Multivariate Statistical Analysis in Environmental Process

POSTECH

Dept. Chem. Eng.

PSE Lab.

Contents

- I. Multivariate Analysis
  - MLR
  - **PCA**
  - PCR
  - **PLS**
- II. Application
  - Slurry-Fed Ceramic Melter (SFCM)

Why is the multivariate analysis important in chemical process?

- From DCS(Distributed Control System) etc. , we obtain many correlated data.

↓

How do we treat these data ?

→ Multivariate Analysis

- Monitoring process condition
- Fault detection
- Diagnosis

⇒

- Obtaining stable condition
- Development of the productivity

Chemical Analysis

- Calibration(training) and Prediction(test) steps
  - Find a model for its behavior (Y=f(X))
  - Test the model
- Mean-centering and scaling of variables
  - To make the calculation easier
  - Scaling
    - no scaling (same unit)
    - variance scaling (different unit) → variance =1

Data structure

Underlying Assumptions

Method	Data shape	Assumptions
Classical methods of statistics - MLR	Long and Lean	- X-variables are independent. - X-variables are exact.
Chemometrics - PCA, PLS, PCR	Short and Fat	- X-variables are not independent. - X-variables may have errors.

MLR (Multiple Linear Regression)

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m + e$$

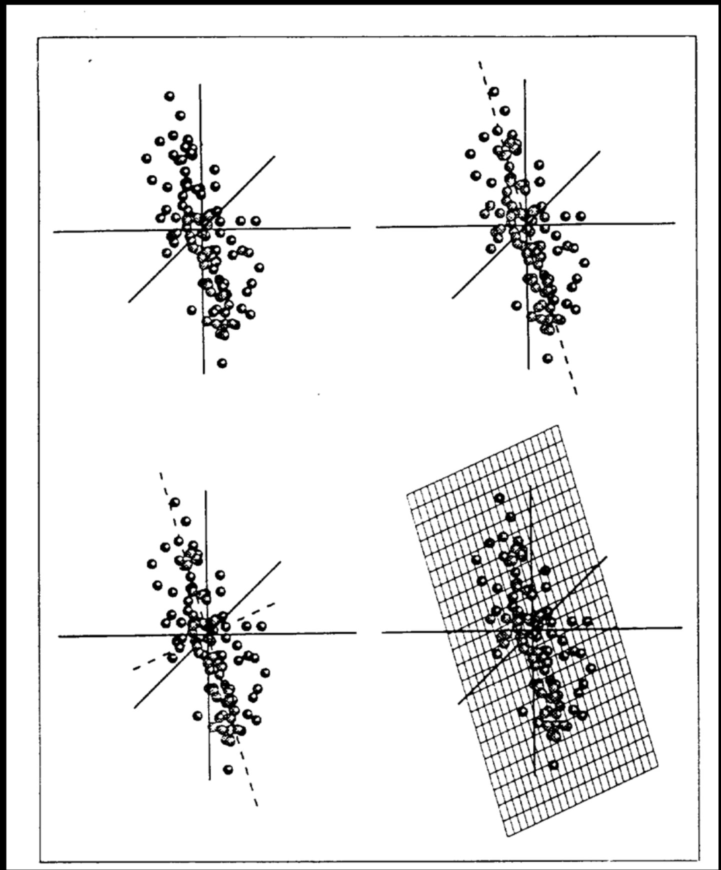
↓ n samples

$$y = Xb + e \quad \Rightarrow \quad \hat{b} = (X'X)^{-1}X'y$$

- Disadvantage
  - For m=n and m<n , the matrix conversion can cause problems

⇒ Multicollinearity of X (zero determinant)
linear function among predictor variables

PCA (Principal Component Analysis)



-Analyze a single block

-Data compression and information extraction

-PCA finds combinations of variables that describe major trends in a data set.

