^{12} cells).

Metagenome Composition of all genes present in an environment (soil, gut, seawater), regardless of species.

Transcriptome These are the mRNA transcripts in the cell, it reflects the genes that are being actively expressed at any given time.

Metabolome The metabolites (small molecules) nucleic or fatty acids, sugars,... present in the sample either endogenous or exogenous (medication, pollution).



Source: YK Lee and SK Mazmanian Science, 2010.

Bacteria etc... and Us

The human microbiome or human microbiota is the assemblage of microorganisms that reside on the surface and in deep layers of skin, in the saliva and oral mucosa, in the conjunctiva, and in the gastrointestinal tracts.

- ▶ They include bacteria, fungi, and archaea.
- ▶ Some of these organisms perform tasks that are useful for the human host. (live in symbiosis)
- ▶ Majority have no known beneficial or harmful effect.

Human Microbiome: What are the data?

DNA The Genomic material present (16sRNA-gene especially, but also shotgun).

RNA What genes are being turned on (gene expression), transcriptomics.

Mass Spec Specific signatures of chemical compounds present (LC/MS, GC/MS).

Clinical Multivariate information about patients' clinical status, medication, weight.

Environmental Location, nutrition, drugs, chemicals, temperature, time.

Domain Knowledge Metabolic networks, phylogenetic trees, gene ontologies.

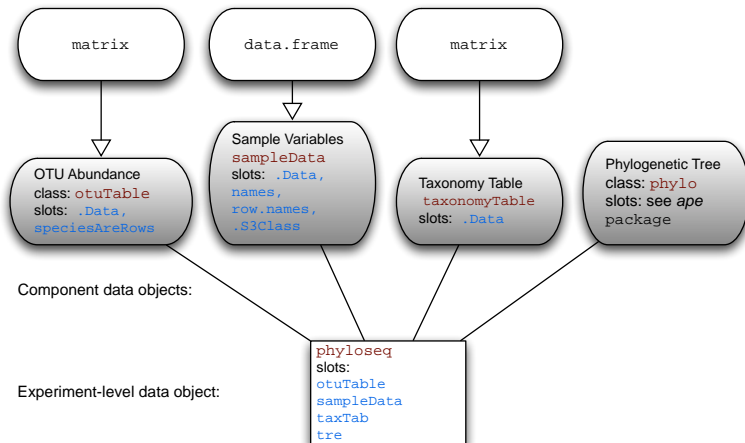
An example of taxa/specimen table.

ASV/OTU	Ctrl1	Ctrl2	Ctrl3	Ctrl4	Ctrl5	IBD1	IBD2	IBD3
Bacteroides	1822	913	147	2988	4616	172	3516	6
Bifidobacterium	0	162	0	0	84	0	85	19
Collinsella	1359	0	0	206	0	327	0	0
Enterococcus	621	0	0	3	40	0	0	0
Streptococcus	75	139	2161	110	97	1820	85	5

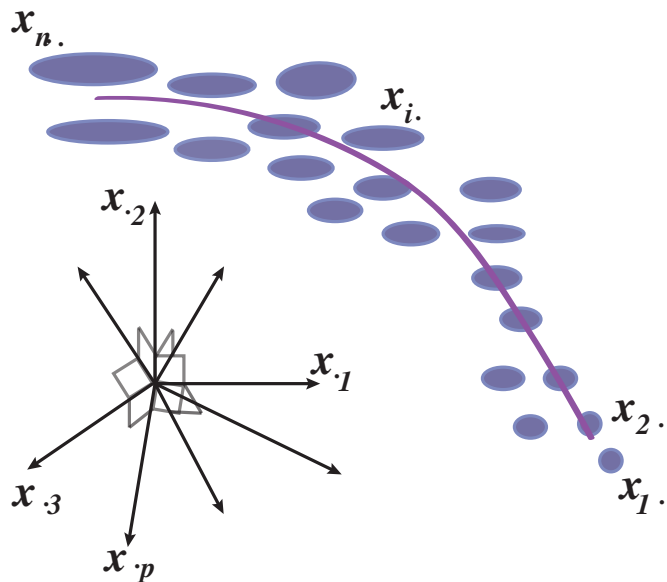
Heterogeneous Data Objects

Object oriented input and data manipulation with phyloseq
(McMurdie and Holmes, 2013, Plos ONE)

Object oriented data in R:

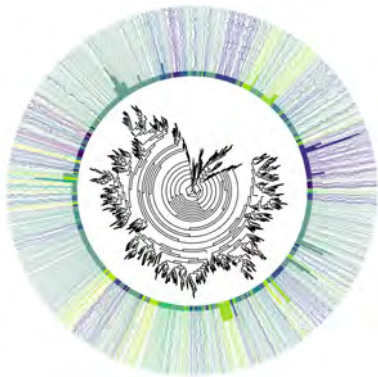


Points are measured with unequal variance



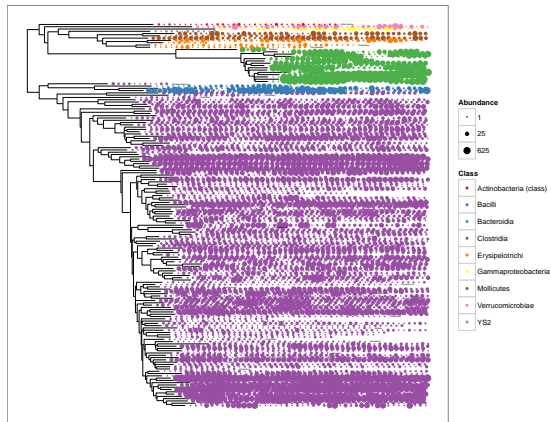
Part IV

Combining a phylogenetic tree with
the count data



A distance on the known tree

Monge-Kantorovich earth mover's distance on the tree.
Used to compare two samples or body sites for instance.
Incorporate taxa abundances and phylogenetic tree



Duality diagram methods that can use any dependency structure.

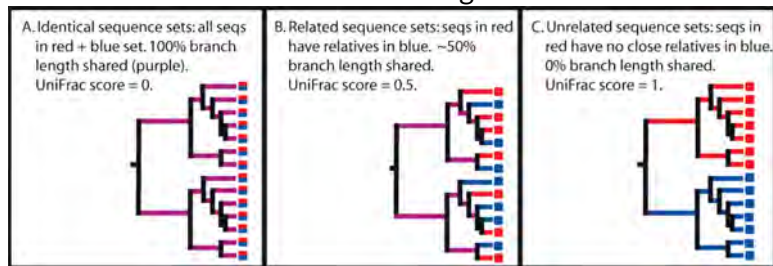
Unifrac Distance (Lozupone and Knight, 2005)

is a distance between groups of organisms that are related to each other by a tree.

