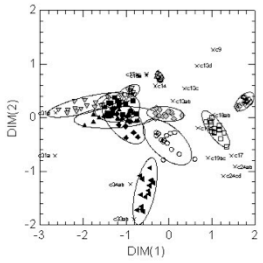
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION

3. PRINCIPAL COMPONENT ANALYSIS

Prof. Eugenio Parente
Scuola di Scienze Agrarie - Università della Basilicata

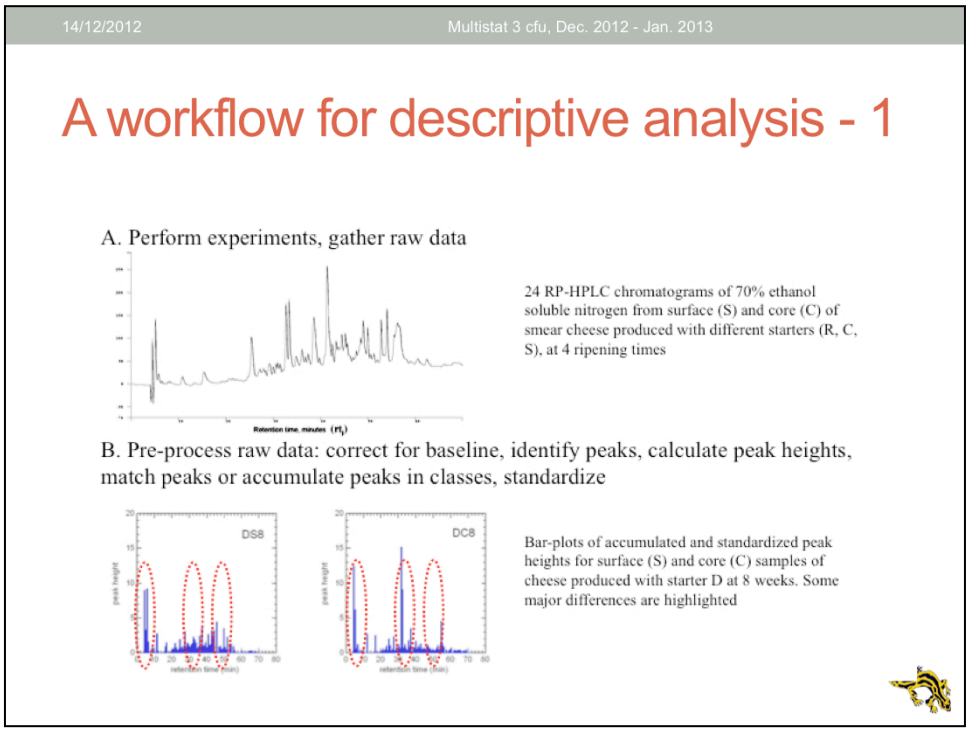


Outline

- **Descriptive multivariate statistical analysis workflow**
- **Principal component analysis (PCA)**
  - objectives of PCA
  - calculation of principal components
  - component loadings and scores
  - examples
  - how to interpret PCA in published papers (practical)
  - how to extract information from the output of PCA (practical)

The objectives of descriptive multivariate statistical analysis

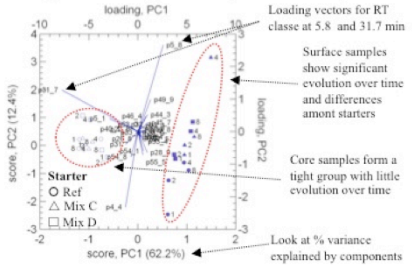
- Explore the data set
- Find out if “natural” groups of observation exist
- Find relationships between variables (if any)
- Find out descriptive relationships between values of variables and groups of observations
- Document the analysis and produce interpretable graphs



The figure is from the Encyclopedia of Dairy Science Chapter 9

A workflow for descriptive analysis - 2

C. Carry out Principal Component Analysis for data reduction.

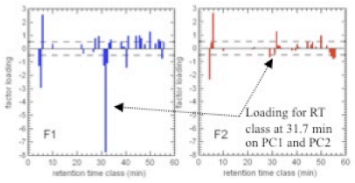


Loading vectors for RT classe at 5.8 and 31.7 min

Surface samples show significant evolution over time and differences amont starters

