Multitables: The Power of Inertia

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

Bio-X and Statistics, Stanford University

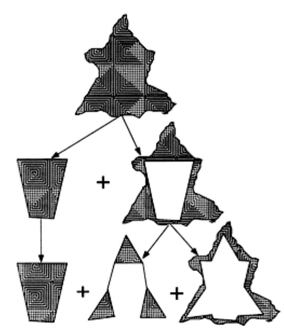
SISMID14-Lecture 12, 2011

"It is not inertia alone that is responsible for human relationships repeating themselves from case to case, indescribably monotonous and unrenewed: it is shyness before any sort of new, unforeseeable experience with which one does not think oneself able to cope.
But only someone who is ready for everything, who excludes nothing, not even the most enigmatical will live the relation to another as something alive."
Rainer Maria Rilke (Letters to a Young Poet)

Heterogeneity

- Status : response/ explanatory.
- Continuous
- Binary, categorical
- Graphs/ Trees
- Maps/ Spatial Information

Iterative Structuration (Tukey, 1977)



A distance-> projection

Multi-table methods

Inertia, Co-Inertia
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.

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

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.

Distances and Inertia

Variance is at the heart of ANOVA: decomposition of variability into parts that can be explained by a factor and a residual part.

$$\text{Test statistic:} \frac{\text{average ss explained by factor}}{\text{average of residual ss}}$$

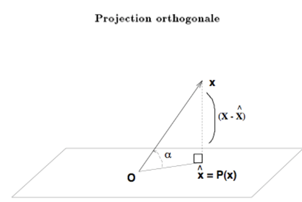
Features

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 $Q = \mathcal{I}_p$, $D = \frac{1}{n}\mathcal{I}_n$, and the variables are centered, the inertia is the sum of the variances of all the variables.
If the variables are standardized (Q is the diagonal matrix of inverse variances), then the inertia is the number of variables p.
For correspondence analysis the inertia is the Chi-squared statistic.

Quality of Representations



- The cosine again, $\cos(x, y) = \frac{<x,y>}{||x||||y||}$
- $$\cos^2\alpha = \frac{||\hat{x}||^2}{||x||^2}$$
  tells us how well x is represented by its projection.

Inertia and Contributions

- $$In(X) = ||X||^2 = \sum_i p_i ||x_{i.}||^2 = \sum_j q_j ||x^{\cdot j}||^2 = \sum_{\ell=1}^{p} \lambda_\ell$$
- Contribution of an observation to the total inertia: $\frac{p_i ||x_{i.}||^2}{||X||^2}$
- Contribution of a variable to the total inertia: $\frac{q_j ||x^{\cdot j}||^2}{||X||^2}$

Inertia and Contributions

- Contribution of the kth axis to variable j: $\frac{\lambda_k v_{kj}^2}{||x^j||_D^2}$
- Contribution of variable j to the kth axis $q_j v_{kj}^2$.
- Contribution of the kth axis to observation i: $\frac{\lambda_k u_{ik}^2}{||x_i||_Q^2}$
- Contribution of observation i to the kth axis $p_i u_{ik}^2$.

Quality of approximations

- $$\frac{||X^{[k]}||^2}{||X||^2} = \frac{\sum_{\ell=1}^{k} \lambda_\ell}{\sum_{\ell=1}^{p} \lambda_\ell}$$

This is like a cosine, we can compare two operators with this.

Part I

Two Table Analyses

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:

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

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

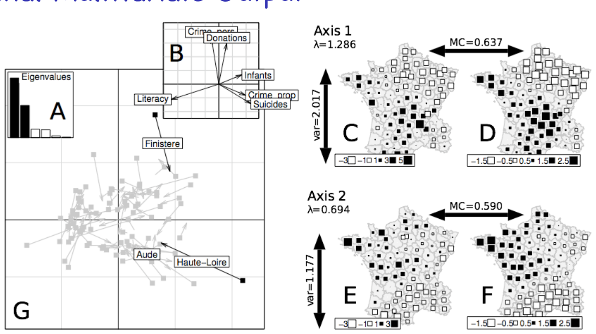
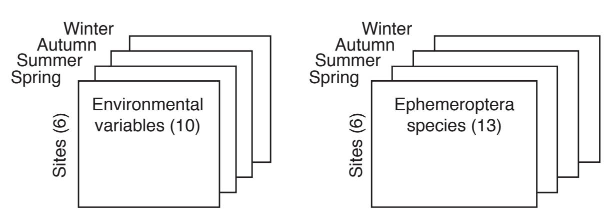


FIG 8. *MULTISPATI of Guerry's data. (A) Barplot of eigenvalues. (B) Coefficients of variables. Mapping of scores of plots on the first (C) and second (E) axis and of lagged scores (averages of neighbors weighted by the spatial connection matrix) for the first (D) and second (F) axis. Representation 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.