-Think our body!! (We can specify our body with two dimension instead of using three dimension)

Sequence of adapting PCA

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

→ $X_d = X - \bar{X}$

Variance scaled data matrix

$$X_s = X_d D^{-1/2}$$

Covariance matrix

$$S = \frac{1}{n-1} X_d' X_d$$

Correlation matrix

$$R = \frac{1}{n-1}(D^{-1/2} X_d' X_d D^{-1/2})$$

PCA application

Meaning of PCA

$X = M_1 + M_2 + M_3 + \ldots + M_r$

where X is rank r, M_h is rank 1

→

$X = t_1 p_1' + t_2 p_2' + \ldots + t_a p_a'$

$= TP'$

where t_h is score vector

and p_h' is loading vector

Caution :

모든 축들(PCs)과 그들에 대한 정사영값(Score vectors)을 이용하여 시스템을 분석하는 것이 아니라 유일한 a개의 축들과 그것들에 투영된 정사영 값들만을 가지고 그들의 linear combination 으로 시스템을 근사하여 분석하게 된다.

Finding principal components

$X = U\Sigma V' = TP'$

$\therefore T = U\Sigma \, , \; V = P$

Correlated variable x

$Z = P'X$

Uncorrelated variable z

$S = PLP'$ or $P'SP = L$

where L is a diagonal matrix containing the ordered eigenvalues of S and P is unitary matrix whose columns are the normalized eigenvectors of S

$t_h = Xp_h$

Example

$$S = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \longrightarrow$$

$$\lambda_1 = 5.83 \quad e_1' = [0.383, -0.924, 0]$$
$$\lambda_2 = 2.00 \quad e_2' = [0, 0, 1]$$
$$\lambda_3 = 0.17 \quad e_3' = [0.924, 0.383, 0]$$

Principal component $Y_1 = e_1'X$, $Y_2 = e_2'X$, … , $Y_p = e_p'X$

∴ PC is

$Y_1 = 0.383X_1 - 0.924X_2$

$Y_2 = X_3$

$Y_3 = 0.924X_1 + 0.383X_2$

각각의 eigenvalue는 corresponding principal component의 variance가 된다

NIPALS (Nonlinear Iterative Partial Least Squares)

(1) take a vector x_j from X and call it t_h : $t_h = x_j$

(2) calculate $p_h' = t_h' X / t_h' t_h$ ⟵ $X = t_h p_h'$

(3) normalize p_h' to length 1:

$p_{h\ new}' = p_{h\ old}' / \| p_{h\ old}' \|$

(4) calculate t_h: $t_h = X p_h / p_h' p_h$

(5) compare the t_h used in step2 with that obtained in step 4. (iteration until they are same)

$$E_1 = X - t_1 p_1', \; E_2 = E_1 - t_2 p_2', \; \ldots, \; E_h = E_{h-1} - t_h p_h'$$

PCR (Principal Component Regression)