Core samples form a tight group with little evolution over time

Look at % variance explained by components



Loading for RT class at 31.7 min on PC1 and PC2

Examine bar-plots of loadings to evaluate relationships between PCs and original variables

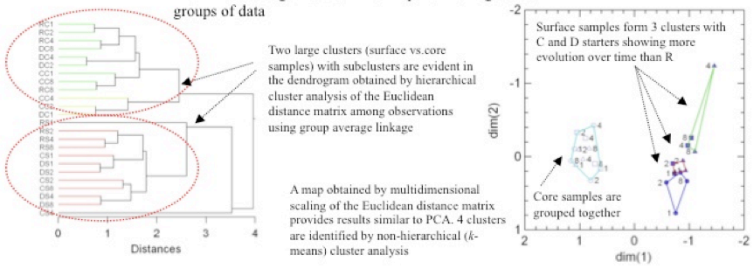
Examine bi-plots to evaluate relationships between PCs and original variables to identify variables which contribute to separation of samples, relationships among variables, relationships among samples

The figure is from the Encyclopedia of Dairy Science

A workflow for descriptive analysis - 3

D. Carry out Multidimensional scaling (MDS) and/or cluster analysis

Examine dendrograms and MDS maps for homogeneous groups of data



The figure is from the Encyclopedia of Dairy Science

Objectives of PCA

"To describe the variation of a set of multivariate data in terms of a set of uncorrelated variables (principal components) each of which is a particular linear correlation of the original variables. The new variables are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the original variation in the data. The second component is chosen to account as much as possible of the remaining variance subject to being uncorrelated with the first component – and so on" (Everitt and Dunn 2001)

Objectives of PCA

- to obtain a **reduction in dimensionality**, by summarizing the variance in a few derived variables (components), in order to aid graphical examination of the results or to derive performances indexes
- to obtain **derived variables** to be used in regression analysis or in analysis of variance (when there are too many variables compared to observations and when the variables are highly correlated)

Eigenvalues and eigenvectors

The eigenvalues of a matrix **A** with dimension *pxp* are the solutions of the equation

$$\left|\mathbf{A} - \lambda\mathbf{I}\right| = 0$$

a. the product of the eigenvalues of **A** is equal to the determinant of **A**
b. the sum of the eigenvalues of **A** is equal to trace(**A**)

Therefore the sum of the eigenvalues of the S matrix are the sum of variances of the original matrix A

Eigenvalues and eigenvectors

Each of the eigenvalues λ_i is associated to a vector $\mathbf{x}_i$ whose elements satisfy the system of equations:

$$\left|\mathbf{A} - \lambda\mathbf{I}\right|\mathbf{x}_i = 0$$

the eigenvectors $\mathbf{x}_i$ and $\mathbf{x}_j$ associated to the eigenvalues l_i and l_j of a simmetrical matrix are orthogonal

PCs extracted from the covariance matrix

The first principal component (a **derived variable**) of the first observation is defined as:
$y_1=a_{11}x_1+a_{12}x_2+...a_{1p}x_p$ and, in matrix notation, $\mathbf{y_1}=\mathbf{a_1}'\mathbf{x}$ with the constraint $\mathbf{a_1}'\mathbf{a_1}=1$
The second component is $\mathbf{y_2}=\mathbf{a_2}'\mathbf{x}$ with the constraint $\mathbf{a_2}'\mathbf{a_2}=1$ and $\mathbf{a_2}'\mathbf{a_1}=0$ (i.e. $\mathbf{a_2}$ and $\mathbf{a_1}$ must be orthogonal)
Further components are defined similarly.
If $\mathbf{a_1}$ must be chosen in order to maximize the variance of y_1 and since $Var(y_1)=Var(\mathbf{a_1}'\mathbf{x})=\mathbf{a_1}'\mathbf{S}\mathbf{a_1}$ then $\mathbf{a_1}$ is the eigenvector associated to the largest eigenvalue of **S** and the variance explained by the component is given by the corresponding eigenvalue.

