Chapter 3

Multivariate Statistics Between Two-Observation Spaces

As mentioned in the previous chapter, industrial data are usually divided into two categories, process data and quality data, belonging to different measurement spaces. The vast majority of smart manufacturing problems, such as soft measurement, control, monitoring, optimization, etc., inevitably require modeling the data relationships between the two kinds of measurement variables. This chapter's subject is to discover the correlation between the sets in different observation spaces.

The multivariate statistical analysis relying on correlation among variables generally include canonical correlation analysis (CCA) and partial least squares regression (PLS). They all perform linear dimensionality reduction with the goal of maximizing the correlation between variables in two measurement spaces. The difference are that CCA maximize **correlation**, while PLS maximize **covariance**.

3.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) was first proposed by Hotelling in 1936 (Hotelling 1936). It is a multivariate statistical analysis method that uses the correlation between two composite variables to reflect the overall correlation between two sets of variables. The CCA algorithm is widely used in the analysis of data correlation and it is also the basis of partial least squares. In addition, it is also used in feature fusion, data dimensionality reduction, and fault detection (Yang et al. 2015; Zhang and Dou 2015; Zhang et al. 2020; Hou 2013; Chen et al. 2016a, b).

3.1.1 Mathematical Principle of CCA

Assuming that there are l dependent variables $\mathbf{y} = (y_1, y_2, \dots, y_l)^T$ and m independent variables $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$. In order to capture the correlation between the dependent variables and the independent variables, n sample points are observed, which constitutes two data sets

$$\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(n)]^T \in R^{n \times m}$$

$$\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n)]^T \in R^{n \times l}$$

CCA draws on the idea of component extraction to find a canonical component u , which is a linear combination of variables x_i ; and a canonical component v , which is a linear combination of y_i . In the process of extraction, the correlation between u and v is required to be maximized. The correlation degree between u and v can roughly reflect the correlation between $\mathbf{X}$ and $\mathbf{Y}$.

Without loss of generality, assuming that the original variables are all standardized, i.e., each column of the data set $\mathbf{X}$ and $\mathbf{Y}$ has mean 0 and variance 1, the covariance matrix of $\text{cov}(\mathbf{X}, \mathbf{Y})$ is equal to its correlation coefficient matrix, in which,

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Y} \\ \mathbf{Y}^T \mathbf{X} & \mathbf{Y}^T \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^T & \boldsymbol{\Sigma}_{yy} \end{bmatrix}$$

PCA is analyzed for $\boldsymbol{\Sigma}_{xx}$ or $\boldsymbol{\Sigma}_{yy}$, while CCA is analyzed for $\boldsymbol{\Sigma}_{xy}$

Now the problem is how to find the direction vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and then use them to construct the canonical components:

$$\begin{aligned} u &= \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_m x_m \\ v &= \beta_1 y_1 + \beta_2 y_2 + \dots + \beta_l y_l, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_m]^T \in R^{m \times 1}$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_l]^T \in R^{l \times 1}$, such that the correlation between u and v is maximized. Obviously, the sample means of u and v are zero, and their sample variances are as follows:

$$\begin{aligned} \text{var}(u) &= \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} \\ \text{var}(v) &= \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} \end{aligned}$$

The covariance of u and v is

$$\text{cov}(u, v) = \alpha^T \Sigma_{xy} \beta.$$

One way to maximize the correlation of u and v is to make the corresponding correlation coefficient maximum, i.e.,

$$\max \rho(u, v) = \frac{\text{cov}(uv)}{\sqrt{\text{var}(u)\text{var}(v)}}. \quad (3.2)$$

In CCA, the following optimization objective is used:

$$\begin{aligned} J_{\text{CCA}} &= \max \langle u, v \rangle = \alpha^T \Sigma_{xy} \beta \\ \text{s.t. } &\alpha^T \Sigma_{xx} \alpha = 1; \beta^T \Sigma_{yy} \beta = 1. \end{aligned} \quad (3.3)$$

This optimization objective can be summarized as follows: to seek a unit vector α on the subspace of X and a unit vector β on the subspace of Y such that the correlation between u and v is maximized. Geometrically, $\rho(u, v)$ is again equal to the cosine of the angle between u and v . Thus, (3.3) is again equivalent to making the angle ω between u and v take the minimum value.