Suppose we have the OTUs present in sample 1 (blue) and in sample 2 (red).

Question: Do the two samples differ phylogenetically?

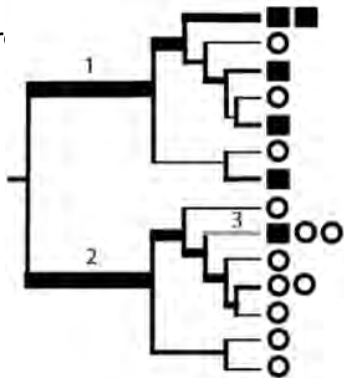
It is defined as the ratio of the sum of the lengths of the branches leading to members of group A or members of group B but not both to the total branch length of the tree.



Weighted UniFrac distance A modification of UniFrac, weighted UniFrac is defined in (Lozupone et al., 2007) as

$$\sum_{i=1}^n b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

- ▶ n = number of branches in the tree
- ▶ b_i = length of the i th branch
- ▶ A_i = number of descendants of i th branch in group A
- ▶ A_T = total number of sequences in group A

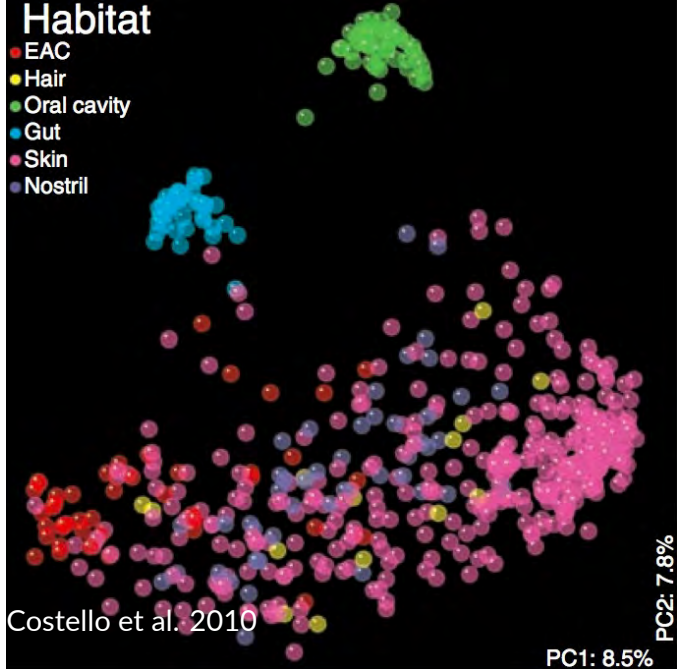


[6].

[7].

Habitat

- EAC
- Hair
- Oral cavity
- Gut
- Skin
- Nostril



Rao's Distance

We start with a distance between individuals.

The heterogeneity of a population (H_i) is the average distance between members of that population.

The heterogeneity between two populations (H_{ij}) is the average distance between a member of population i and a member of population j .

The distance between two populations is

$$D_{ij} = H_{ij} - \frac{1}{2}(H_i + H_j)$$

Decomposition of Diversity

If we have populations $1, \dots, k$ with frequencies π_1, \dots, π_k , then the diversity of all the populations together is

$$H_0 = \sum_{i=1}^k \pi_i H_i + \sum_i \sum_j \pi_i \pi_j D_{ij} = H(w) + D(b)$$

Double Principal Coordinate Analysis

Pavoine, Dufour and Chessel (2004), Purdom (2010) and Fukuyama et al. (2011). .

Suppose we have n species in p locations and a (euclidean) matrix Δ giving the squares of the pairwise distances between the species. Then we can

- ▶ Use the distances between species to find an embedding in $n - 1$ -dimensional space such that the euclidean distances between the species is the same as the distances between the species defined in Δ .
- ▶ Place each of the p locations at the barycenter of its species profile. The euclidean distances between the locations will be the same as the square root of the Rao dissimilarity between them.
- ▶ Use PCA to find a lower-dimensional representation of the locations.

Give the species and communities coordinates such that the inertia decomposes the same way the diversity does.

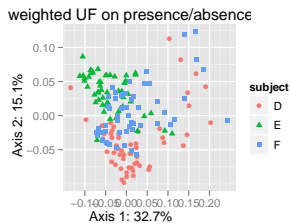
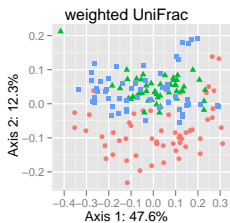
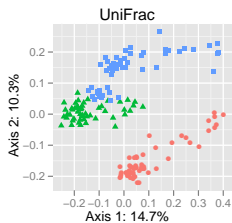
Fukuyama and Holmes, PSB, 2012.

Method	Original description	New formula	Properties
DPCoA	square root of Rao's distance based on the square root of the patristic distances	$[\sum_i b_i (A_i/A_T - B_i/B_T)^2]^{1/2}$	Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations
wUniFrac	$\sum_i b_i A_i/A_T - B_i/B_T $	$\sum_i b_i A_i/A_T - B_i/B_T $	Less sensitive to outliers/more sensitive to noise than DPCoA
UniFrac	fraction of branches leading to exactly one group	$\sum_i b_i \mathbf{1}\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

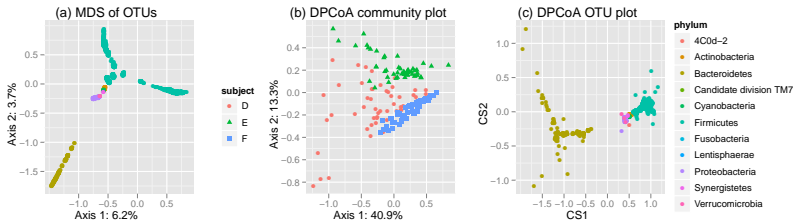
Summary of the methods under consideration. "Outliers" refers to highly abundant taxa, and noise refers to noise in detecting low-abundance taxa.

Antibiotic Time Course Data

Measurements of about 2500 different bacterial OTUs from stool samples of three patients (D, E, F)
Each patient sampled ~ 50 times during the course of treatment with ciprofloxacin (an antibiotic).
Times categorized as Pre Cp, 1st Cp, 1st WPC (week post cipro), Interim, 2nd Cp, 2nd WPC, and Post Cp.



