Chemometrics and Intelligent Laboratory Systems 161 (2017) 118–129

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

On-The-Fly Processing of continuous high-dimensional data streams

Raffaele Vitale[a,*], Anna Zhyrova[b], João F. Fortuna[c], Onno E. de Noord[d], Alberto Ferrer[a], Harald Martens[c,e]

[a] Grupo de Ingeniería Estadística Multivariante, Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
[b] FFPW and CENAKVA, Institute of Complex Systems, University of South Bohemia in Ceske Budejovice, Zàmek 136, 37333 Novè Hrady, Czech Republic
[c] Department of Engineering Cybernetics, Faculty of Information Technology, Mathematics and Electrical Engineering, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[d] Shell Global Solutions International B.V., Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN Amsterdam, The Netherlands
[e] Idletechs AS, NTNU Innovation Centre, Richard Birkelansvei 2B, 7491 Trondheim, Norway

ARTICLE INFO

Keywords:
On-The-Fly Processing (OTFP)
Bilinear modelling
High-dimensional data streams
Generalised Taylor expansion
Singular Value Decomposition (SVD)
BIG DATA analytics

ABSTRACT

A novel method and software system for rational handling of time series of multi-channel measurements is presented. This quantitative learning tool, the *On-The-Fly Processing* (OTFP), develops reduced-rank bilinear subspace models that summarise massive streams of multivariate responses, capturing the evolving covariation patterns among the many input variables over time and space. Thereby, a considerable data compression can be achieved without significant loss of useful systematic information.

The underlying proprietary OTFP methodology is relatively fast and simple – it is linear/bilinear and does not require a lot of raw data or huge cross-correlation matrices to be kept in memory. Unlike conventional compression methods, the approach allows the high-dimensional data stream to be graphically interpreted and quantitatively utilised – in its compressed state. Unlike adaptive moving-window methods, it allows all past and recent time points to be reconstructed and displayed simultaneously.

This new approach is applied to four different case-studies: (i) multi-channel Vis-NIR spectroscopy of the Belousov–Zhabotinsky reaction, a complex, ill understood chemical process; (ii) quality control of oranges by hyperspectral imaging; (iii) environmental monitoring by airborne hyperspectral imaging; (iv) multi-sensor process analysis in the petrochemical industry. These examples demonstrate that the OTFP can automatically develop high-fidelity subspace data models, which simplify the storage/transmission and the interpretation of more or less continuous time series of high-dimensional measurements – to the extent there are covariations among the measured variables.

1. Introduction

1.1. The modern data issue

Many modern measurement technologies generate massive amounts of data in a very short time – e.g. continuous streams of high-dimensional data via one-step analytical procedures.[1] For instance:

- modern spectrometers can deliver hundreds of informative, high-dimensional spectra per second;
- hyperspectral cameras produce multivariate spatially resolved images. In addition, when configured in a time-lapse mode, they can yield continuous streams of high-dimensional spatiotemporal recordings;
- industrial monitoring for condition-based maintenance, as well as the control of complex dynamic processes, requires high-dimensional inputs to be sufficiently informative;
- computer experiments, needed in order to study the behaviour of complex mathematical models, involve advanced workstations performing thousands of simulations, each one possibly characterised by just as many input and output properties.

Hence, a measurement revolution (recently termed *data tsunami* [1]) is currently taking place in numerous fields of applied science, ranging from analytical chemistry and medicine to environmental

* Corresponding author.
E-mail address: rvitale86@gmail.com (R. Vitale).
[1] Contrary to unstructured data from e.g. free text, they are systematically recorded and are here referred to as quantitative data.

http://dx.doi.org/10.1016/j.chemolab.2016.11.003
Received 31 May 2016; Received in revised form 5 November 2016; Accepted 14 November 2016
Available online 06 December 2016
0169-7439/

surveillance, informatics and industrial *Internet of Things* (IoT). However, these incredibly quick advances run the risk of being practically useless for three reasons:

- The human ability to grasp content of interest from data remains fairly constant, and data simplification is therefore desirable for interpretative purposes. Here, one possible solution could be the removal of irrelevant descriptors among the available ones. Nevertheless, for most applications their identification is not straightforward which makes such a simplification risky and complicated;
- Despite Moore's first law [2], which predicts a continuous exponential increase for both computer processing speed and storage capacity along time, it is estimated that in the near future they will not be sufficient for coping with this ongoing *data explosion*. For instance, IoT threatens to flood both communication channels and the users' cognitive capacity with overwhelming torrents of repetitive, more or less redundant data;
- Traditional computing systems are generally not capable of performing analytics on constantly streaming data, typical of today's world of multimedia communication [3].

In a scenario like this, if it were possible to simultaneously compress and model high-dimensional measurement series as they flow from e.g. an analytical platform and without significant loss of useful information content, their storage, transfer, retrieval, visualisation and interpretation would be radically eased. The present paper illustrates a feasible approach to achieve this goal.

