

# HETEROGENEOUS DATA CHALLENGE COMBINING COMPLEX DATA

*Susan Holmes*

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

*Bio-X and Statistics, Stanford University*

*NSF grant #0241246 and NIH-RO1GM086884-2*



*'Homogeneous data are all alike;*

*all heterogeneous data are heterogeneous*

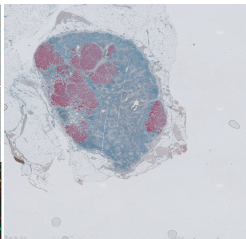
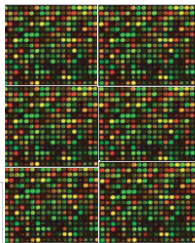
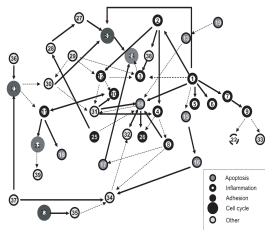
*in their own way.'*

# Heterogeneity

- ▶ *Status* : response/ explanatory.
- ▶ *Hidden (latent)/ measured.*
- ▶ *Type* :
  - ▶ *Continuous*
  - ▶ *Binary, categorical*
  - ▶ *Graphs/ Trees*
  - ▶ *Images*
  - ▶ *Maps/ Spatial Information*
  - ▶ *Rankings*
- ▶ *Amounts of dependency: independent/time series/spatial.*

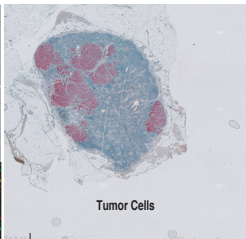
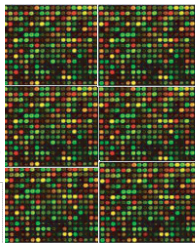
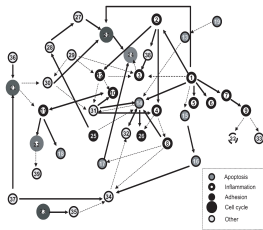
# Goals in Modern Biology: Systems Approach

Look at the data/ all the data: data integration



# Goals in Modern Biology: Systems Approach

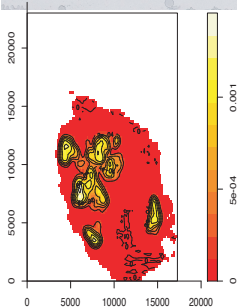
Look at the data/ all the data: data integration



$$\begin{pmatrix} 0110-11000-1 \\ 01100000001 \\ 01-10-10000-1 \\ 01100000101 \\ 01100-11011 \end{pmatrix}$$

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

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

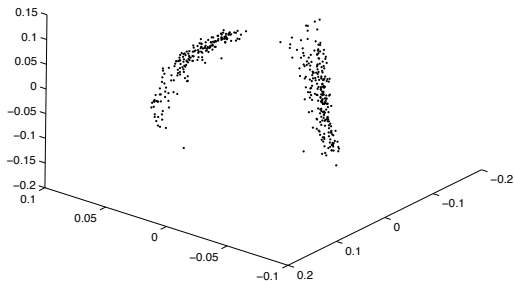


# Taking Categorical Data and Making it into a Continuum

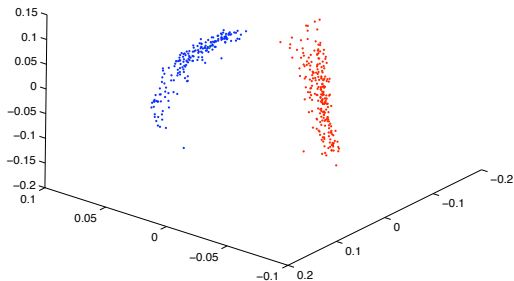
Horseshoe Example: Joint with Persi Diaconis and Sharad Goel. Data from 2005 U.S. House of Representatives roll call votes. We further restricted our analysis to the 401 Representatives that voted on at least 90% of the roll calls (220 Republicans, 180 Democrats and 1 Independent) leading to a  $401 \times 669$  matrix of voting data.