It can be seen from (3.3) that the goal of the CCA algorithm is finally transformed into a convex optimization process. The maximum value of this optimization goal is the correlation coefficient of X and Y , and the corresponding α and β are projection vectors, or linear coefficients. After the first pair of canonical correlation variables are obtained, the second to k th pair of canonical correlation variables that are not correlated with each other can be similarly calculated.

The following Fig. 3.1 shows the basic principle diagram of the CCA algorithm:

At present, there are two main methods which include eigenvalue decomposition and singular value decomposition for optimizing the above objective function to obtain α and β .

Fig. 3.1 Basic principle diagram of the CCA algorithm

3.1.2 Eigenvalue Decomposition of CCA Algorithm

Using the Lagrangian function, the objective function of (3.3) is transformed as follows:

$$\max J_{\text{CCA}}(\alpha, \beta) = \alpha^T \Sigma_{xy} \beta - \frac{\lambda_1}{2} (\alpha^T \Sigma_{xx} \alpha - 1) - \frac{\lambda_2}{2} (\beta^T \Sigma_{yy} \beta - 1). \quad (3.4)$$

Set $\frac{\partial J}{\partial \alpha} = 0$ and $\frac{\partial J}{\partial \beta} = 0$, then

$$\begin{aligned} \Sigma_{xy} \beta - \lambda_1 \Sigma_{xx} \alpha &= 0 \\ \Sigma_{xy}^T \alpha - \lambda_2 \Sigma_{yy} \beta &= 0. \end{aligned} \quad (3.5)$$

Let $\lambda = \lambda_1 = \lambda_2 = \alpha^T \Sigma_{xy} \beta$, and multiply (3.5) to the left by Σ_{xx}^{-1} and Σ_{yy}^{-1} , respectively, and get:

$$\begin{aligned} \Sigma_{xx}^{-1} \Sigma_{xy} \beta &= \lambda \alpha \\ \Sigma_{yy}^{-1} \Sigma_{yx} \alpha &= \lambda \beta. \end{aligned} \quad (3.6)$$

Substituting the second formula in (3.6) into the first formula, we can get

$$\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \alpha = \lambda^2 \alpha \quad (3.7)$$

From (3.7), we can get the largest eigenvalue λ and the corresponding maximum eigenvector α only by eigenvalue decomposition of the matrix $\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx}$. In the similar way, the vector β can be obtained. At this time, the projection vectors α and β of a set of canonical correlation variables can be obtained.

3.1.3 SVD Solution of CCA Algorithm

Let $\alpha = \Sigma_{xx}^{-1/2} a$, $\beta = \Sigma_{yy}^{-1/2} b$, and then we can get

$$\begin{aligned} \alpha^T \Sigma_{xx} \alpha &= 1 \rightarrow a^T \Sigma_{xx}^{-1/2} \Sigma_{xx} \Sigma_{xx}^{-1/2} a = 1 \rightarrow a^T a = 1 \\ \beta^T \Sigma_{yy} \beta &= 1 \rightarrow b^T \Sigma_{yy}^{-1/2} \Sigma_{yy} \Sigma_{yy}^{-1/2} b = 1 \rightarrow b^T b = 1 \\ \alpha^T \Sigma_{xy} \beta &= a^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} b. \end{aligned} \quad (3.8)$$

In other words, the objective function of (3.3) can be transformed into as follows:

$$\begin{aligned} J_{\text{CCA}}(\mathbf{a}, \mathbf{b}) &= \arg \max_{\mathbf{a}, \mathbf{b}} \mathbf{a}^T \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{b} \\ \text{s.t. } \mathbf{a}^T \mathbf{a} &= \mathbf{b}^T \mathbf{b} = 1. \end{aligned} \quad (3.9)$$

A singular value decomposition for matrix $\mathbf{M}$ yields

$$\mathbf{M} = \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} = \boldsymbol{\Gamma} \boldsymbol{\Sigma} \boldsymbol{\Psi}^T, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{A}_\kappa & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (3.10)$$

where κ is the number of principal elements or non-zero singular values, and $\kappa \leq \min(l, m)$, $\mathbf{A}_\kappa = \text{diag}(\lambda_1, \dots, \lambda_\kappa)$, $\lambda_1 \geq \dots \geq \lambda_\kappa > 0$.

