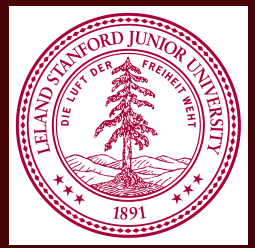
Statistics using Geometry to show uncertainties and integrate graph information

Susan Holmes
http://www-stat.stanford.edu/~susan/

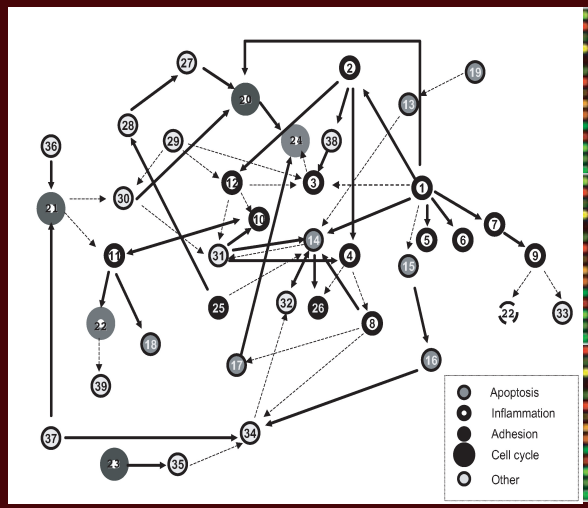
Bio-X and Statistics, Stanford University

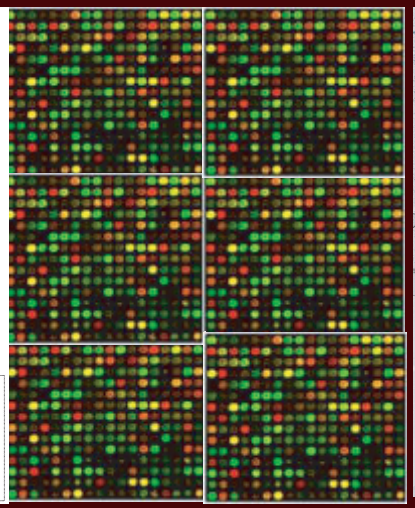
January 12, 2018, stancon, Asilomar

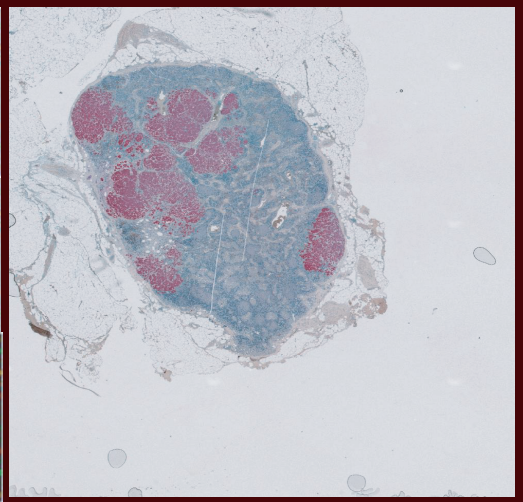


Goals in Modern Biology: Systems Approach

Look at the data/ all the data: data integration

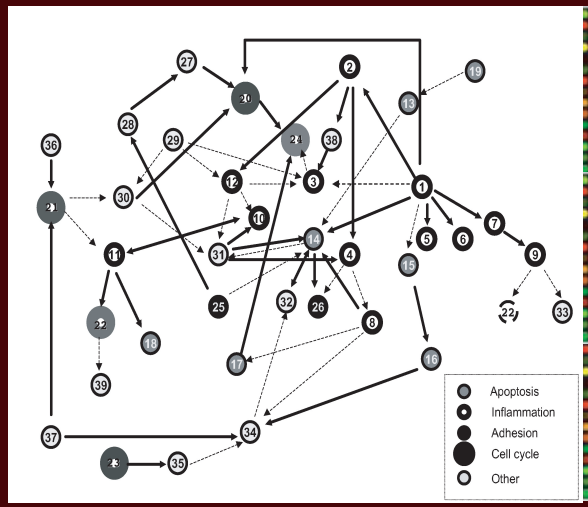


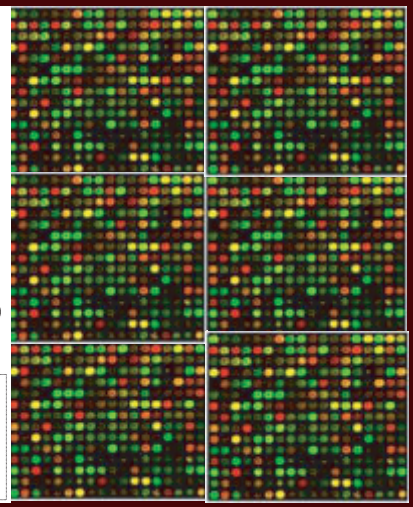


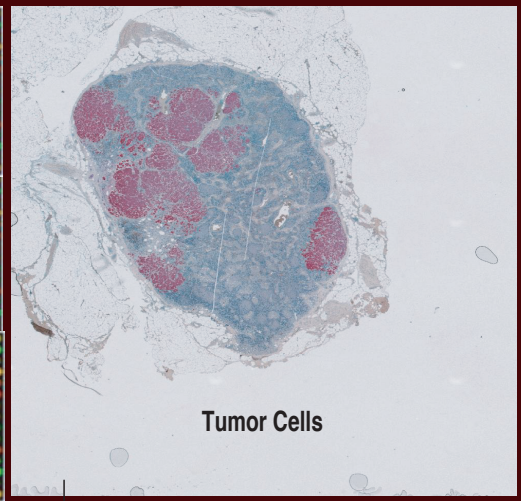


Goals in Modern Biology: Systems Approach

Look at the data/ all the data: data integration

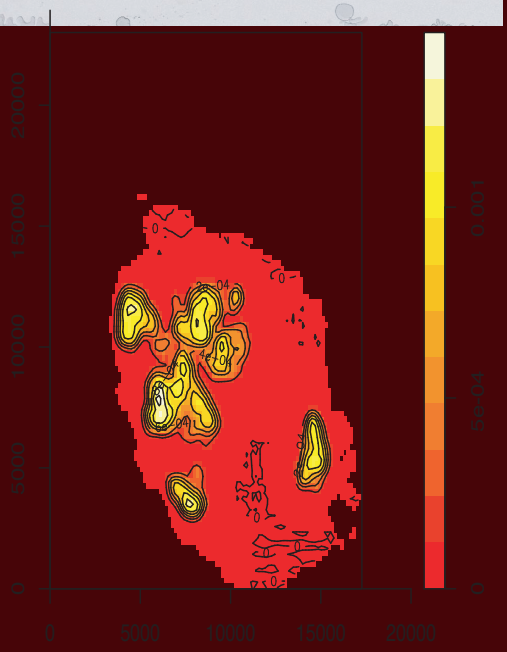






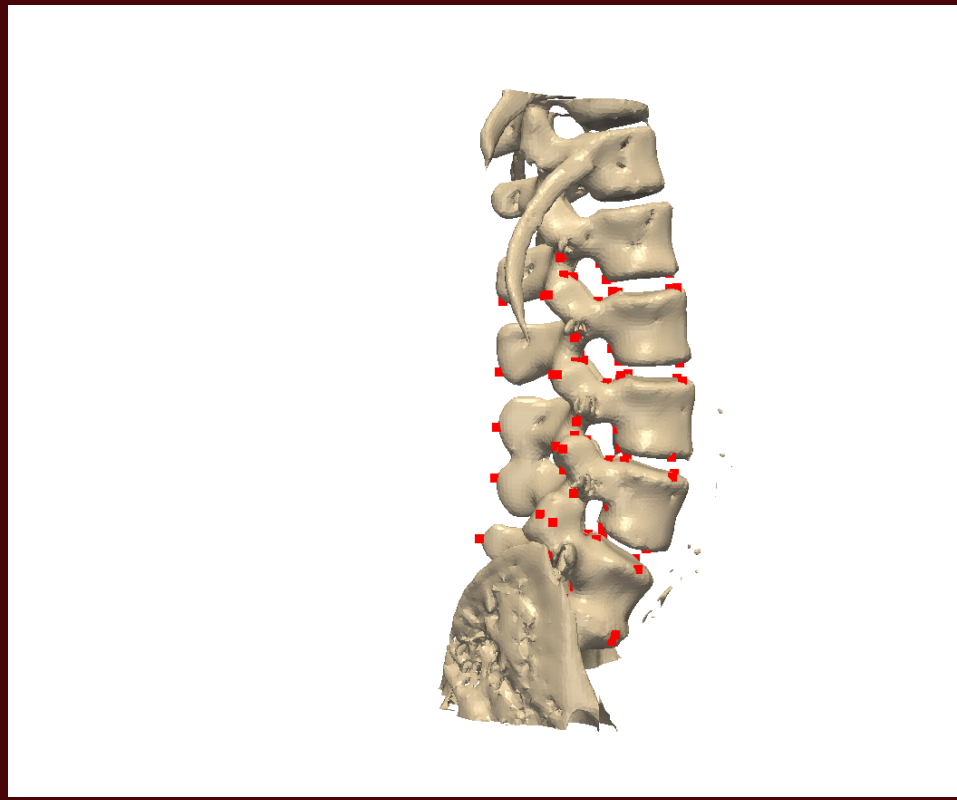
$$X_{Blood} = \begin{pmatrix} 0.5 & 1.1 & 1.6 & 1.2 & \dots \\ 0.3 & 1.9 & 2.2 & 1.1 & \dots \\ 1.1 & 0 & 3.2 & 0.4 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 2.7 & 2.3 & 1.2 & 1.1 & \dots \end{pmatrix}$$

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

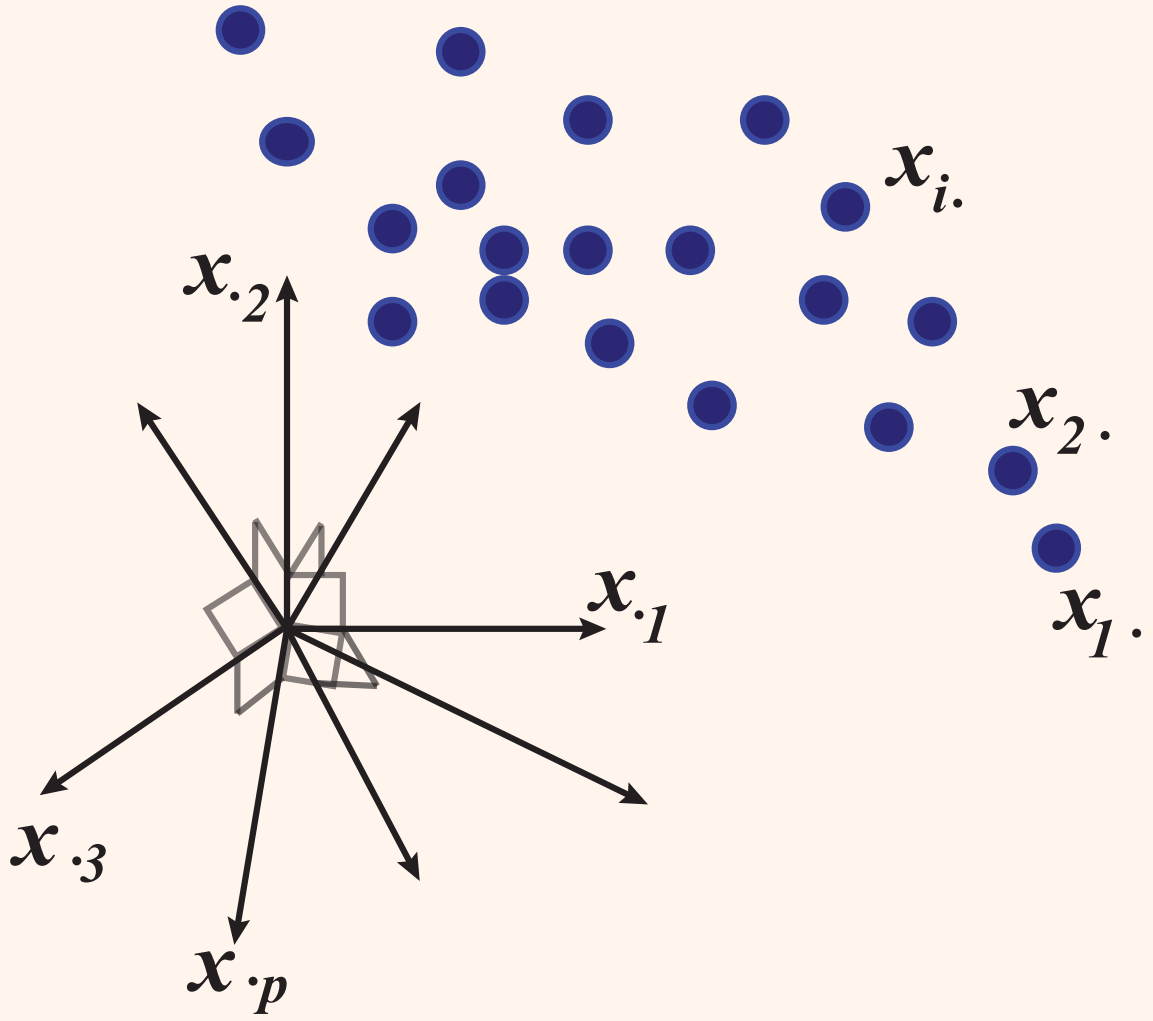


Today's challenge

- Data are not uniformly distributed from some manifold.
- Data are not an identically distributed random sample.
- Data are not independent.
- Heteroscedasticity everywhere.



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.
- 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=$\frac{1}{K-1}\sum_{k=1}^{K} d^2(T_0, T_k) = \hat{s}^2$. 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.

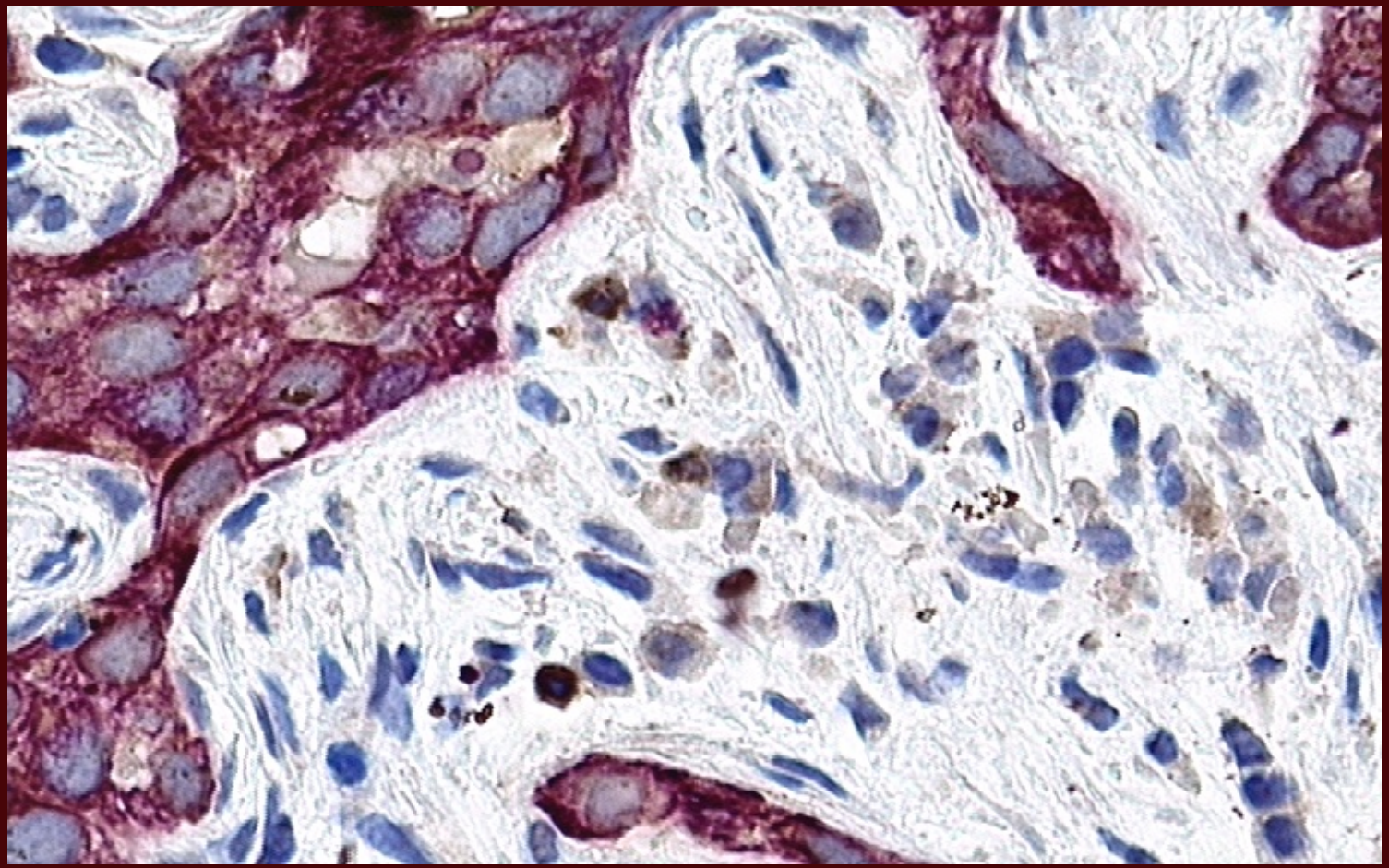
What do statisticians use distances for?