## The Data

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

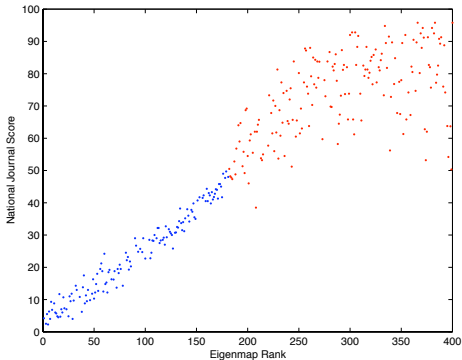


*3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes.*



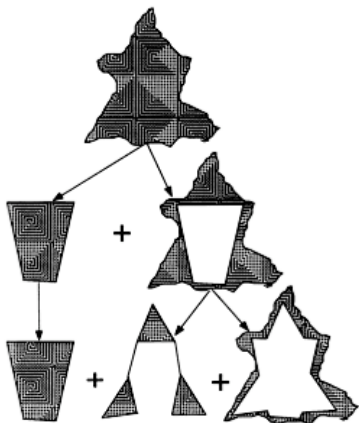
*3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. Color has been added to indicate the party affiliation of each representative.*





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

# Iterative Structuration (Tukey, 1977)



*A distance  $\rightarrow$  projection*

# Phylogenetic Trees

```

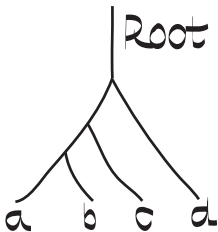
Pre1      GTACTTGTTA GGCCTTATAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pme2      GTATCTGTTA AGCCTTATAA AAAGATAGT- T-TAAATTTA AGGAATTATA
Pma3      GTATTTGTTA AGCCTTATAA GAGAAAAGTA TATTAACCTA AGGA-TTATA
Pfa4      GTATTTGTTA GGCCTTATAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pbe5      GTATTTGTTA AGCCTTATAA GAAAAA--T- TTTTAAATTA AGGAATTATA
Plo6      GTATTTGTTA AGCCTTATAA GAAAAAAGT- TACTAACTAA AGGAATTATA
Pfr7      GTACTTGTTA AGCCTTATAA GAAAAGAGT- TATTAACCTA AGGAATTATA
Pkm8      GTACTTGTTA AGCCTTATAA GAAAAGAGT- TATTAACCTA AGGAATTATA
Pcy9      GTACTCGTTA AGCCTTTTAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pvi10     GTACTTGTTA AGCCTTTTAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pga11     GTATTTGTTA AGCCTTATAA GAAAAAAGT- TATTAATTTA AGGAATTATA

```

```

ACAAAGAAGT AACACGTAAT AA--ATTAT TTTATTT--- -AGTGTGAT
ACAAAGAAGT AACACGTAAT AA--ATTATA TTTATTA--- -AGTGTGAT
ACAAAGAAGT AACACATAAT AAA-TTTCGA -ATATTT--- -AGTGTGAT
ACAAAGAAGT AACACGTAAT AA--ATTAT TTTATTT--- -AGTGTGAT
ACAAAGAAGT AACACATAAT AT--ATTAC TATATTT--- -AGTGTGAT
ACAAAGAAGC AACACATAAT AAAGCTGCGT CTTATTT--- -AGTGTGAT
ACAAAGAAGT AACACGTGAA ATGGATTAACTCCATTTTT TAGTGTGAT
ACAAAGAAGT AACACGTAAT --GGATTCT- TCCATTTTT-- TAGTGTGAT
ACAAAGAAGT AACACGTAAT --GGATCCG- TCCATTTTT-- TAGTGTGAT
ACAAAGAAGC GACACGTAAT --GGATCCG- TCCATTTTT-- TAGTGTGAT
ACAAAGAAGC AACACATAAT AAAACTTTGT TTTATTT--- -AGTGTGAT

```

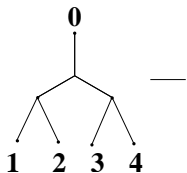


## Distances between Trees

- ▶ Nearest Neighbor Interchange (NNI).

