LUT University

**BM20A6100Advanced Data Analysis and Machine Learning** 23.10.2026

Only writing instruments are allowed in the exam.

1. **(5p total)** Choose the correct answer:

a) The rank of **X** is equal to the number of non-zero [select] **1) singular values or 2) eigenvalues or 3) both.**

b) Latent variables are **1) principal components, or 2) PLS components, or 3) both.**

c) In autoscaling the variance of **1) samples, or 2) variables, or 3) both** is set to one.

d) PCA is computed using SVD of mean-centered data: **1) always, 2) never, 3) there are alternative ways.**

e) In PLS the covariance between **1) score vectors *t* and *u*, 2) matrices X and Y,** or **3) eigen matrices** is maximized.

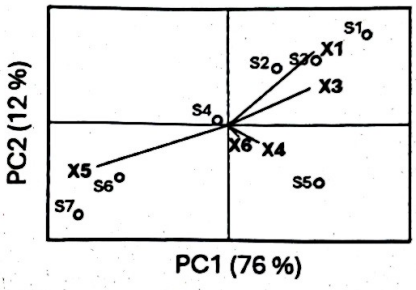
2. **(5p total) Principal Component Analysis.**
In biplot, there are loadings of 5 X variables, and scores of 7 samples.

a) Sample 4 (S4) is near the origin. Does it mean that it is an average sample?

b) Variables X1 and X2 seem to be positively correlated. Is there any other variable correlated with them?

c) X6 is near the origin. Does it mean that it is an average variable or that its variation is not captured by PC1 and PC2?

d) S5 is projected into lower right corner. Can you draw any conclusion about it?




3. **(10p) Regression model validation.**

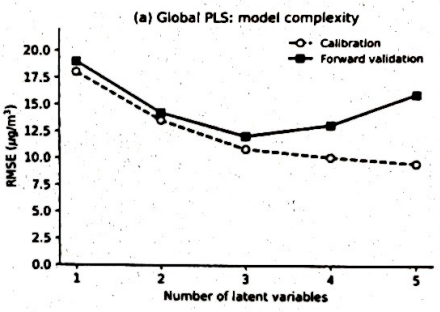
a) The aim is to **keep models as simple as possible** to gain a global model. Explain how this affects various steps of modelling and selection of methods.

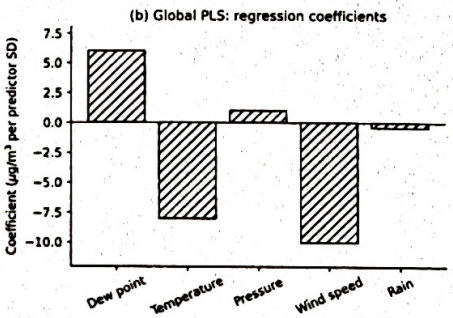
b) Number of latent variables can be defined with cross-validation. Describe the method.

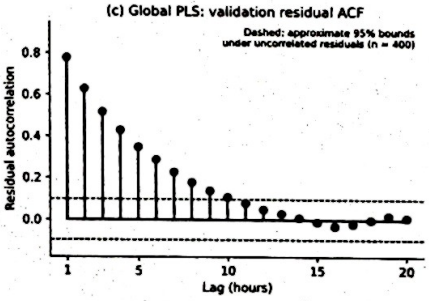
c) Describe validation and considerations with external data.

d) List 3-5 indicators for "a good model".

4. **(10p total) Partial Least-Squares.** An air-quality station records hourly $PM_{2.5}$ concentrations and five meteorological variables. The objective is to forecast $PM_{2.5}$ one hour ahead. The figure presents results from a global PLS model(a-c) and dynamic PLS models (d).