Comparing the UniFrac variants. From left to right: PCoA/MDS with unweighted UniFrac, with weighted UniFrac, and with weighted UniFrac performed on presence/absence data extracted from the abundance data used in the other two plots

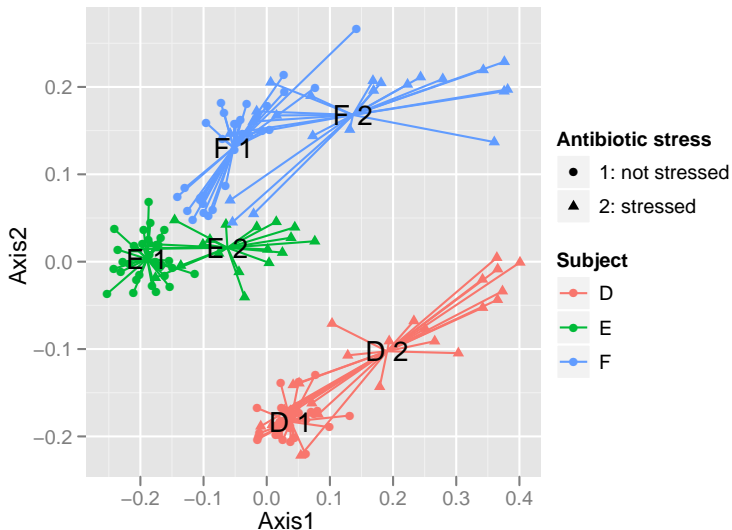


(a)

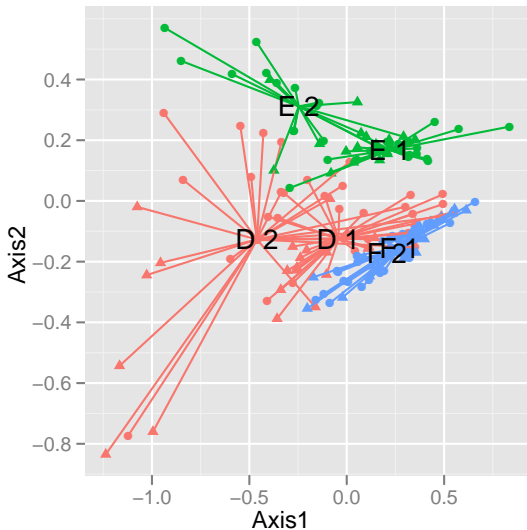
PCoA/MDS of the OTUs based on the patristic distance, (b) community and (c) species points for DPCoA after removing two outlying species.

Antibiotic Stress

We next want to visualize the effect of the antibiotic. Ordinations of the communities due to DPCoA and UniFrac with information about the whether the community was stressed or not stressed (pre cipro, interim, and post cipro were considered “not stressed”, while first cipro, first week post cipro, second cipro, and second week post cipro were considered “stressed”). We see that for UniFrac, the first axis seems to separate the stressed communities from the not stressed communities. DPCoA also seems to separate the out the stressed communities along the first axis (in the direction associated with *Bacteroidetes*), although only for subjects D and E.



PCoA/MDS with unweighted UniFrac. The labels represent subject plus antibiotic condition.



Community points as represented by DPCoA. The labels represent subject plus antibiotic condition.

Conclusions for Antibiotic Stress

Since UniFrac emphasizes shallow differences on the tree and since PCoA/MDS with UniFrac seems to separate the subjects from each other better than the other two methods, we can conclude that the differences between subjects are mainly shallow ones.

However, DPCoA also separates the subjects and the stressed versus non-stressed communities, and examining the community and OTU ordinations can tell us about the differences in the compositions of these communities.

Modulating the tree-based distances

We would like the axes to be both smooth on the tree and for which the projections of the samples have a large variance.

We can design an inner product on the rows which will pull out axes with these properties.

One extreme will be PCA without a tree, the other is DPCoA.

We create a family of gPCAs interpolating between DPCoA and standard PCA or as giving us a tunable parameter controlling how smooth we want the principal axes to be.

Adaptive gPCA



Fukuyama, Julia (2019), Ann. of Appl. Statistics.

We want to incorporate the prior (tree-like) information about the structure of the variables.

The intuition is that the variables which are similar to each other should behave in similar ways (in the case of microbiome data the idea is that species close together on the tree will behave similarly).

Perform generalized PCA on the posterior estimate of each sample given the data, taking into account the variance structure of the posterior.

Varying the scalings of the prior and noise variances gives a one-dimensional family of generalized PCAs which favor progressively smoother solutions according to the structure of the variables.

Data

Suppose we have a positive definite similarity matrix $Q \in \mathbb{R}^{p \times p}$ (a kernel matrix) between the variables.

To prevent scaling issues, assume that $\text{tr}(Q) = p$.

Note that since Q is positive definite, it is also a covariance matrix, and a random vector with covariance Q will have stronger positive correlations between variables which are more similar to each other.

Special case of the phylogenetic tree

Q is the matrix where Q_{ij} represents the amount of shared ancestral branch length between species i and j .

This is the kernel implicit in DPCoA; it is also related to the covariance of a Brownian motion run along the branches of the tree.

With this in mind, consider the following model for our data matrix X :

$$\mathbf{x}_i \stackrel{\text{iid}}{\sim} N(\mu_i, \sigma_2^2 I), \quad i = 1, \dots, n \quad (2)$$

$$\mu_i \stackrel{\text{iid}}{\sim} N(0, \sigma_1^2 Q), \quad i = 1, \dots, n \quad (3)$$

We are simply including prior knowledge into our model. The prior incorporates information about the structure in our variables: since the μ_i 's have covariance equal to a scalar multiple of Q , inference using this prior will allow us to regularize towards this structure, or to smooth the data towards our expectation that similar variables will behave in similar ways.

PCA on Bayes estimates

We are interested in the “true” values given in μ_i and not the observed data \mathbf{x}_i , and so the appropriate next step is to compute the posterior distribution of the the μ_i 's and then perform PCA on these posteriors. We can compute the posterior distribution $\mu_i | \mathbf{x}_i$ using Bayes' rule, which is

$$\mu_i | \mathbf{x}_i = x \sim N(\sigma_2^{-2} Sx, S) \quad (4)$$

with

$$S = (\sigma_1^{-2} Q^{-1} + \sigma_2^{-2} I)^{-1} \quad (5)$$

Now we want to perform PCA on the posterior estimates of the μ_i 's. We need to take into account the fact that the posterior distributions for each μ_i have non-spherical variance, and so we need to use gPCA instead of standard PCA.

Theorem

The row scores from gPCA on the posterior estimates $\mu_i \mid \mathbf{x}_i$ from the model are the same, up to a scaling factor, to the row scores from gPCA on (X, S, I_n) . The principal axes from gPCA on the posterior estimates are the same, up to a scaling factor, as the principal axes from gPCA on (X, S, I_n) pre-multiplied by S .

From this theorem, we see that when we perform gPCA on the posterior estimates obtained from the model, different scalings of the prior and the noise variances simply lead to gPCAs with different row inner product matrices.

