Aalto University

Chemometric data analysis, fundamental methods (III)

Advanced crystallization and characterization techniques
June 1-5, 2020

Francesco Corona

Chemical and Metallurgical Engineering
School of Chemical Engineering

Principal component methods

We can start by assuming that both data blocks $\mathbf{X}$ and $\mathbf{Y}$ have been previously centred

$$\mathbf{X} \leftsquigarrow \mathbf{X} - \mathbf{1}\bar{\mathbf{X}} \quad (1a)$$

$$\mathbf{Y} \leftsquigarrow \mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} \quad (1b)$$

We then discuss a general method for the analysis of multivariate data

- The principal components analysis (PCA)
- It will be extended for regression (PCR)

To appreciate PCA, we need to overview a matrix factorisation method

- The singular value decomposition (SVD)

Singular value decomposition

Consider a $(N \times K)$ matrix $\mathbf{X}$ and let $\mathtt{t} = \min\{N, K\}$ (the dimension of the matrix)

The singular value decomposition (SVD) of $\mathbf{X}$ is a factorisation of matrix $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{P}^T, \quad \text{with} \quad \begin{cases} \mathbf{U} \text{ is an orthogonal } (N \times \mathtt{t}) \text{ matrix} \\ \mathbf{P} \text{ is an orthogonal } (K \times \mathtt{t}) \text{ matrix} \\ \mathbf{D} \text{ is an diagonal } (N \times N) \text{ matrix} \end{cases} \tag{2}$$

That is,

$$\mathbf{X} = \underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1\mathtt{t}} & p_{2N} & \cdots & p_{K\mathtt{t}} \end{bmatrix}}_{\mathbf{P}^T}$$

A matrix $\mathbf{A}$ is said to be orthogonal if its columns are orthonormal vectors, $\mathbf{A}^T\mathbf{A} = \mathbf{I}$

- Two vectors are orthogonal if their inner product is zero, $\mathbf{a}_i^T\mathbf{a}_j = 0$

Singular value decomposition (cont.)

Example

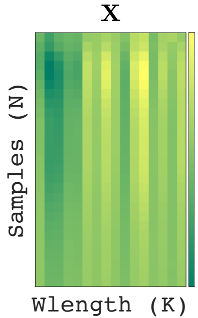
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Consider the (centred) spectral block, $\mathbf{X}$

$$\mathbf{X} = \mathbf{UDP}^T$$

$N = 16$ and $K = 27$, we have that $\mathtt{t} = 16$

- $\mathbf{U}$ is an orthogonal $(N \times \mathtt{t})$ matrix
- $\mathbf{P}$ is an orthogonal $(K \times \mathtt{t})$ matrix
- $\mathbf{D}$ is an diagonal $(N \times N)$ matrix



Singular value decomposition (cont.)

$$\mathbf{X} = \underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1\mathtt{t}} & p_{2N} & \cdots & p_{K\mathtt{t}} \end{bmatrix}}_{\mathbf{P}^T}$$

Let us first consider matrix $\mathbf{D}$, it is a diagonal matrix whose dimension is $(\mathtt{t} \times \mathtt{t})$

$$\underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{r}} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}=\mathtt{diag}\{d_1, d_2, \dots, d_{\mathtt{t}}\}}$$

There are $\mathtt{r} \leq \mathtt{t}$ non-negative values d_i

- The **singular values** of $\mathbf{X}$

The zero-valued d_i can be neglected

- There are $\mathtt{t} - \mathtt{r}$ of them

$$\underbrace{d_1 \geq d_2 \geq \cdots \geq d_{\mathtt{r}}}_{\text{non-zeros}} \geq \underbrace{d_{r+1} \geq \cdots \geq d_{\mathtt{t}}}_{\text{zeros}}$$

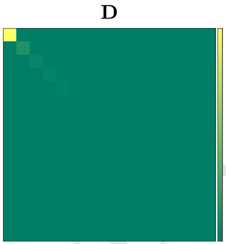
Singular value decomposition (cont.)

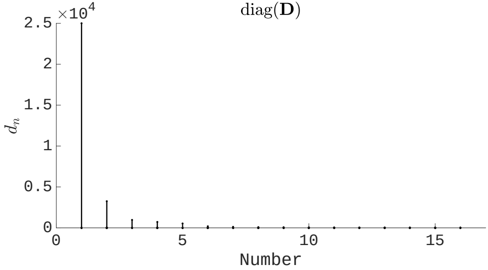
Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Consider the (centred) spectral block, $\mathbf{X}$ ($N = 16$ and $K = 27$, we have that $\mathtt{t} = 16$)

$$\mathbf{X} = \mathbf{U} \underbrace{\mathbf{D}} \mathbf{P}^T$$





- $\mathbf{U}$ is an orthogonal $(N \times \mathtt{t})$ matrix
- $\mathbf{P}$ is an orthogonal $(K \times \mathtt{t})$ matrix
- ⇝ $\mathbf{D}$ is an diagonal $(N \times N)$ matrix