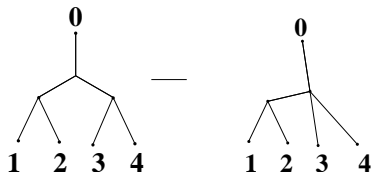
# Distances between Trees

- ▶ Nearest Neighbor Interchange (NNI). *Rotation Moves*



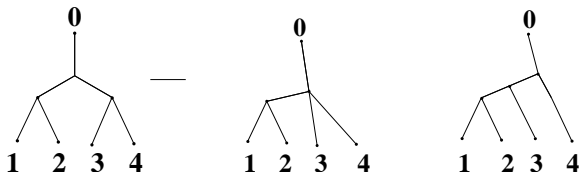
# Distances between Trees

- ▶ Nearest Neighbor Interchange (NNI). *Rotation Moves*



# Distances between Trees

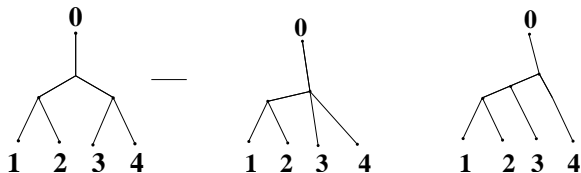
- ▶ Nearest Neighbor Interchange (NNI). *Rotation Moves*





# Distances between Trees

- ▶ Nearest Neighbor Interchange (NNI). *Rotation Moves*

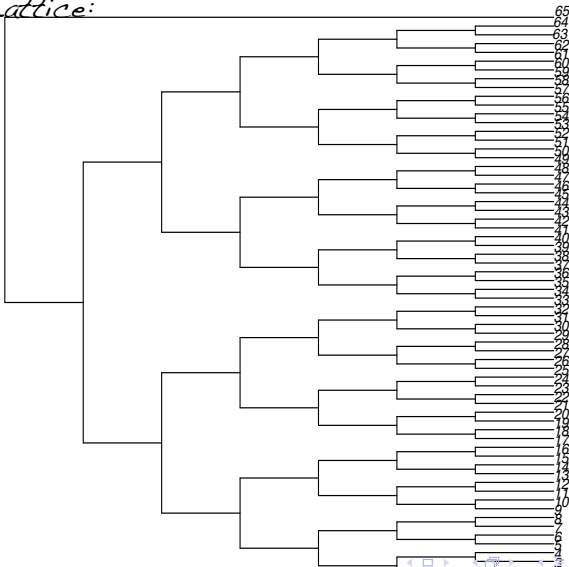


- ▶ Fill-in of NNI moves: Billera, Holmes, Vogtmann (BHM).

The boundaries between regions represent an area of uncertainty about the exact branching order. In biological terminology this is called an 'unresolved' tree.

# Empirical Evidence on Mixing on Bethe Lattice

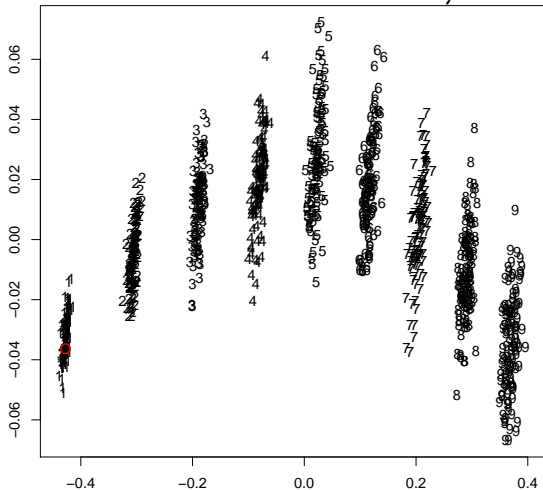
E. Mossel noticed that one of the extreme points of tree space with regards to predicting the root was the Bethe Lattice:



# Seeing the Mutation Rate Gradient

We generated 9 sets of trees with mutation rates set from  $\alpha = 0.01$  to  $\alpha = 0.09$  and we generated the data according to the Bethe lattice tree.

Here are the results in the first plane of the MDS:



## Nontechnical description of Multi-table methods

**Variance, Inertia, Co-Inertia** The study of variability of one continuous variable is done through the use of the variance. We generalize it in several directions through the idea of inertia.

As in physics, we define inertia as a weighted sum of distances of weighted points.

This enables us to use abundance data in a contingency table and compute its inertia which in this case will be the weighted sum of the squares of distances between observed and expected frequencies, such as is used in computing the chisquare statistic.

Another generalization of variance-inertia is the useful Phylogenetic diversity index. (computing the sum of the squares of distances between a subset of taxa through the tree).

We also have such generalizations that cover variability of points on a graph taken from standard spatial statistics.

## Co-Inertia

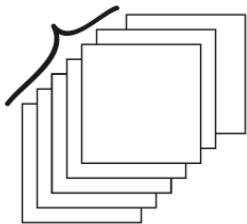
When studying two variables measured at the same locations, for instance PH and humidity the standard quantification of covariation is the covariance. A simple generalization to this when the variability is more complicated to measure as above is done through Co-Inertia analysis (CIA).