(d) Moving-window PLS: validation RMSE

| Window (h) | 1 LV | 2 LVs | 3 LVs | 4 LVs |
|---|---|---|---|---|
| 24 | 12.0 | 11.1 | 12.4 | 14.0 |
| 72 | 10.1 | 8.5 | 9.0 | 10.8 |
| 168 | 11.3 | 9.6 | 9.4 | 10.0 |

All RMSE values in µg/m³.
Global PLS (3 LVs): 12.0
Persistence forecast: 9.5

a) Based on figure (a), how many latent variables would you select for the global PLS model?

b) Figure (b): Which predictor has the strongest positive coefficient, and which has the strongest negative coefficient? Explain what these coefficients describe and whether they establish causal effects.

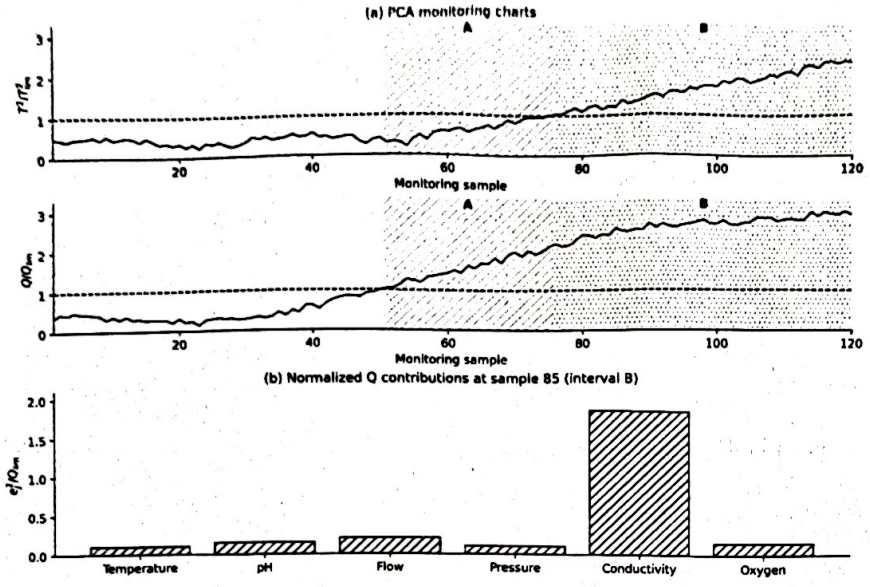
c) Figure (c): what does it suggest about the information remaining in the residuals? Propose one modification to the predictor set, giving a specific example.

d) Select a configuration from figure (d) and compare its validation performance with the global model and the persistence forecast, which predicts that the next concentration equals the current concentration.

LUT Group Confidential - Other information (3Y)

5. **(10p total) Fault Diagnostics.**

a) Explain what $T^2$ and Q represent. Interpret how the monitoring pattern changes from interval A to interval B, including the difference in alarm timing. Explain why monitoring both statistics is useful.




b) Which variable would you investigate first at sample 85, and why? Does the contribution plot establish the physical cause of the alarm? Describe one additional check you would perform.

c) Describe how you would construct the reference PCA model and establish its control limits. Include the choice of reference observations, preprocessing, retained components, and the intended false-alarm rate.

d) Suppose similar Q alarms appear repeatedly when the process moves from summer to autumn, although treatment performance remains acceptable. Give one plausible explanation and explain what evidence you would require before updating the reference model.

6. **(10p total) Advanced methods.** Choose either Scenario A or Scenario B.

**Scenario A – Kernel PCA**

You have measured physical and chemical properties of plant samples. You wish to visualize a map of similarity between different groups of samples, but you do not have the labels for the groups. The variables show nonlinear relationships that the linear PCA does not adequately capture. Explain how kernel PCA could provide a useful low-dimensional representation.

Explain the method and its advantage over the corresponding linear method. Describe its main mathematical principles. Outline the pre-processing, parameters selection, and model assessment.

**OR**

**Scenario B – Kernel PLS**

You want to predict a biochemical property from highly correlated spectral variables. Linear PLS produces systematic curved patterns in validation residuals, even after increasing the number of latent variables. Explain how kernel PLS could improve the model.

Explain the method and its advantage over the corresponding linear method. Describe its main mathematical principles. Outline the pre-processing, parameters selection, and model assessment.

LUT Group Confidential - Other information (3Y)