Singular value decomposition (cont.)

$$\mathbf{X} = \underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1\mathtt{t}} & p_{2N} & \cdots & p_{K\mathtt{t}} \end{bmatrix}}_{\mathbf{P}^T}$$

Let us now consider matrix $\mathbf{P}$, it is an orthogonal matrix whose dimension is $(K \times \mathtt{t})$

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1\mathtt{t}} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K\mathtt{t}} & p_{2N} & \cdots & p_{K\mathtt{t}} \end{bmatrix} \quad (3)$$

Matrix $\mathbf{P}$ is called the **loadings matrix**

- Its columns $\mathbf{p}_i$ are the **loadings**

$$\mathbf{p}_i = \begin{bmatrix} p_{i1} & p_{i2} & \cdots & p_{i\mathtt{t}} \end{bmatrix}$$

Matrix $\mathbf{P}\mathbf{P}^T$ is an identity matrix, the inner product between the columns of $\mathbf{P}$ is zero

Singular value decomposition (cont.)

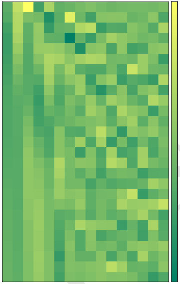
Example

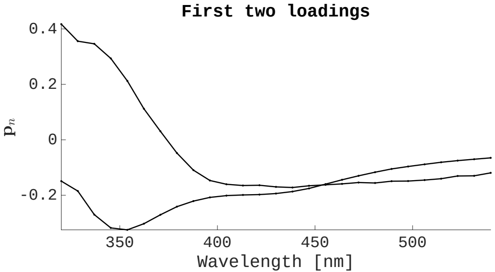
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Consider the (centred) spectral block, $\mathbf{X}$ ($N = 16$ and $K = 27$, we have that $\mathtt{t} = 16$)

$$\mathbf{X} = \mathbf{U}\mathbf{D}\underbrace{\mathbf{P}^T}$$

P





- $\mathbf{U}$ is an orthogonal ($N \times \mathtt{t}$) matrix
- $\mathbf{P}$ is an orthogonal ($K \times \mathtt{t}$) matrix
- $\mathbf{D}$ is an diagonal ($N \times N$) matrix

Singular value decomposition (cont.)

$$\mathbf{X} = \underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1\mathtt{t}} & p_{2N} & \cdots & p_{K\mathtt{t}} \end{bmatrix}}_{\mathbf{P}^T}$$

Let us now consider matrix $\mathbf{U}$, it is an orthogonal matrix whose dimension is $(N \times \mathtt{t})$

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N\mathtt{t}} & u_{N,2} & \cdots & u_{N\mathtt{t}} \end{bmatrix} \quad (4)$$

The columns of $\mathbf{U}$ are orthonormal

- Matrix $\mathbf{U}\mathbf{U}^T = \mathbf{I}$

Singular value decomposition (cont.)

$$\mathbf{X} = \underbrace{\underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}}}_{\mathbf{UD}} \underbrace{\begin{bmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1\mathtt{t}} & p_{2N} & \cdots & p_{K\mathtt{t}} \end{bmatrix}}_{\mathbf{P}^T}$$

The matrix $\mathbf{T} = \mathbf{UP}$ is called **scores matrix**

$$\mathbf{T} = \underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\mathtt{t}} \\ u_{21} & u_{22} & \cdots & u_{2\mathtt{t}} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}} \tag{5}$$

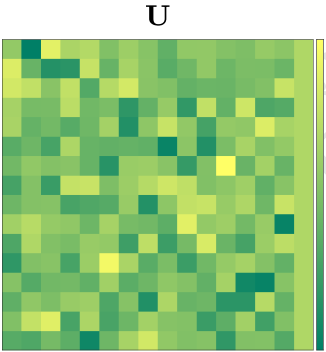
The columns $\mathbf{t}_i$ of matrix $\mathbf{UD}$ are called the **scores**

Singular value decomposition (cont.)

Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Consider the (centred) spectral block, $\mathbf{X}$ ($N = 16$ and $K = 27$, we have that $\mathtt{t} = 16$)

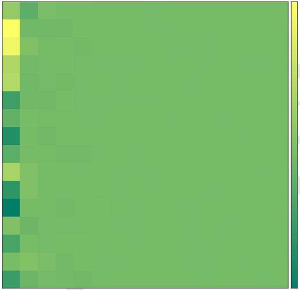


$$\mathbf{X} = \underbrace{\mathbf{U}}\mathbf{D}\mathbf{P}^T$$

- $\rightsquigarrow$ $\mathbf{U}$ is an orthogonal ($N \times \mathtt{t}$) matrix
- $\mathbf{P}$ is an orthogonal ($K \times \mathtt{t}$) matrix
- $\mathbf{D}$ is an diagonal ($N \times N$) matrix

Singular value decomposition (cont.)