$$Y = XB + E \longrightarrow Y = TB_r + E_r = TP'B + E$$

$$\therefore \hat{B}_r = (T'T)^{-1} T'Y = P'B$$

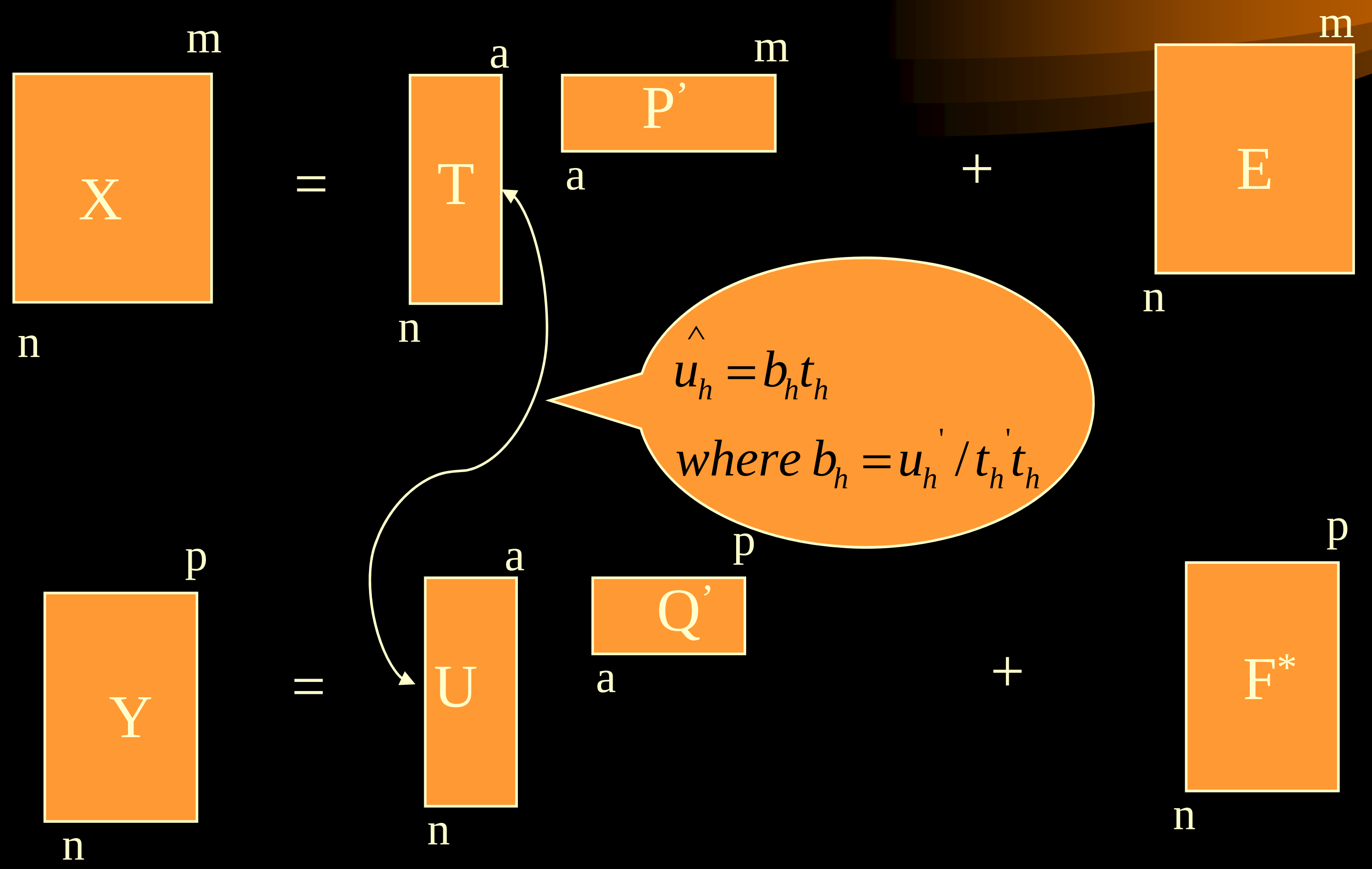
$$\hat{B} = P(T'T)^{-1} T'Y$$

The inversion of T'T gives no problem.