- 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 I

The Geometries of Data

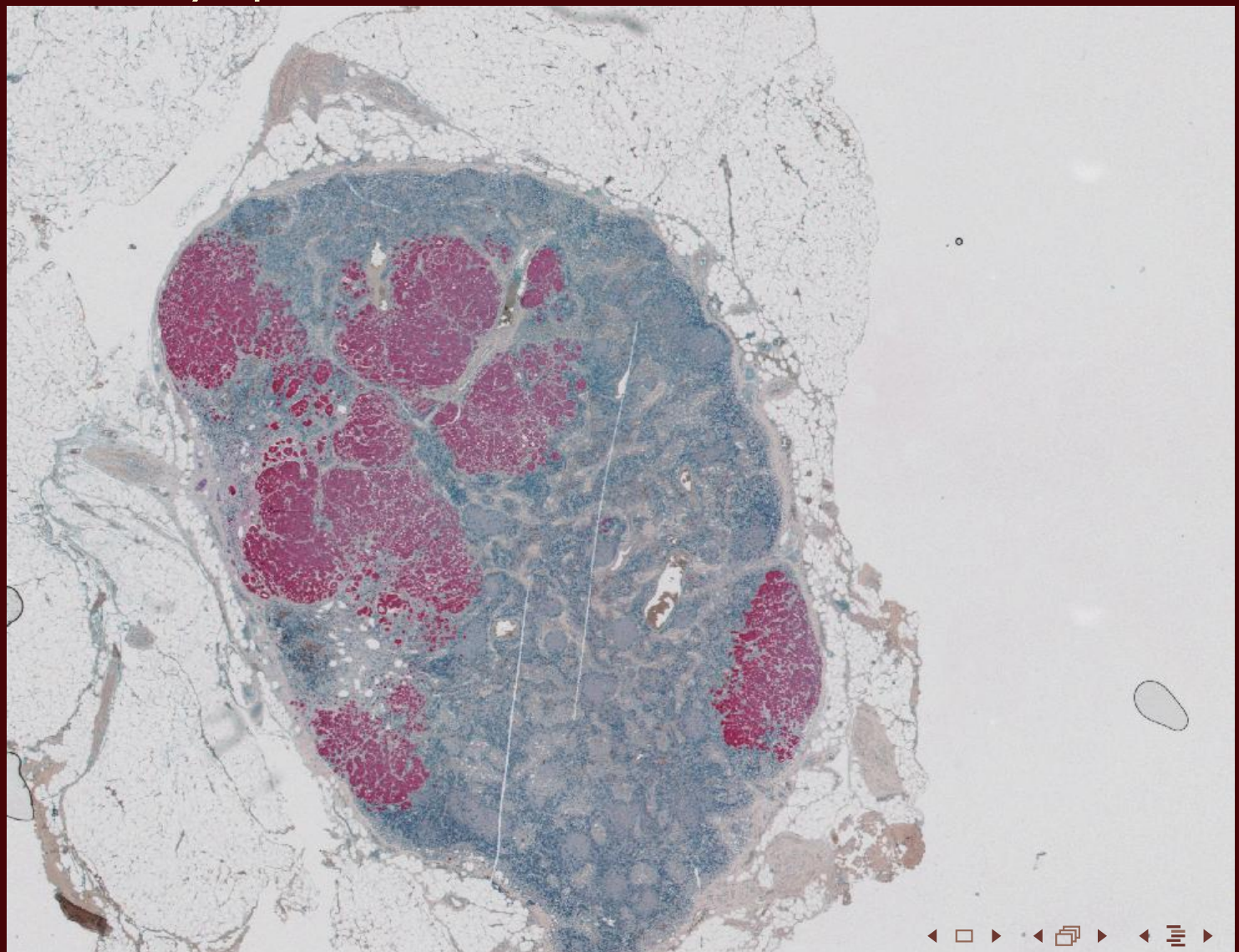


First example: cell segmentation

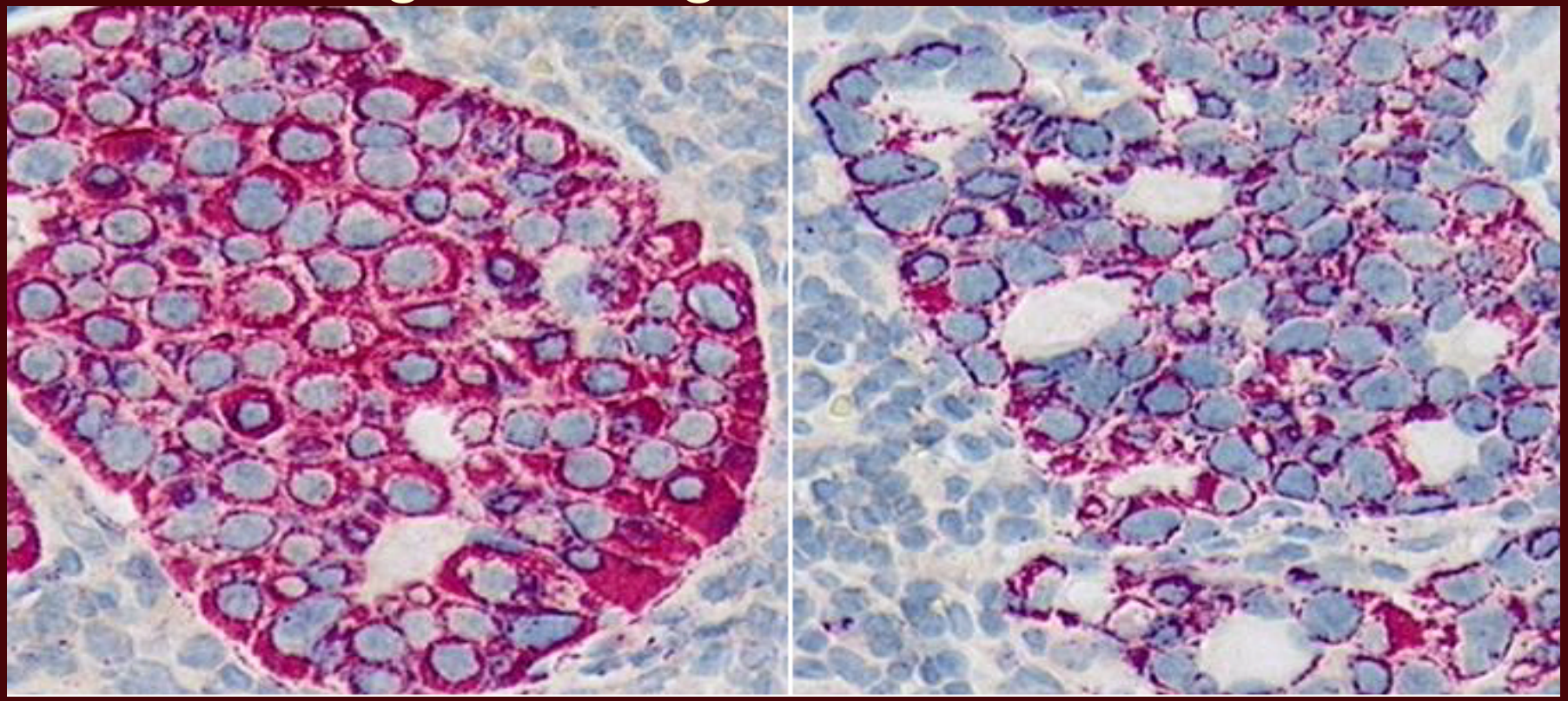
Joint work with Adam Kapelner and PP Lee.

Stained biopsy slides. Multispectral imaging (8 levels/wavelengths).

Stained Lymph Node Aim to identify cell.



Problem : Staining is heterogeneous



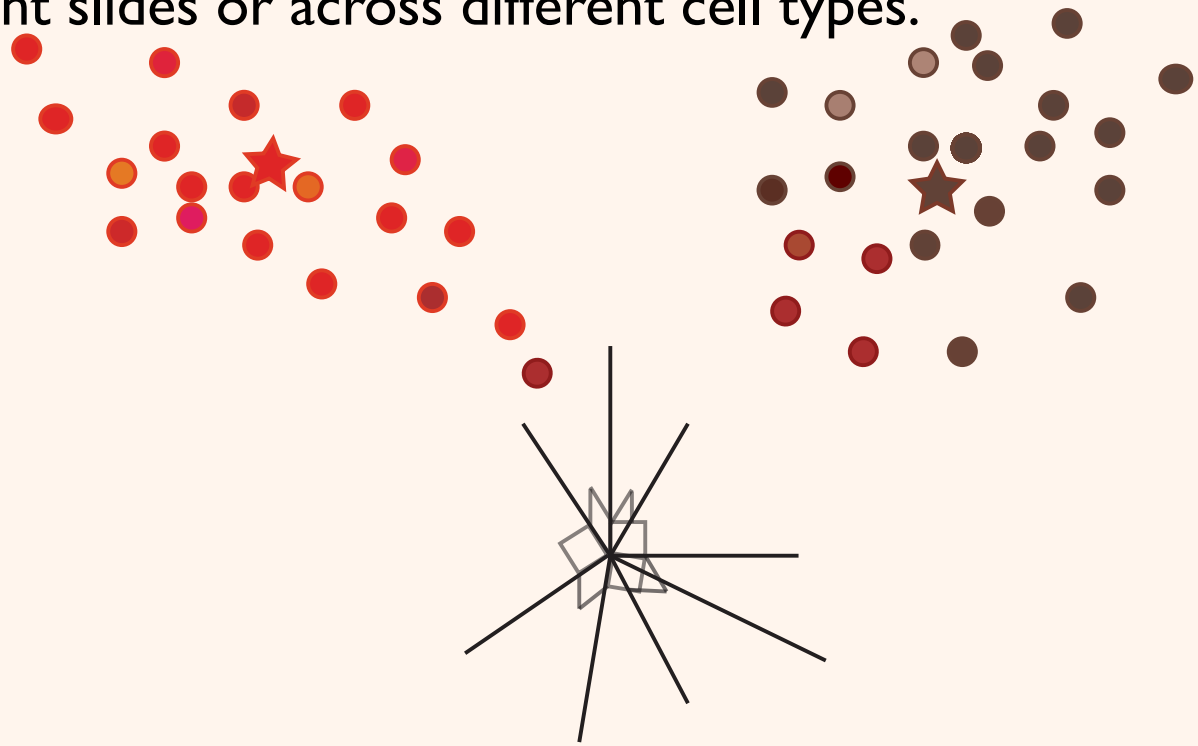
Both images are from the same image set. The stained cells are cancer cells stained with Fast Red red.

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

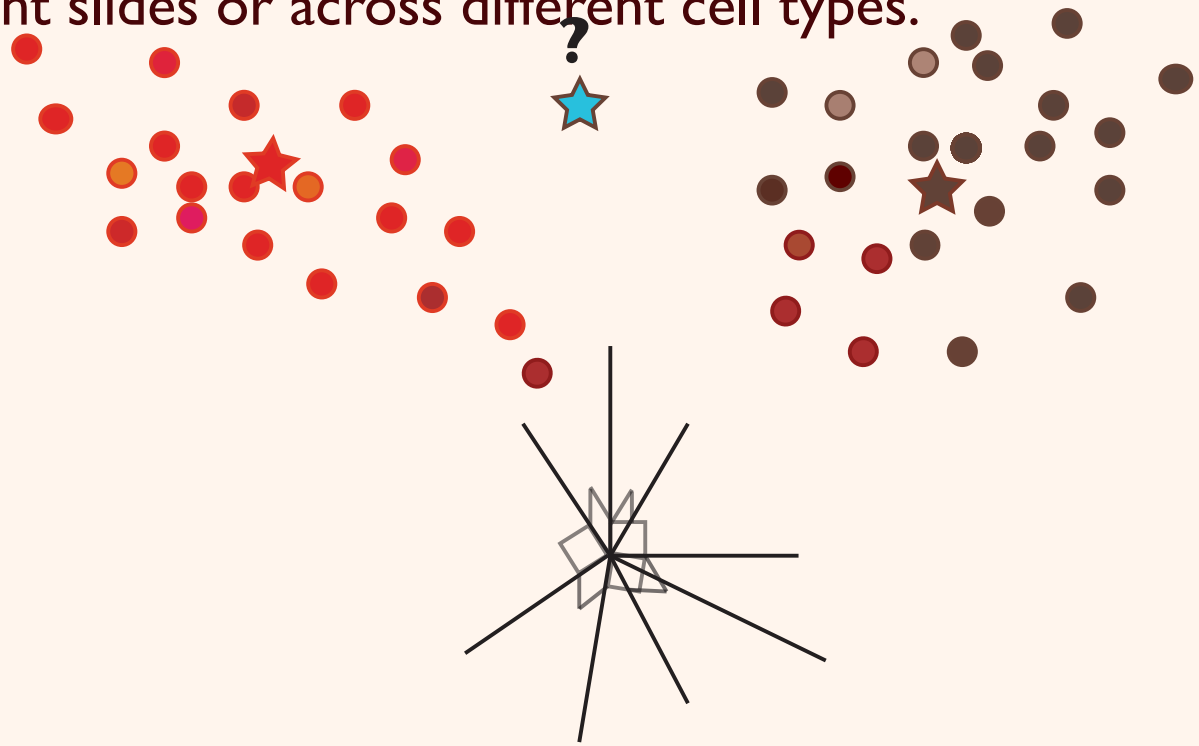
Extreme variability in the image colors/intensity/contrast.
Pixels from a same cell not independent and identically distributed across the different slides or across different cell types.



Simple nearest neighbor approach:
- Take 8 dimensional pixels points.
- Assigning the point to the closest neighbor

