Η χρήση της μεθοδου PCA στην βιοτεχνολογια

Κωνσταντινα Κυριακιδου

May 17, 2025

1 Εισαγωγή

hemometrics as a separate field has grown considerably in the past 4 decades. One of the popular definitions is that from Wold—"How to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data.1" Chemometrics is being increasingly used in both basic research and applied scientific fields and has enabled the diagnostic evaluation of parameter interactions that were previously undefined. For complex datasets, univariate or bivariate analysis is often inefficient and is likely to result in misinterpretation of data.2-4 Use of projection methods can effectively deal with challenges such as multidimensionality of the data set, multicollinearity, missing data, and variability from experimental error and noise.5

The chemical industry was early in recognizing and adopting chemometrics as a quick and economical method of extracting real-time information from data and, thus, leading to improved process monitoring and control. Visible spectroscopy, near-infrared (NIR) spectroscopy, mid-infrared (MIR) spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, and Fourier transform infrared (FTIR) spectroscopy are some of the commonly used process analyzers that have been used in the chemical industry. Principal component analysis (PCA), partial least squares (PLS) regression, principal component regression (PCR), canonical variable analysis (CVA), and modified soft independent modeling of the class analogy (SIMCA) are some of the statistical tools that have been used to facilitate analysis and modeling of the abundant data that are provided by the aforementioned process analyzers. PCA has been used to transform 12 cross-correlated variables of an activated sludge wastewater treatment into three uncorrelated principal components, which were able to account for

$$\text{Total Variance Explained} = 78$$

of the system variability. This enabled easier analysis, monitoring, and diagnosis of the system.6 Combination of visible (Vis) and NIR spectroscopy and chemometrics has been used for discrimination between samples of Australian commercial white wines of different varietal origins—Chardonnay and Riesling.7 Models developed using PCA, PCR, and discriminant PLS regression gave an excellent discrimination between samples of the two varietal origins under consideration with an accuracy up to 98

The use of chemometrics in the pharmaceutical industry has been relatively more recent with a lot of work being done in the last 2 decades. NIR spectroscopy has emerged as the analyzer of choice for a wide range of applications. NIR, in combination with PCA and SIMCA, has been used for routine testing of packed pharmaceutical substances directly in warehouses using a fiber optic probe.13 The proposed approach allowed for measurements through closed polyethylene bags, and a trichotomy classification procedure was proposed:

$$\text{Class}(x) = \begin{cases} \text{Acceptable}, & \text{if } d(x) \leq \tau_1 \\ \text{Uncertain}, & \text{if } \tau_1 < d(x) \leq \tau_2 \\ \text{Reject}, & \text{if } d(x) > \tau_2 \end{cases}$$

which allowed an operator to identify raw materials of satisfactory quality. NIR, in combination with

PCA and SIMCA, has also been used for the identification of counterfeit drugs.14 The tool was capable of identifying subtle alterations in drug composition with 100 accuracy and could be used as a tool for rapid, nondestructive testing. Other NIR-based applications include analysis of high-shear granulation, characterization and determination of azithromycin polymorphs, and determination of the crystalline form present in amorphous miokamycin.15-17 UV–vis spectroscopy, attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy, infrared imaging, and Raman spectroscopy are some of the other commonly used process analyzers in the pharmaceutical industry. NIR transmittance and Raman spectroscopy with PLS have been used for the prediction of active substance content in tablets.18 Successful calibrations were developed with prediction errors less than 3.7 and comparable to the error of the chromatographic reference method. PLS of absorbance spectra has been used for simultaneous determination of benzyl alcohol and diclofenac in pharmaceutical formulations.19 The proposed method was found to be simple, precise, inexpensive, requiring no complex pretreatment, and thus an effective means for quality control of pharmaceutical products.

