In this paper we examine 2- and 3-way chemometric methods for analysis of Arctic and Antarctic water samples. Standard CTD (conductivity–temperature–depth) sensor devices were used during two oceanographic expeditions (July 2007 in the Arctic; February 2009 in the Antarctic) covering a total of 174 locations. The output from these devices can be arranged in a 3-way data structure (according to sea water depth, measured variables, and geographical location). We used and compared 2- and 3-way statistical tools including PCA, PARAFAC, PLS, and N-PLS for exploratory analysis, spatial patterns discovery and calibration. Particular importance was given to the correlation and possible prediction of fluorescence from other physical variables. MATLAB’s mapping toolbox was used for geo-referencing and visualization of the results. We conclude that: 1) PCA and PARAFAC models were able to describe data in a satisfactory way, but PARAFAC results were easier to interpret; 2) applying a 2-way model to 3-way data raises the risk of flattening the covariance structure of the data and losing information; 3) the distinction between Arctic and Antarctic seas was revealed mostly by PC1, relating to the physico-chemical properties of the water samples; and 4) we confirm the ability to predict fluorescence values from physical measurements when the 3-way data structure is used in N-way PLS regression.
Automatic scatter detection in fluorescence landscapes by means of spherical principal component analysis

In this paper, we introduce a new method, based on spherical principal component analysis (S-PCA), for the identification of Rayleigh and Raman scatters in fluorescence excitation–emission data. These scatters should be found and eliminated as a prestep before fitting parallel factor analysis models to the data, in order to avoid model degeneracies. The work is inspired and based on a previous research, where scatter removal was automatic (based on a robust version of PCA called ROBPCA) and required no visual data inspection but appeared to be computationally intensive. To overcome this drawback, we implement the fast S-PCA in the scatter identification routine. Moreover, an additional pattern interpolation step that complements the method, based on robust regression, will be applied. In this way, substantial time savings are gained, and the user's engagement is restricted to a minimum, which might be beneficial for certain applications. We conclude that the subsequent parallel factor analysis models fitted to excitation–emission data after scatter identification based on either ROBPCA or S-PCA are comparable; however, the modified method based on S-PCA clearly outperforms the original approach in relation to computational time.

General information
Publication status: Published
Organisations: Department of Applied Mathematics and Computer Science, Statistics and Data Analysis, National Food Institute, Division of Industrial Food Research
Contributors: Kotwa, E. K., Jørgensen, B. M., Brockhoff, P. B., Frosch, S.
Pages: 3-11
Publication date: 2013
Peer-reviewed: Yes

Publication information
Journal: Journal of Chemometrics
Volume: 27
Issue number: 1-2
ISSN (Print): 0886-9383
Ratings:
BFI (2013): BFI-level 2
Scopus rating (2013): CiteScore 2.2 SJR 0.757 SNIP 1.307
Web of Science (2013): Impact factor 1.803
ISI indexed (2013): ISI indexed yes
Original language: English
Keywords: S-PCA, Raman and Rayleigh scatters, Robustness, PARAFAC, Fluorescence
DOIs:
10.1002/cem.2485
Source: dtu
Source ID: n:oai:DTIC-ART:wiley/380254085::34756
Robust procedures in chemometrics

The general aim of the thesis was to contribute to the improvement of data analytical techniques within the chemometric field. Regardless of the multivariate structure of the data, it is still common in some fields to perform uni-variante data analysis using only simple statistics such as sample mean and variance. Recent instrumental developments in chemometrics often result in high-order data, for which uni-variante tools do not suffice and multivariate data analysis is required. Moreover, many multivariate models assume normality of the residuals (which in many cases is far from reality) and are not resistant towards outliers (which are known to be more the rule than the exception for empirical data). That is the reason for robust methods being a valuable tool for both semi-automated detection of outliers and model building.

The approach adapted in this thesis, can be split in two main parts: 1. applying a multivariate and multi-way data analytical frame-work in fields where less sophisticated data analysis methods are currently used, and 2. developing new, more robust alternatives to already existing multivariate tools.

The first part of the study was realised by applying two- and three-way chemometrical methods, such as PCA and PARAFAC models for analysing spatial and depth profiles of sea water samples, defined by three data modes: depth, variables and geographical location. Emphasis was also put on predicting fluorescence values, as being a natural measure of biological activity, by applying and comparing the Partial Least Squares (PLS) regression technique with its multi-way alternative, N-PLS. Results of the analysis indicated superiority of the three-way frame-work, potentially constituting a novel assessment of the sea water measurements. Particularly in the case of regression models there is a clear preference towards the more complex model, delivering more reliable predictions than a classical 2-way PLS. Therefore, using multi-way data analysis tools is recommended, in order to extract the full information from multi-way data structures.

The second part of the thesis targeted qualitative properties of the analysed data. The broad theoretical background of robust procedures was given as a very useful supplement to the classical methods, and a new tool, based on robust PCA, aiming at identifying Rayleigh and Raman scatters in excitation-mission (EEM) data was developed. The results show clearly that robust methods can significantly contribute to the improvement of existing analytical techniques used commonly in chemometrics, for example by providing excellent outlier detection tools. It is therefore advised to apply robust and classical procedures simultaneously, at least to determine if contamination in the data is present. For this becoming a standard procedure, further work is required, aiming at implementing reliable robust algorithms into standard statistical programs.

General information
Publication status: Published
Organisations: Mathematical Statistics, Department of Informatics and Mathematical Modeling
Contributors: Kotwa, E.
Number of pages: 174
Publication date: 2012

Publication information
Place of publication: Kgs. Lyngby
Publisher: Technical University of Denmark
Original language: English
(IMM-PHD-2012; No. 284).
Electronic versions:
phd284_Kotwa_E.pdf

Research output: Contribution to journal › Journal article – Annual report year: 2013 › Research › peer-review

Spatio-temporal analysis and modeling of short-term wind power forecast errors

Forecasts of wind power production are increasingly being used in various management tasks. So far, such forecasts and related uncertainty information have usually been generated individually for a given site of interest (either a wind farm or a group of wind farms), without properly accounting for the spatio-temporal dependencies observed in the wind generation field. However, it is intuitively expected that, owing to the inertia of meteorological forecasting systems, a forecast error made at a given point in space and time will be related to forecast errors at other points in space in the following period. The existence of such underlying correlation patterns is demonstrated and analyzed in this paper, considering the case-study of western Denmark. The effects of prevailing wind speed and direction on autocorrelation and cross-correlation patterns are thoroughly described. For a flat terrain region of small size like western Denmark, significant correlation between the various zones is observed for time delays up to 5 h. Wind direction is shown to play a crucial role, while the effect of wind speed is more complex. Nonlinear models permitting capture of the interdependence structure of wind power forecast errors are proposed, and their ability to mimic this structure is discussed. The best performing model is shown to explain 54% of the variations of the forecast errors observed for the individual forecasts used today. Even though focus is on 1-h-ahead forecast errors and on western Denmark only, the methodology proposed may be similarly tested on the cases of further look-ahead times, larger areas, or more complex topographies. Such generalization may not be straightforward. While the results presented here comprise a first step only, the revealed error propagation principles may be seen as a basis for future related work.
Projects:

**Analysis and Modelling of Chain Data**
Kotwa, E. K., PhD Student, Department of Informatics and Mathematical Modeling
Brockhoff, P. B., Main Supervisor
Kulahci, M., Examiner
Rinnan, Å., Examiner
Westad, F. O., Examiner
Globaliseringsmidler
01/10/2007 → 05/09/2014
Award relations: Analysis and Modelling of Chain Data
Project: PhD