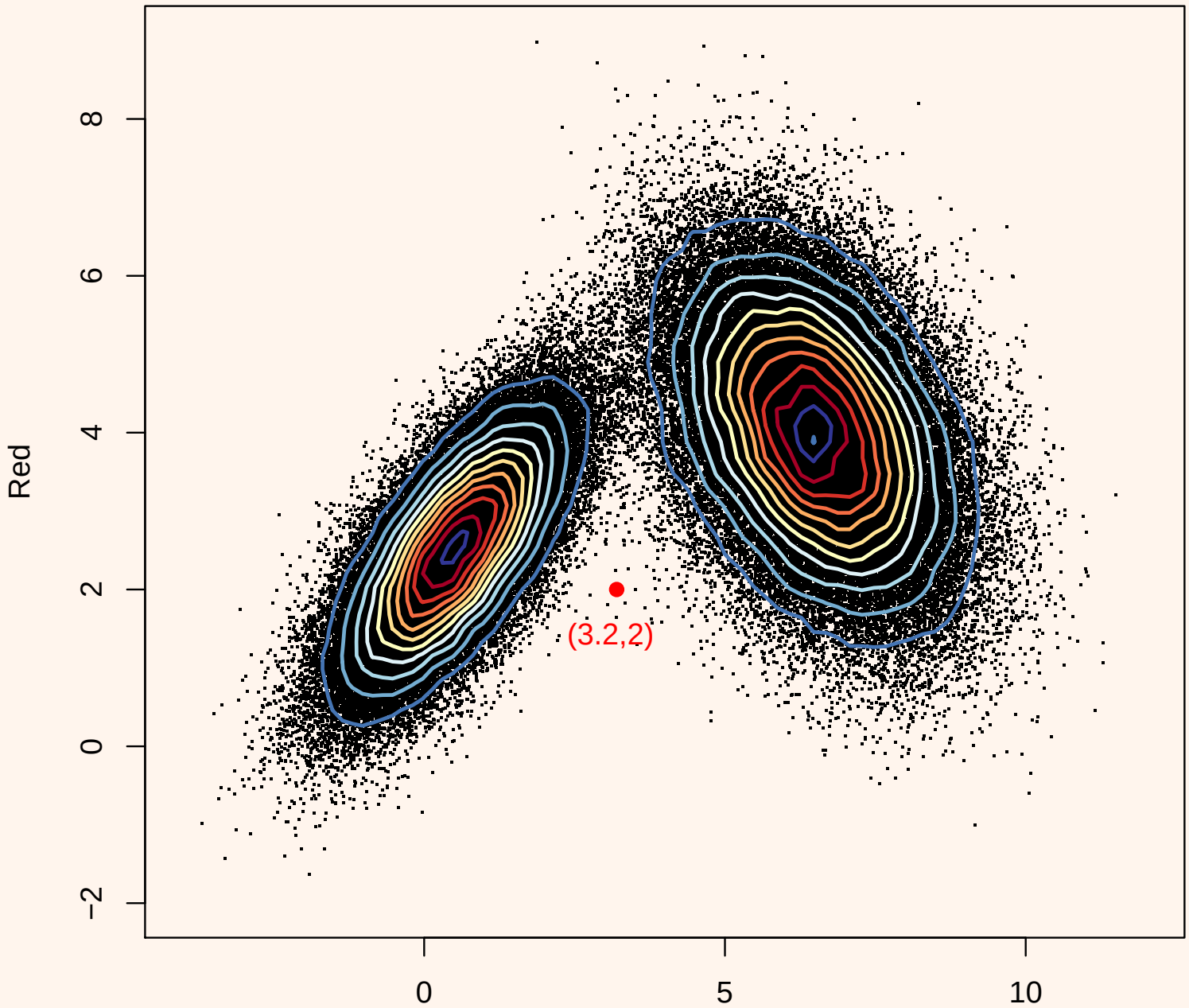
Problem : Staining is heterogeneous

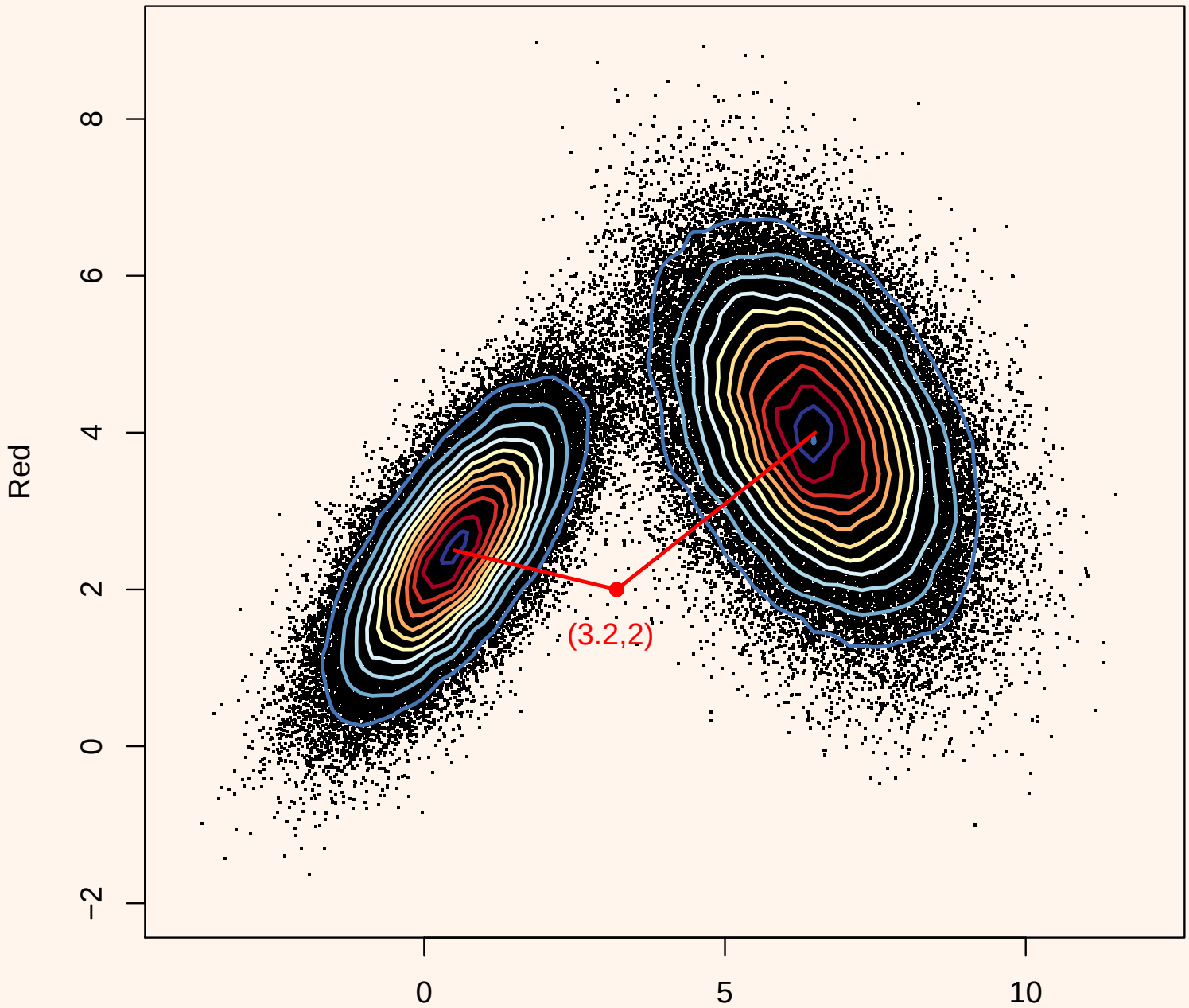
Extreme variability in the image colors/intensity/contrast.
Pixels from a same cell not independent and identically distributed across the different slides or across different cell types.

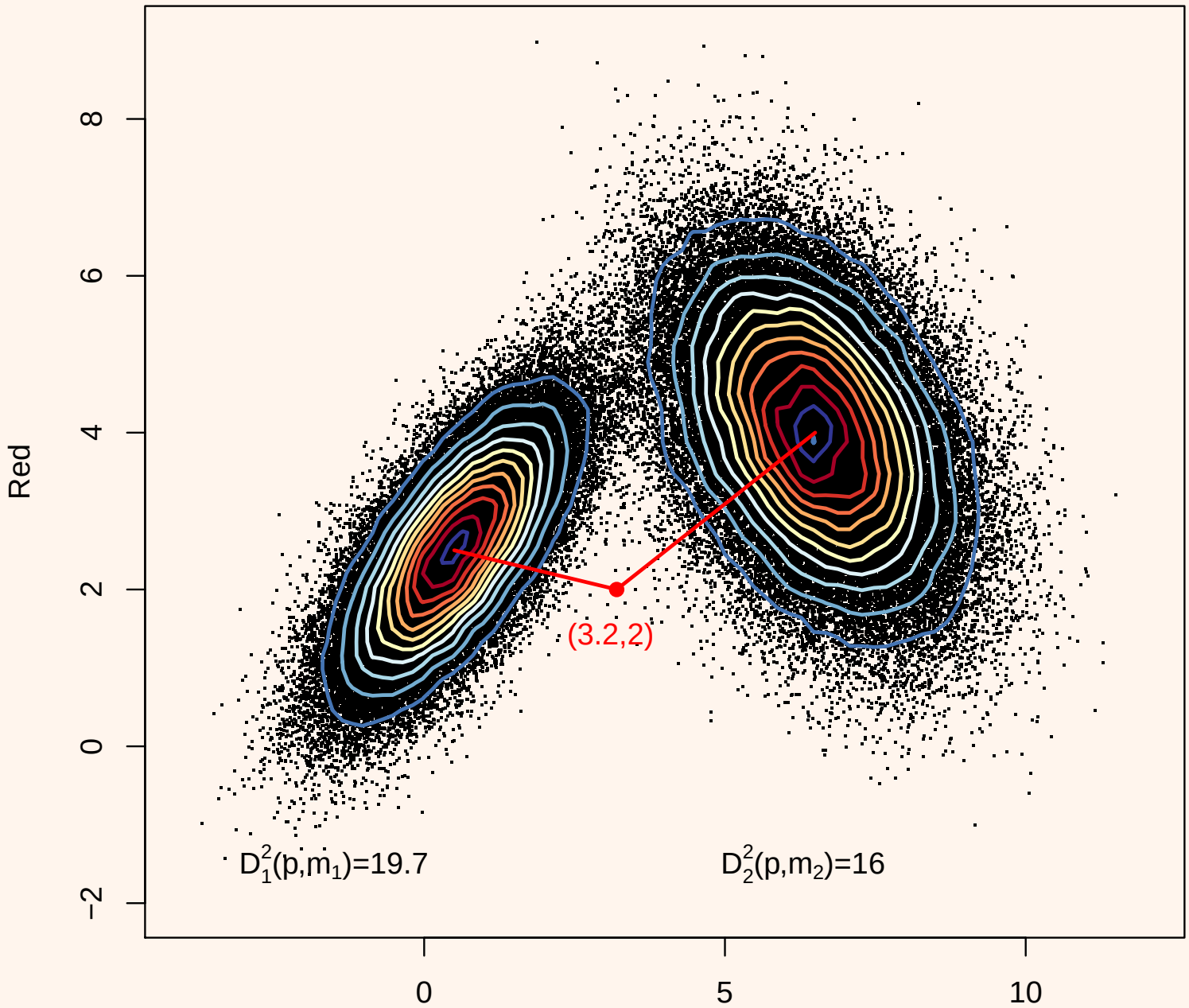


Simple nearest neighbor approach:
- Take 8 dimensional pixels points.
- Assigning the point to the closest neighbor

Red
8
6
4
2
0
−2
0
5
10


Red
8
6
4
2
0
−2
0
5
10
(3.2,2)


Red
8
6
4
2
0
−2
0
5
10
(3.2,2)


Red
8
6
4
2
0
−2
0
5
10
(3.2,2)
$D_1^2(p,m_1)=19.7$
$D_2^2(p,m_2)=16$

Multivariate Normal Data

Mahalanobis Transformation.
Several different clusters with different variance-covariance matrices and different means.
(μ_1, Σ_1) (μ_2, Σ_2)

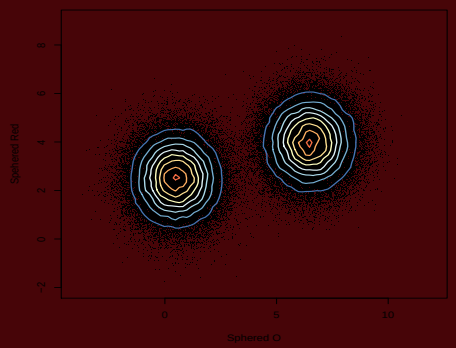
$$D_1^2(x, \mu_1) = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$$
$$D_2^2(x, \mu_2) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$

Corresponding Data Transformation

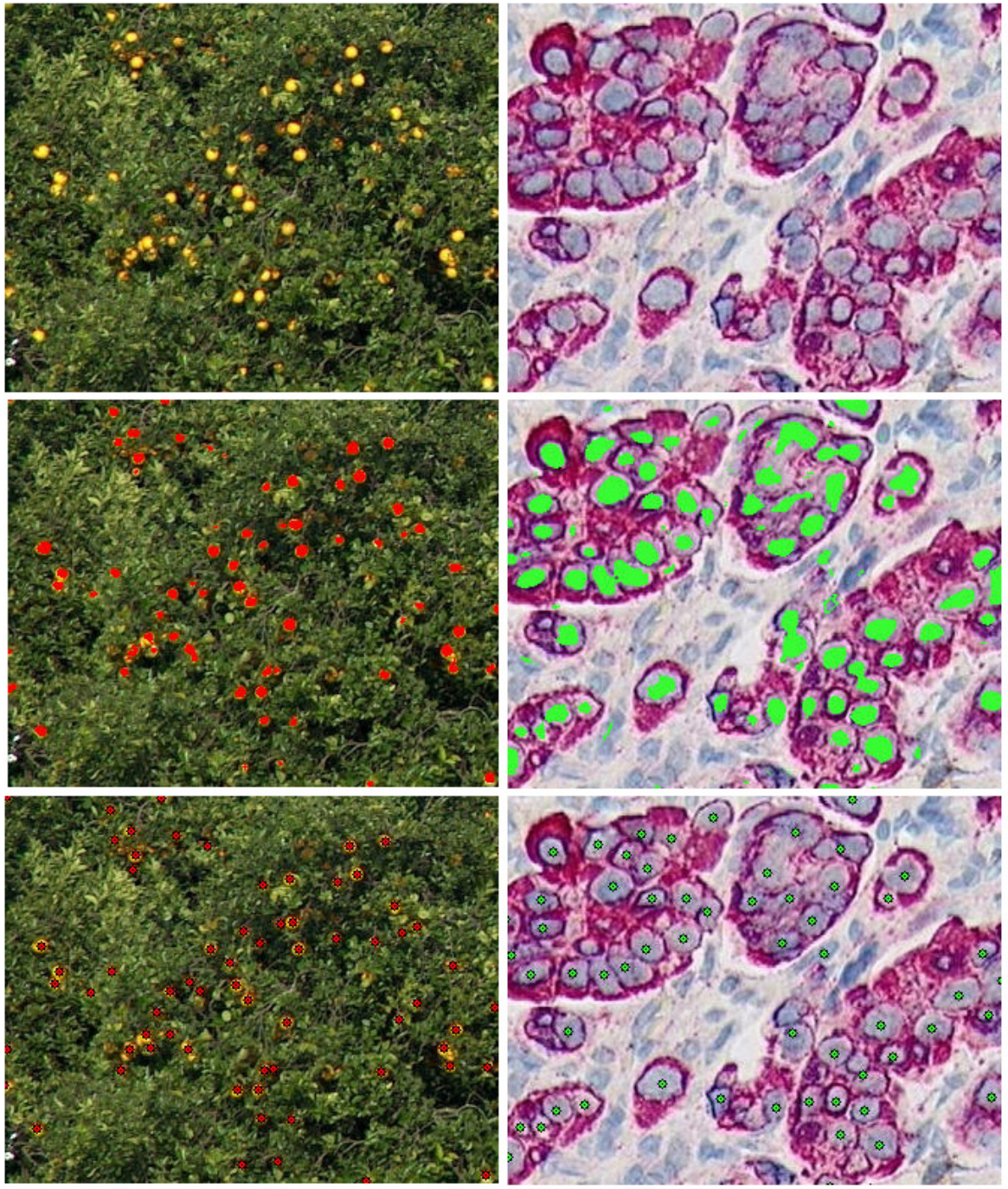
$$H = I - 1D_n1^T, \qquad S = X'HD_nHX$$

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

This is sometimes called ‘data sphering’.

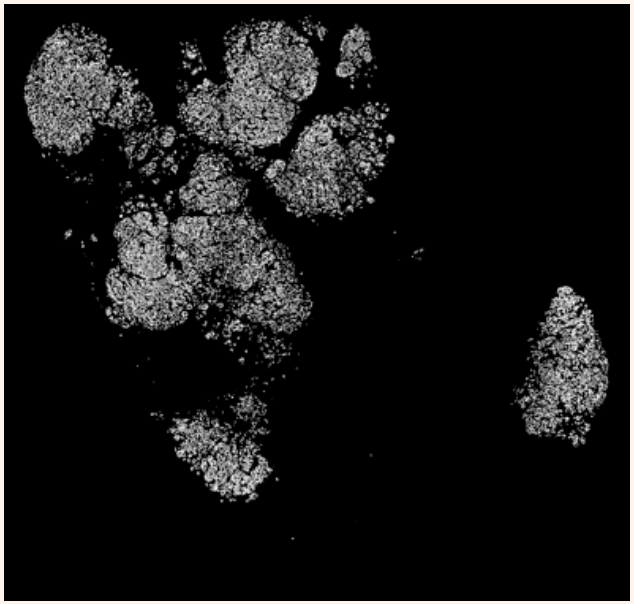


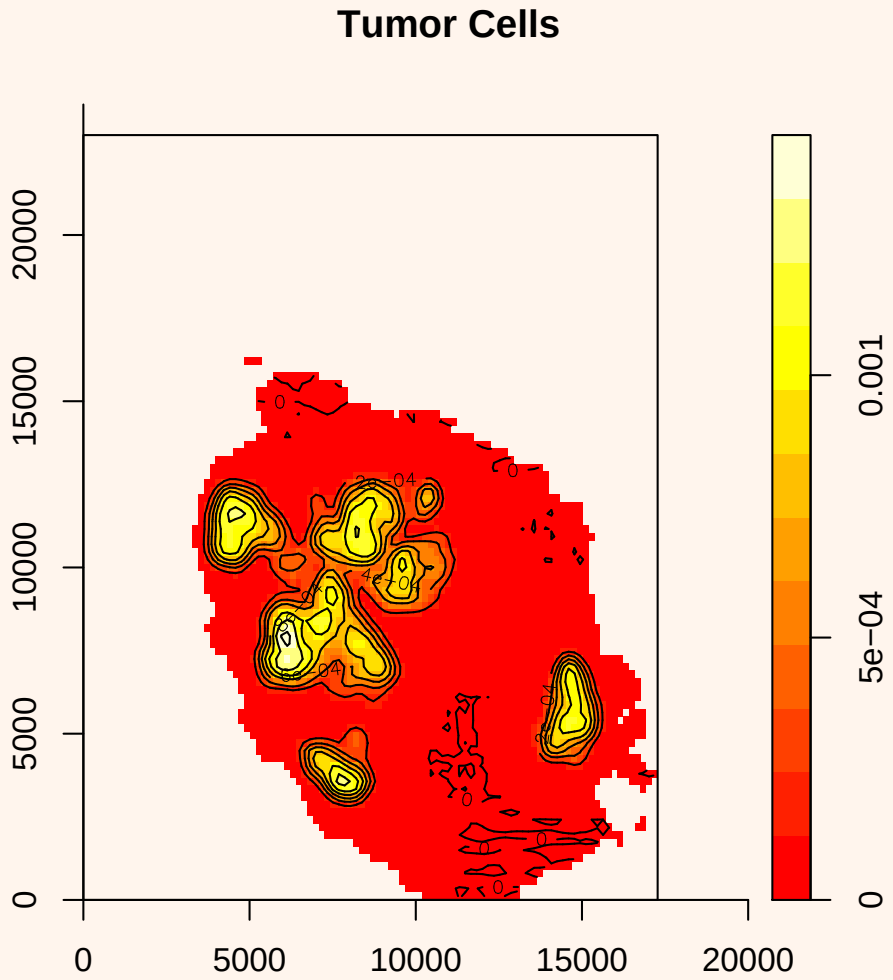
Change of Measure more generally see Diaconis, Holmes, Shahshahani, 2014.