Since all columns of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ are standard orthogonal basis, $\mathbf{a}^T \boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}^T \mathbf{b}$ are vectors with only one scalar value of 1, and the remaining scalar value of 0. So, we can get

$$\mathbf{a}^T \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{b} = \mathbf{a}^T \boldsymbol{\Gamma} \boldsymbol{\Sigma} \boldsymbol{\Psi}^T \mathbf{b} = \sigma_{ab}. \quad (3.11)$$

From (3.11), it can be seen that $\mathbf{a}^T \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{b}$ maximizes actually the left and right singular vectors corresponding to the maximum singular values of $\mathbf{M}$. Thus, using the corresponding left and right singular vectors $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$, we can obtain the projection vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for a set of canonical correlation variables, namely,

$$\begin{aligned} \boldsymbol{\alpha} &= \boldsymbol{\Sigma}_{xx}^{-1/2} \mathbf{a} \\ \boldsymbol{\beta} &= \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{b}. \end{aligned} \quad (3.12)$$

3.1.4 CCA-Based Fault Detection

When there is a clear input-output relationship between the two types of data measurable online, CCA can be used to design an effective fault detection system. The CCA-based fault detection method can be considered as an alternative to PCA-based fault detection method, and an extension of PLS-based fault detection method (Chen et al. 2016a).

Let

$$\begin{aligned} \mathbf{J}_s &= \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Gamma}(:, 1 : \kappa) \\ \mathbf{L}_s &= \boldsymbol{\Sigma}_{yy}^{-1/2} \boldsymbol{\Psi}(:, 1 : \kappa) \\ \mathbf{J}_{\text{res}} &= \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Gamma}(:, \kappa + 1 : l) \\ \mathbf{L}_{\text{res}} &= \boldsymbol{\Sigma}_{yy}^{-1/2} \boldsymbol{\Psi}(:, \kappa + 1 : m). \end{aligned}$$

According to CCA method, $\mathbf{J}_s^T \mathbf{x}$ and $\mathbf{L}_s^T \mathbf{y}$ are closely related. However, in actual systems, measurement variables are inevitably affected by noise, and the correlation between $\mathbf{J}_s^T \mathbf{x}$ and $\mathbf{L}_s^T \mathbf{y}$ can be expressed as

$$\mathbf{L}_s^T \mathbf{y}(k) = \mathbf{A}_\kappa^T \mathbf{J}_s^T \mathbf{x}(k) + v_s(k), \quad (3.13)$$

where v_s is the noise term and weakly related to $\mathbf{J}_s^T \mathbf{x}$. Based on this, the residual vector is

$$\mathbf{r}_1(k) = \mathbf{L}_s^T \mathbf{y}(k) - \mathbf{A}_\kappa^T \mathbf{J}_s^T \mathbf{x}(k). \quad (3.14)$$

Assume that the input and output data obey the Gaussian distribution. It is known that linear transformation does not change the distribution of random variables, so the residual signal $\mathbf{r}_1$ also obeys the Gaussian distribution and its covariance matrix is

$$\boldsymbol{\Sigma}_{r1} = \frac{1}{N-1} (\mathbf{L}_s^T \mathbf{Y} - \mathbf{A}_\kappa^T \mathbf{J}_s^T \mathbf{X}) (\mathbf{L}_s^T \mathbf{Y} - \mathbf{A}_\kappa^T \mathbf{J}_s^T \mathbf{X})^T = \frac{\mathbf{I}_\kappa - \mathbf{A}_\kappa^2}{N-1}. \quad (3.15)$$

Similarly, another residual vector can be obtained

$$\mathbf{r}_2(k) = \mathbf{J}_s^T \mathbf{x}(k) - \mathbf{A}_\kappa \mathbf{L}_s^T \mathbf{y}(k). \quad (3.16)$$

Its covariance matrix is

$$\boldsymbol{\Sigma}_{r2} = \frac{1}{N-1} (\mathbf{J}_s^T \mathbf{U} - \mathbf{A}_\kappa \mathbf{L}_s^T \mathbf{Y}) (\mathbf{J}_s^T \mathbf{U} - \mathbf{A}_\kappa \mathbf{L}_s^T \mathbf{Y})^T = \frac{\mathbf{I}_\kappa - \mathbf{A}_\kappa^2}{N-1}. \quad (3.17)$$

It can be seen from formula (3.15)–(3.16) that the covariance of residual $\mathbf{r}_1$ and $\mathbf{r}_2$ are the same. For fault detection, the following two statistics can be constructed:

$$\mathbf{T}_1^2(k) = (N-1) \mathbf{r}_1^T(k) (\mathbf{I}_\kappa - \mathbf{A}_\kappa^2)^{-1} \mathbf{r}_1(k) \quad (3.18)$$

$$\mathbf{T}_2^2(k) = (N-1) \mathbf{r}_2^T(k) (\mathbf{I}_\kappa - \mathbf{A}_\kappa^2)^{-1} \mathbf{r}_2(k). \quad (3.19)$$