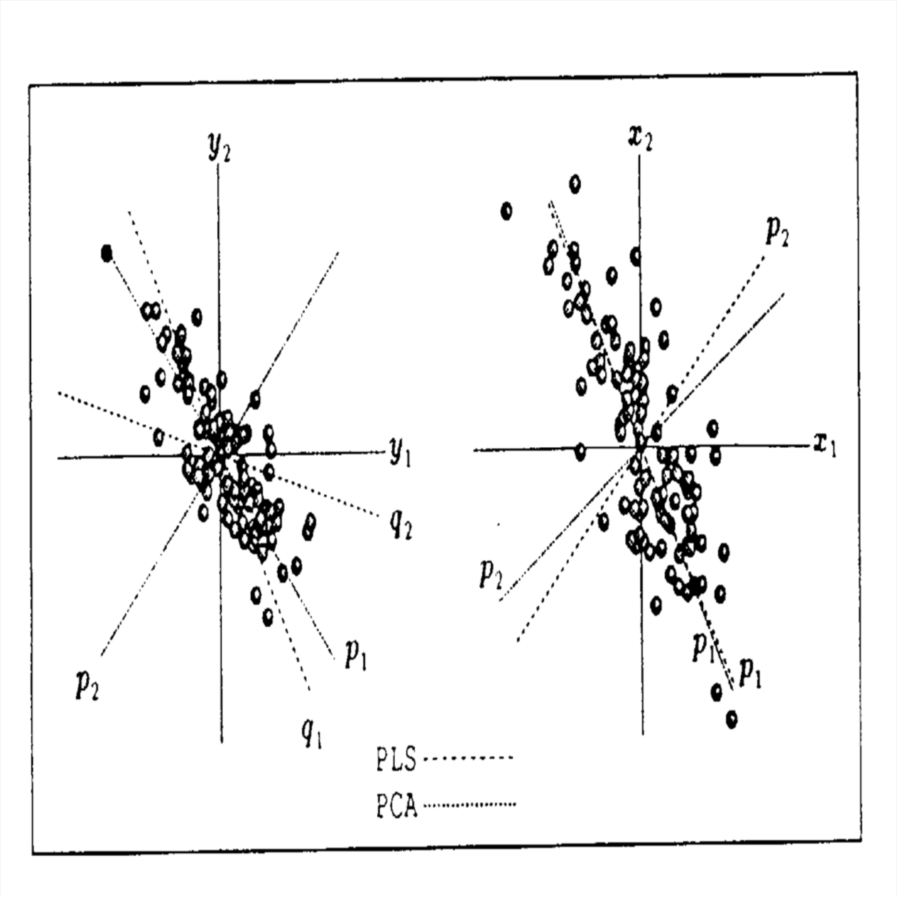
⟶ Solve collinearity problem in MLR

But, we can not say that score vector corresponding first PCs explain Y well, also.

PLS (Partial Least-Squares regression)



Comparison PCA with PLS



- Loading vectors in PCA are orthogonal.
- In PLS, the orthogonality is lost.
- The rotation allows a better model for the relation between two data matrices.

The PLS algorithm

Assume X and Y are mean-centered and scaled

For each component: (1) take u_{start}= some y_j

In the X bolck: (2) $w' = u'X / u'u$ (regress columns of X on u)

(3) $w'_{new} = w'_{old} / ||w'_{old}||$ (normalization)

(4) $t = Xw / w'w$

In the Y block: (5) $q' = t'Y / t't$ (regress columns of Y on u)

(6) $q'_{new} = q'_{old} / ||q'_{old}||$ (normalization)

(7) $u = Y_q / q'q$

Check convergence: (8) compare the t in step 4 with the one from the preceding iteration. If they are equal go to step(9), else go to step(2)

The PLS algorithm (continued)

Calculate the X loadings and rescale the scores and weights accordingly:

(9) $p' = t'X/ t't$ (p' are replaced by weights w')

(10) $p'_{new} = p'_{old}/ ||p'_{old}||$ (normalization)

(11) $t_{new} = t_{old} ||p'_{old}||$

(12) $w'_{new} = w'_{old} ||p'_{old}||$

Find the regression coefficient b for the inner relation:

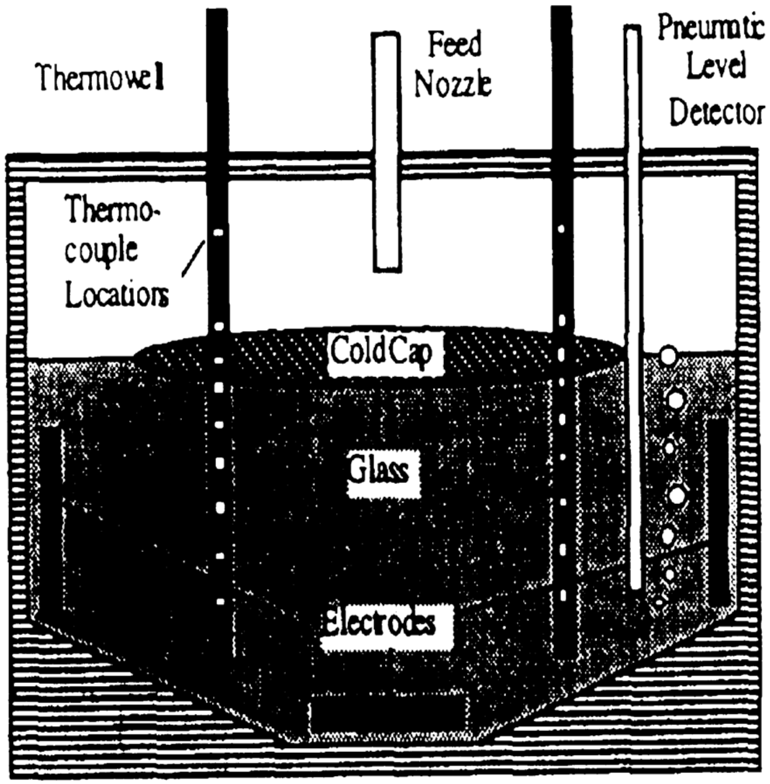
(13) $b= u't/ t't$

Calculation of the residuals

$E_h = E_{h-1} - t_h p_h'$; $X=E_0$

$F_h= F_{h-1}- b_h t_h q_h'$; $Y=F_o$

Application of PCA to chemical process