Output Data

Tumor

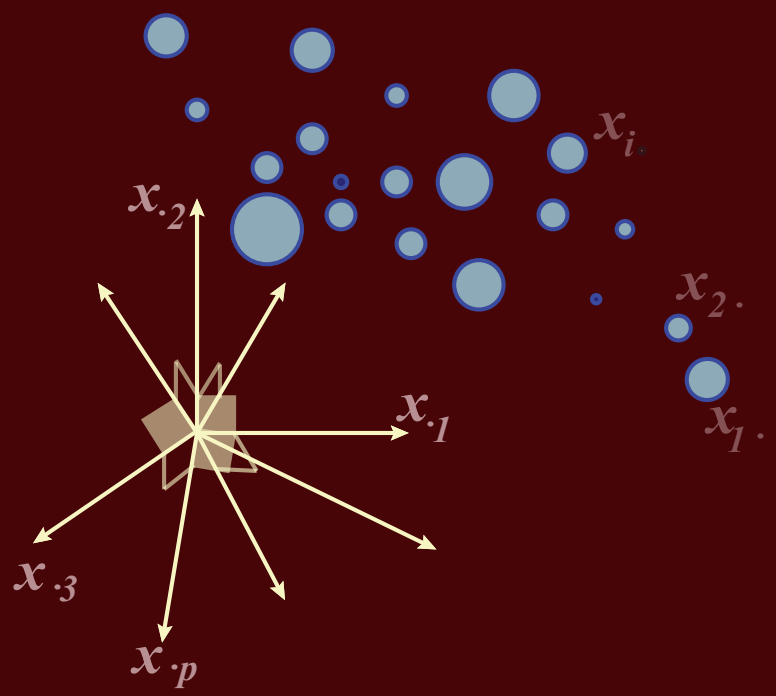




Number of Tumor cells: 27,822

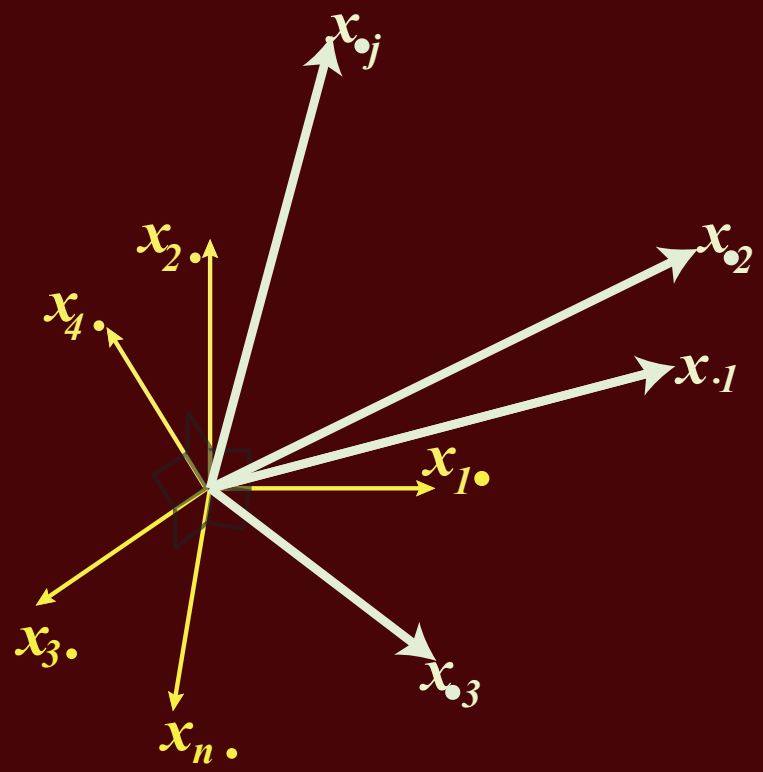
We can add information through choice of distances

Sample data can often be seen as points in a state space.
$\mathbb{R}^p$



$x^t Q y = < x, y >_Q$
Duality : Transposable data.

Variables are 'vectors' in data point space
$\mathbb{R}^n$



$x^t D y = < x, y >_D$

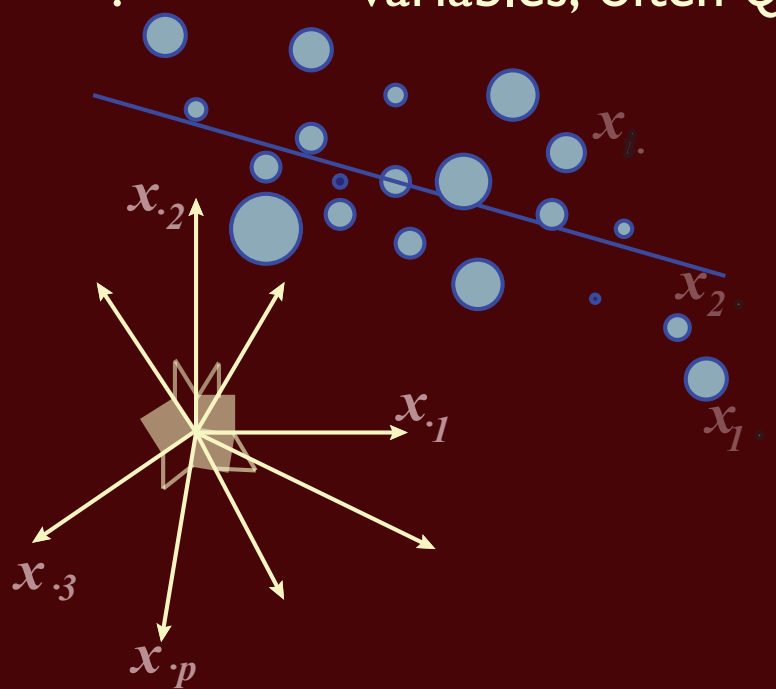
Data Analysis: Geometrical Approach

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



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

$$x^t Q y = < x, y >_Q \qquad x, y \in \mathbb{R}^p$$
$$x^t D y = < x, y >_D \qquad x, y \in \mathbb{R}^n.$$

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.

Comparing Two Diagrams: the RV coefficient

Many problems can be rephrased in comparison of two "duality diagrams".
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 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
$covV(O_1, O_2) = Tr(O_1^t O_2) =< O_1, O_2 >$ or a vector correlation (see Escoufier, 1977 and Josse and Holmes, 2016 [6])

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

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

Comparing two distance matrices

Similar to Mantel (Henry Daniels' method):

$$
\begin{array}{ccc}
D & & D' \\
D_{ij} & & D'_{ij} \\
vec(D_{ij}) & & vec(D'_{ij}) \\
as.dist(D) & & as.dist(D') \\
 & cor(D_{ij}, D'_{ij}) &
\end{array}
$$

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

$$Y^t DX = \Delta_Y A \qquad \text{where } \Delta_Y = Y^t DY = diag(w_1, w_2, \ldots, w_g),$$

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 $T = X^t DX$, 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 $W = (X - YA)^t D(X - YA)$.

A generalized Huyghens' formula:

$$T = B + W$$

Proof: Expanding W gives

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

□

Part II

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[1]

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

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

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* $A_0A_1 \cdots A_n$? *More general, what are the conditions that they be the lengths of the edges of a n-"simplex"*[2] $A_0A_1 \cdots A_n$ *lying in a euclidean space* R_r $(1 \leqq r \leqq n)$ *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

$$X = HY, \qquad \text{with } H = (I - \frac{1}{n}\mathbf{1}\mathbf{1}'), \text{and } \mathbf{1}' = (1, 1, 1 \ldots, 1)$$

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 $\frac{-1}{2} H D_2 H = B$.
- Diagonalize B to find the principal coordinates SV'.
- Choose a number of dimensions by inspecting the eigenvalue's screeplot.

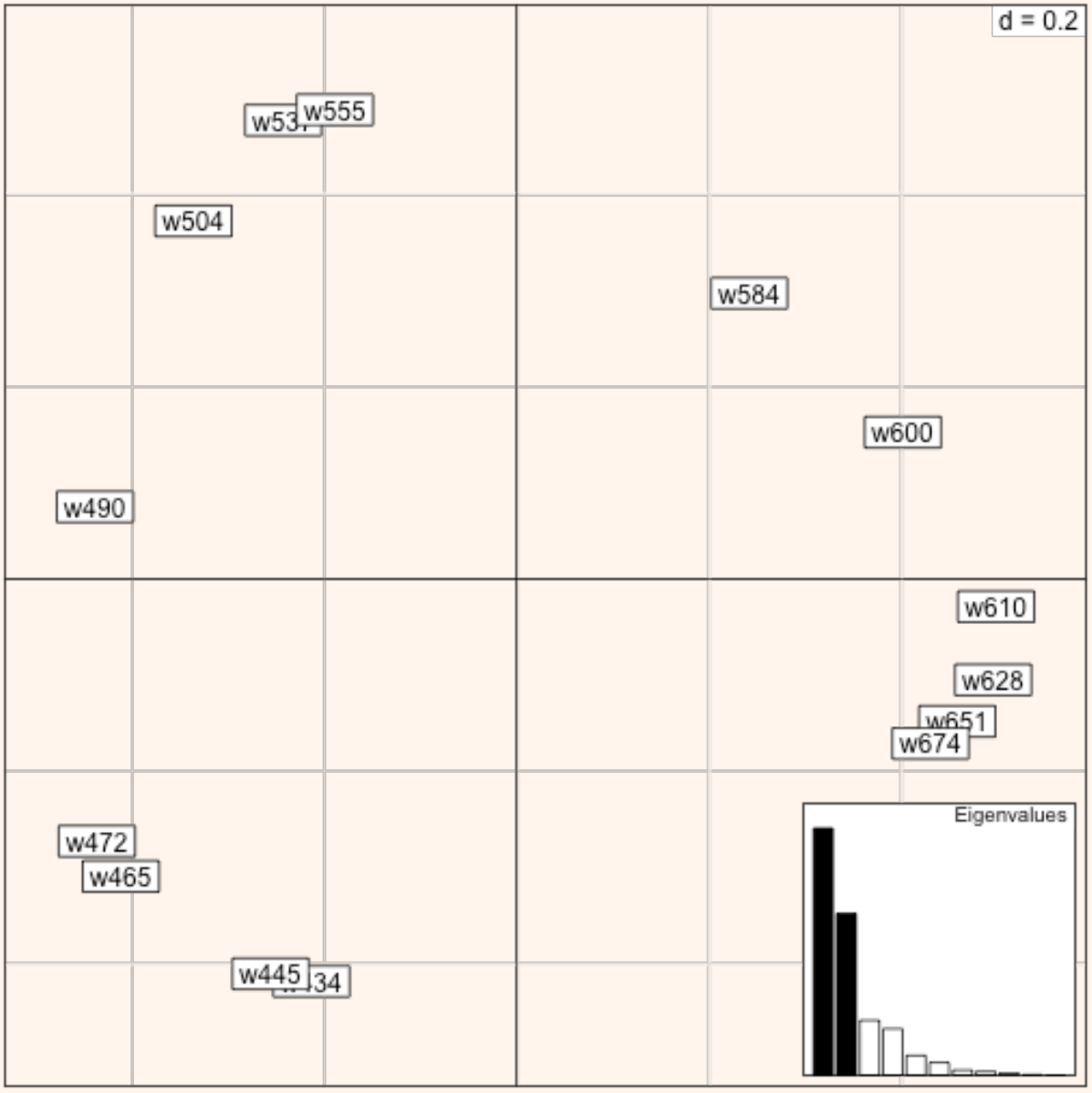
The advantage is that the original distances don't have to be Euclidean.

Psychological Data

Color confusion data (Eckman, 1954):

	w434	w445	w465	w472	w490	w504	w537	w555	w584	w600	w61
1	0.00	0.86	0.42	0.42	0.18	0.06	0.07	0.04	0.02	0.07	0.0
2	0.86	0.00	0.50	0.44	0.22	0.09	0.07	0.07	0.02	0.04	0.0
3	0.42	0.50	0.00	0.81	0.47	0.17	0.10	0.08	0.02	0.01	0.0
4	0.42	0.44	0.81	0.00	0.54	0.25	0.10	0.09	0.02	0.01	0.0
5	0.18	0.22	0.47	0.54	0.00	0.61	0.31	0.26	0.07	0.02	0.0
6	0.06	0.09	0.17	0.25	0.61	0.00	0.62	0.45	0.14	0.08	0.0
7	0.07	0.07	0.10	0.10	0.31	0.62	0.00	0.73	0.22	0.14	0.0
8	0.04	0.07	0.08	0.09	0.26	0.45	0.73	0.00	0.33	0.19	0.0
9	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	0.00	0.58	0.3
10	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	0.00	0.7
11	0.09	0.07	0.02	0.00	0.02	0.02	0.05	0.04	0.37	0.74	0.0
12	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.50	0.7
13	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.20	0.41	0.6
14	0.16	0.14	0.03	0.04	0.00	0.01	0.00	0.02	0.23	0.28	0.5

Results: Color Confusion



Planar configuration

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

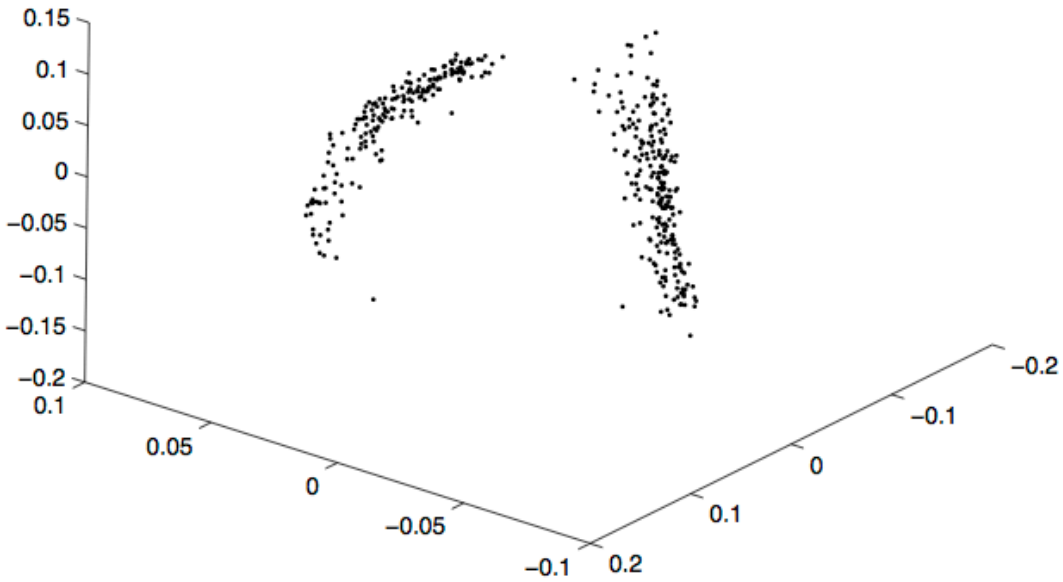
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
R1	-1	-1	1	-1	0	1	1	1	1	1	...
R2	-1	-1	1	-1	0	1	1	1	1	1	...
R3	1	1	-1	1	-1	1	1	-1	-1	-1	...
R4	1	1	-1	1	-1	1	1	-1	-1	-1	...
R5	1	1	-1	1	-1	1	1	-1	-1	-1	...
R6	-1	-1	1	-1	0	1	1	1	1	1	...
R7	-1	-1	1	-1	-1	1	1	1	1	1	...
R8	-1	-1	1	-1	0	1	1	1	1	1	...
R9	1	1	-1	1	-1	1	1	-1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

L_1 distance

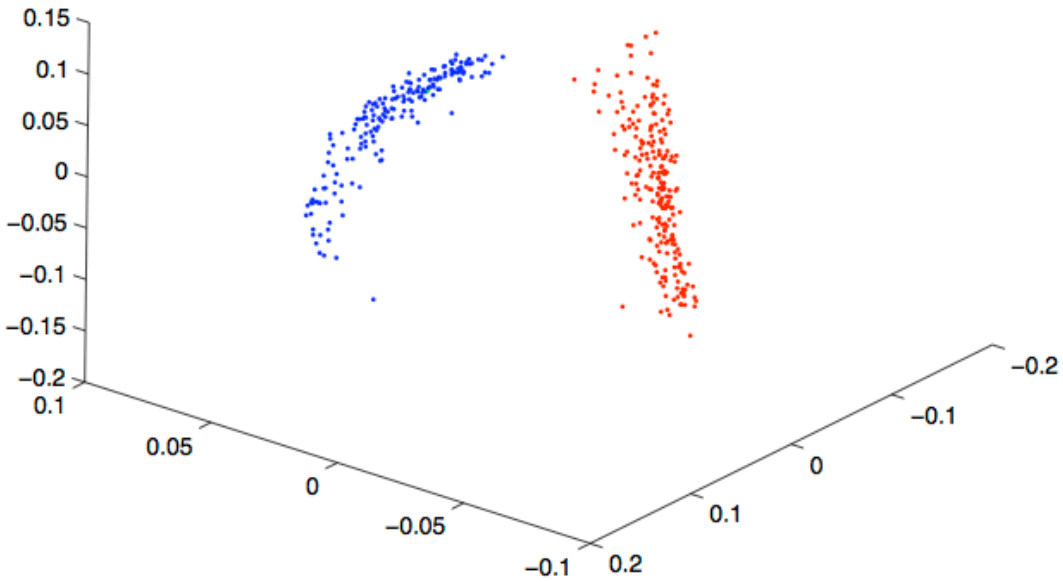
We define a distance between legislators as

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

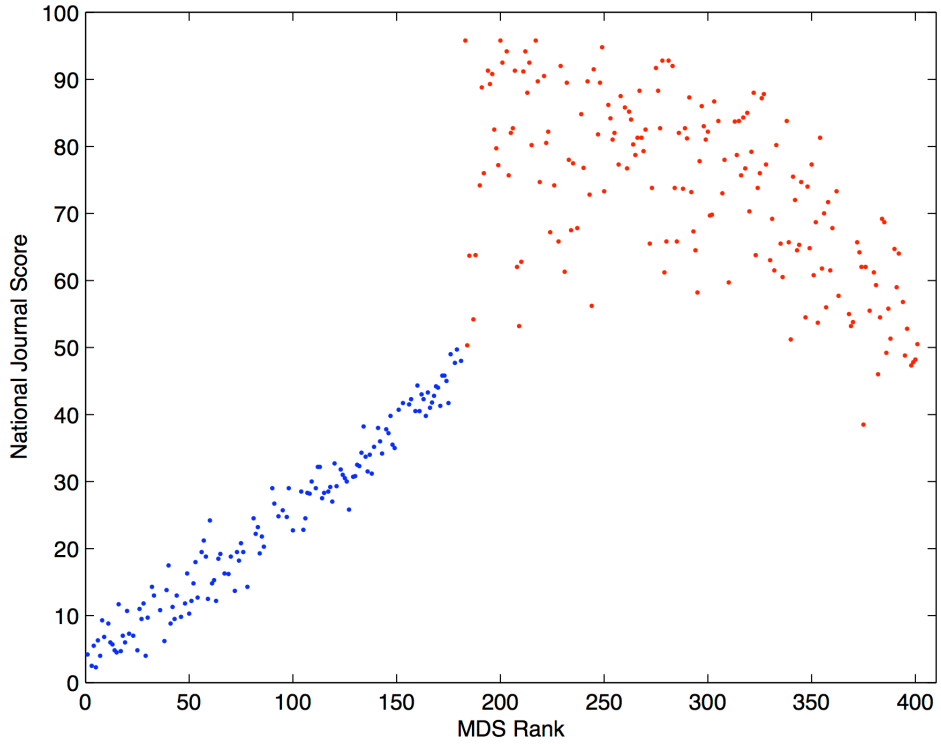
Roughly, $\hat{d}(l_i, l_j)$ is the percentage of roll calls on which legislators l_i and l_j disagreed.