3.2 Partial Least Squares

Multiple linear regression analysis is relatively common and the least square method is generally used to estimate the regression coefficient in this type of regression method. But the least square technique often fails when there is multiple correlation between the independent variables or the number of samples is less than the number of variables. So the partial least square technique is developed to resolve this problem. S. Wold and C. Albano et al. proposed the partial least squares method for the first time and applied it to the field of chemistry (Wold et al. 1989). It aims at

the regression modeling between two sets of multi-variables with high correlation and integrates the basic functions of multiple linear regression analysis, principal component analysis, and canonical correlation analysis. PLS is also called the second-generation regression analysis method due to its simplification model in the data structure and correlation (Hair et al. 2016). It has developed rapidly and widely used in various fields recent years (Okwuashi et al. 2020; Ramin et al. 2018).

3.2.1 Fundamental of PLS

Suppose there are l dependent variables $(y_1, y_2, \dots, y_l)$ and m independent variables $(x_1, x_2, \dots, x_m)$. In order to study the statistical relationship between the dependent variable and the independent variable, n sample points are observed, which constitutes a data set $(\mathbf{X} = [x_1, x_2, \dots, x_m] \in R^{n \times m}, \mathbf{Y} = [y_1, y_2, \dots, y_l] \in R^{n \times l})$ of the independent variables and the dependent variables.

To address the problems encountered in least squares multiple regression between $\mathbf{X}$ and $\mathbf{Y}$, the concept of component extraction is introduced in PLS regression analysis. Recall that principal component analysis, for a single data matrix $\mathbf{X}$, finds the composite variable that best summarizes the information in the original data. The principal component $\mathbf{T}$ in $\mathbf{X}$ is extracted with the maximum variance information of the original data:

$$\max \text{var}(\mathbf{T}), \quad (3.20)$$

PLS extracts component vectors $\mathbf{t}_i$ and $\mathbf{u}_i$ from $\mathbf{X}$ and $\mathbf{Y}$, which means $\mathbf{t}_i$ is a linear combination of $(x_1, x_2, \dots, x_m)$, and $\mathbf{u}_i$ is a linear combination of $(y_1, y_2, \dots, y_l)$. During the extracting of components, in order to meet the needs of regression analysis, the following two requirements should be satisfied:

- (1) $\mathbf{t}_i$ and $\mathbf{u}_i$ carry the variation information in their respective data set as much as possible, respectively;
- (2) The correlation between $\mathbf{t}_i$ and $\mathbf{u}_i$ is maximized.

The two requirements indicate that $\mathbf{t}_i$ and $\mathbf{u}_i$ should represent the data set $\mathbf{X}$ and $\mathbf{Y}$ as well as possible and the component $\mathbf{t}_i$ of the independent variable has the best ability to explain the component $\mathbf{u}_i$ of the dependent variable.

3.2.2 PLS Algorithm

The most popular algorithm used in PLS to compute the vectors in the calibration step is known as nonlinear iterative partial least squares (NIPALS). First, normalize the data to achieve the purpose of facilitating calculations. Normalize $\mathbf{X}$ to get matrix $\mathbf{E}_0$ and normalize $\mathbf{Y}$ to get matrix $\mathbf{F}_0$:

$$\mathbf{E}_0 = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{mn} \end{bmatrix}, \quad \mathbf{F}_0 = \begin{bmatrix} y_{11} & \cdots & y_{1l} \\ \vdots & \vdots & \vdots \\ y_{n1} & \cdots & x_{nl} \end{bmatrix} \quad (3.21)$$

In the first step, set $\mathbf{t}_1$ ($\mathbf{t}_1 = \mathbf{E}_0 \mathbf{w}_1$) to be the first component of $\mathbf{E}_0$, and $\mathbf{w}_1$ is the first direction vector of $\mathbf{E}_0$, which is a unit vector, $\|\mathbf{w}_1\| = 1$. Similarly, set $\mathbf{u}_1$ ($\mathbf{u}_1 = \mathbf{F}_0 \mathbf{c}_1$) to be the first component of $\mathbf{F}_0$, and $\mathbf{c}_1$ is the first direction vector of $\mathbf{F}_0$, which is a unit vector too, $\|\mathbf{c}_1\| = 1$.