A family of gPCAs

Now we can explore the family of inner product matrices which our model gives rise to. Up to a scaling factor, the matrix $S = (\sigma_1^{-2}Q^{-1} + \sigma_2^{-2}I)^{-1}$ depends only on the relative sizes of σ_1 and σ_2 , the scalings for the prior and the noise. We therefore have a one-dimensional family of gPCAs determined by the relative sizes of σ_1 and σ_2 . To get some insight into this family, we can first consider the endpoints.

As $\sigma_1/\sigma_2 \rightarrow 0$, that is, as the noise becomes very small compared to the prior structure, S becomes more and more like a scalar multiple of the identity, and so we approach a scalar multiple of gPCA on the triple (X, I, I) , or standard PCA. At the other end, as $\sigma_2/\sigma_1 \rightarrow 0$, we approach a scalar multiple of gPCA on the triple (X, Q, I) . The gPCA on (X, Q, I) turns out to be very closely related to double principal coordinates analysis (DPCoA), which is another method for incorporating information about the variables into the analysis.

Automatic selection of family member

If we do not want to assume σ_1 and σ_2 are known, we can estimate the values σ_1 and σ_2 from the data itself by maximum marginal likelihood. To be more concrete, according to our data model we have

$$\mathbf{x}_i \stackrel{\text{iid}}{\sim} N(0, \sigma_1^2 Q + \sigma_2^2 I) \quad (6)$$

The overall log likelihood of the data is therefore (up to a constant factor)

$$\ell(X; \sigma_1, \sigma_2) = -\frac{n}{2} \log |\sigma_1^2 Q + \sigma_2^2 I| - \sum_{i=1}^n \frac{1}{2} \mathbf{x}_i^T (\sigma_1^2 Q + \sigma_2^2 I)^{-1} \mathbf{x}_i \quad (7)$$

Maximizing this likelihood is not a convex problem: we transform it into a one parameter problem over the unit interval. Let $r = \sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$, and let $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Let $Q = V\Lambda V^T$ be the eigendecomposition of Q where V is an orthogonal matrix and Λ is diagonal containing the eigenvalues $\lambda_1, \dots, \lambda_p$. Finally, let $\mathbf{x}_i = V^T \mathbf{x}_i$ and \tilde{x}_{ij} be the j th element of $\tilde{\mathbf{x}}_i$. The log likelihood in the new parameterization is

$$\ell(X; r, \sigma) = -\frac{np}{2}\sigma^2 \log |rQ + (1-r)I| - \sigma^{-2} \sum_{i=1}^n \frac{1}{2} \mathbf{x}_i^T (rQ + (1-r)I) \mathbf{x}_i \quad (8)$$

$$= -\frac{np}{2}\sigma^2 \sum_{j=1}^p \log(r\lambda_j + 1 - r) - \sigma^{-2} \sum_{i=1}^n \sum_{j=1}^p \frac{1}{2} \frac{\tilde{x}_{ij}^2}{r\lambda_j + 1 - r} \quad (9)$$

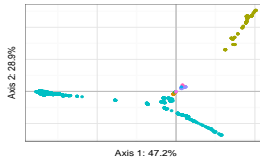
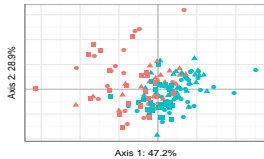
Based on the expression above, we can find a closed-form solution for the maximizing value of σ^2 for any fixed r .

This gives us

$$\sigma^{2*}(r) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \tilde{x}_{ij}^2 / (r\lambda_i + 1 - r) \quad (10)$$

We re-write the likelihood as a function of r only. This is still not convex but only has one parameter which lies on the unit interval, the optimization can be performed numerically.

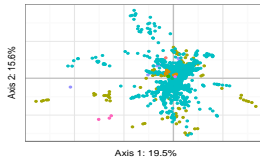
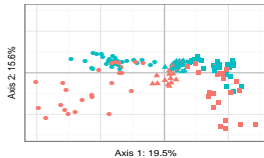
DPCoA



Adaptive gPCA

Antibiotic
● abx
● no abx

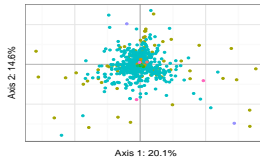
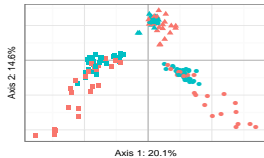
Subject
● D
▲ E
■ F



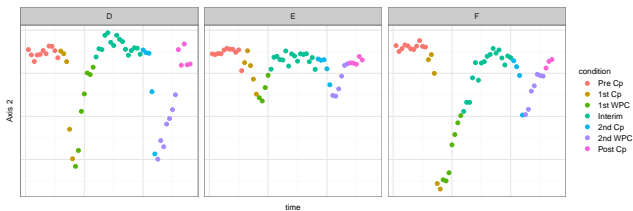
Phylum

- 4C0d-2
- Actinobacteria
- Bacteroidetes
- Candidate division TM7
- Cyanobacteria
- Firmicutes
- Lentisphaerae
- Proteobacteria
- Synergistetes
- Verrucomicrobia

PCA



Sample (left) and species (right) plots for DPCoA (top), adaptive gPCA (middle), and standard PCA (bottom). Colors in the sample plots represent a binning of the sample points into abx (either when the subject was on antibiotics or the week following immediately)



A plot of the scores along the second axis from adaptive gPCA by time, plotted for each of the three individuals. We see very clearly that this axis is capturing species that change during the administration of the antibiotic but which are stable otherwise. The corresponding plots for PCA and DPCoA are much less compelling.

Alternatives

We could add a ridge penalty to Q , resulting in gPCA on $(X, Q + \lambda I, I)$. This family has the same endpoints as the family we have described: when $\lambda = 0$ we have gPCA on (X, Q, I) , and as $\lambda \rightarrow \infty$ we get standard PCA.

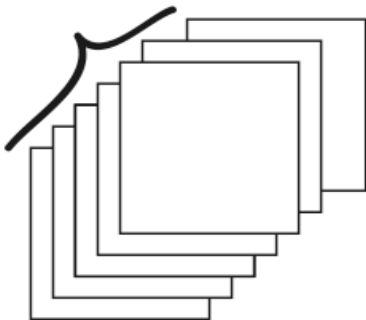
Very roughly, when we add a ridge penalty to Q , the main effect is to increase the small eigenvalues, but when we add a ridge penalty to Q^{-1} we make the large eigenvalues more similar to each other.

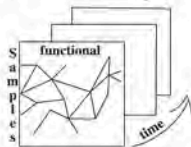
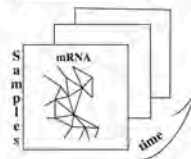
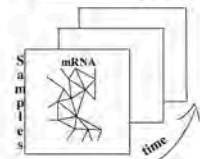
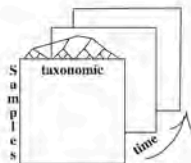
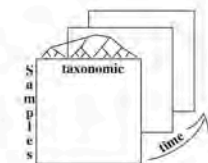
Small eigenvalues of Q correspond to eigenvectors that are very rough, while the large eigenvalues correspond to eigenvectors that are smooth.