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.*

## Multiple table methods



In PCA we compute the variance-covariance matrix, in multiple table methods we can take a cube of tables and compute the RV coefficient of their characterizing operators.

We then diagonalize this and find the best weighted 'ensemble'.

This is called the 'compromise' and all the individual tables can be projected onto it.

## Data Matrix: Geometrical Approach

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

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



# Euclidean Spaces

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

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

$$x^t Q y = \langle x, y \rangle_Q \quad x, y \in \mathbb{R}^p$$

$$x^t D y = \langle x, y \rangle_D \quad x, y \in \mathbb{R}^n.$$

This simple type of inner product has been generalized to kernels in Elizabeth Purdom's thesis (2008).

# An Algebraic Approach

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



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

# An Algebraic Approach

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

Duality diagram

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

# Notes

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

$$(x, y) \in \mathbb{R}^p \times \mathbb{R}^p \mapsto x^t Q y = \langle x, y \rangle_Q \in \mathbb{R}$$

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

$$\|x\|_Q^2 = \langle x, x \rangle_Q = \sum_{j=1}^p q_j (x^j)^2 \quad \|x\|_D^2 = \langle x, x \rangle_D = \sum_{j=1}^p p_j (x_i)^2$$

(2) We say an operator  $O$  is  $B$ -symmetric if

$\langle x, O y \rangle_B = \langle O x, y \rangle_B$ , or equivalently  $BO = O^t B$ .

The duality diagram is equivalent to  $(X, Q, D)$  such that  $X$  is  $n \times p$ .

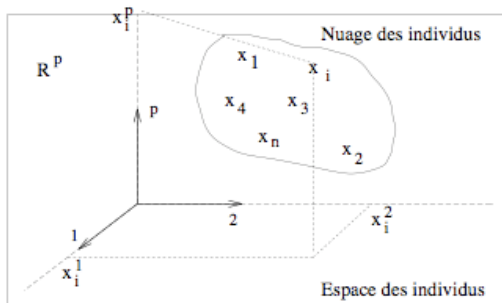
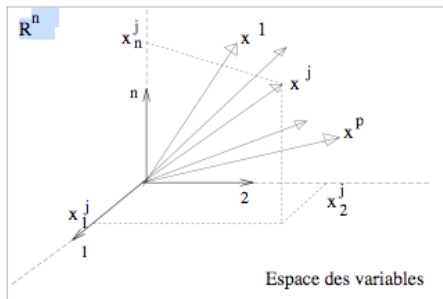
Escoufier (1977) defined as  $X Q X^t D = W D$  and  $X^t D X Q = V Q$  as the characteristic operators of the diagram.

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

## Transposable Data

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

# Two Dual Geometries



## Properties of the Diagram

Rank of the diagram:  $X, X^t, \sqrt{Q}$  and  $WD$  all have the same rank.

For  $Q$  and  $D$  symmetric matrices,  $\sqrt{Q}$  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:  $\sqrt{Q}$  is  $Q$  symmetric, thus we can find  $Z$  such that

$$\sqrt{Q}Z = Z\Lambda, Z^t Q Z = I_p, \text{ where } \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p). \quad (i)$$



# Features

1. *Inertia* :  $\text{Trace}(VQ) = \text{Trace}(WD)$   
(inertia in the sense of Huyghens inertia formula for instance). Huygens, C. (1657),

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

*Inertia* with regards to a point  $a$  of a cloud of  $p_i$ -weighted points.

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

If the variables are standardized ( $Q$  is the diagonal matrix of inverse variances), then the inertia is the number of variables  $p$ .

For correspondence analysis the inertia is the Chi-squared statistic.

## 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.

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

$\text{cov}(O_1, O_2) = \text{Tr}(O_1 O_2) = \langle O_1, O_2 \rangle$  or a vector correlation (Escoufier, 1977)

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

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: Special case

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

$$RV(X_{n \times p}, Y_{n \times g}) = \frac{\text{Tr}(XQX^t \mathcal{D} Y Y^t \mathcal{D})}{\sqrt{\text{Tr}(XQX^t \mathcal{D})^2 \text{Tr}(Y Y^t \mathcal{D})^2}}$$

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

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

Of course, classical PCA has  $D = \frac{1}{n}I$ ,  $Q = I$ , 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  $ZMZ^tD$ .