1.2. Data compression strategies

Data compression plays a central role in telecommunications and many other scientific and technological branches of interest [4]. According to the nature and features of the algorithmic procedure through which it is performed, it can be defined as either *lossless* or *lossy*. Lossless methods utilise statistical distribution properties and simple patterns in the data for compression, converting the inputs into compressed bit series.[2]

Lossy compression techniques – e.g. the various dedicated versions of JPEG and MPEG methods used for digital image, video and sound compression – approximate the main, perceptible variations in the input data by local *ad hoc* patterns, filtering out less perceptible variation types and noise. Lossy approaches are commonly much more efficient (in terms of compression rate) than lossless ones, like *algebraic* zipping, but allow the original input to be only roughly restored. Moreover, when set to compress too much, they not only cause loss of valid information (resulting in e.g. image blurring or loss of high-frequency sound), but can also introduce undesired decoding artefacts (e.g. visible block effects or audible errors).

Whether lossless or lossy compression methods are used, the compressed data are represented by *per se* meaningless streams that cannot be directly used for quantitative calculations, mathematical modelling or graphical representation.

The novelty of the developed *On-The-Fly Processing* (OTFP) tool is represented by the fact that a hitherto under-utilised source of redundancy (the intercorrelation usually evolving in multi-channel data streams) is mathematically modelled to prevent significant loss of useful systematic information carried by the original measurements. Based on the model's automatically estimated parameters the data stream may be interpreted and utilised for prediction, forecasting and fault detection in the compressed state. The idea behind this strategy was recently outlined in [5]. Here, more algorithmic details will be given and its applicability to different types of high-dimensional data streams demonstrated.

Conceptually, the OTFP system may be motivated by the following thought experiment: assume that a space probe has to be constructed and sent out to explore – for the first time – the unknown geological properties of the hidden back side of a remote planet, using a multi-wavelength camera. Prior to the launch, scarce knowledge about this planet is available to design the ideal instrument, and after the probe has landed, it is too late to change anything. Which wavelength should be chosen, and how should the imaging data be transmitted back to Earth? Some individual wavelengths distinguishing between already known, earthly rock types might be included. But possible geological *surprises* should also be taken into account. Therefore, it is decided to equip the probe camera with a wide spectral range detector, capable of measuring e.g. 1000 different wavelength channels. However, the limited communication bandwidth then becomes a problem: the probe cannot transmit all those measurements for every point in time and space. What would be the best way to send spectral data back to Earth? Perhaps, could that be automatically settled on-the-fly by the space probe's computer itself, based on what its camera measures? The on-board computer could be programmed to discover, compress and transmit the essence of all the recorded images, in a continuous learning-and-communicating process that never sends the same information twice. But how to quantify this compact spectral essence comprehensively? To understand the unknown geological landscape, a reliable approximation of the spectral profile of every pixel in every image, with as many spectral and spatial details and as few artefacts as possible, is needed. A lossless multivariate spectral preprocessing followed by a continuously developing bilinear compression/classification model could deliver a compact summary of the sequence of hyperspectral image data, which would yield maximal insight here on Earth from the limited quantity of received data. The first three application examples described below will illustrate this, albeit in more mundane settings.

1.3. Subspace compression

The OTFP is based on evolving bilinear subspace modelling. The software automatically detects systematic patterns of covariation in the data and use these to model the data mathematically. Subspace projection and dimensionality reduction techniques based on bilinear models, e.g. Principal Component Analysis (PCA), constitute one of the possible ways to compress and approximate a certain set of data, removing simultaneously both statistical redundancy and uninformative noise. Their basic principles can be summarised as follows: let $j = 1, 2, \ldots, J$ be the number of input channels (e.g. J wavelengths of light per pixel in a hyperspectral camera, J sensor variables monitored during a dynamic industrial process or J metabolites quantified in biological samples) recorded for each of $n = 1, 2, \ldots, N$ measurements performed, for instance, on N objects on a conveyor belt, at N spatial locations, N time steps or N different experimental conditions. In the present-day instrumental context, outlined in Section 1.1, where J might be very large, the useful information carried by such data structures ($N \times J$ matrices) is usually intercorrelated among various input channels over the continuously growing set of registered measurements. In these circumstances, for a chosen degree of acceptable accuracy (e.g. depending on the amount of data variance explained), it is possible to reduce the J-dimensional space of the original descriptors to an A-dimensional subspace, onto which all the N objects under study can be projected and represented as new points. *Prima facie*, as $A < J$, this projection can be regarded as a compression operation, whose efficiency is related to the ratio $\frac{A}{J}$.

[2] Most of the lossless compression approaches, such as standard file *zipping*, recodes the original input by using shorter bit sequences for *probable* (e.g. often encountered) data and larger ones for *improbable* (e.g. rare) data.