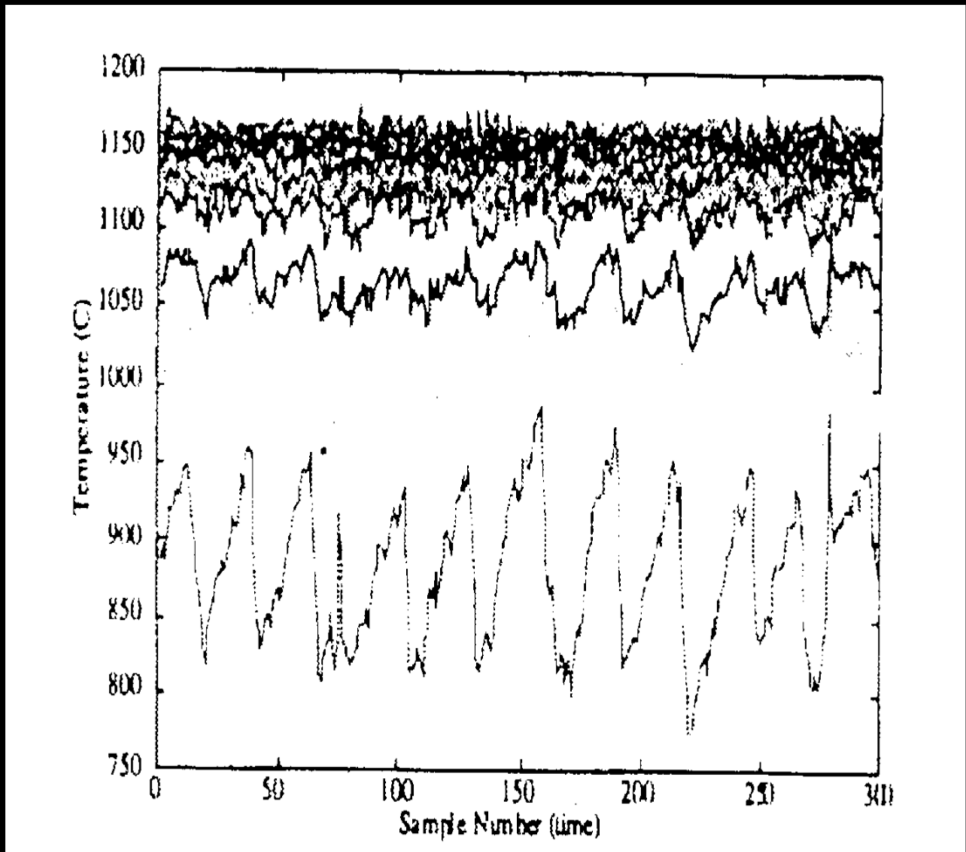


Slurry-Fed Ceramic Melter

nuclear fuel reprocessing wastes

⟶ stable borosilicate glass

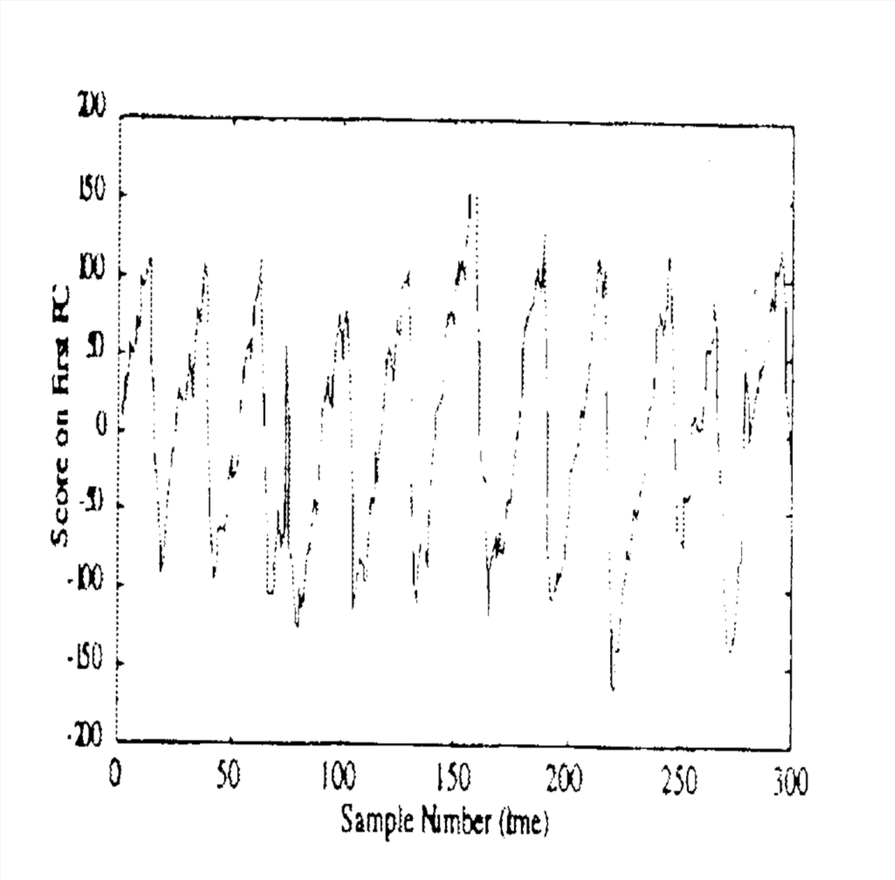
Application of PCA to chemical process (continued)