According to the principle of principal component analysis, $\mathbf{t}_1$ and $\mathbf{u}_1$ should meet the following conditions in order to be able to represent the data variation information in $\mathbf{X}$ and $\mathbf{Y}$ well:

$$\begin{aligned} \max \text{var}(\mathbf{t}_1) \\ \max \text{var}(\mathbf{u}_1) \end{aligned} \quad (3.22)$$

On the other hand, $\mathbf{t}_1$ is further required to have the best explanatory ability for $\mathbf{u}_1$ due to the needs of regression modeling. According to the thinking of canonical correlation analysis, the correlation between $\mathbf{t}_1$ and $\mathbf{u}_1$ should reach the maximum value:

$$\max r(\mathbf{t}_1, \mathbf{u}_1). \quad (3.23)$$

The covariance of $\mathbf{t}_1$ and $\mathbf{u}_1$ is usually used to describe the correlation in partial least squares regression:

$$\max \text{Cov}(\mathbf{t}_1, \mathbf{u}_1) = \sqrt{\text{var}(\mathbf{t}_1) \text{var}(\mathbf{u}_1)} r(\mathbf{t}_1, \mathbf{u}_1) \quad (3.24)$$

Converting to the normal mathematical expression, $\mathbf{t}_1$ and $\mathbf{u}_1$ is solved by the following optimization problem:

$$\begin{aligned} \max_{\mathbf{w}_1, \mathbf{c}_1} & \langle \mathbf{E}_0 \mathbf{w}_1, \mathbf{F}_0 \mathbf{c}_1 \rangle \\ \text{s.t.} & \begin{cases} \mathbf{w}_1^T \mathbf{w}_1 = 1 \\ \mathbf{c}_1^T \mathbf{c}_1 = 1. \end{cases} \end{aligned} \quad (3.25)$$

Therefore, it needs to calculate the maximum value of $\mathbf{w}_1^T \mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1$ under the constraints of $\|\mathbf{w}_1\|^2 = 1$ and $\|\mathbf{c}_1\|^2 = 1$.

In this case, the Lagrangian function is

$$s = \mathbf{w}_1^T \mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1 - \lambda_1 (\mathbf{w}_1^T \mathbf{w}_1 - 1) - \lambda_2 (\mathbf{c}_1^T \mathbf{c}_1 - 1). \quad (3.26)$$

Calculate the partial derivatives of s with respect to $\mathbf{w}_1$, $\mathbf{c}_1$, λ_1 , and λ_2 , and let them be zero

$$\frac{\partial s}{\partial \mathbf{w}_1} = \mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1 - 2\lambda_1 \mathbf{w}_1 = 0, \quad (3.27)$$

$$\frac{\partial s}{\partial \mathbf{c}_1} = \mathbf{E}_0^T \mathbf{F}_0 \mathbf{w}_1 - 2\lambda_2 \mathbf{c}_1 = 0, \quad (3.28)$$

$$\frac{\partial s}{\partial \lambda_1} = -(\mathbf{w}_1^T \mathbf{w}_1 - 1) = 0, \quad (3.29)$$

$$\frac{\partial s}{\partial \lambda_2} = -(\mathbf{c}_1^T \mathbf{c}_1 - 1) = 0. \quad (3.30)$$

It can be derived from the above formulas that

$$2\lambda_1 = 2\lambda_2 = \mathbf{w}_1^T \mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1 = \langle \mathbf{E}_0 \mathbf{w}_1, \mathbf{F}_0 \mathbf{c}_1 \rangle \quad (3.31)$$

Let $\theta_1 = 2\lambda_1 = 2\lambda_2 = \mathbf{w}_1^T \mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1$, so θ_1 is the value of the objective function of the optimization problem (3.25). Then (3.27) and (3.28) are rewritten as

$$\mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1 = \theta_1 \mathbf{w}_1, \quad (3.32)$$

$$\mathbf{F}_0^T \mathbf{E}_0 \mathbf{w}_1 = \theta_1 \mathbf{c}_1. \quad (3.33)$$

Substitute (3.33) into (3.32),

$$\mathbf{E}_0^T \mathbf{F}_0 \mathbf{F}_0^T \mathbf{E}_0 \mathbf{w}_1 = \theta_1^2 \mathbf{w}_1. \quad (3.34)$$