When we do structured dimensionality reduction, we want to dampen any variance along rough eigenvectors, but we don't necessarily prefer variance in the direction of an extremely smooth eigenvector over variance in the direction of a mostly-smooth eigenvector. When we use $Q + \lambda I$, we remove the dampening on the rough directions, but when we use $S = (\sigma_1^1 Q^{-1} + \sigma_2^{-2} I)^{-1}$ we keep the eigenvalues of the rough directions small and decrease the difference between eigenvalues of smooth eigenvectors.

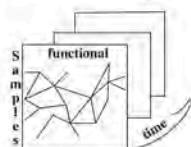
Part V

Multitable methods for
heterogeneous data





Unperturbed



Perturbed

Multi-table methods: use Inertia/Co-Inertia

Generalize variance and covariance \rightarrow moments of inertia.
weighted (p_i) sum of distances.

Abundance data in a contingency table \rightarrow weighted sum of the squares
weighted frequencies (chisquare).

Co-Inertia

When studying two variables measured at the same locations, for instance PH and humidity the standard quantification of covariation is the *covariance*.

$$\text{sum}(x1 * y1 + x2 * y2 + x3 * y3)$$

if x and y co-vary -in the same direction this will be big.

A simple generalization to this when the variability is more complicated to measure as above is done through Co-Inertia analysis (CIA).

Co-inertia analysis (CIA) is a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same samples or the same time points.

That is the rows or columns of the matrix have to be weighted similarly and thus must be matchable.

RV coefficient

The global measure of similarity of two data tables as opposed to two vectors can be done by a generalization of covariance provided by an inner product between tables that gives the RV coefficient, a number between 0 and 1, like a correlation coefficient, but for tables.

$$RV(A, B) = \frac{Tr(A'B)}{\sqrt{Tr(A'A)}\sqrt{Tr(B'B)}}$$

Survey on RV: Josse, Holmes (2015) Statistics Surveys, [arXiv link](#).

Example

Combining different types of data (antibiotic study).

Taxa Read counts (3 patients taking cipro: two time courses) : .

Mass-Spec Positive and Negative ion Mass Spec features and their intensities: .

RNA-seq Metagenomic data on genes :.

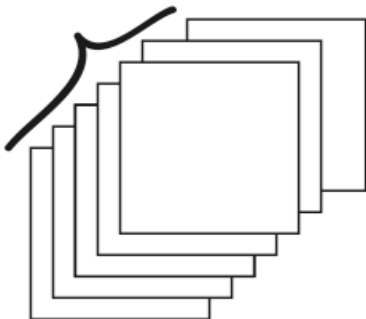
Here is the RV table of the three array types:

```
> fourtable$RV
```

	Taxa	Kegg	MassSpec+	MassSpec-
Taxa	1	0.565	0.561	0.670
Kegg	0.565	1	0.686	0.644
MassSpec+	0.561	0.686	1	0.568
MassSpec-	0.670	0.644	0.568	1

Part VI

Distances between "aligned" graphs



Bacteria 'sharing' between mice

Using the Jaccard index that measures the co-occurrence or co-occurrence of species between mice.

$$\text{Jaccard Similarity} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

$$\text{Jaccard Disimilarity} = \frac{f_{01} + f_{10}}{f_{01} + f_{10} + f_{11}}$$

```
mouse1
```

```
0 0 0 1 0 1 0 1 0 0 0 0 0 0 1
```

```
mouse4
```

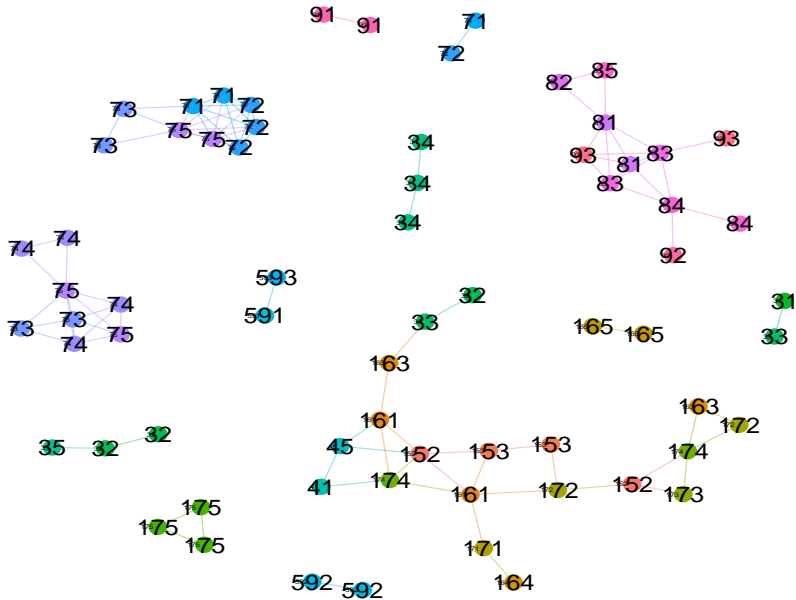
```
1 0 0 0 0 0 0 0 0 0 0 0 0 0 1
```

```
vegdist(rbind(mouse1,mouse4),method="jaccard")
```

```
0.8
```

Bacteria 'sharing' between mice as a network

```
netbaseline=make_network(phy_pifn_glom)
p=plot_network(netbaseline,phy_pifn_glom,
color="mousenames",label="mousenames",point_size=7)
+geom_text(aes(label=mousenames),size=7)
p+scale_colour_hue(guide="none")
```



Does the network relate to 'communities'?

Friedman and Rafsky (1979) devised a nonparametric test for multivariate data using the minimum spanning tree with any metric.

Then compute the number of 'pure' edging connecting labels from the same groups compared to the mixed edges connecting labels from different groups, call F_o the observed statistic.