Plese note that here a is a row vector; increasing arbitrarily a1i would increase the variance of yi, therefore the constraint; thevector has unit length

Components in matrix notation

$$\mathbf{Y} = \mathbf{XA}$$

- **Y** is the matrix of values for the components (there are *p* component vectors); it has *n* (number of observations) elements
- **A** is the matrix of component coefficients (a square matrix with *pxp* elements*)*
- **X** is the transposed data matrix (*nxp*)
- Remember matrix multiplication:
  - **X** can be postmultiplied by **A** because it has n rows and p columns
  - The resulting matrix has the same numer of rows as **X** and the same number of columns as **A**

Principal components extracted from the covariance matrix

$$\sum_{i=1}^{p} \lambda_i = trace(\mathbf{S})$$

the proportion P_j of the variance explained by the jth component and the proportion of the variance accounted for by the first p* principal components (p*<p) are:

$$P_j = \frac{\lambda_j}{trace(\mathbf{S})} \qquad P^* = \frac{\sum_{i=1}^{p} \lambda_i}{trace(\mathbf{S})}$$

Principal components from the covariance matrix (**S**) or from the correlation matrix (**R**)?

When using **S**:

- the proportion of the variance explained by the first components is high
- scales of the variables are important and changing the scale results in a different set of components
- the variables with the highest variance will dominate the first components

Using **R**:

- is equivalent to using **S** after standardization of the variables to 0 mean and unit variance
- is useful when standard deviations are not thought to be theoretically significant
- "involves an arbitrary decision in making variables equally important"

if scales are important, variables should at least be on a common scale

Component loadings

Component loadings are the covariances of the original variables with the components (if components were extracted from **S**; if they were extracted from **R** they are the correlations).

Signs of the components are arbitrary since $\mathbf{Ax}=\lambda\mathbf{x}$ and $-\mathbf{Ax}=-\lambda\mathbf{x}$ are equivalent.

To give component loadings, elements of the eigenvectors are rescaled so that their sum of squares is equal to the corresponding eigenvalue rather than to 1. In this way the coefficients of the more important components are scaled up compared to those of the less important components

Component scores

$$y_{i1} = \mathbf{a}_1'(\mathbf{x}_i - \bar{\mathbf{x}})$$

$$\vdots$$

$$y_{ip} = \mathbf{a}_p'(\mathbf{x}_i - \bar{\mathbf{x}})$$

yip are the scores for individual 1, with p components; x sub i are the original vector of variable values for individual i, x hat is the vector of the means of the original variables. The transformed variables have 0 mean and variances corresponding to the eigenvalues

How many components?

1. use a specified (arbitrary) proportion of the variance to choose the components to retain. For example retain the first j components that explain 70% (or 90%) of the original variance (use smaller values if p increases)
2. exclude components whose eigenvalues are less than average; using **R** this means retaining only components with eigenvalues>1 (the average variance is 1); lower values (0.7-0.8 may also be appropriate)
3. look for elbows in a graph plotting λ_i against i (scree plot); the number of components to select corresponds to the value of i located in the elbow of the curve; an alternative is using a modified scree plot with $\log(\lambda_i)$

Examples of PCA

- open file technolab.syo for a PCA on technological properties of LAB (correlation matrix)
- open file RPHPLC.syo of a PCA on RP-HPLC of 70% ethanol soluble N in smear cheese (covariance matrix)

Look at:

- the output of the analysis
- the scree plot
- the loading plot (note that when R is used interpretation of factors should be based on eigenvectors)
- the bi-plot
- PCA after rotation

interpretation of relationships between original variables and PCs; using rotations to make factors more interpretable in that they are more associated with one of the original variables; after rotation with p variables and m loadings: each component should have at least m near-zero loadings; few components should have high loadings on the same variable.