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



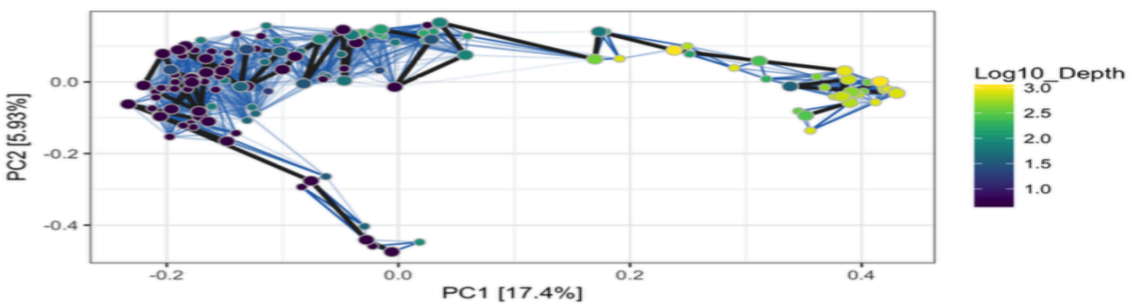
3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. Color has been added to indicate the party affiliation of each representative.



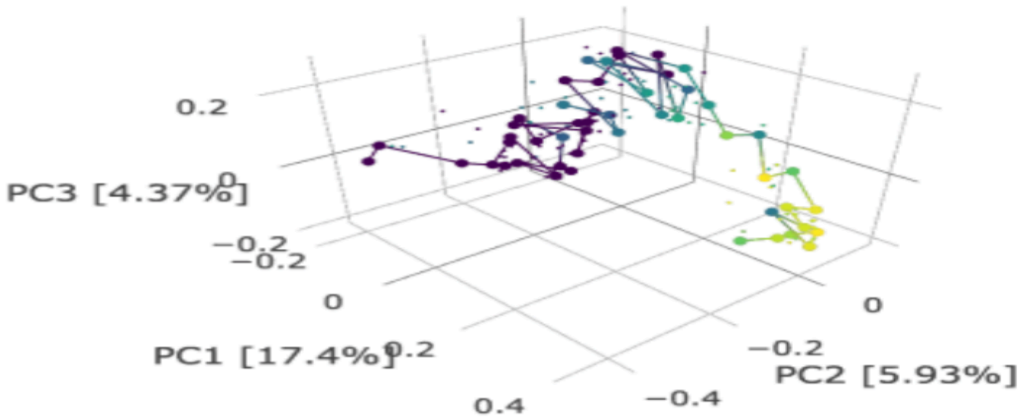
Comparison of the MDS derived rank for Representatives with the National Journal's liberal score

Uncertainty Quantification for rankings and gradients

Bayesian Unidimensional Scaling (Lan Huong Nguyen and Susan Holmes, 2017, BMC Bioinformatics).



(a) Trajectory in 2D



(b) Trajectory in 3D

Bayesian model for distances

$$
\begin{aligned}
d_{ij}|\delta_{ij} &\sim \text{Gamma}[\mu_{ij} = \delta_{ij},\, \sigma^2_{ij} = s^2_{ij}\sigma^2_\epsilon], \qquad (1)\\
\delta_{ij} &= |\tau_i - \tau_j|,\\
\tau_i|\alpha_\tau, \beta_\tau &\sim \text{Beta}(\alpha_\tau, \beta_\tau),\\
\alpha_\tau &\sim \text{Cauchy}^+(1, \gamma_\tau),\\
\beta_\tau &\sim \text{Cauchy}^+(1, \gamma_\tau),\\
\sigma_\epsilon &\sim \text{Cauchy}^+(0, \gamma_\epsilon),
\end{aligned}
$$

Modeling the heteroscedasticity

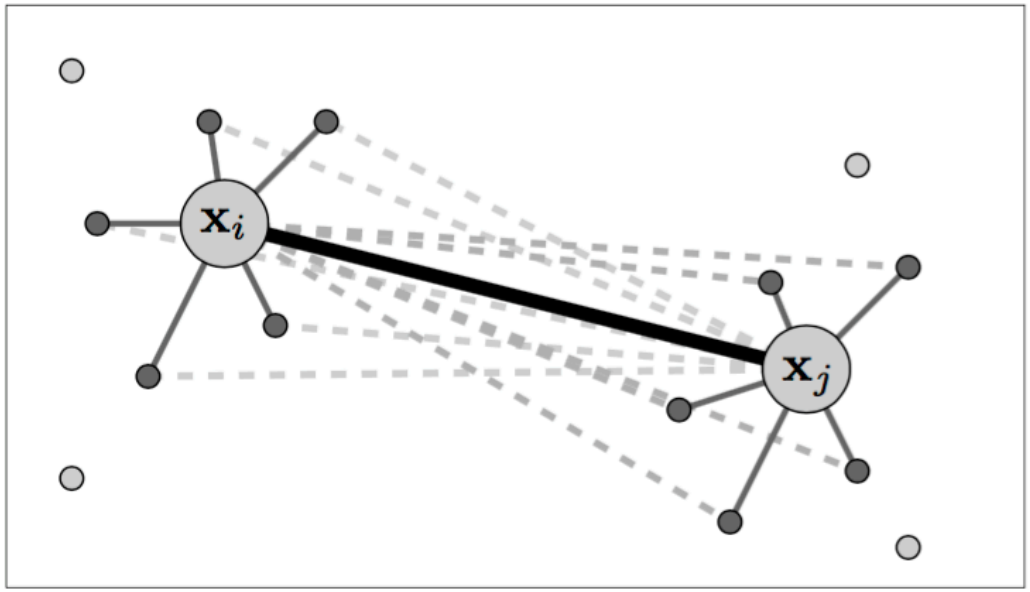
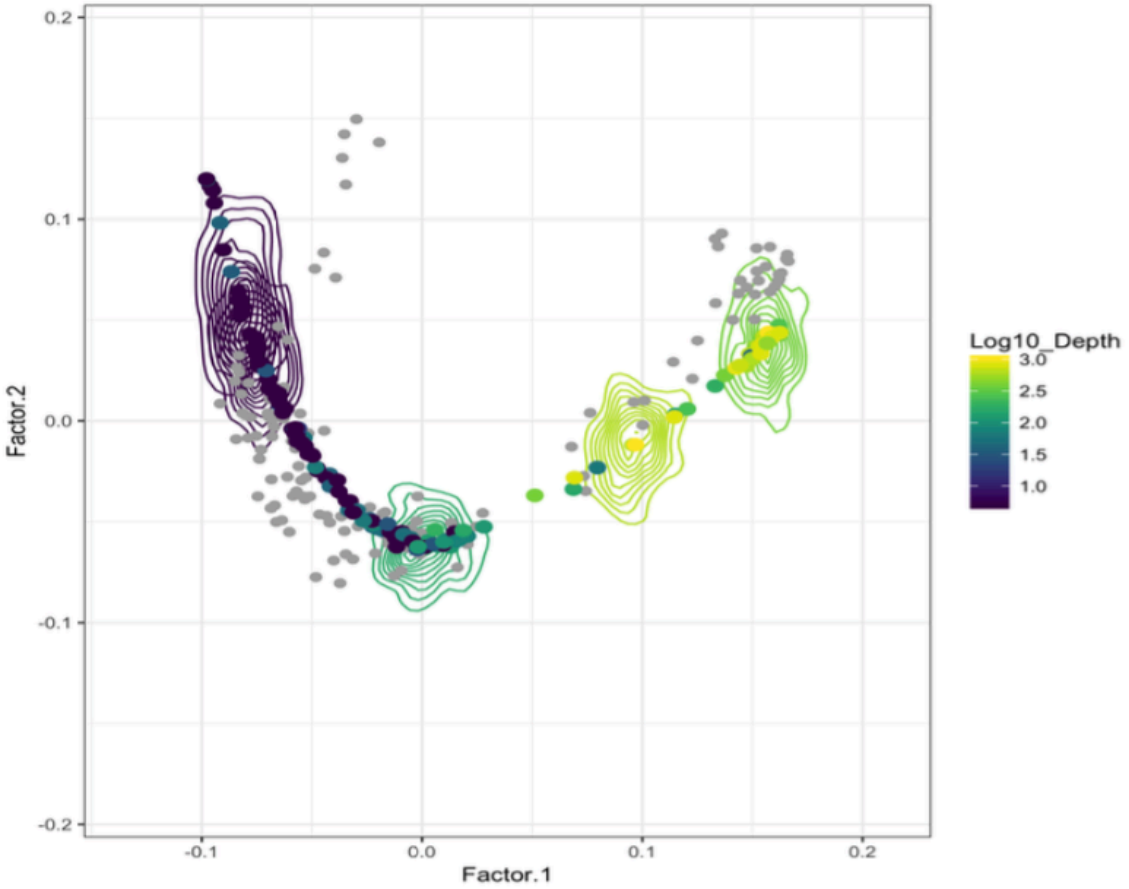


Figure 1 Graphical representation of points $\mathbf{x}_i$ and $\mathbf{x}_j$ together with their neighbors. The set of (dashed) distances from $\mathbf{x}_i$ to the K-nearest-neighbors of $\mathbf{x}_j$, and from $\mathbf{x}_i$ to the K-nearest-neighbors of $\mathbf{x}_i$ is used to compute $\hat{s}^2(d_{ij})$, the estimate of the variance of d_{ij}. Here we chose $K = 5$.

$$s(\hat{d}_{ij}) = \frac{1}{|D_{ij}^K|} \sum_{d \in D_{ij}^K} (d - \bar{d}_{ij}^K)^2$$

Scale parameter for the error term: $s_{ij}^2 = s(\hat{d}_{ij}) / s(\bar{\hat{d}}_{ij})$.



(b) Datapoint location confidence contours

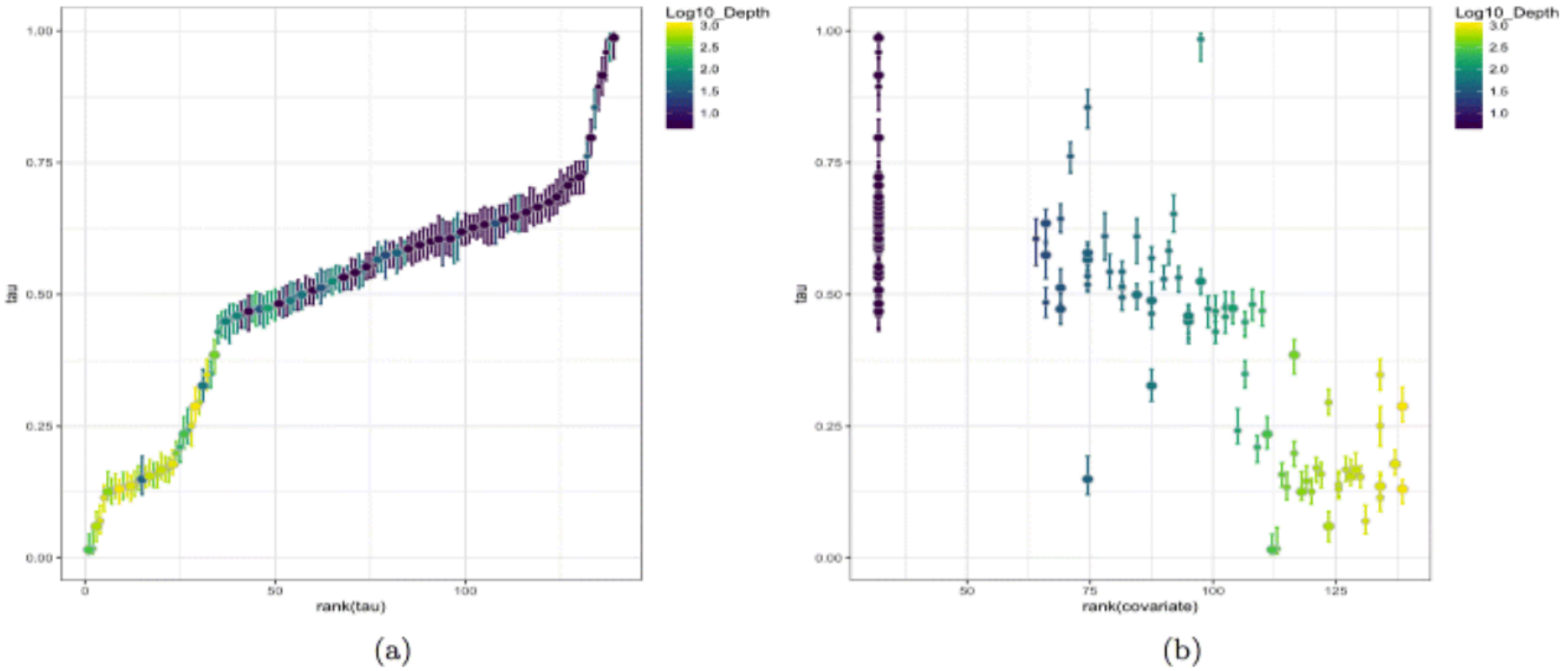


Fig. 4
Latent ordering in TARA Oceans dataset shown with uncertainties. The differences in the slope of plot (**a**) indicate varying data coverage along the underlying gradient. Correlation between the water depth and the latent ordering in microbial composition data is shown in (**b**). Coloring corresponds to log10 of the water depth (in meters) at which the ocean sample was collected

Code using stan

```
fit_buds <- function(D, K = NULL,
                     method = c("vb", "mcmc"),
                     hyperparams = list(
                       "gamma_tau"= 2.5,
                       "gamma_epsilon" = 2.5,
                       "gamma_bias" = 2.5,
                       "gamma_rho" = 2.5,
                       "min_sigma" = 0.03),
                     init_from = c("random", "princ
                     seed = 1234, max_trials = 20,
```

buds package on github:
https://github.com/nlhuong/buds.
See BMC paper [10].

Part III

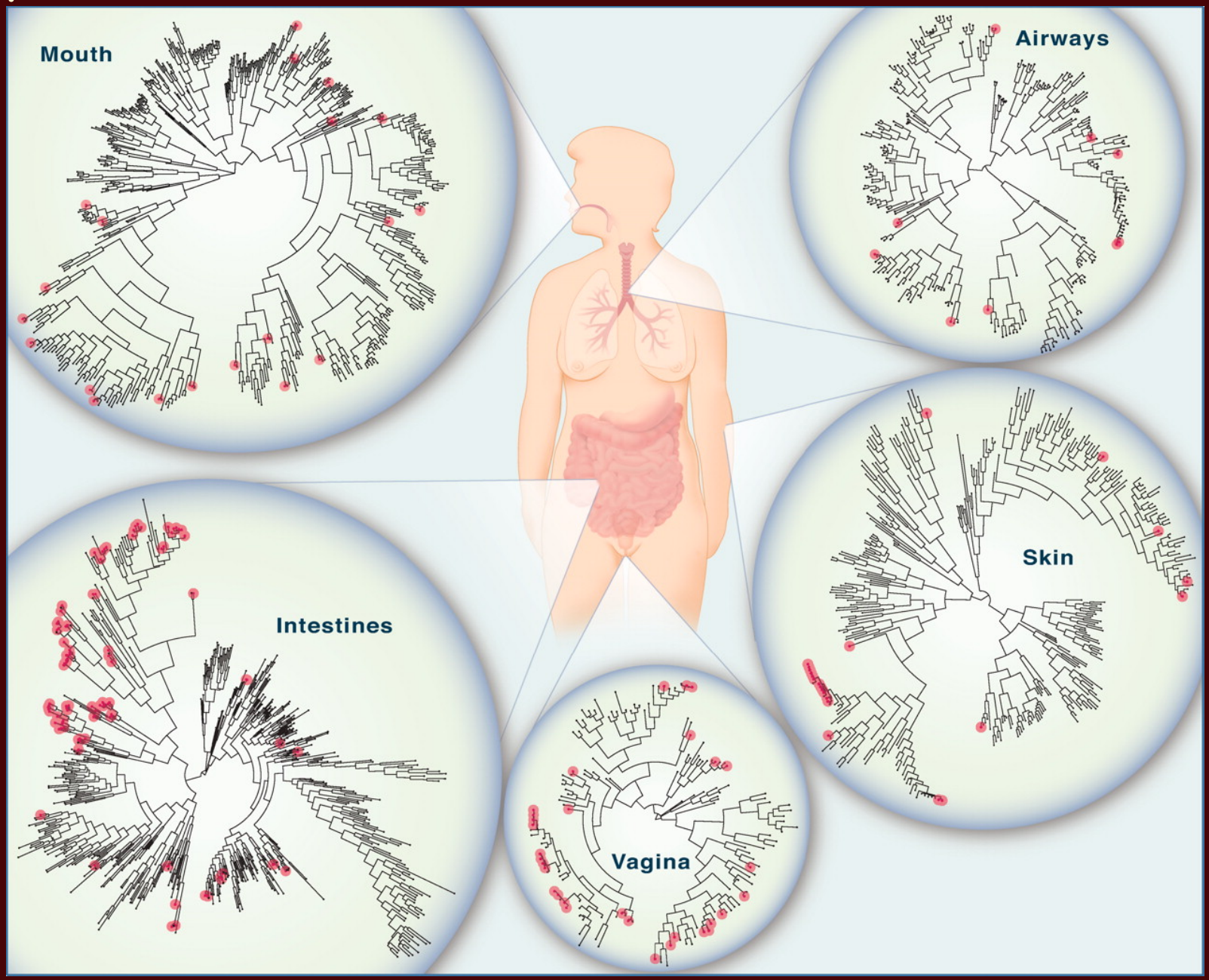
Combine and Compare Trees, Graphs and Contingent Count Data


Homogeneous data are all alike;
all heterogeneous data are
heterogeneous
in their own way.

