

# Multivariate Data Reduction

## General Tools

Vectors:

Two vectors  $\mathbf{x}$  and  $\mathbf{y}$  can be added giving the sum  $\mathbf{x}+\mathbf{y}$ .

Any vector  $\mathbf{x}$  and real number  $\alpha$  can be multiplied giving  $\alpha\mathbf{x}$ .

Properties:

- 1)  $\mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$ ;
- 2)  $(\mathbf{x}+\mathbf{y})+\mathbf{z} = \mathbf{x}+(\mathbf{y}+\mathbf{z})$ ;
- 3) There exists a  $\mathbf{0}$  such that  $\mathbf{x}+\mathbf{0} = \mathbf{x}$ ;
- 4) Every vector  $\mathbf{x}$  has a negative vector  $\mathbf{y} = -\mathbf{x}$  such that  $\mathbf{x}+\mathbf{y} = \mathbf{0}$ ;
- 5)  $1\cdot\mathbf{x}=\mathbf{x}$ ;
- 6)  $(\alpha\mathbf{x}+\mathbf{y}) = \alpha\mathbf{x} + \mathbf{y}$ .

Let  $\mathbf{y} = \sum a_i \mathbf{x}_i$ .

The vector  $\mathbf{y}$  is called a linear combination of the  $\mathbf{x}$ 's. If all the  $a_i$ 's are zero then  $\mathbf{y} = \mathbf{0}$ . However it may be that  $\mathbf{y} = \mathbf{0}$  and not all of the  $a_i$ 's = zero. Then the  $\mathbf{x}$ 's are called linearly dependent. If they are not dependent then they are independent.

Dimension of a vector space;

The dimension is said to be the maximum number of linearly independent vectors.

Matrices:

$$\{a_{ij}\} = A_{p \times k}$$

$$\{b_{ij}\} = B_{k \times q}$$

$$AB_{p \times q} = \sum_t a_{it} b_{tj} .$$

$A^t = A' = \{a_{ji}\}$ . It is a  $k$  by  $p$  matrix.

If  $k=p$  the  $A$  is called a square matrix.

$$\text{Let } ij = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise} \end{cases} .$$

Then any square matrix  $I = \{ ij \}$  is called an identity matrix.

If  $A$  has full column rank (row rank) then  $A$  has an inverse denoted as  $A^{-1}$ .  $AA^{-1} = A^{-1}A = I$ .

Two vectors  $\mathbf{x}$  and  $\mathbf{y}$  are orthogonal if  $\mathbf{x} \cdot \mathbf{y} = 0$ . They are called orthonormal if  $\mathbf{x} \cdot \mathbf{x} = \mathbf{y} \cdot \mathbf{y} = 1$ .

Orthogonal matrices:

A square matrix  $O$  is called orthogonal if  $O^t O = O O^t = I$ .

Eigenvalues and eigenvectors:

Every matrix  $A$  having rank  $k$  has  $k$  orthonormal eigenvectors  $\mathbf{v}_i$  and eigenvalues  $\lambda_i$  satisfying:

$A \mathbf{v}_i = \lambda_i \mathbf{v}_i$ . In particular there exists an Orthonormal matrix  $O$  such that  $O^t A O = \Lambda = \{ \lambda_{ij} \}$ .

Determinants:

Principal Components

Factor Analysis

Canonical Correlations

Partial Least Squares

Principal Components:

Linear combinations of components or variables.

The principal components have special properties in terms of variance.

$$X \sim (0, \Sigma)$$

Let  $\mathbf{v}$  be a  $p$ -component column vector such that  $\mathbf{v}'\mathbf{v} = 1$ .

$$E(\mathbf{v}'\mathbf{X})^2 = E\mathbf{v}'\mathbf{X}\mathbf{X}'\mathbf{v} = \mathbf{v}'\Sigma\mathbf{v}.$$

To maximize this we use Lagrange multipliers.

$$L = \mathbf{v}'\Sigma\mathbf{v} - \lambda(\mathbf{v}'\mathbf{v} - 1).$$

$$\frac{\partial L}{\partial \mathbf{v}} = 2\Sigma\mathbf{v} - 2\lambda\mathbf{v} = 0.$$

Solving  $F(\mathbf{v}, \lambda) = 0$  implies  $(\Sigma - \lambda\mathbf{I})\mathbf{v} = 0$ .

Since  $\lambda = 0$  we must have  $|\lambda I - C| = 0$ .

$|\lambda I - C|$  is a polynomial of degree  $p$  in  $\lambda$ , and the  $p$  roots are

$$\lambda_1, \dots, \lambda_p.$$

The successive vectors  $v_i$  that solve  $|\lambda_i I - C| = 0$  are orthogonal and are called the principal components of  $X$ .

**Theorem:** An orthogonal transformation  $V = CX$  of a random vector  $X$  leaves invariant the generalized variance and the sum of the variance components.

$$|C' C| = |C| |C'| = 1 \cdot |C'| = 1.$$

$$\text{Trace}(C' C) = \text{Trace}(C' C) = \text{Trace}(I) = \text{Trace}(C' C) = \sum \text{Var}(X_j)$$

Variation of principal components.

Sometimes the correlation matrix is used. Usually the sample covariance or correlation matrices are used.