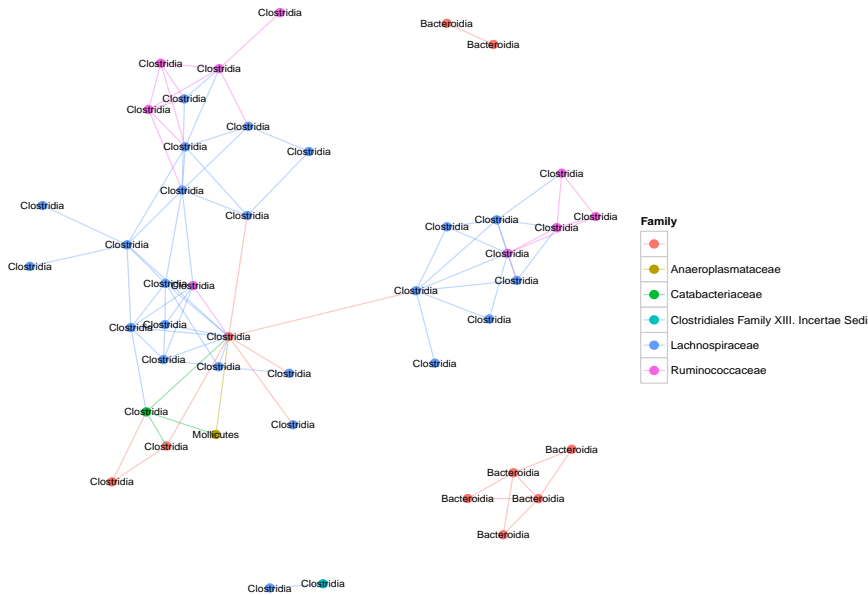
In our example: $F_o = 82$

Keeping the graph fixed, permute the labels and recompute the number of pure edges.

All 1000 simulated values had $F_s < 82$ so $p < 0.001$.

Co-occurrence networks for taxa of the baseline mice

```
p=plot_network(netbasetaxa,phy_pifn_glom,color="Family",  
type="taxa",label=NULL)  
p+geom_text(aes(label=Class),size=3)
```

Changes of the network over time?

OTU Network Plot

Type:
Taxa

Mouse 71 (infected)
 Mouse 72 (infected)
 Mouse 51 (uninfected)
 Mouse 52 (uninfected)
 Mouse 62 (uninfected)
 Mouse 64 (uninfected)

Color Attribute:
Genus

Distance:
jaccard

Time Interval G1:
0 2 27

Time Interval G2:
0 10 27

Decimal:
0 0.5 1

Coloring taxa based on Genus



Genus

- Adierocytia
- Akkermansia
- Anaerostipes
- Bifida
- Butyrivibrio
- Clostridium
- Coprococcus
- Epulopiscium
- Escherichia
- Eubacterium
- Lachnospira
- Lachnobacterium
- Lactobacillus
- Moryella
- Oscillospira
- Roseburia
- Ruminococcus
- Tributella
- Weissella



Genus

- Ad
- Ac
- Bu
- Cl
- Co
- De
- Ep
- Es
- Eu
- La
- La
- La
- M
- O
- Ru
- Ru
- Tr

Differences between two graphs?

Edges Added

Edges Removed



Family

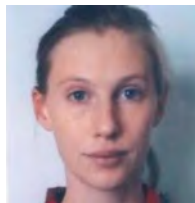
- Catabacteraceae
- Clostridiaceae
- Clostridiales Family XIII_Incertae Sedis
- Conobacteraceae
- Dehalobacteriaceae
- Enterobacteriaceae
- Erysipelotrichaceae
- Lachnospiraceae
- Lactobacillaceae
- Ruminococcaceae



Family

- Catabacter
- Enterobact
- Erysipelotr
- Lachnospir
- Lactobacilli
- Leuconost
- Ruminococ

Distances between (node-identified graphs)



Claire Donnat, SH, Ann. of Applied Stat., 2018.

Example:

Each graph corresponds to a cuisine (French, American, Greek, etc...).

Each of 1,530 ingredients constitutes a node in the graph and each of the 49 cuisines is assigned to a weighted graph.

The weight on the edge is the frequency of co-occurrence of the two ingredients for that particular cuisine. Some graphs includes a collection of disconnected nodes (ingredients that never co-occur in a single recipe) and a weighted connected component.

Graphs with identified vertices

$G = (\mathcal{V}, \mathcal{E})$ the graph with vertices \mathcal{V} and edges \mathcal{E} . $N = |\mathcal{V}|$, $i \sim j$ if nodes i and j are neighbors. A refers to the adjacency matrix of the graph, and D to its degree matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \quad \text{and } D = \text{Diag}(d_i)_{i=1 \dots N} \text{ s.t. } d_i = \sum_{j=1}^N A_{ij}$$

Restricting ourselves to undirected graphs, the matrix A is symmetric: $A^T = A$.

Hamming distance

It measures the number of edge deletions and insertions necessary to transform one graph into another.

$$d_H(G, \tilde{G}) = \sum_{i,j} \frac{|A_{ij} - \tilde{A}_{ij}|}{N(N-1)} = \frac{1}{N(N-1)} \|A - \tilde{A}\|_1 \quad (11)$$

This defines a metric between graphs, since it is a scaled version of the L_1 norm between the adjacency matrices A and \tilde{A} . It defines a distance bounded between 0 and 1 over all graphs of size N .

The Jaccard distance

$$d_{\text{Jaccard}}(G, \tilde{G}) = \frac{|G \cup \tilde{G}| - |G \cap \tilde{G}|}{|G \cup \tilde{G}|} = \frac{\sum_{i,j} |A_{ij} - \tilde{A}_{ij}|}{\sum_{i,j} \max(A_{i,j}, \tilde{A}_{i,j})} = \frac{\|A - \tilde{A}\|_1}{\|A + \tilde{A}\|_*} \quad (12)$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix.

Eq. 12 is known to define a proper distance between the graphs. A straightforward way to see this is to use the Steinhaus Transform: for (X, d) a metric and c a fixed point, the transformation $\delta(x, y) = \frac{2d(x,y)}{d(x,c)+d(y,c)+d(x,y)}$ produces a metric. Apply this transformation, with d the Hamming distance and c the empty graph, to see:

$$\begin{aligned} \delta(G, \tilde{G}) &= \frac{2\|A - \tilde{A}\|_1}{\|A\|_1 + \|\tilde{A}\|_1 + \|A - \tilde{A}\|_1} = \frac{2(|G \cup \tilde{G}| - |G \cap \tilde{G}|)}{2|G \cup \tilde{G}|} \quad (*) \\ &= d_{\text{Jaccard}}(G, \tilde{G}). \end{aligned}$$

The recipes graphs

Each cuisine-graph has nodes that represent ingredients; edges are co-occurrence frequencies.

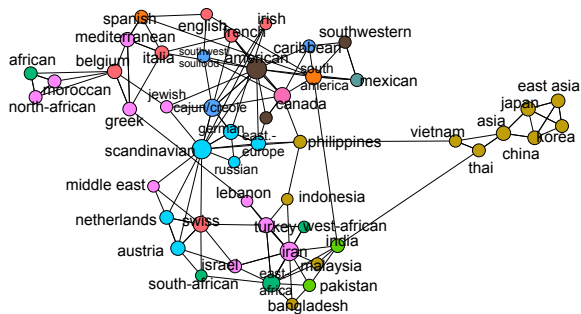
Cuisines can be better characterized by typical associations of ingredients.