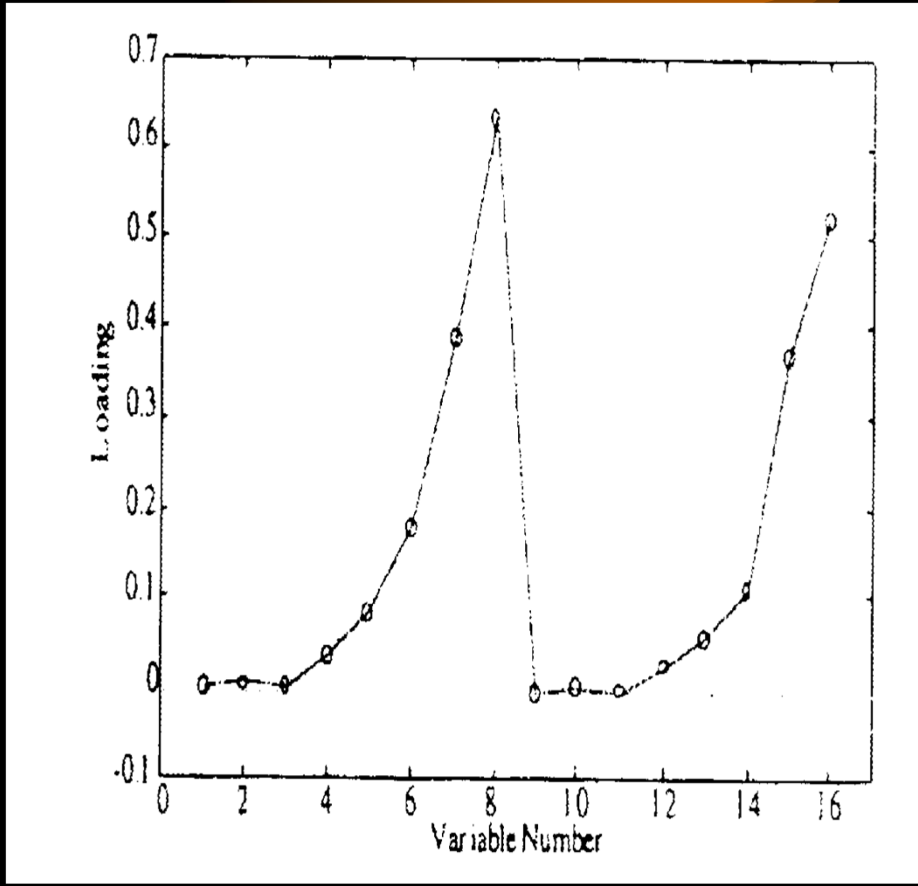


Variance captured by PCA model of SFCM data

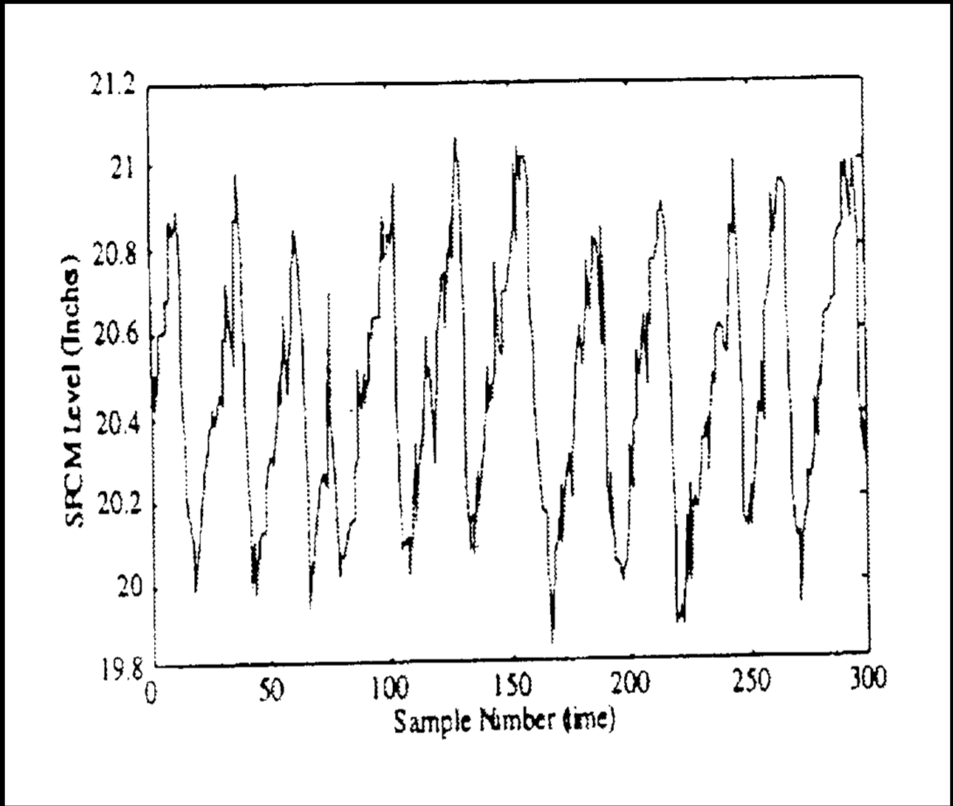
PC number	This PC	Percent variance captured Total
1	88.0711	88.0711
2	6.6974	94.7686
3	2.0442	96.8127
4	0.9122	97.7249
5	0.6693	98.3942
6	0.5503	98.9445
7	0.3614	99.3059
8	0.2268	99.5327

Application of PCA to chemical process (continued)





Application of PCR and PLS



Develop a regression model that relates the temperature to the level of the molten glass

Application of PCR and PLS (continued)