## One Diagram to replace Two Diagrams

Canonical correlation analysis was introduced by Hotelling to find the common structure in two sets of variables  $X_1$  and  $X_2$  measured on the same observations. This is equivalent to merging the two matrices columnwise to form a large matrix with  $n$  rows and  $p_1 + p_2$  columns and taking as the weighting of the variables the matrix defined by the two diagonal blocks  $(X_1^t D X_1)^{-1}$  and  $(X_2^t D X_2)^{-1}$

$$Q = \left( \begin{array}{c|c} (X_1^t D X_1)^{-1} & 0 \\ \hline 0 & (X_2^t D X_2)^{-1} \end{array} \right)$$

$$\begin{array}{ccc}
 \mathbb{R}^{p_1^*} & \xrightarrow{x_1} & \mathbb{R}^n \\
 \uparrow I_{p_1} & & \downarrow V_1 \\
 \mathbb{R}^{p_1} & \xleftarrow{x_1^c} & \mathbb{R}^{n^*} \\
 & & \downarrow D \\
 & & \uparrow \omega_1
 \end{array}
 \qquad
 \begin{array}{ccc}
 \mathbb{R}^{p_2^*} & \xrightarrow{x_2} & \mathbb{R}^n \\
 \uparrow I_{p_2} & & \downarrow V_2 \\
 \mathbb{R}^{p_2} & \xleftarrow{x_2^c} & \mathbb{R}^{n^*} \\
 & & \downarrow D \\
 & & \uparrow \omega_2
 \end{array}$$
  

$$\begin{array}{ccc}
 \mathbb{R}^{p_1+p_2^*} & \xrightarrow{[x_1; x_2]} & \mathbb{R}^n \\
 \uparrow Q & & \downarrow V \\
 \mathbb{R}^{p_1+p_2} & \xleftarrow{[x_1; x_2]^c} & \mathbb{R}^{n^*} \\
 & & \downarrow D \\
 & & \uparrow \omega
 \end{array}$$

This analysis gives the same eigenvectors as the analysis of the triple

$(x_2^c D x_1, (x_1^c D x_1)^{-1}, (x_2^c D x_2)^{-1})$ , also known as the canonical correlation analysis of  $x_1$  and  $x_2$ .

# PCA with regards to Instrumental Variables

CR Rao, 1964: Explain one matrix by another (one matrix is a response, the other explanatory). It is the extension of PCA and regression. If  $Z$  is the explanatory table and  $X$  is the response, we take the projector:

$$P_Z = Z(Z'DZ)^{-1}Z'D, \quad \hat{X} = P_Z X \text{ are the predicted values}$$

Take the triplet  $(\hat{X}, Q, D)$  and do the PCA.

See [6] for more details.

# Integrating Spatial Information into the triplet

If we make  $Z$  the explanatory table contain the spatial information, we are integrating the spatial information into the multivariate analysis.

Another solution explained in Dray and Jombart's paper is to study the coinertia of  $X$  and  $WX$ , the spatially lagged version of  $X$ .



# Spatial Multivariate Output

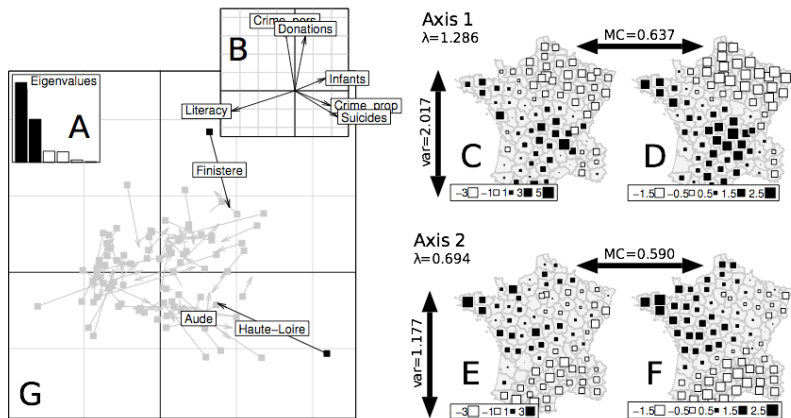
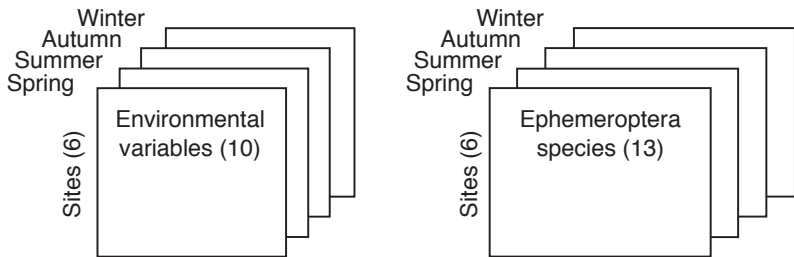