PCA of techological properties of LAB

ScienceDirect

INTERNATIONAL DAIRY JOURNAL

Acid production, proteolysis, autolytic and inhibitory properties of lactic acid bacteria isolated from pasta filata cheeses: A multivariate screening study

Paolo Pinnoa, Teresa Zottaa, Annamaria Ricciardia, Paul L.H. McSweeneyb, Eugenio Parenteb,*

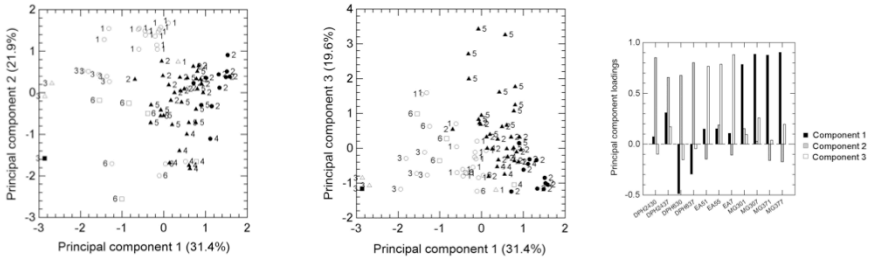
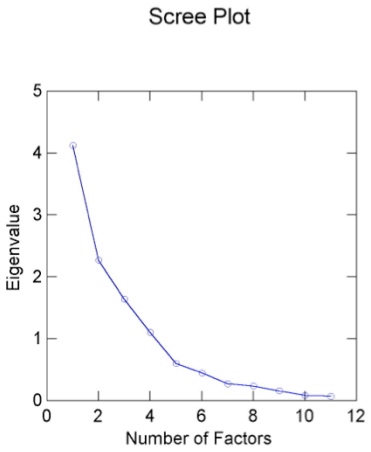
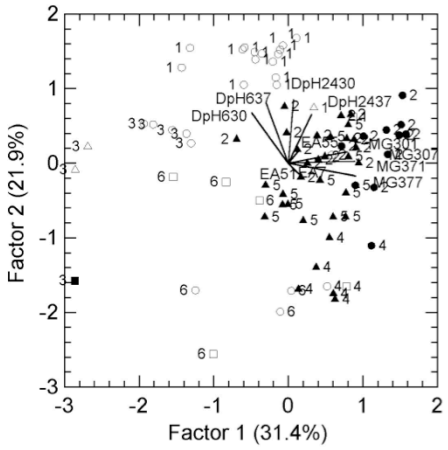


Fig. 1. Principal component analysis score for some technological properties of lactic acid bacteria isolated from pasta filata cheeses (Table 1): (a) Score for principal components 1 and 2 and (b) score for principal components 1 and 3. (○) *Lc. lactis*, (△) *Str. thermophilus*, (□) enterococci, (●) *Lb. helveticus*, (▲) NSLAB, and (■) *Lb. fermentum*. Numbers show cluster memberships (see the Section 3.6). Non-starter lactic acid bacteria (NSLAB) include *Lb. casei*, *Lb. paracasei*, *Lb. rhamnosus* and *Lb. curvatus*. Bars in (c) represent principal component loadings for decrease in pH in skim milk after 6 h (DPH630 and DPH637) and 24 h (DPH2430 and DPH2437) of incubation at 30 and 37 °C; free aminoacids, as mg Leu L^{-1}, at 30 °C after 1 day (MG301 and MG371) and 7 days (MG307 and MG3737) of incubation at 30 and 37 °C, respectively; extent of autolysis in: 88.5 mmol L^{-1} Na-lactate, 0.7 mol L^{-1} NaCl, pH 5.1 (EA51), 0.2 mol L^{-1} NaCl, pH 5.5 (EA55), and 50 mmol L^{-1} Na-phosphate, pH 7.0 (EA7).