For instance, the Japanese cuisine might be characterized by a higher associativity of ingredients such as “rice” and “nori” than Greek cuisine.

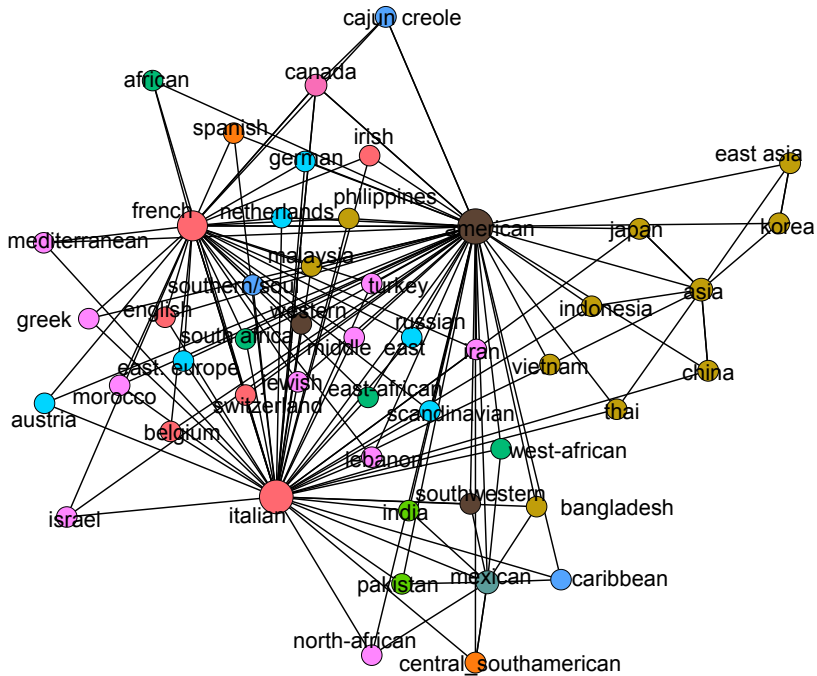
We use the co-occurrence counts of 1,530 different ingredients for 49 different cuisines (Chinese, American, French, etc.) Each cuisine is then characterized by its own co-occurrence graph.

The weight on the edge is the frequency of co-occurrence of the two ingredients in a given cuisine. The final graph for a given cuisine thus consists in a collection of disconnected nodes (ingredients that never appear in a single recipe for that cuisine) and a weighted connected component.

Hamming: metagraph



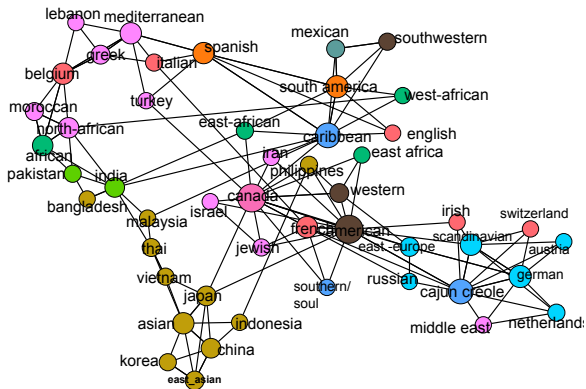
3-nearest neighbor from Hamming distance between graphs



3 nearest neighbors from distances between cuisine-graphs

Distances between networks (node-identified graphs)

Claire Donnat, Susan Holmes, *Annals of Applied Statistics*, 2018.



3 nearest neighbors Polynomial kernel distance between graphs.

Ingredient comparisons (heat-wavelet based distances)		
Cuisine	Neighbor	top changes (char. distance)
Middle Eastern	Indian	mustard, dill, bread, thyme, oregano, feta cheese, walnut sesame seed, coconut, olive
	Moroccan	chive, nut, red wine, feta cheese, cane molasses, yogurt, rose, oregano, fennel, walnut
	Spanish	apricot, lentil, mint, zucchini walnut, pork sausage, feta cheese, sesame seed, lamb, yogurt
Chinese	Asian	black bean, oyster, turmeric, cumin, lime juice, nira, coconut, basil, beef broth, lime
	Japanese	lemon, oyster, salmon, buckwheat enokidake, tuna, radish, barley, kelp, katsuobushi
	Thai	peanut butter, mint, roasted peanut, fenugreek, turmeric, lime juice, cumin, coconut, basil, lime

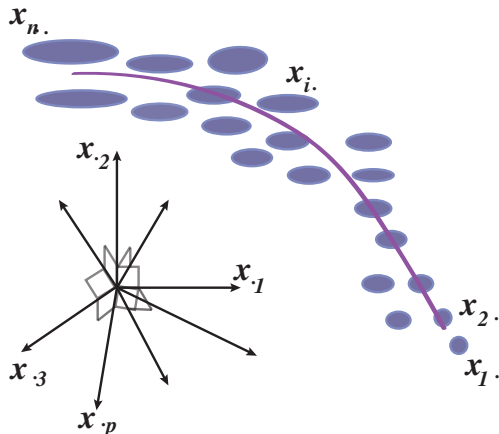
Identification of the ingredients that change the most from one graph to another

Distances enable statisticians to....

- ▶ Summarize data with medians, means and principal directions.
- ▶ Encode some variations in uncertainty.
- ▶ Make comparisons of heterogeneous sources of information.
- ▶ Integrate network and tree information.
- ▶ Measure diversity, inertia and generalize the notion of variance.

Questions for mathematicians

- ▶ How to build distances between images that account for unequal measurement errors, even locally?



Work by Adler, Taylor and Worsley (2003,2005,2007) using Random Fields.

Questions for mathematicians






- ▶ How well can the Euclidean embedding approximations do compared to the inherent noise?
- ▶ Are there better ways of approximating the commutative diagrams?

This is also an important point of contact with the use of Stein's method in probability theory.

Questions for mathematicians

- ▶ How to distinguish between the effect of the curvature of a state space and the effect of the unequal sampling?

References

-  L. Billera, S. Holmes, and K. Vogtmann.
The geometry of tree space.
Adv. Appl. Maths, 771–801, 2001.
-  J. Chakerian and S. Holmes.
distory:Distances between trees, 2010.
-  Daniel Chessel, Anne Dufour, and Jean Thioulouse.
The ade4 package - i: One-table methods.
R News, 4(1):5–10, 2004.
-  P. Diaconis, S. Goel, and S. Holmes.
Horseshoes in multidimensional scaling and kernel methods.
Annals of Applied Statistics, 2007.
-  Y. Escoufier.
Operators related to a data matrix.
In J.R. et al. Barra, editor, *Recent developments in Statistics.*,
pages 125–131. North Holland,, 1977.