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" $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ defining a multivariate data analysis.
Q and D define geometries or inner products in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively, through

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

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.

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r \geq 0 \geq \cdots \geq 0.$$

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

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

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
$covV(O_1, O_2) = Tr(O_1 O_2) = < O_1, O_2 >$ or a vector correlation (Escoufier, 1977)

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

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 $RV = \rho^2$.

PCA: Special case

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

$$RV\left(XQX^t D, YY^t D\right) = \frac{Tr\left(XQX^t DYY^t D\right)}{\sqrt{Tr\left(XQX^t D\right)^2 Tr\left(YY^t D\right)^2}}.$$

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

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

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 $XQX^t D$ and $ZMZ^t D$.

PCA IV, multiresponse regression (RDA):special case

They can also be seen as the PCA with regards to instrumental variables of (Y, Δ_Y^{-1}, D) with regards to (X, M, D).

Canonical Correlation Analysis

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. Maximizing Covariance

$$cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \qquad \begin{pmatrix} v(x) & cov(x,y) \\ cov(y,x) & v(y) \end{pmatrix}$$

$$cor(x,y) = \frac{cov(x,y)}{\sqrt{v(x)}\sqrt{v(y)}} = \frac{<x,y>}{\sqrt{<x,x>}\sqrt{<y,y>}}$$

Two tables: different constraints

Maximization of $cov^2(\mathbf{XQu}_i, \mathbf{YRv}_i) = corr^2(\mathbf{XQu}_i, \mathbf{YRv}_i) \times var(\mathbf{XQu}_i) \times var(\mathbf{YRv}_i)$		
with the constraints		
$var(\mathbf{XQu}_i) = 1$, $var(\mathbf{XRv}_i) = 1$	$var(\mathbf{XQu}_i) = 1$	No constraint
Canonical correlation analysis	Analysis with respect to instrumental variables (CCA, RDA)	Co-inertia analysis
Y **X**	**Y** **X**	**Y** **X**

Canonical/Constrained Correspondence Analysis

"Constraints" are explanatory variables.
Interpretability of the variability in terms of "environmental variables".
Similar to a regression but for the inertia instead of the variance.
uses the formula notation as linear models do in R

```
require(vegan)
plot(tmp <- cca(varespec ~ Al + P + K), dis=c("bp","lc
```

Part II

Confirmatory Analysis: Nonparametric

Nonparametric Tests

- CCA tests on a factor.
- Mantel's Test between distance matrices
- Multiple testing correction.
- Bootstrap tests.

Everything is done by shuffling labels

```
 Example:
Is there a shedding effect in Yana's data?
> shed = scan("shed.txt")
> shedf = as.factor(shed)[-(70:71)]
> resca.shed = cca(t(pib.nz) ~ shedf)
> anova(resca.shed)
Permutation test for cca under reduced model

Model: cca(formula = t(pib.nz) ~ shedf)
         Df  Chisq      F N.Perm Pr(>F)
Model     6 0.3825 2.0627    199  0.005 **
Residual 87 2.6886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

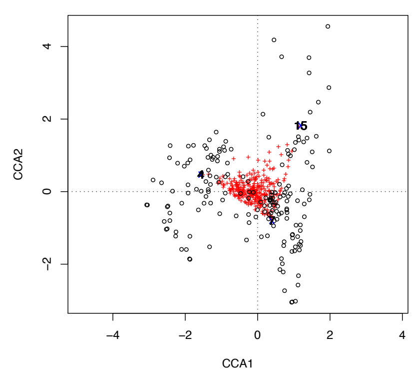
```
Example: Is there a subject effect in Katie's data?
 subjcc = vegan::cca(tnorepnz ~ subject)
 anova(subjcc)

Permutation test for cca under reduced model

Model: cca(formula = tnorepnz ~ subject)
          Df   Chisq       F N.Perm Pr(>F)
Model      7  1.2177 10.7999    199  0.005 **
Residual 472  7.6027
---
Signif. codes:  0  ***  0.001 **  0.01  * 0.05 . 0.1
```

```
> cca.cage = vegan::cca(t(tcmall) ~ cagef)
> plot(cca.cage)
> text(cca.cage, choices = c(1, 2), label = cagef, dis
> anova(cca.cage)
Permutation test for cca under reduced model

Model: cca(formula = t(tcmall) ~ cagef)
          Df  Chisq      F N.Perm Pr(>F)
Model      2 0.3722 7.6501    199  0.005 **
Residual 178 4.3305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```




```
> plot(cca.cage, scaling = 1)
> text(cca.cage, scaling = 1, choices = c(1, 2), displ
+     cex = 0.6)
> title("Species, Cage effect, 1,2")
```