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 $\times 10^{12}$ cells).
Metagenome	Composition of all genes present in an environment (soil, gut, seawater), regardless of species.
Transciptome	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.

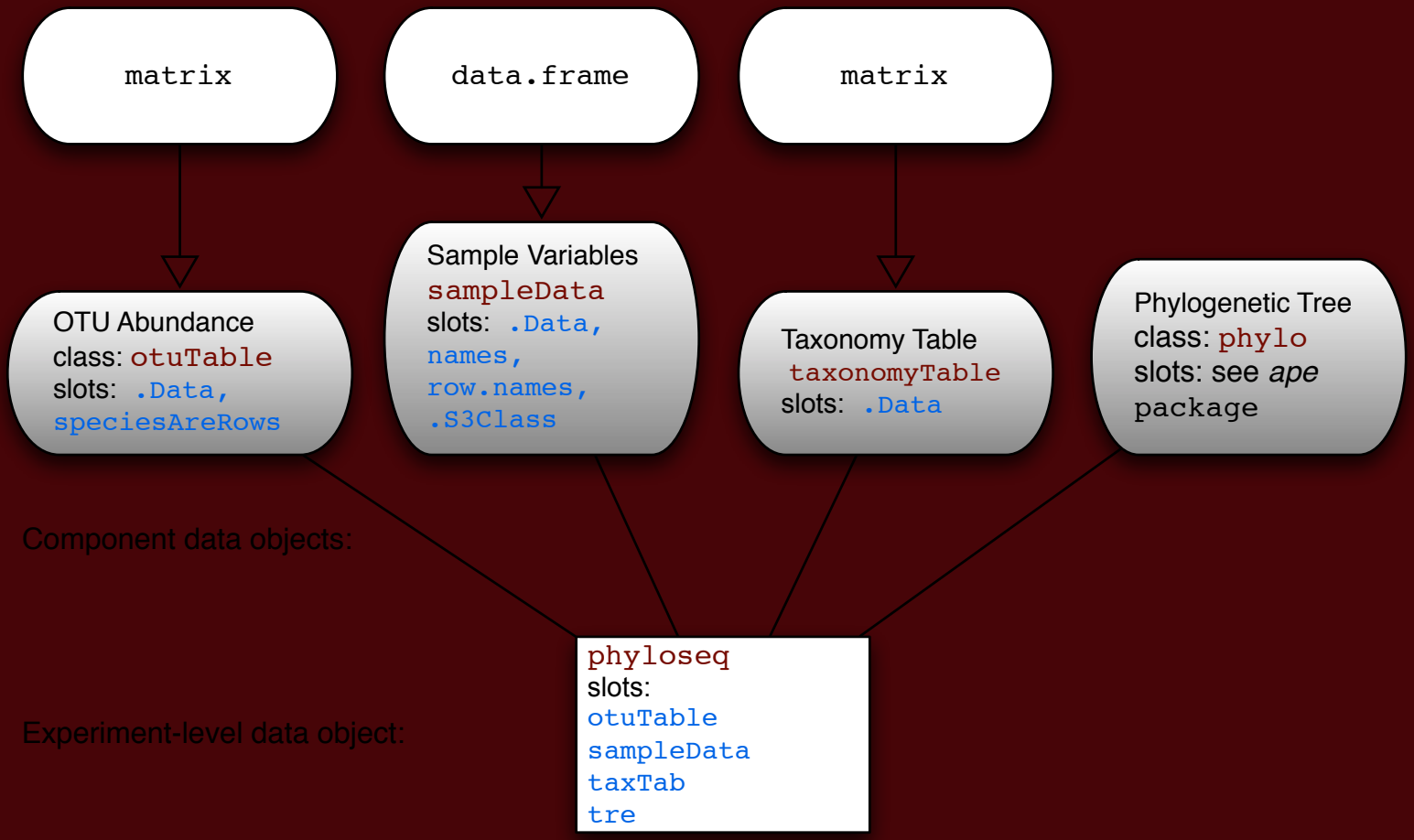
Human Microbiome: What are the data?

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.

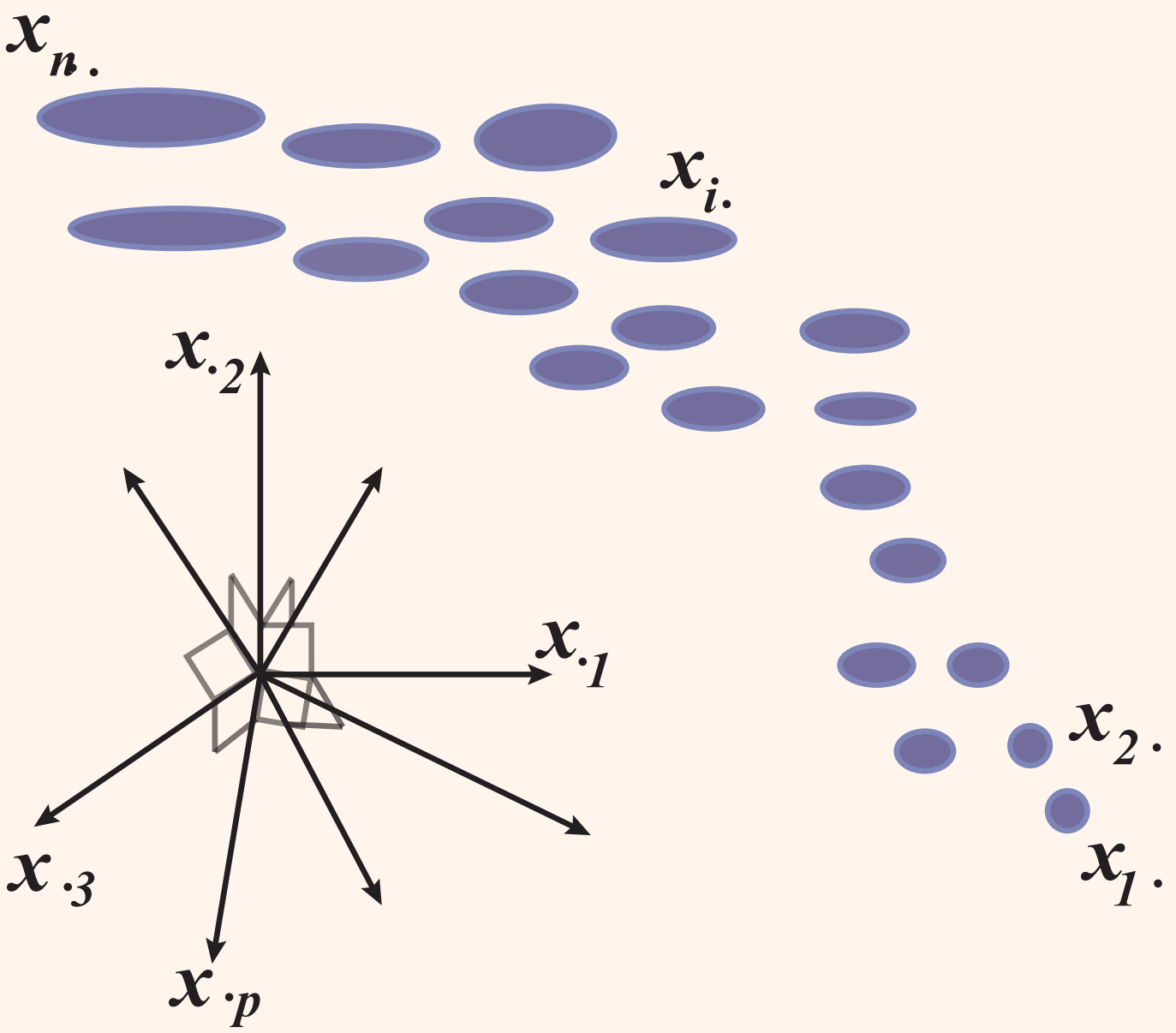
Heterogeneous Data Objects

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



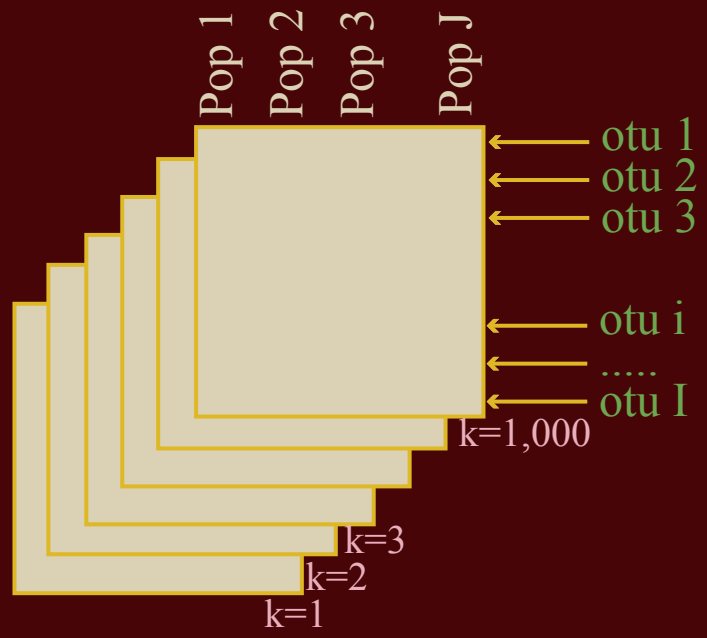
```
Some real data (Caporoso et al, 2011)
> GlobalPatterns

phyloseq-class experiment-level object
otu_table()  OTU Table:         [ 19216 taxa and 26 sampl
sample_data()Sample Data:       [ 26 samples by 7 sample
tax_table()Taxonomy Table:    [ 19216 taxa by 7 taxonomic
phy_tree() Phylogenetic Tree:[ 19216 tips and 19215 inte

> sample_sums(GlobalPatterns)
     CL3       CC1       SV1  M31Fcsw  M11Fcsw  M31Plmr  M
  864077   1135457    697509  1543451  2076476   718943
 .....
     NP3       NP5  TRRsed1  TRRsed2  TRRsed3     TS28
 1478965   1652754    58688   493126   279704   937466  1

> summary(sample_sums(GlobalPatterns))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58690  567100 1107000 1085000 1527000 2357000
```

An example of OTU table.

OTU	Ctrl1	Ctrl2	Ctrl3	Ctrl4	Ctrl5	IBD1	IBD2	IBD3
Bacteroides	1822	913	147	2988	4616	172	3516	657
Bifidobacterium	0	162	0	0	84	0	85	1927
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	58

Output showing Bayesian posterior uncertainty measures

The methods that we consider here are all related to PCA and use the normalized Gram matrix **S** between biological samples. **S** is the Gram operator matrix of $(Q_{i,1}, \ldots, Q_{i,J})$. Based on a single posterior instance of **S**, we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.



Bayesian nonparametric ordination for the analysis of microbial communities
Ren, Bacallado, Favaro, Holmes, Trippa (2017)
JASA, February .

Full Bayesian nonparametric model

- We do not know the number of OTUs.
- We suppose underlying low dimensional latent variables for the sample P^j's.
- We use dependent microbial distributions, marginal priors of discrete distributions are built using manipulation of a Gaussian process and then extending this to multiple correlated distributions.

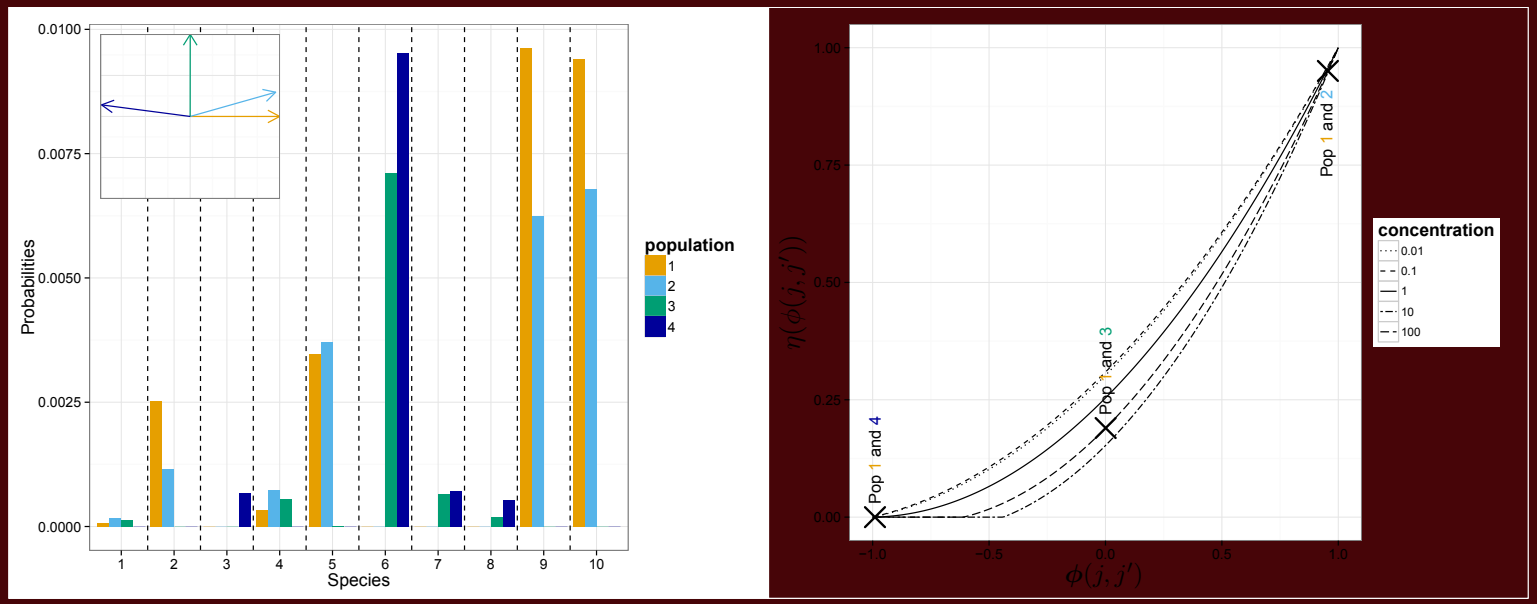


Figure: **Left panel**: realization of 4 microbial distributions from a dependent Dirichlet processes with 10 OTUs **Right panel**: correlation of two random probability measures when the cosine $\phi(j, j')$ between $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$ varies from -1 to 1. (Ren et al, JASA, 2017).

Parameters for samples

$$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$$

Define a joint prior on these factors through the Gram matrix

$$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$$

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of OTUs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \tag{1}$$

where the $\epsilon_{i,j}$ are independent Normal variables.