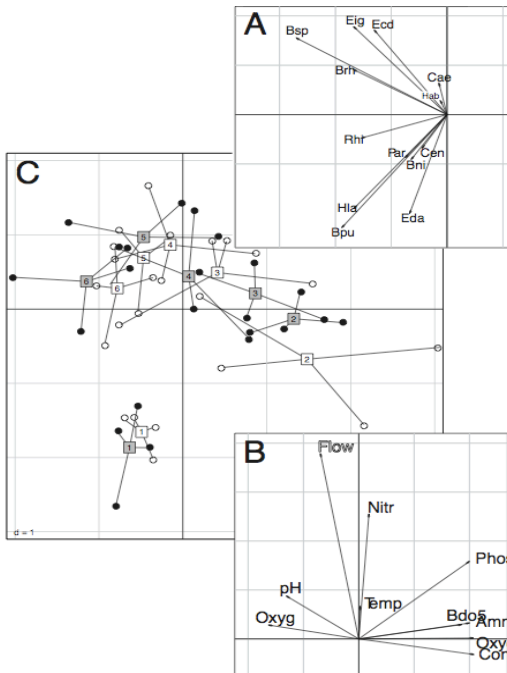
FIG 8. MULTISPATI of Guerry's data. (A) Barplot of eigenvalues. (B) Coefficients of variables. Mapping of scores of plots on the first (C) and second (E) axis and of lagged scores (averages of neighbors weighted by the spatial connection matrix) for the first (D) and second (F) axis. Representation of scores and lagged scores (G) of plots (for each département, the arrow links the score to the lagged score). Only the départements discussed in the text are indicated by their labels.

Jean Thioulouse uses the generalized notion of co-inertia to analyze these complex data:



An example data set consists of two data cubes. The first one contains 10 environmental variables that have been measured four times (in Spring, Summer, Autumn and Winter) along six sampling sites. The second one contains the numbers of Ephemeroptera belonging to 13 species, collected in the same conditions.

# Complex Output



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


Thanks to the R community, in particular Chessel, Jombart, Dray, Thioulouse (ade4 group in Lyon) and Emmanuel Paradis.


## Collaborators:


Persi Diaconis, Sharad Goel, John Chakarian, Adam Guetz, Adam Kapelner, Elisabeth Purdom, Omar delaCruz, Nelson Ray, Yves Escoufier.


## Data:

Alfred Spormann, Peter Lee, Francesca Setiadi.  
Funding from NIH/ NIGMS RO1 and NSF-DMS.

 L. Billera, S. Holmes, and K. Vogtmann.  
The geometry of tree space.  
Adv. Appl. Maths, 771--801, 2001.

 J. Chakerian and S. Holmes.  
Computational tools for evaluating phylogenetic and hierarchical clustering trees  
computational tools for evaluating phylogenetic and hierarchical clustering trees.  
Technical Report 1006.1015, arXiv, 2010.

 P. Diaconis, S. Goel, and S. Holmes.  
Horseshoes in multidimensional scaling and kernel methods.  
Annals of Applied Statistics, 2007.

 S. Dray and T. Jombart.  
Revisiting guerry's data: Introducing spatial constraints in multivariate analysis.  
Annals of Applied Statistics, 2010.



Y. Escoufier.

Operators related to a data matrix.

In J.R. Barra and coll., editors, *Recent developments in Statistics.*, pages 125--131. North Holland,, 1977.



Susan Holmes.

Multivariate analysis: The french way.

Festschrift for David Freedman, IMS, 2006.



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

Multivariate Analysis.

Academic Press, NY., 1979.



E. Mossel.

Phase transitions in phylogeny.

Trans. Amer. Math. Soc., 356(6):2379--2404  
(electronic), 2004.



C. R. Rao.

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

*Sankhya A*, 26:329--359., 1964.



*J. Thioulouse.*

*Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods.*

*Annals of Applied Statistics*, 2010.

*to appear.*