The last decade has seen a flurry of activities in the area of development of chemometrics applications for the biopharmaceutical processes.20, 21 They include analysis of NIR spectral information for an antibiotic production process, multivariate statistical process monitoring for processing of pharmaceutical granules, the assessment of seed inoculum quality from a manufacturing process, and development of an integrated on-line multivariate statistical process monitoring, product attributes prediction, and fault diagnosis framework for a fed-batch penicillin fermentation.22-25 A flexible process monitoring method has been applied for analysis of pilot plant cell culture data for fault detection and diagnosis.26 A PCA model was constructed from 19 batches, and the model was shown to successfully detect abnormal process conditions and diagnose root causes. Feasibility of using chemometrics for supporting key activities required for successful manufacturing of biopharmaceutical products, including scale-up, process comparability, process characterization, and fault diagnosis, has been examined.27 Representative data from small-scale (2 L) and large-scale (2,000 L) cell culture batches were analyzed in this study. Scores plots, loadings plots, and variable importance for the projection (VIP) plots were utilized for assessing scale-up and comparability of the cell culture process. Batch control charts were also shown to be useful for fault diagnosis during routine manufacturing. A subsequent publication examined the usefulness of chemometrics for root cause analysis to identify scale-up differences and parameter interactions that adversely impacted cell culture process performance.28 We will discuss this application in more detail in the next section. In an even more recent publication, NIR-PCA was effectively used for screening of lots of basal medium powders based on their impact on process performance and product attributes for a cell culture process.29 These lots had identical composition as per the supplier and were manufactured at different scales using an identical process. A combined NIR-PCA approach made it possible to fingerprint the raw materials and to distinguish between the good and poor performing media lots.

In the following sections, we review chemometrics applications in upstream and downstream processing of biotech products. It will be evident that because of the complexity present in most biotech unit operations, implementation of quality by design (QbD) and process analytical technology (PAT) initiatives will require an increased use of chemometric tools and approaches.30-35 We wish to provide to the readers a summary of major research that has been done on this topic recently.

2 Related Work

Chemometric tools have been used in the cell culture operations in the last decade as summarized in Table 1. PCA has been used for detection and diagnosis of abnormal process conditions in an industrial fed-batch cell culture process. The model was successfully able to detect abnormal process conditions, which resulted from three known fault types, namely

irregular thermal heating, elevated dissolved oxygen values, and large variation in agitation.26 PLS calibration models of NIR spectra have been utilized for the measurement of glucose, lactate, glutamine, and ammonia in undiluted serum-based cell culture media.37 Robust, analyte-specific models were generated, and the low values of standard errors of prediction for each analyte demonstrate that the models can be used to (off-line) determine the important nutrient and byproduct content in a serum-based cell culture medium. A novel PLS approach called evolving PLS has been compared with the traditional PLS using data from an industrial fed-batch mammalian cell culture process for prediction of intermediate and final quality variable values.40 Use of in situ 2D fluorometry in combination with chemometrics has been evaluated for monitoring the concentration of viable cells and the concentration of recombinant proteins in mammalian cell culture.42 PCA was used to filter the large volumes of redundant spectral data, while PLS correlated the reduced data with the target state variables. Both viable cells density and glycoprotein concentration were accurately estimated, which strongly suggests that the combination of 2D fluorometry with suitable chemometric techniques is a consistent technique for monitoring of a cell culture medium.