Substitute (3.32) into (3.33) simultaneously,

$$\mathbf{F}_0^T \mathbf{E}_0 \mathbf{E}_0^T \mathbf{F}_0 \mathbf{c}_1 = \theta_1^2 \mathbf{c}_1. \quad (3.35)$$

Equation (3.34) shows that $\mathbf{w}_1$ is the eigenvector of matrix $\mathbf{E}_0^T \mathbf{F}_0 \mathbf{F}_0^T \mathbf{E}_0$ with the corresponding eigenvalue θ_1^2 . Here, θ_1 is the objective function. If we want to get its maximum value, $\mathbf{w}_1$ should be the unit eigenvector of the maximum eigenvalue of matrix $\mathbf{E}_0^T \mathbf{F}_0 \mathbf{F}_0^T \mathbf{E}_0$. Similarly, $\mathbf{c}_1$ should be the unit eigenvector of the largest eigenvalue of the matrix $\mathbf{F}_0^T \mathbf{E}_0 \mathbf{E}_0^T \mathbf{F}_0$.

Then the first components $\mathbf{t}_1$ and $\mathbf{u}_1$ are calculated from the direction vectors $\mathbf{w}_1$ and $\mathbf{c}_1$:

$$\begin{aligned} \mathbf{t}_1 &= \mathbf{E}_0 \mathbf{w}_1 \\ \mathbf{u}_1 &= \mathbf{F}_0 \mathbf{c}_1. \end{aligned} \quad (3.36)$$

The regression equations of $\mathbf{E}_0$ and $\mathbf{F}_0$ is found by $\mathbf{t}_1$ and $\mathbf{u}_1$:

$$\begin{aligned} \mathbf{E}_0 &= \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{E}_1 \\ \mathbf{F}_0 &= \mathbf{u}_1 \mathbf{q}_1^T + \mathbf{F}_1^* \\ \mathbf{F}_0 &= \mathbf{t}_1 \mathbf{r}_1^T + \mathbf{F}_1. \end{aligned} \quad (3.37)$$

The regression coefficient vectors in (3.37) are

$$\begin{aligned} p_1 &= \frac{\mathbf{E}_0^T \mathbf{t}_1}{\|\mathbf{t}_1\|^2} \\ q_1 &= \frac{\mathbf{F}_0^T \mathbf{u}_1}{\|\mathbf{u}_1\|^2} \\ r_1 &= \frac{\mathbf{F}_0^T \mathbf{t}_1}{\|\mathbf{t}_1\|^2}. \end{aligned} \quad (3.38)$$

$\mathbf{E}_1$, $\mathbf{F}_1^*$ and $\mathbf{F}_1$ are the residual matrices of the three regression equations.

Second step is to replace $\mathbf{E}_0$ and $\mathbf{F}_0$ with residual matrices $\mathbf{E}_1$ and $\mathbf{F}_1$, respectively. Then find the second pair of direction vectors $\mathbf{w}_2$, $\mathbf{c}_2$, and the second pair of components $\mathbf{t}_2$ and $\mathbf{u}_2$:

$$\begin{aligned} \mathbf{t}_2 &= \mathbf{E}_1 \mathbf{w}_2 \\ \mathbf{u}_2 &= \mathbf{F}_1 \mathbf{c}_2 \\ \theta_2 &= \mathbf{w}_2^T \mathbf{E}_1^T \mathbf{F}_1 \mathbf{c}_2. \end{aligned} \quad (3.39)$$

Similarly, $\mathbf{w}_2$ is the unit eigenvector corresponding to the largest eigenvalue of matrix $\mathbf{E}_1^T \mathbf{F}_1 \mathbf{F}_1^T \mathbf{E}_1$, and $\mathbf{c}_2$ is the unit eigenvector of the largest eigenvalue of matrix $\mathbf{F}_1^T \mathbf{E}_1 \mathbf{E}_1^T \mathbf{F}_1$. Calculate the regression coefficient

$$\begin{aligned} p_2 &= \frac{\mathbf{E}_1^T \mathbf{t}_2}{\|\mathbf{t}_2\|^2} \\ r_2 &= \frac{\mathbf{F}_1^T \mathbf{t}_2}{\|\mathbf{t}_2\|^2}. \end{aligned} \quad (3.40)$$