The degree of similarity between the discrete distributions $\{P^j; j \in \mathcal{J}\}$ is summarized by the Gram matrix $(\phi(j,j') = \langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle; j,j' \in \mathcal{J})$.
The dependent Dirichlet processes is defined by setting

$$P^j(A) = \frac{\sum_i \mathbb{I}(Z_i \in A) \times \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}, \qquad \forall j \in \mathcal{J}, \tag{2}$$

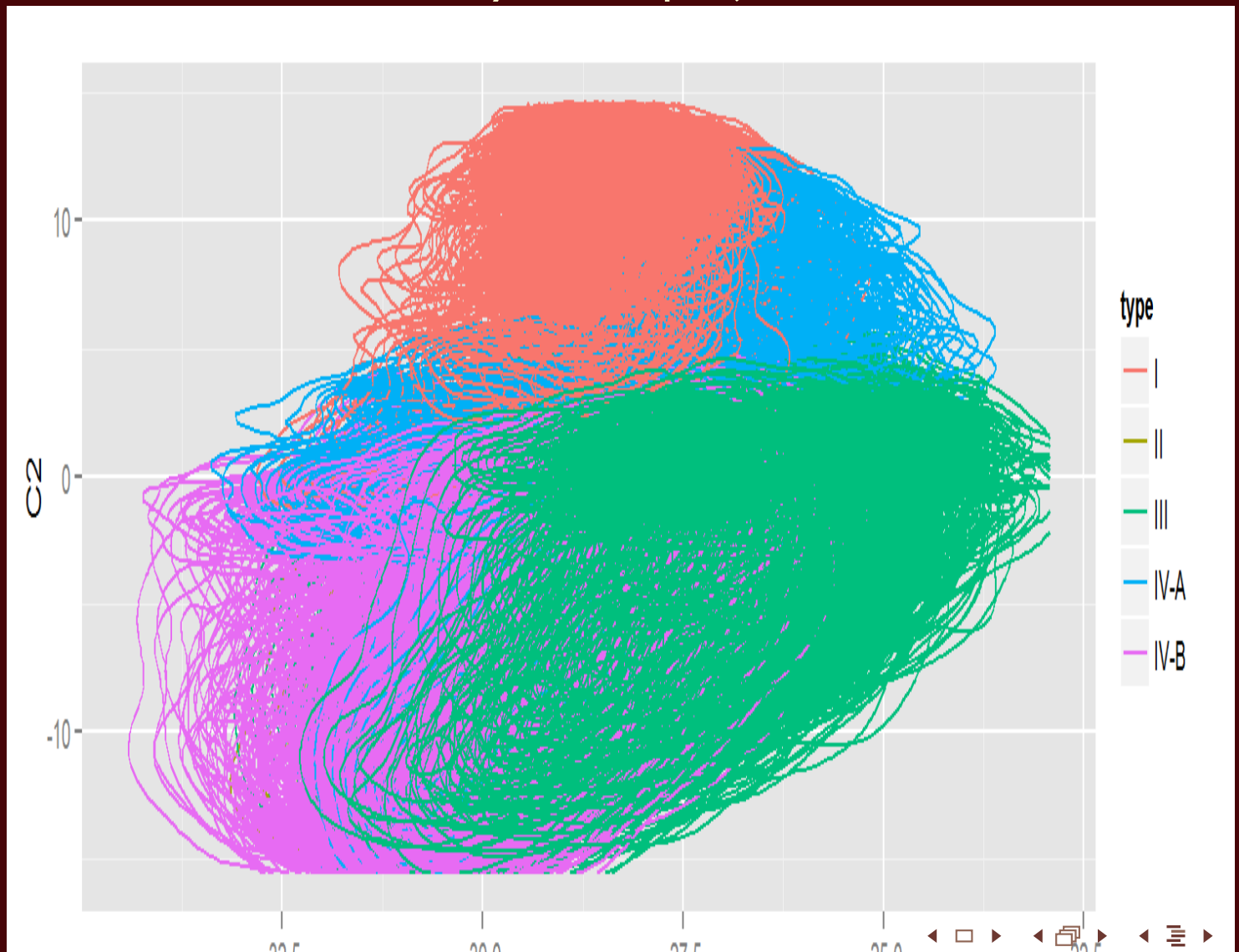
for every $A \in \mathcal{F}$. Here the sequence $(Z_1, Z_2, \ldots)$ and the array $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$, contain independent and identically distributed random variables, while σ is a Poisson process on the unit interval defined by using a prior on $\sigma = (\sigma_1, \sigma_2, \ldots)$, the distribution of ordered points $(\sigma_i > \sigma_{i+1})$ in a Poisson process on $(0,1)$ with intensity

$$\nu(\sigma) = \alpha \sigma^{-1} (1-\sigma)^{-1/2}, \tag{3}$$

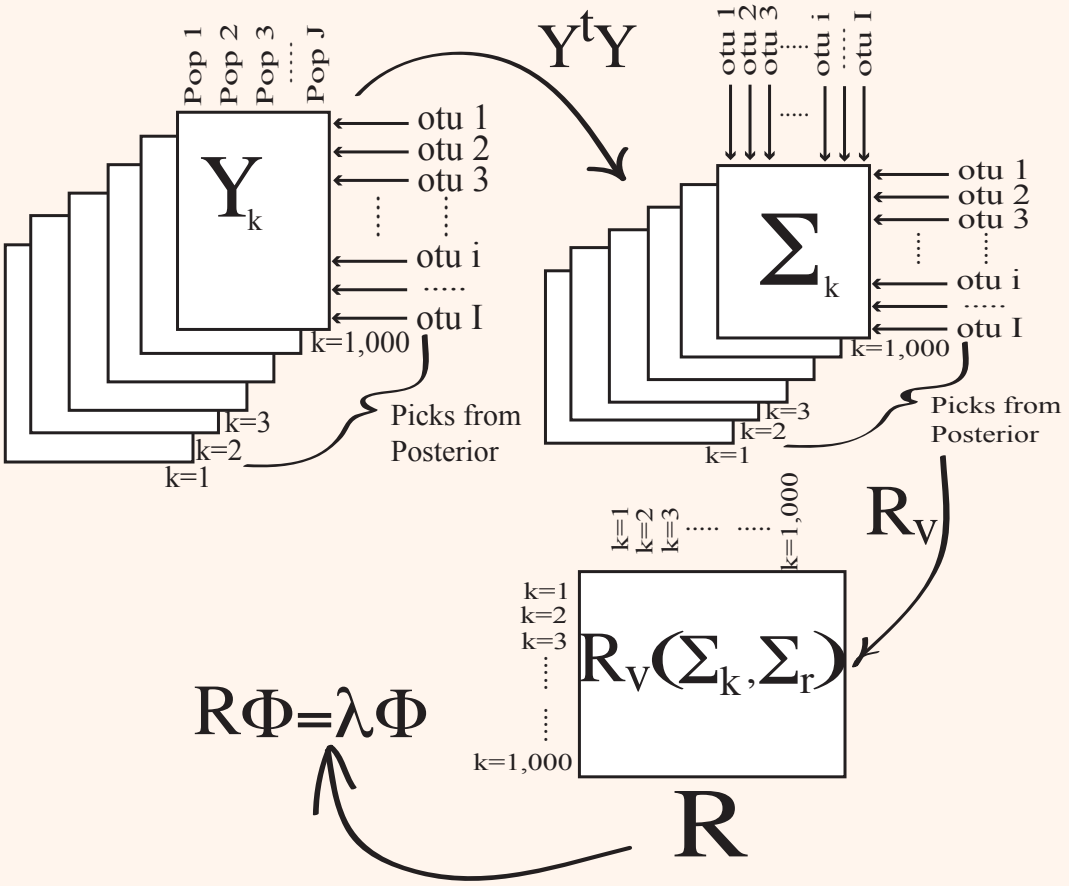
where $\alpha > 0$ is a concentration parameter.
We will use the notation $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$.

The Naive projection approach

Naively overlaying projections of the principal coordinate loadings generated from different posterior samples of **S** on the same plot *could* show the variability of the projections.

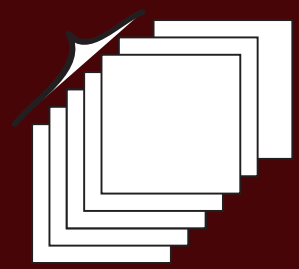


Alternatively



We identify a consensus lower dimensional space for all posterior samples using STATIS (Escoufier, 1980, see Holmes, 2005). We list the three main steps used to visualize the variability of **S**.

Registration: Find $\mathbf{S}_0$



Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.
We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$.
We use the **RV** similarity metric between two symmetric square matrices $\mathbf{A}$ and $\mathbf{B}$

$$\mathrm{RV}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{AB}) / \sqrt{\mathrm{Tr}(\mathbf{AA})\mathrm{Tr}(\mathbf{BB})}$$

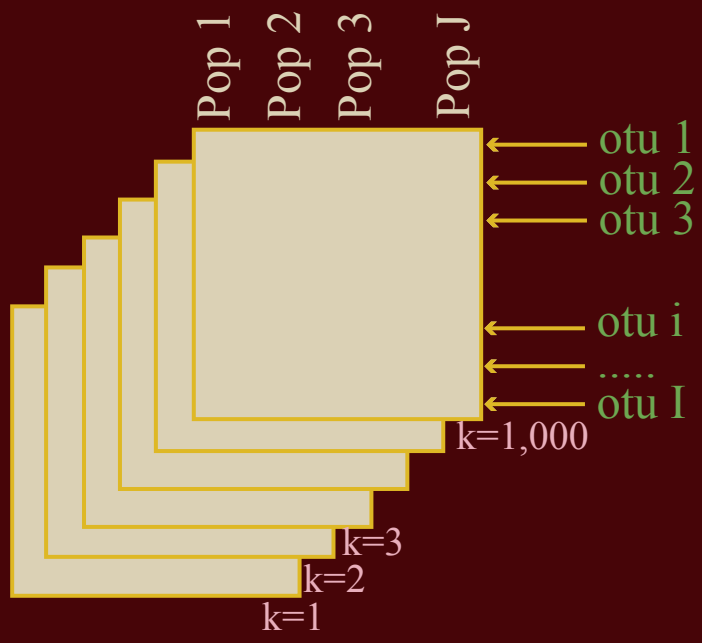
We diagonalize the **RV** matrix to obtain $\mathbf{S}_0$.

Find lower dimensional consensus space *V*

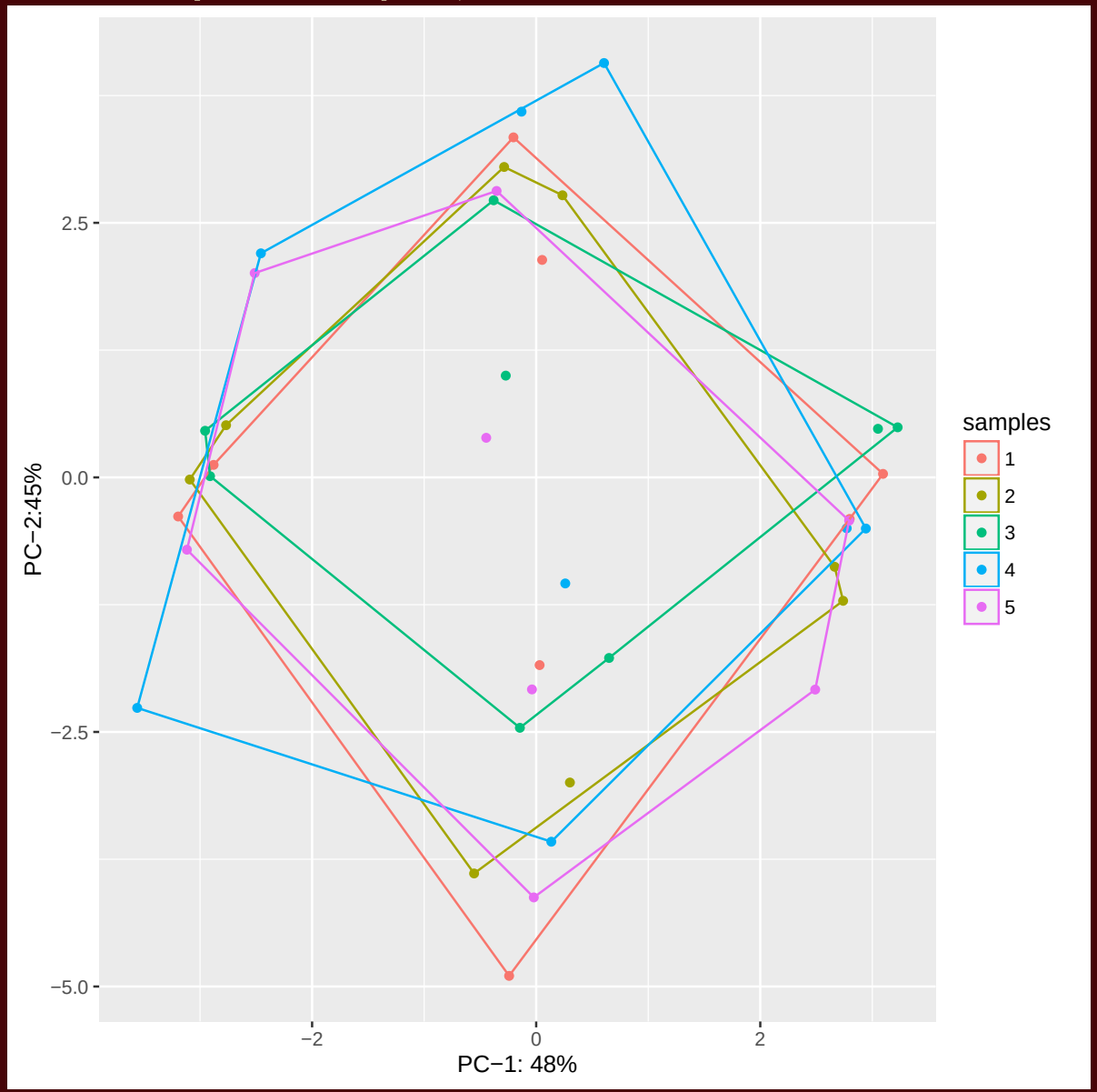
For dim 2, $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\mathbf{S}_0$ corresponding to the largest eigenvalues λ_1 and λ_2. All biological samples in *V* are visualized by projecting rows of $\mathbf{S}_0$ onto *V*: $(\psi_1^0, \psi_2^0) = \mathbf{S}_0(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$.

Project the rows of posterior sample $\mathbf{S}_k$ onto V by $(\psi_1^k, \psi_2^k) = \mathbf{S}_k(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$. Overlaying all the ψ^k displays uncertainty of $\mathbf{S}$ in the same linear subspace. Posterior variability of the biological samples' projections is visualized in V by plotting each row of the matrices (ψ_1^k, ψ_2^k), $k = 1, \ldots, K$, in the same figure.

Posterior distribution of MDS plots of taxa abundance contingency tables

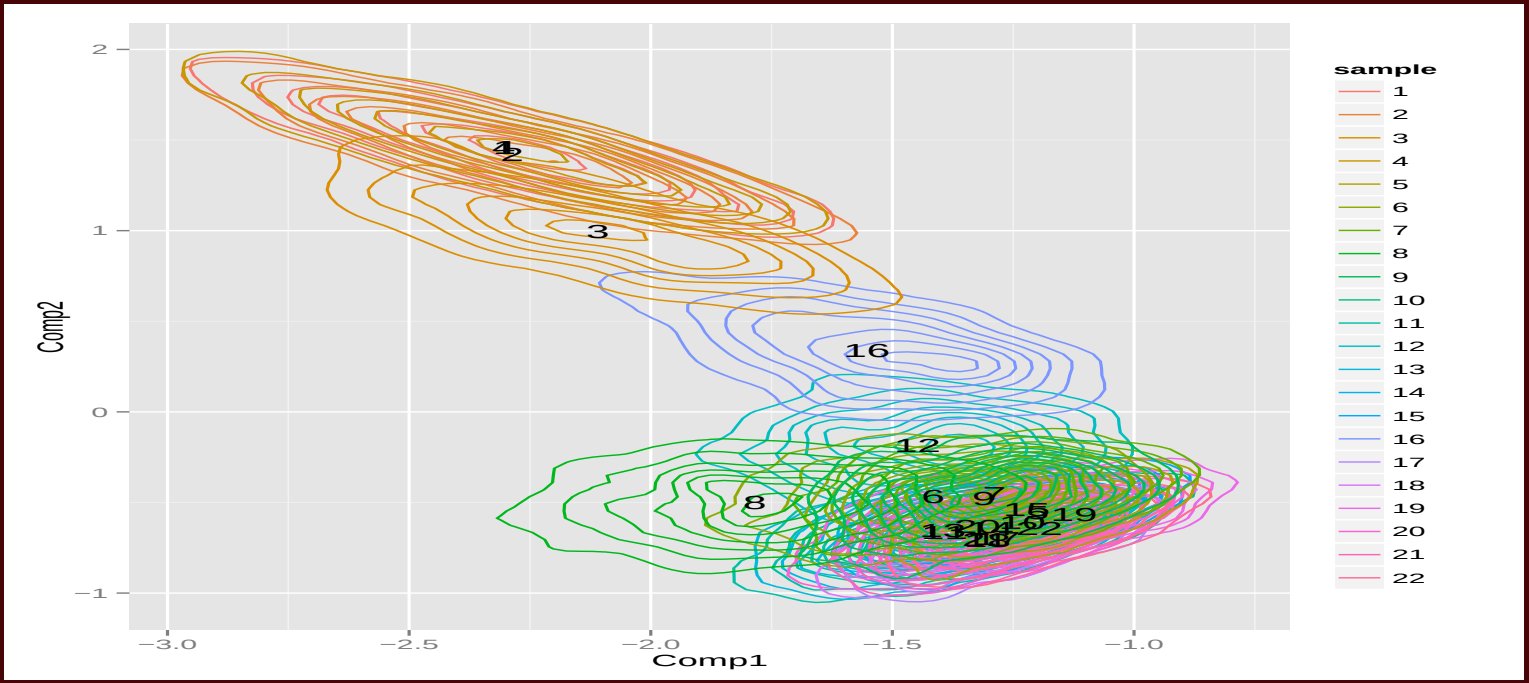


Naively overlaying projections of the principal coordinate loadings generated from different resamples on the same plot *could* show the variability of the projections.