Scree plot



PCA of technological properties of LAB



PCA of RP-HPLC data of surface ripened cheese

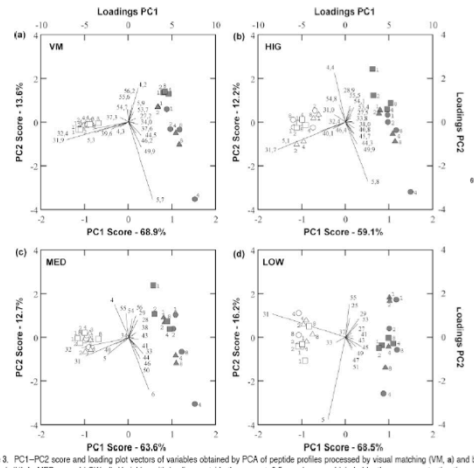


Figure 3. PC1–PC2 score and loading plot vectors of variables obtained by PCA of peptide profiles processed by visual matching (VM, **a**) and by fuzzy approach (HI, **b**; MED, **c**; and LOW, **d**). Variables with loadings outside the range ± 0.5 are shown and labeled by the average retention time of peaks (**a**) or by retention time of class center (**b–d**). Score symbols refer to reference smear R (■) and to defined-strain smear mix D (▲) or mix C (●); open and solid symbols are for core and surface samples, respectively. Score numbers refer to ripening time (weeks).

6904 J. Agric. Food Chem. 2004, 52, 6904–6911

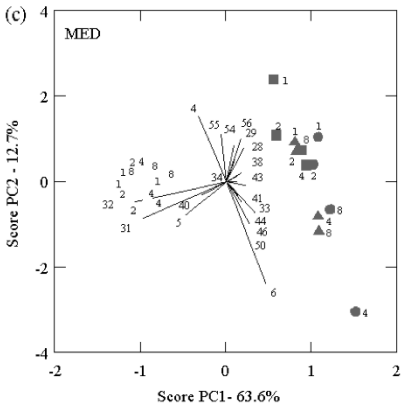
JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

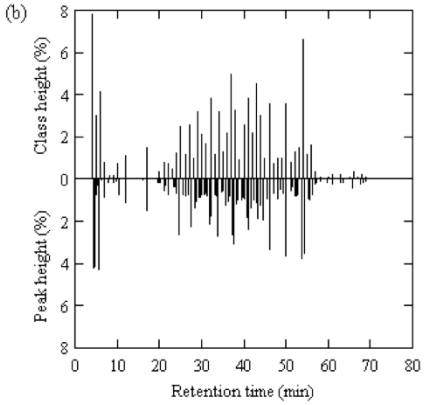
Processing of Chromatographic Data for Chemometric Analysis of Peptide Profiles from Cheese Extracts: A Novel Approach

Paolo Piraino,*,† Eugenio Parente,† and Paul L. H. McSweeney‡

Dipartimento di Biologia D.B.A.F., Università Basilicata, 85100 Potenza, Italy, and Department of Food and Nutritional Sciences, University College, Cork, Ireland

PCA of RP-HPLC data of surface ripened cheese





Multivariate analysis of 2-DE gels

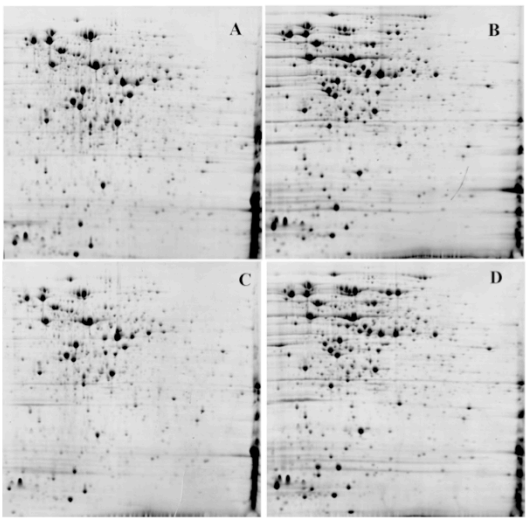
- PCA can be used as a complementary tool (i.e. in combination with univariate analysis of spot volumes) to identify spots which are significantly affected by treatments
- As an example I will show results form a proteomic analysis of whole-cell proteins of *L. plantarum* subsp.*plantarum* WCFS1 and its mutant WCFS1-2 (in which *ccpA* was inactivated by insertion); both strains were grown in batch culture at controlled pH and temperature in a complex medium in anaerobiosis (N_2 sparging) and aerobiosis (air sparging). Here we are interested in:
  - The effect of *ccpA* inactivation
  - The effect of aerobic growth