The regression equation is updated:

$$\begin{aligned} \mathbf{E}_1 &= \mathbf{t}_2 \mathbf{p}_2^T + \mathbf{E}_2 \\ \mathbf{F}_1 &= \mathbf{t}_2 \mathbf{r}_2^T + \mathbf{F}_2. \end{aligned} \quad (3.41)$$

Repeat the calculation according to the above steps. If the rank of $\mathbf{X}$ is R , the regression equation can be obtained:

$$\begin{aligned} \mathbf{E}_0 &= \mathbf{t}_1 \mathbf{p}_1^T + \cdots + \mathbf{t}_R \mathbf{p}_R^T \\ \mathbf{F}_0 &= \mathbf{t}_1 \mathbf{r}_1^T + \cdots + \mathbf{t}_R \mathbf{r}_R^T + \mathbf{F}_R. \end{aligned} \quad (3.42)$$

If the number of feature vectors used in the PLS modeling is large enough, the residuals could be zero. In general, it only needs to select a ($a \ll R$) components among them to form a regression model with better prediction. The number of principal components required for modeling is determined by cross-validation discussed

in Sect. 3.2.3. Once the appropriate component number is determined, the external relationship of the input variable matrix X as

$$X = T P^T + \bar{X} = \sum_{h=1}^a t_h p_h^T + \bar{X}. \quad (3.43)$$

The external relationship of the output variable matrix Y can be written as

$$Y = U Q^T + \bar{Y} = \sum_{h=1}^a u_h q_h^T + \bar{Y}. \quad (3.44)$$

The internal relationship is expressed as

$$\hat{u}_h = b_h t_h, \quad b_h = t_h^T u_h / t_h^T t_h. \quad (3.45)$$

3.2.3 Cross-Validation Test

In many cases, the PLS equation does not require the selection of all principal components for regression modeling, but rather, as in principal component analysis, the first d ($d \leq l$) principal components can be selected in a truncated manner, and a better predictive model can be obtained using only these d principal components. In fact, if the subsequent principal components no longer provide more meaningful information to explain the dependent variable, using too many principal components will only undermine the understanding of the statistical trend and lead to wrong prediction conclusions. The number of principal components required for modeling can be determined by cross-validation.

Cross-validation is used to prevent over-fitting caused by complex model. Sometimes referred to as the circular estimation, it is a statistically useful method for cutting data sample into smaller subset. This is done by first doing the analysis on a subset, while the other subset is used for subsequent confirmation and validation of this analysis. The subset used for analysis is called the training set. The other subset is called validation set and generally separated from the testing set. Two cross-validation methods often used in practice are K -fold cross-validation (K-CV) and leave-one-out cross-validation (LOO-CV).

K-CV divides the n original data into K groups (generally evenly divided), makes each subset of data into a validation set once separately. The rest of the $K - 1$ subsets are considered as the training set, so K-CV will result in K models. In general, K is selected between 5 and 10. LOO-CV is essentially N-CV. The process of determining the number of principal components will be described in detail using LOO-CV as an example.

All n samples are divided into two parts: the first part is the set of all samples excluding a certain sample i (containing a total of $n - 1$ samples) and a regression equation is fitted with this data set using d principal components; The second part is to substitute the i th sample that was just excluded into the fitted regression equation to obtain the predicted value $\hat{y}_{(i)j}(d)$, $j = 1, 2, \dots, l$ of y_j . Repeating the above test for each $i = 1, 2, \dots, n$, the sum of squared prediction errors for y_j can be defined as $\text{PRESS}_j(d)$.

$$\text{PRESS}_j(d) = \sum_{i=1}^n (y_{ij} - \hat{y}_{(i)j}(d))^2, j = 1, 2, \dots, l. \quad (3.46)$$

The sum of squared prediction errors of $Y = (y_1, \dots, y_l)^T$ can be obtained as

$$\text{PRESS}(d) = \sum_{j=1}^l \text{PRESS}_j(d). \quad (3.47)$$

Obviously, if the robustness of the regression equation is not good, the error is large and thus it is very sensitive to change in the samples, and the effect of this perturbation error will increase the $\text{PRESS}(d)$ value.