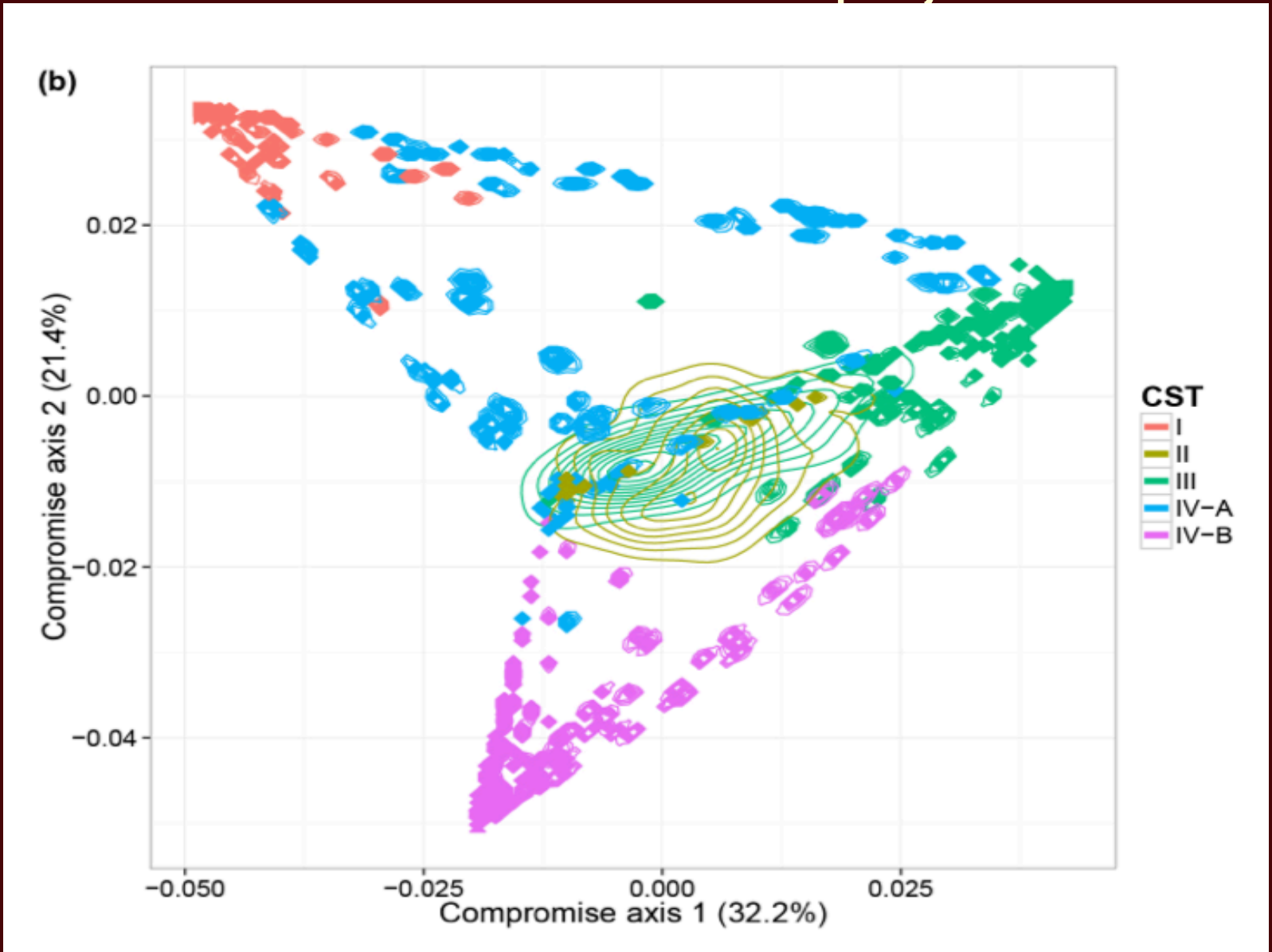
Why?

- Principal coordinate directions are only defined up to a sign.
- Principal coordinates, 1 and 2 or 2 and 3 can be permuted.
- We need to do **registration** first.



A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space *V*.

Posterior distribution of ordination projections



Given posterior samples of the model parameters, we use a procedure to plot credible regions in visualizations.

Distances and probabilistic weights enable statisticians to....

- Summarize data with medians, means and principal directions.
- Encode some measures of uncertainty (posterior probability regions, bootstrap).
- Make comparisons of heterogeneous sources of information.

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.

Summary

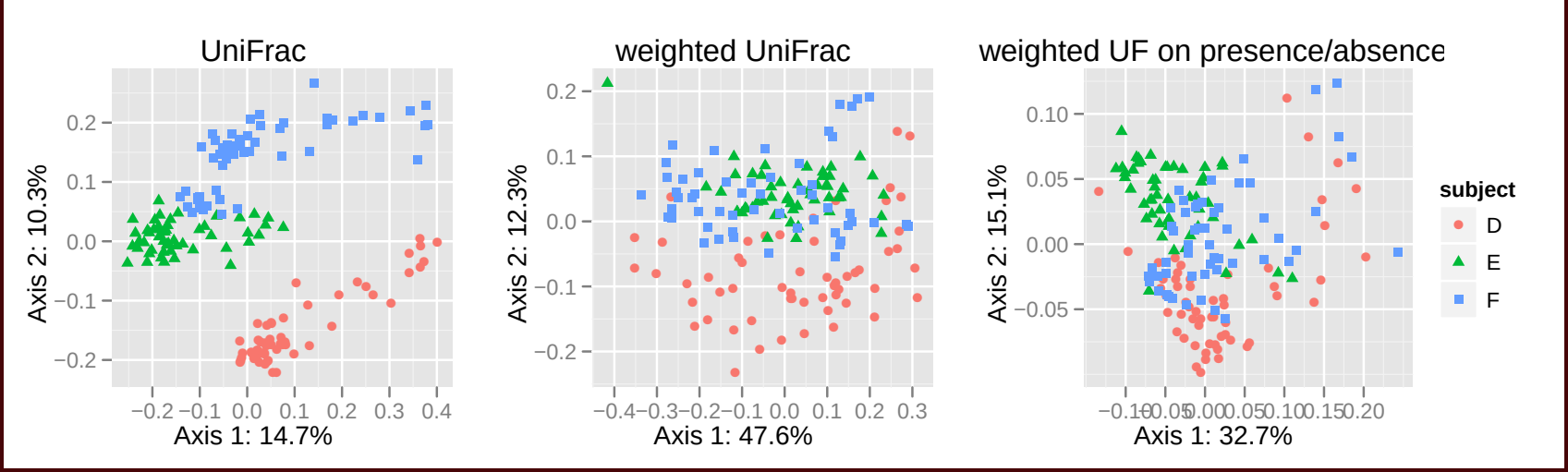
Fukuyama and Holmes, 2012.

Original description	New formula	Properties
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
$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	Less sensitive to outliers/more sensitive to noise than DPCoA
fraction of branches leading to exactly one group	$\sum_i b_i \mathbf{I}\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

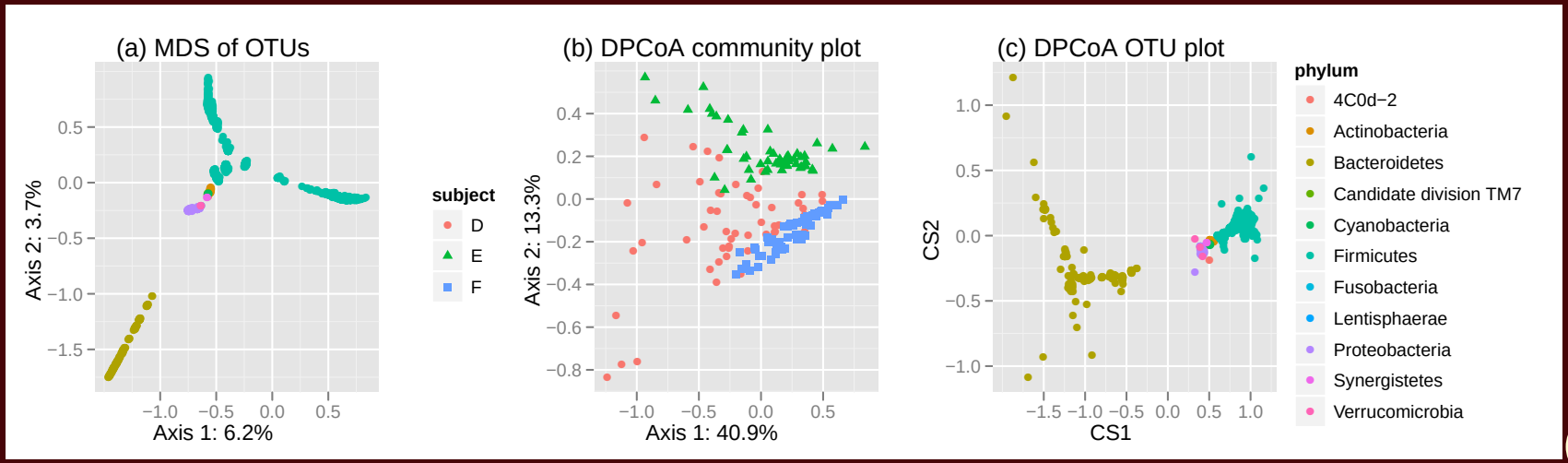
of the methods under consideration. "Outliers" refers to highly abundant OTUs, and noise refers to noise in detecting low-abundance OTUs (see the text for more detail).

Antibiotic Time Course Data

Measurements of about 2500 different bacterial OTUs from stool samples of three patients (D, E, F)
Each patient sampled $\sim$ 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

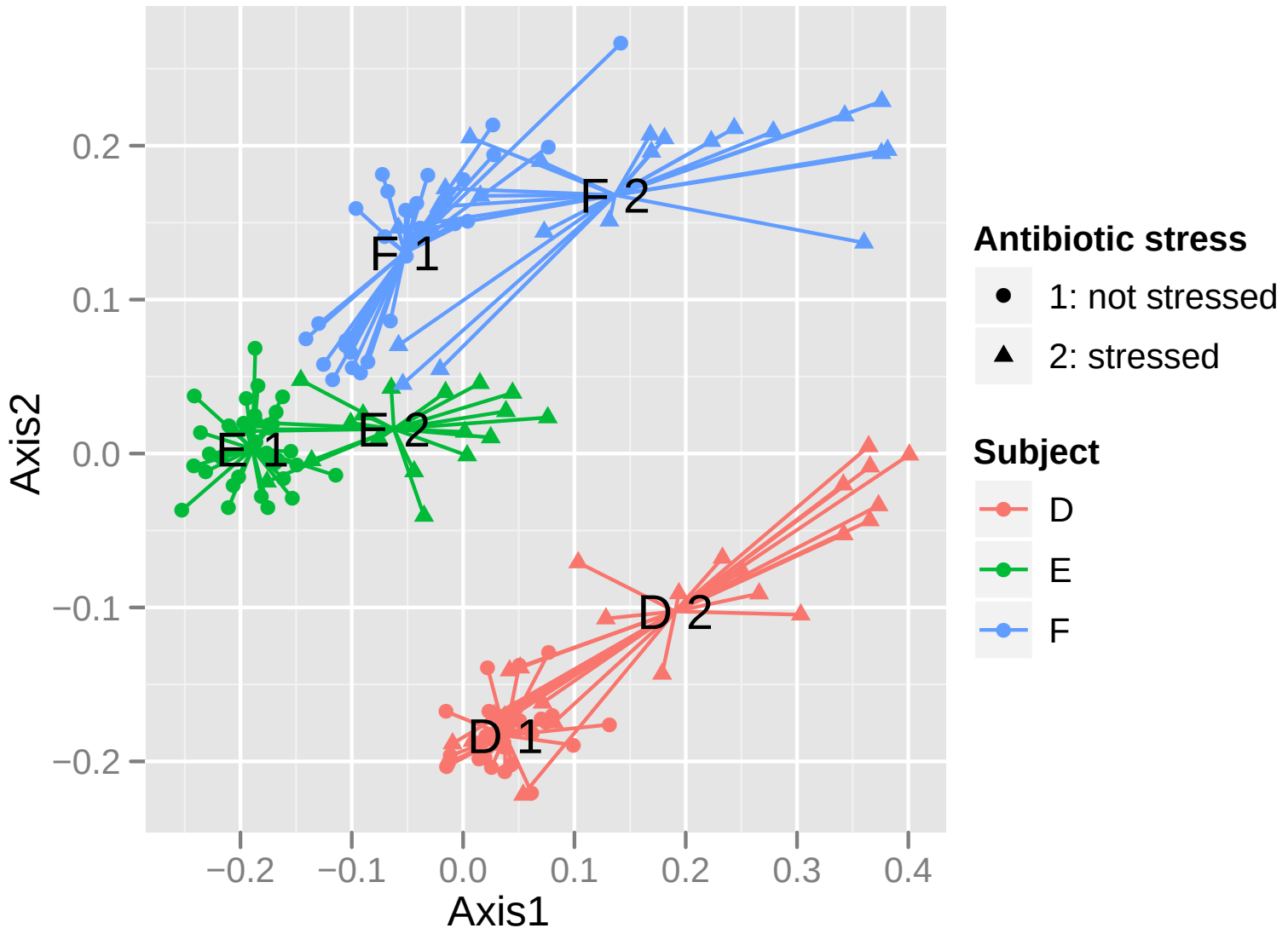


(a)

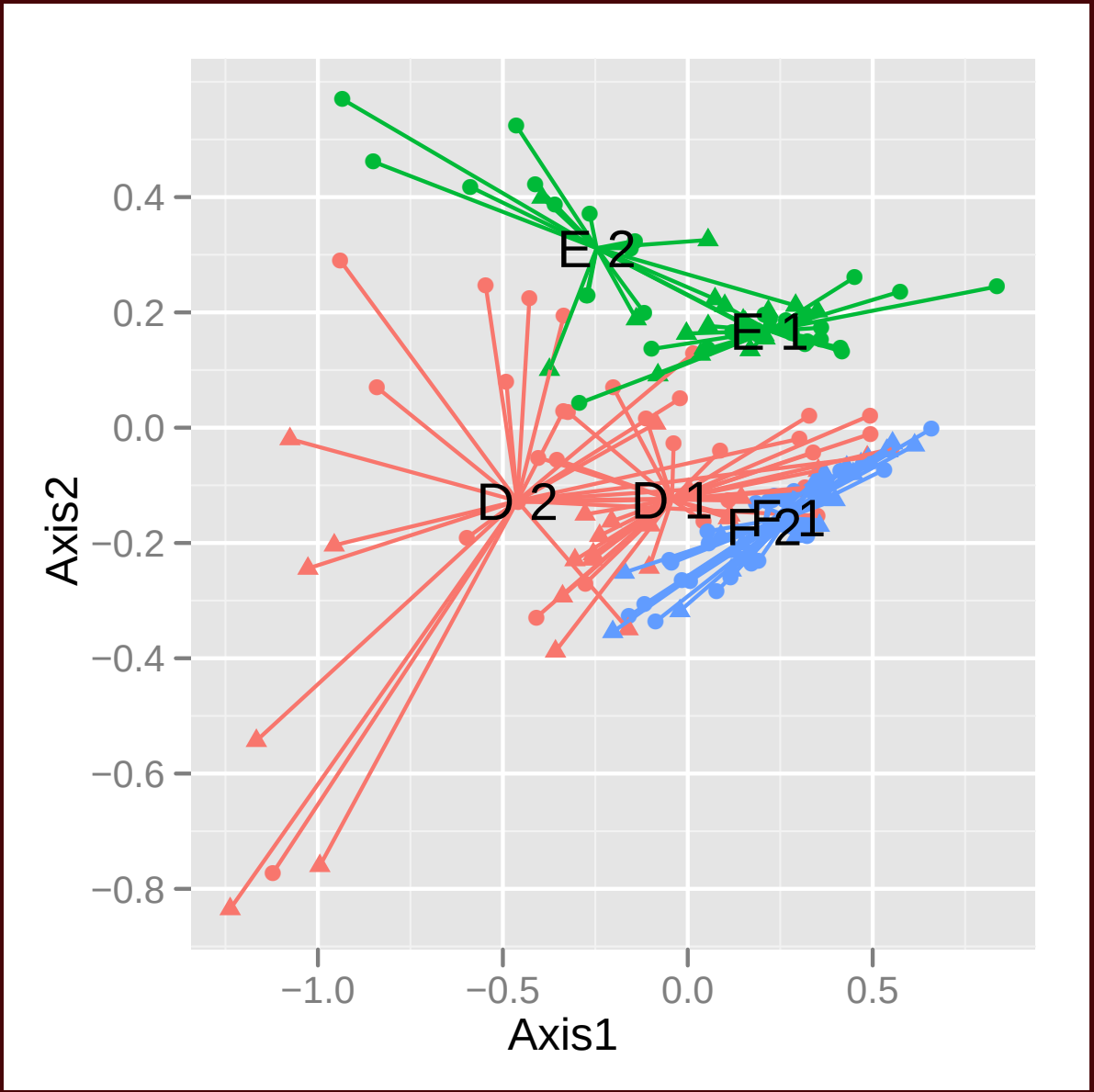
PCoA/MDS of the OTUs based on the patristic distance, (b) community and (c) species points for DPCoA after removing two outlying species.

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.

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.

Benefitting from the tools and schools of Statisticians.....

Thanks to the `R` and `stan` community:

- RStudio for tools for reproducible research and ggplot2.
- `stan` for good implementations of HMC and VB.
- Ecologists and biologists: Wolfgang Huber (EMBL), Chessel, Jombart, Dray, Thioulouse `ade4` and Emmanuel Paradis `ape`.

Collaborators: David Relman, Alfred Spormann, Yves Escoufier, Les Dethfelsen, Justin Sonnenburg, Persi Diaconis, Sergio Baccallado, Elisabeth Purdom.

Lab Group

Postdoctoral Fellows

Paul (Joey) McMurdie, Pratheepa Jeganathan, Ben Callahan, Simon Rubinstein-Salzado, Christof Seiler.

Students: John Chakerian, Diana Proctor, Julia Fukuyama, Kris Sankaran, Lan Nguyen, Claire Donnat.

Funding from NIH/ NIGMS R01, NSF-VIGRE and NSF-DMS.

References


J. Chakerian and S. Holmes.
`distory`:Distances between trees, 2010.


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.


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.


Julie Josse, Susan Holmes, et al.
Measuring multivariate association and beyond.
Statistics Surveys, 10:132–167, 2016.


K. Mardia, J. Kent, and J. Bibby.
Multiariate Analysis.
Academic Press, NY., 1979.


P. J. McMurdie and S. Holmes.
Phyloseq: Reproduible 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.


Lan Huong Nguyen and Susan Holmes.
Bayesian unidimensional scaling for visualizing uncertainty in high dimensional datasets with latent ordering of observations.
BMC bioinformatics, 18(10):394, 2017.


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.


Boyu Ren, Sergio Bacallado, Stefano Favaro, Susan Holmes, and Lorenzo Trippa.
Bayesian nonparametric ordination for the analysis of microbial communities.
Journal of the American Statistical Association, (just-accepted), 2017.


Christof Seiler, Xavier Pennec, and Susan Holmes.
Random spatial structure of geometric deformations and Bayesian nonparametrics.
In *Geometric science of information*, pages 120–127. Springer, 2013.


Christof Seiler, Simon Rubinstein-Salzedo, and Susan Holmes.
Positive curvature and Hamiltonian Monte Carlo.
In *Advances in Neural Information Processing Systems*, pages 586–594, 2014.