Steven N Evans and Frederick A Matsen.

The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples.

arXiv, q-bio.PE, Jan 2010.



M Hamady, C Lozupone, and R Knight.

Fast unifracs: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data.

The ISME Journal, Jan 2009.



Susan Holmes.

Multivariate analysis: The French way.

In D. Nolan and T. P. Speed, editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 56 of *IMS Lecture Notes–Monograph Series*. IMS, Beachwood, OH, 2006.



Ross Ihaka and Robert 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.



P. J. McMurdie and S. Holmes.
Phyloseq: Reproducible research platform for bacterial
census data.
PlosONE, 2013.
April 22,.



Serban Nacu, Rebecca Critchley-Thorne, Peter Lee, and
Susan Holmes.
Gene expression network analysis and applications to
immunology.
Bioinformatics, 23(7):850–8, Apr 2007.



Sandrine Pavoine, Anne-Béatrice Dufour, and Daniel
Chessel.

From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis.

Journal of Theoretical Biology, 228(4):523–537, 2004.



Elizabeth Purdom.

Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree.

Annals of Applied Statistics, Jul 2010.



C. R. Rao.

The use and interpretation of principal component analysis in applied research.

Sankhya A, 26:329–359., 1964.

Part VIII

Dimension Reduction: the
Euclidean embedding workhorse:
MDS

Metric Multidimensional Scaling

Schoenberg (1935)

ANNALS OF MATHEMATICS
Vol. 36, No. 3, July, 1935

REMARKS TO MAURICE FRÉCHET'S ARTICLE "SUR LA DÉFINITION AXIOMATIQUE D'UNE CLASSE D'ESPACE DISTANCIÉS VECTORIELLEMENT APPLICABLE SUR L'ESPACE DE HILBERT"

BY I. J. SCHOENBERG

(Received April 16, 1935)

1. Fréchet's developments in the last section of his article suggest an elegant solution of the following problem.

Let

$$a_{ik} = a_{ki} \quad (i \neq k; i, k = 0, 1, \dots, n)$$

be $\frac{1}{2}n(n+1)$ given positive quantities. What are the necessary and sufficient conditions that they be the lengths of the edges of a n -simplex $A_0A_1 \dots A_n$? More general, what are the conditions that they be the lengths of the edges of a n -"simplex" $A_0A_1 \dots A_n$ lying in a euclidean space R_r ($1 \leq r \leq n$) but not in a R_{r-1} ?

This problem is fundamental in K. Menger's metric investigation of euclidean spaces ([6] and [7], particularly his third fundamental theorem in [7], pp. 737-743). It was solved by Menger by means of equations and inequalities involving certain determinants. Theorem 1 below furnishes a complete and independent solution of this problem. Theorem 2 solves the similar problem for spherical spaces previously treated by Menger's methods by L. M. Blumenthal and G. A. Garrett ([1]) and Laura Klanfer ([5]); it may be conveniently applied (Theorems 3 and 3') to prove and extend a theorem of K. Gödel ([4]). The method of Theorem 1 is finally applied to solve the corresponding problem for spaces with indefinite line element recently considered by A. Wald ([8]) and H. S. M. Coxeter and J. A. Todd ([2]).

From Coordinates to Distances and Back

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

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

Schoenberg's result: exact Euclidean distance If B is positive semi-definite then D can be seen as a distance between points in a Euclidean space.

Reverse engineering an Euclidean embedding

We can go backwards from a matrix D to X by taking the eigendecomposition of $B = -\frac{1}{2}HD^{(2)}H$ 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' .

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

Multidimensional Scaling (MDS)

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 SV' .
- ▶ Choose a number of dimensions by inspecting the eigenvalue's screeplot.

The advantage is that the original distances don't have to be Euclidean.

Taking Categorical Data and Making it into a Continuum

Horseshoe Example: Joint with Persi Diaconis and Sharad Goel (Annals of Applied Stats, 2005). 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 of voting data.

The Data

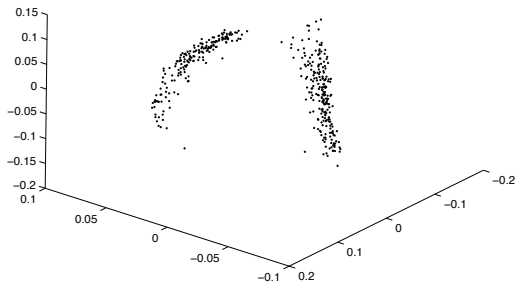
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
R1	-1	-1	1	-1	0	1	1	1	1	1	...
R2	-1	-1	1	-1	0	1	1	1	1	1	...
R3	1	1	-1	1	-1	1	1	-1	-1	-1	...
R4	1	1	-1	1	-1	1	1	-1	-1	-1	...
R5	1	1	-1	1	-1	1	1	-1	-1	-1	...
R6	-1	-1	1	-1	0	1	1	1	1	1	...
R7	-1	-1	1	-1	-1	1	1	1	1	1	...
R8	-1	-1	1	-1	0	1	1	1	1	1	...
R9	1	1	-1	1	-1	1	1	-1	-1	-1	...
R10	-1	-1	1	-1	0	1	1	0	0	0	...

L_1 distance

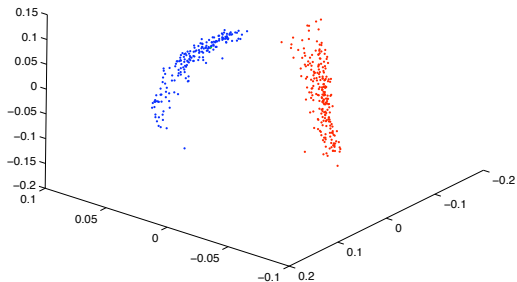
We define a distance between legislators as

$$\hat{d}(l_i, l_j) = \frac{1}{669} \sum_{k=1}^{669} |v_{ik} - v_{jk}|.$$

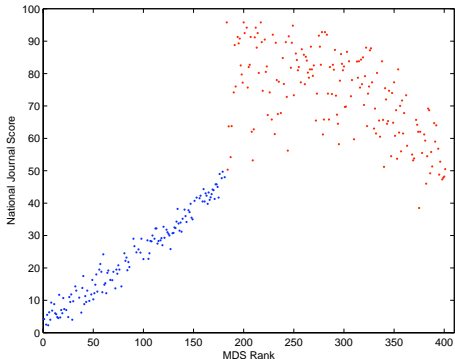
Roughly, $\hat{d}(l_i, l_j)$ is the percentage of roll calls on which legislators l_i and l_j disagreed.



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. We used dissimilarity indices $1 - \exp(-\lambda d(R_1, R_2))$



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.



Comparison of the MDS derived rank for Representatives with the National Journal's liberal score