2 DE maps (three replicates per threatment available)

WCFS1 WCFS1-2

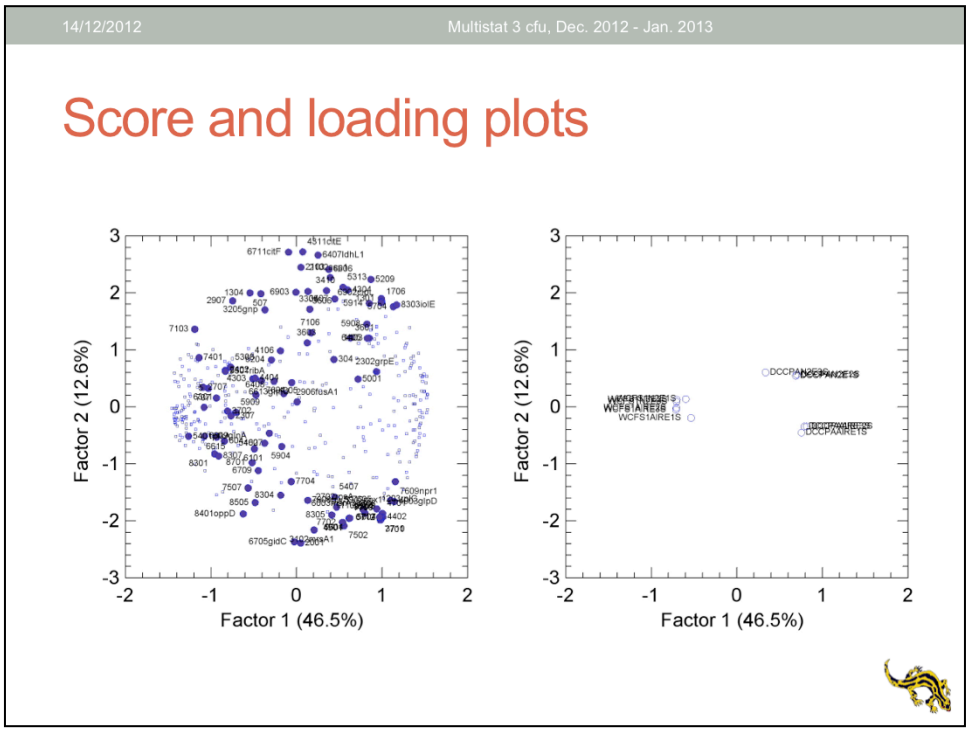
N_2

Air

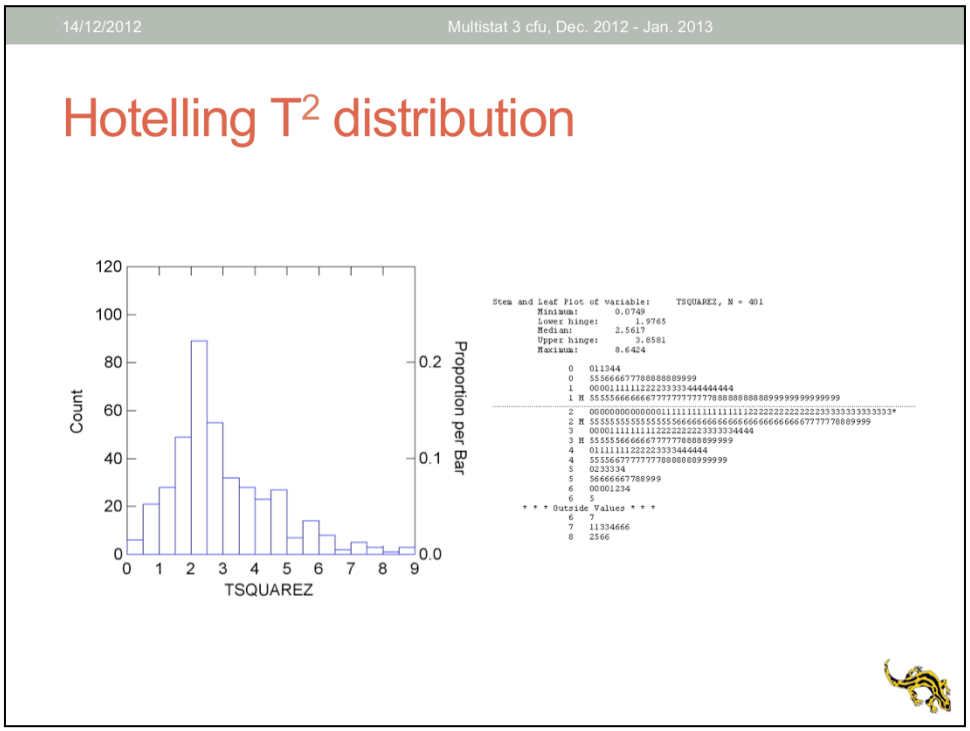


Steps in the analysis

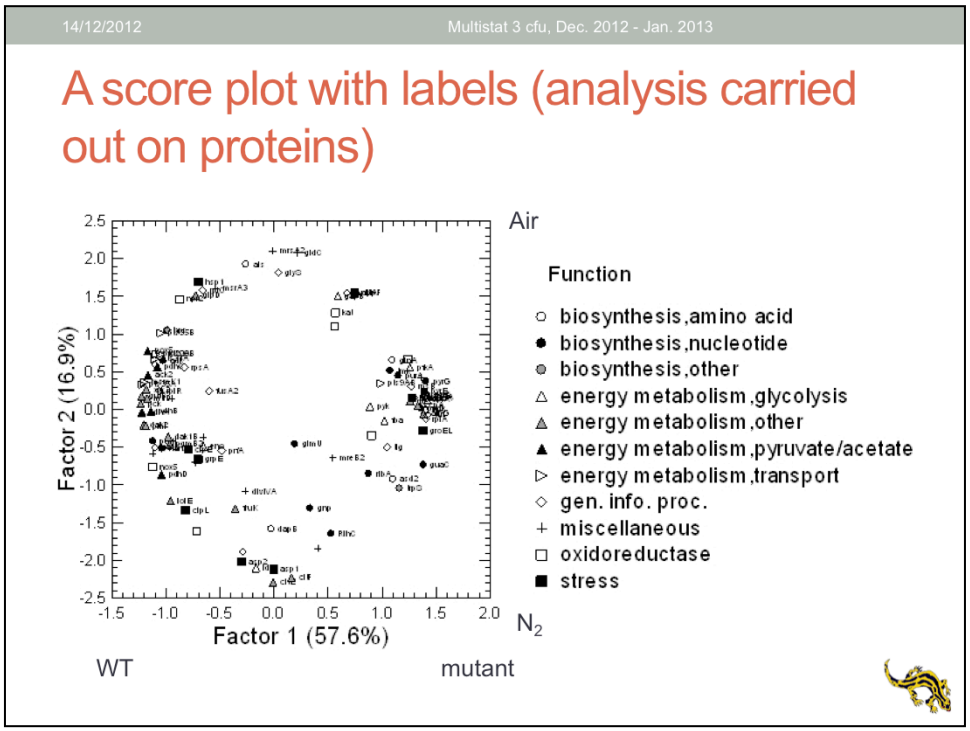
- Obtain a file with normalized volumes for protein spots
- Rearrange the file with spots/proteins as variables and gels as observations
- Run PCA
- Save scores, loadings
- Look at score and loading plots
- Look at the distribution of Hotelling's T^2 values for the identification of spots which are significantly affected by treatments



The spots with higher Hotelling T2 values



The Hotelling T2 statistics is the square of the standardized distance of each case from the centroid of the distribution; to identify observations with unusually high values you can look at the upper tail probability associated with T2. An arbitrary cutpoint can also be chosen



Proteins which are far from the centre are affected by treatments; interpretation of which treatment is affecting which protein can be made on the basis of loadings on the basis of loadings

Some rights reserved