3 Περιγραφή της μεθόδου

Figure 1 illustrates an application involving use of chemometrics for root cause analysis to identify scale-up differences and parameter interactions that adversely impacted cell culture process performance.29 Data from a total of 171 batches run at small scale (2 L), pilot scale (2,000 L), and commercial scale (15,000 L) were used in this analysis. Input parameters examined included continuous on-line measurements of operating parameters (e.g., pH, dissolved oxygen, and temperature), daily measurements of dissolved CO2, metabolic indicators, and cell growth parameters. The output parameters included product attributes such as product titer, viable cell density, cell viability, and osmolality. Time course performance variables (daily, initial, peak, and end point) were also evaluated. A total of 119 output variables from raw materials, product attributes, time course, and seed inocula trains were evaluated using chemometrics. The VIP plot in Figure 1A summarizes the observations made from the score and loading plots by showing the relative importance of each included variable in the analysis. It is observed that large-scale raw material type and culture metabolism evolution (e.g., elevated lactate and osmolality) exert the greatest influence on product attributes. These two parameters were further examined for their impact and interactions. Next, a score scatter plot (shown in Figure 1B) was generated based on data analysis. It provides a visual summary of the process behavior over time, with the score vectors for the first two principal components, t[1] and t[2], plotted against each other. It is seen in Figure 1B that a correlation exists between raw material type and product attributes. On the basis of the chemometric analysis, the authors were able to conclude that the root cause of process underperformance was the combination of large-scale media and high pCO2 conditions in the large-scale bioreactor. This hypothesis was later confirmed by experiments performed at pilot scale using the different types of raw materials and CO2 concentrations

4 Συμπεράσματα

This article reviews chemometric applications in biotech processes. We hope to have demonstrated that chemometric tools can be very useful and powerful in extracting useful process information through analysis of the readily available data in order to maximize process understanding. In view of the complexity of biotech processes and products, we expect chemometrics to continue to act as a significant enabler of the QbD and PAT initiatives.

αναφορες απο [8] [6] [2] [7] [5] [3] [1] [4] [9] [10] διαφορες ερευνες

References

[1] M Usman Ali, Shahzad Ahmed, Javed Ferzund, Atif Mehmood, and Abbas Rehman. Using pca and factor analysis for dimensionality reduction of bio-informatics data. *arXiv preprint arXiv:1707.07189*, 2017.

[2] Nitish Bhushan, Sandip Hadpe, and Anurag S Rathore. Chemometrics applications in biotech processes: assessing process comparability. *Biotechnology progress*, 28(1):121–128, 2012.

[3] Fu-Hsiang Chen, Der-Jang Chi, and Yi-Cheng Wang. Detecting biotechnology industry's earnings management using bayesian network, principal component analysis, back propagation neural network, and decision tree. *Economic Modelling*, 46:1–10, 2015.

[4] Yajun Chen. Multivariate data analysis for biotechnology & bio-processing. *Journal of Commercial Biotechnology*, 27(4):197–204, 2022.

[5] Ying Hou, Canping Jiang, Abhinav A Shukla, and Steven M Cramer. Improved process analytical technology for protein a chromatography using predictive principal component analysis tools. *Biotechnology and bioengineering*, 108(1):59–68, 2011.

[6] Rob Johnson, Oliver Yu, Alime Ozlem Kirdar, Alagappan Annamalai, Sanjeev Ahuja, Kripa Ram, and Anurag S Rathore. Applications of multivariate data analysis in biotech processing. *BioPharm International*, 20(10):130, 2007.

[7] Alime Ozlem Kirdar, Jeremy S Conner, Jeffrey Baclaski, and Anurag S Rathore. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnology progress*, 23(1):61–67, 2007.

[8] Anurag S Rathore, Nitish Bhushan, and Sandip Hadpe. Chemometrics applications in biotech processes: a review. *Biotechnology progress*, 27(2):307–315, 2011.

[9] Anurag S Rathore, Shachi Mittal, Scott Lute, and Kurt Brorson. Chemometrics applications in biotechnology processes: predicting column integrity and impurity clearance during reuse of chromatography resin. *Biotechnology progress*, 28(5):1308–1314, 2012.

[10] Jong Il Rhee, Tae-Hyoung Kang, Kum-Il Lee, Ok-Jae Sohn, Sun-Yong Kim, and Sang-Wook Chung. Application of principal component analysis and self-organizing map to the analysis of 2d fluorescence spectra and the monitoring of fermentation processes. *Biotechnology and bioprocess engineering*, 11:432–441, 2006.