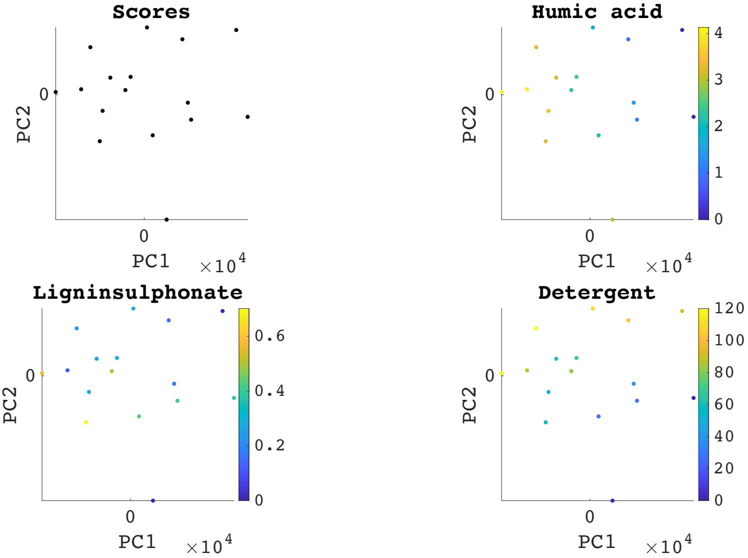
$$\mathbf{T} = \mathbf{U}\mathbf{D}$$



The scores matrix $\mathbf{T}$ and the scores

- Its columns $\{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{\mathtt{t}}\}$

Singular value decomposition (cont.)



Singular value decomposition (cont.)

$$\mathbf{X} = \underbrace{\begin{bmatrix} u_{11} & \cdots & u_{1\mathtt{d}} & \cdots & u_{1\mathtt{t}} \\ \vdots & & \vdots & & \vdots \\ u_{N1} & \cdots & u_{N\mathtt{d}} & \cdots & u_{N\mathtt{t}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & d_{\mathtt{r}} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & d_{\mathtt{t}} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} p_{11} & \cdots & p_{K1} \\ \vdots & & \vdots \\ p_{1\mathtt{r}} & \cdots & p_{K\mathtt{r}} \\ \vdots & & \vdots \\ p_{1\mathtt{t}} & \cdots & p_{K\mathtt{t}} \end{bmatrix}}_{\mathbf{P}^T}$$

$$= \underbrace{\begin{bmatrix} u_{11} & \cdots & u_{1\mathtt{d}} \\ \vdots & \ddots & \vdots \\ u_{N1} & \cdots & u_{N\mathtt{d}} \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{\mathtt{r}} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} p_{11} & \cdots & p_{K1} \\ \vdots & \ddots & \vdots \\ p_{1\mathtt{r}} & \cdots & p_{K\mathtt{r}} \end{bmatrix}}_{\mathbf{P}^T}$$

The $\mathtt{t} - \mathtt{r}$ zero-valued singular values d_i can be discarded

- Together with the last $\mathtt{t} - \mathtt{r}$ columns of $\mathbf{U}$ and $\mathbf{P}$

SVD in reduced form

Eigenvalue decomposition

Consider any square N-matrix $\mathbf{A}$, a number λ, a non-zero N-vector $\mathbf{p}$ and the identity

$$\mathbf{A}\mathbf{p} = \lambda \mathbf{p} \tag{7}$$

λ is an **eigenvalue** of $\mathbf{A}$ and $\mathbf{p}$ is the corresponding **eigenvector**

- (Also any multiple of $\mathbf{p}$ is an eigenvector of λ)

There exist N (not necessarily unique) such numbers λ and associated vectors $\mathbf{p}$

Consider now a symmetric square N-matrix $\mathbf{A}$, and its eigenvectors $\mathbf{p}_1, \dots, \mathbf{p}_N$

- The eigenvectors can be chosen to be orthonormal

For each eigenvalue-eigenvector pair, the eigenequation is $\mathbf{A}\mathbf{p}_n = \lambda_n \mathbf{p}_n$ $(n = 1, \dots, N)$

Eigenvalue decomposition (cont.)