On the other hand, use all sample points to fit a regression equation containing d components. In this case, the fitted value of the i th sample point is $\hat{y}_{ij}(d)$. The fitted error sum of squares for y_j is defined as $\text{SS}_j(d)$ value

$$\text{SS}_j(d) = \sum_{i=1}^n (y_{ij} - \hat{y}_{ij}(d))^2. \quad (3.48)$$

The sum of squared errors of Y is

$$\text{SS}(d) = \sum_{j=1}^l \text{SS}_j(d) \quad (3.49)$$

Generally, $\text{PRESS}(d)$ is greater than $\text{SS}(d)$ because $\text{PRESS}(d)$ contains an unknown perturbation error and the fitting error decreases with the increase of components, i.e., $\text{SS}(d)$ is less than $\text{SS}(d - 1)$. Next, compare $\text{SS}(d - 1)$ and $\text{PRESS}(d)$. $\text{SS}(d - 1)$ is the fitting error of the regression equation that is fitted with all samples with d components; $\text{PRESS}(d)$ contains the perturbation error of the samples but with one more component. If the d component regression equation with perturbation error can be somewhat smaller than the fitting error of the $d - 1$ component regression equation, it is considered that adding one component t_d will result in a significant improvement in prediction accuracy. Therefore, it is always expected that the ratio of $\frac{\text{PRESS}(d)}{\text{SS}(d-1)}$ is as small as possible. The general setting

$$\frac{\text{PRESS}(d)}{\text{SS}(d-1)} \leq (1 - 0.05)^2 = 0.95^2. \quad (3.50)$$

IF $\text{PRESS}(d) \leq 0.95^2 \text{SS}(d-1)$, the addition of the component is considered beneficial. And conversely, if $\text{PRESS}(d) > 0.95^2 \text{SS}(d-1)$, the new addition of components is considered to have no significant improvement in reducing the prediction error of the regression equation.

In practice, the following cross-validation index is used. For each dependent variable y_j , define

$$Q_{dj}^2 = 1 - \frac{\text{PRESS}_j(d)}{\text{SS}_j(d-1)}. \quad (3.51)$$

For the full dependent variable $\mathbf{Y}$, the cross-validation index of component $\mathbf{t}_d$ is defined as

$$Q_d^2 = 1 - \frac{\text{PRESS}(d)}{\text{SS}(d-1)}. \quad (3.52)$$

The marginal contribution of component $\mathbf{t}_d$ to the predictive accuracy of the regression model has the following two scales (cross-validation index).

- (1) $Q_d^2 > 1 - 0.95^2 = 0.0975$, the marginal contribution of $\mathbf{t}_d$ component is significant; and
- (2) For $k = 1, 2, \dots, l$, there is at least one k such that $Q_{dj}^2 > 0.0975$ holds, at which point the addition of component $\mathbf{t}_d$ leads to a significant improvement in the prediction accuracy of at least one dependent variable y_k . Therefore it can also be argued that adding component $\mathbf{t}_d$ is clearly beneficial.

References

- Chen Z, Ding SX, Zhang K, Li Z, Hu Z (2016) Canonical correlation analysis-based fault detection methods with application to alumina evaporation process. *Control Eng Pract* 46:51–58
- Chen Z, Zhang K, Ding SX, Shardt YAW, Hu Z (2016) Improved canonical correlation analysis-based fault detection methods for industrial processes. *J Process Control* 41:26–34
- Hair JF, Hult GTM, Ringle C, Sarstedt M (2016) A primer on partial least squares structural equation modeling (PLS-SEM). Sage Publications, Thousand Oaks
- Hotelling H (1936) Relations between two sets of variates. *Biom Trust* 28(3/4):321–377
- Hou B (2013) Research on canonical correlation analysis algorithm based on sparse representation. Nanjing University of Science and Technology, Nanjing
- Li W, Yang J, Zhang J et al (2015) CCA-based algorithm for personalized privacy preservation in trajectory. *J Jilin Univ (Eng Technol Ed)* 45:630–638
- Li C, Zhang K, Liu W et al (2020) A person re-identification method based on feature fusion of canonical correlation analysis. *J Optoelectron Laser* 31:500–508

- Okwuashi O, Ndehedehe C, Attai H (2020) Tide modeling using partial least squares regression. *Ocean Dyn* 70(8):1089–1101
- Ramin N, Werner Z, Edwin L, Susanne S (2018) Domain-invariant partial-least-squares regression. *Anal Chem* 90(11):6693–6701
- Zhang K, Dou J (2015) An image feature-matching algorithm based on CCA. *J Yunnan Natl Univ (Nat Sci Ed)* 24(3):244–247
- Wold S, Kettaneh-Wold N, Skagerberg B (1989) Nonlinear PLS modeling. *Chemom Intell Lab Syst* 7:53–65

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.