Proportion of standardized population variance due to k-th

principal component  $= \frac{k}{p}$ ,  $k = 1, \dots, p$ .

Geometrical interpretation:

Data can be plotted in p dimensional space

If  $(x - \bar{x})'S^{-1}(x - \bar{x}) = c^2$  defines a hyperellipsoid whose center is at  $\bar{x}$ , whose axes are given by the eigenvectors of  $S^{-1}$  or equivalently of  $S$ .

The absolute value of the i-th principal component  $|X_i|$  is the length of the projection of  $X$  on  $e_i$ .

Factor Analysis:

$$\mathbf{X}_{p \times 1} = \boldsymbol{\mu}_{p \times 1} + \mathbf{L}_{p \times m} \mathbf{F}_{m \times 1} + \boldsymbol{\epsilon}_{p \times 1}$$

$$\mu_i = E X_i.$$

$x_i$  = i-th specific factor

$F_j$  = j-th common factor

F and  $\epsilon$  are independent  $E F = 0$ ,  $\text{Cov}(F) = I$  and  $E \epsilon = 0$  and  $\text{Cov}(\epsilon) = \Sigma$  a diagonal matrix.

A covariance structure is implied.  $\text{Cov}(X) = LL' + \Sigma$ .

$$\text{Cov}(X_i, X_k) = L_{ij} L_{kj}$$

The sample covariance matrix S is an estimator of the unknown population covariance matrix  $\Sigma$ . If the off diagonal elements of S are small or those of the sample correlation matrix are small then the specific factors dominate.

## Principal Component (and Principal Factor) Method

$$[\sqrt{\lambda_1}, \dots, \sqrt{\lambda_m}] [\sqrt{\lambda_1}, \dots, \sqrt{\lambda_m}]' + \dots$$

Interpretation:

From PLS\_Toolbox Handbook

(Using handbooks notation)

$Z = X + \epsilon$ , or for calibration measurements  $Y = X + \epsilon$

Where the  $Z$ ,  $Y$  are the observed spectra,  $X$ ,  $Y$  are observed concentrations,  $X$  are the so called pure spectra and  $\epsilon$  is our error or fudge factor.

We'd like to know  $X$  from  $Z$ .

$$X = (X^t X)^{-1} X^t Y$$

$$X_{\text{est}} = (X^t X)^{-1} X^t Y$$

**DO NOT CALCULATE THIS AS THE PLS TOOLBOX INDICATES. IT COULD BE A DISASTER.**

**PCR REGRESSION IS DONE BACKWARDS**

## Canonical Correlations

We have two sets of variables:  $X$  and  $Y$ .

Goal to find sets of "loadings"  $(a_1, \dots, a_k)$   $(b_1, \dots, b_k)$  such that  $a_i'X$  and  $b_i'Y$  are highly correlated. Further we hope that successive pairs of loadings represent new knowledge.

Let:

$$U_i = a_i'X$$

$$V_i = b_i'Y.$$

Let us assume that variance matrix of  $X$  is  $\Sigma_{11}$

and of  $Y$  is  $\Sigma_{22}$ . We assume that the  $\text{cov}(X, Y)$  is  $\Sigma_{12}$ .

We seek loadings  $a_1$  and  $b_1$  such that  $a_1' \Sigma_{12} b_1 / (a_1' \Sigma_{11} a_1 b_1' \Sigma_{22} b_1)^{1/2}$  is as large as possible.

The solutions are:

$a_1 = e_1 \Sigma_{11}^{-1/2} X$  and  $b_1 = e_1 \Sigma_{22}^{-1/2} Y$ . Here  $e_1$  is the first eigenvector of  $\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{11}^{-1/2}$ . The first eigenvalue of this matrix is the correlation of  $U_1$  with  $V_1$ . Other loadings are found similarly using other eigenvectors.

Note that "Applied Multivariate Statistical Analysis" Say that if  $x_{11}$  or  $x_{22}$  are singular variables may be deleted and the optimal solutions found from the reduced set. This is not satisfactory for us.

PLS Formal Algorithm:

Usually mean center and scale the spectral data first.

Select a column of Y (chemists Y statisticians X) and call it  $u_1$ , then starting with the X data block:

$$w_1 = \frac{X' u_1}{\|X' u_1\|}.$$

$$t_1 = X w_1$$

In the Y data:

$$q_1 = \frac{u_1' t_1}{\|u_1' t_1\|}.$$

$$u_1 = Y q_1$$

Then we iterate till convergence.

Then:

$$p_1 = \frac{X' t_1}{\|t_1' t_1\|}$$

$$p_{1new} = \frac{p_{1old}}{\|p_{1old}\|} \quad .$$

$$t_{1new} = t_{1old} \|p_{1old}\|$$

$$w_{1new} = w_{1old} \|p_{1old}\|$$

We then find the regression coefficient for the inner relation:

$$b_1 = \frac{u_1' t_1}{t_1' t_1}.$$

Then we replace the X and Y data with

$$E_1 = X - t_1 p_1'$$

$$F_1 = Y - b_1 u_1 q_1',$$

and continue the process.

An easier way to see what is happening:

$$X' Y Y' X w = \lambda w$$

$$X X' Y Y' t = \lambda t$$

$$Y' X X' Y q = \lambda q$$

$$Y Y' X X' u = \lambda u$$