Let $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N$ be the columns of an orthogonal matrix $\mathbf{P}$ ($\mathbf{P}^T\mathbf{P} = \mathbf{I}$)

$$\rightsquigarrow \quad \mathbf{P} = \begin{bmatrix} | & | & & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_n \\ | & | & & | \end{bmatrix}$$

Let $\lambda_1, \dots, \lambda_N$ be the elements of a diagonal matrix $\mathbf{\Sigma} = \texttt{diag}(\lambda_1, \dots, \lambda_N)$

$$\rightsquigarrow \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{bmatrix} \quad (\lambda_1 \geq \cdots \geq \lambda_N)$$

We can write the collection of eigenequations $\mathbf{A}\mathbf{p}_n = \lambda_n \mathbf{p}_n$ in matrix form

$$\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{\Lambda}$$

As for orthogonal matrices $\mathbf{P} = \mathbf{P}^{-1}$, we get the **eigendecomposition** of $\mathbf{A}$

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T \tag{8a}$$

$$= \sum_{n=1}^{N} \lambda_n \mathbf{p}_n \mathbf{p}_n^T \tag{8b}$$

Eigenvalue decomposition (cont.)

Given these definitions, we consider the singular value decomposition of $\mathbf{X}$ (centred)

Let $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, we can write

$$\begin{aligned}
\mathbf{A} &= \mathbf{X}^T\mathbf{X} && \text{(9a)}\\
&= \left(\mathbf{P}\mathbf{D}\mathbf{U}^T\right)\left(\mathbf{U}\mathbf{D}\mathbf{P}^T\right) && \text{(9b)}\\
&= \mathbf{P}\mathbf{D}\underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}}\mathbf{D}\mathbf{P}^T && \text{(9c)}\\
&= \mathbf{P}\mathbf{D}^2\mathbf{P}^T && \text{(9d)}
\end{aligned}$$

We have that the eigenvalues of $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ are the diagonal elements of matrix $\mathbf{D}^2$

- (The squared singular values of matrix $\mathbf{X}$)

Moreover, columns of $\mathbf{P}$ are the eigenvectors of $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ (and the loadings of $\mathbf{X}$)

Matrix $\mathbf{X}^T\mathbf{X}/(N-1)$ estimates the (variance)-covariance matrix from centred $\mathbf{X}$-block

- **Principal components analysis**, eigendecomposition of a covariance matrix

Principal component analysis, PCA

Principal components analysis, PCA is a method for reducing data dimensionality

- Low-dimensional representation of the data
- Visual discovery of data structures

The eigenvectors of the data covariance matrix are directions in original data space

- The loadings $\mathbf{p}_n$ embed the relevance of the columns $\mathbf{X}$ (original directions)
- Interest in retaining only eigenvectors that associate with large variations
- They correspond to the largest eigenvalues of the data covariance matrix

---

The spectral data $\mathbf{X}$, absorbances, are characterised by redundant information

- Absorbances at adjacent wavelength are highly correlated
- (Peaks of pure components are spread over a range)

The objective is to find whether there are data directions of high variability

- These direction will be linear compositions of the original directions
- They will also be orthogonal to each other, thus non-redundant

Principal component regression

Principal component regression uses a suitable value $\mathtt{t}$ to select features of $\mathbf{X}$

Then, the retained features $\mathbf{T}_{\mathtt{t}}$ are used to perform MLR against $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{T}_{\mathtt{t}}\mathbf{C} + \mathbf{F} \tag{10}$$

By the least squares methods, we get the estimates

$$\widehat{\mathbf{C}} = \left(\mathbf{T}_{\mathtt{t}}^{T}\mathbf{T}_{\mathtt{t}}\right)^{-1}\mathbf{T}_{\mathtt{t}}^{T}\mathbf{Y} \tag{11}$$

Since matrix $\mathbf{T}_{\mathtt{t}}^{T}\mathbf{T}_{\mathtt{t}}$ is diagonal, its inverse is trivial

Principal component regression (cont.)

Prediction

$$\mathbf{Y} = \mathbf{T}_{\mathrm{g}}\mathbf{C} + \mathbf{F} \quad (12a)$$

$$= \mathbf{X}\mathbf{P}_{\mathrm{g}}\mathbf{C} + \mathbf{F} \quad (12b)$$

Consider a new sample spectrum $\mathbf{z}$ and the predicted value $\widehat{\mathbf{y}}$, uncentered with $\hat{\mathbf{x}}$ and $\overline{\mathbf{x}}$ be the learning sample means, the prediction

$$\widehat{\mathbf{y}} = \overline{\mathbf{y}} + (\mathbf{z} - \overline{\mathbf{x}})\mathbf{P}_{\mathrm{t}}\widehat{\mathbf{C}} \quad (13)$$

Matrix $\mathbf{P}_{\mathrm{g}}\widehat{\mathbf{C}}$ is called the **regression matrix**

- (Similar to matrix $\widehat{\mathbf{B}}$ in MLR)

Consider the case where $\texttt{rank}(\mathbf{X}) = K$ and $\texttt{t} = K$

- PCR and MLR give the same result

Consider the case where $\texttt{rank}(\mathbf{X}) = K$ but $\texttt{t} < K$

- PCR and MLR give different results