Vitale, R.; Zhyrova, A.; Fortuna, JF.; De Noord, OE.; Ferrer, A.; Martens, H. (2017). On-The-Fly Processing of continuous high-dimensional data streams. Chemometrics and Intelligent Laboratory Systems. 161:118-129. doi:10.1016/j.chemolab.2016.11.003

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, 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*

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.

Keywords: On-The-Fly Processing (OTFP), bilinear modelling, high-dimensional data streams, generalised Taylor expansion, Singular Value Decomposition (SVD), big data analytics

*Corresponding author:
Telephone number: +34684099819
Email address: rvitale86@gmail.com (Raffaele Vitale)

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[i]. 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 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.

[i]Contrary to unstructured data from e.g. free text, they are systematically recorded and are here referred to as quantitative data.

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[ii].

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 should 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

[ii] 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.

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 $\frac{A}{J}$.

1.4. PCA bilinear structure model

The well known PCA bilinear approximation of a generic $N \times J$ matrix of observed data, $\mathbf{X}$, can be described by the following structure model:

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{1}$ ($N \times 1$) is a vector of ones, $\mathbf{m}$ ($J \times 1$) contains a typical profile, e.g. the mean values of the J input variables in $\mathbf{X}$, $\mathbf{P}$ ($J \times A$) is a matrix of *loadings* associated to such input variables, which determine the A basis vectors or *components* of the PCA subspace, $\mathbf{T}$ ($N \times A$) defines the projection coordinates or *scores* of all the N considered objects (locations, time points or experimental conditions) on this lower-dimensional space and $\mathbf{E}$ ($N \times J$) represents the matrix of unmodelled residuals, i.e. the portion of $\mathbf{X}$ not *explained* by the model at the chosen rank, A [6]. The PCA solution may be formulated in different, equivalent ways. Here, it is assumed to show the following properties:

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I} \tag{2}$$

$$\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathrm{diag}(\boldsymbol{\lambda}_A) \tag{3}$$

where $\mathbf{I}$ is an identity matrix of dimensions $A \times A$, while the a-th element of $\boldsymbol{\lambda}_A$ $(A \times 1)$ corresponds to the eigenvalue of the a-th PCA component.
One of the most critical point when deriving the PCA approximation of a set of data is how to choose the A components of its subspace to prevent losing important portions of useful information and to filter out uniquely statistical redundancy and uninteresting noise. Some of these A dimensions may sometimes be defined according to prior knowledge of the investigated system. For instance, the number of known chemical constituents of mixtures characterised by spectroscopic methods might be appealed to for this purpose. However, in cases like this, also more or less *unexpected* constituents and/or physical phenomena may affect the performed measurements, generating new patterns of variation and thus new subspace dimensions which need to be retained for a proper data approximation and interpretation. Therefore, at least to a certain extent, the identification of the new basis vectors associated to these unforeseen sources of variability has to be carried out through a preliminary exploratory analysis of the available empirical records.
If a continuous data stream is dealt with and N rapidly grows over time, correctly determining new possible subspace dimensions is even more complex: new, unexpected patterns of covariation may spring up in the information flow. Therefore, in such situations, it becomes crucial to automatically recognise when the set of initial basis vectors needs to be reestimated and extended, and to address this task in a statistically valid and computationally efficient way.

1.5. PCA as a multivariate series expansion of the underlying data generation mechanism

As outlined in [7–9], the bilinear PCA model can be thought of as a Taylor expansion of the function $\boldsymbol{f}$ defining how the measurement descriptors are jointly related to their common structure. For instance, for each of the J aforementioned input channels, one can envision a local linear approximation of the underlying (unknown) causal phenomena driving their evolution. Mathematical summary modelling of such J local approximations (achieved by PCA or related methods e.g. Partial Least Squares Regression - PLSR - Independent Component Analysis - ICA - or non-linear versions of these) can detect and display their main patterns of covariation. This can unveil the underlying causalities of the data generation mechanism.

1.6. Algorithms for PCA decomposition

The PCA approximation of a certain dataset can be efficiently attained by a variety of algorithms, among which the most widespread and popular one is certainly Singular Value Decomposition (SVD) [10]. However, if N is very high, standard SVD may be very demanding in terms of both CPU load and memory requirements. In the last few years, several variants of classical SVD have been proposed for performing PCA on very large matrices without entirely keeping them in the computer memory (*out-of-core*) [11–15]. *Out-of-core* PCA can be carried out by different procedures e.g.:

- a $J \times 1$ cumulative sum vector and a $J \times J$ cross-product matrix may be accumulated over time, combined and used for eigen-analysis of the covariance in $\mathbf{X}$, which yields the PCA loadings. That is appropriate for parallelisation but then the scores for the past time or space samples are lost;

- if also J is very high (e.g. thousands of wavelengths in an hyperspectral camera monitoring a certain scene or process), the $J \times J$ covariance matrix cannot be easily handled. Evolving moving-window/recursive PCA approaches may then be used instead, working on the most recent subset of observations. But that gives problems when comparing past and present records, e.g. in graphical *scores plots*.

In the attempt of overcoming all these limitations, the OTFP tool is here proposed. The purpose of the OTFP is to identify systematic trends and patterns in high-dimensional data flows, compress these and display them graphically, in addition to automatically detect outliers - key points to be addressed when continuous quantitative data streams are dealt with [16]. Based on what detailed before, it rather represents an extension of classical bilinear PCA, specifically developed for processing multi-channel records as soon as they are collected. It extracts patterns of covariation between the input variables by comparing previous and new observations and thereby identifying and modelling new variation phenomena, without needing large amounts of data or parameters to be retained in memory. Given for instance a continuously growing stream of high-dimensional data, the OTFP modelling system gradually develops a minimal bilinear summary model of the input data stream. For each point in space and/or time, already established components are quantified as spatiotemporal scores by projection of their multi-channel loadings. Furthermore, new, unmodelled patterns of covariation are automatically detected, refined and quantified in terms of additional spatiotemporal scores and multi-channel loadings, then appended to the OTFP model. Hence, unlike bilinear moving-window solutions, this dynamic model extension is executed so that the system preserves the quantitative connection between all past and present records. Yet it does not need to retain all past inputs or bilinear scores in memory - for long-lasting processes the memory usage would grow prohibitively high. Besides, the OTFP system does not require to hold and update a huge $J \times J$ covariance matrix - for many applications that would also be of a prohibitive size. Instead, it repeatedly stores the necessary scores and loadings, avoiding an excessive memory consumption during the process.

2. System overview

The present OTFP algorithm (schematically outlined in Figure 1) is characterised by three fundamental aspects: i) its self-learning and ii) adaptive nature and iii) its stabilising modelling principles. It allows massive amounts of data collected along time to be compressed and modelled with minimal loss of significant information content. The algorithm is initiated with the preliminary choice of a typical input vector, $\mathbf{m}$, and the best guess of which weights to give to the different input channels for balancing their variances, $\mathbf{c}$. In addition, a set of predefined component loadings, $\mathbf{P}$, derived for instance from an initial exploratory investigation of the system under study and representing systematic variation patterns expected to affect the incoming data stream, may or may not be supplied. Then, various system parameters such as the desired modelling fidelity (e.g. the fraction of data variance the OTFP model has to capture, also known as amount of *explained* data variance) need to be specified. As the multi-channel data starts to flow it may deliver a more or less continuous stream of individual J-dimensional input records, e.g. a set of measurements collected by the same set of J simple channels or sensors during the evolution of an industrial

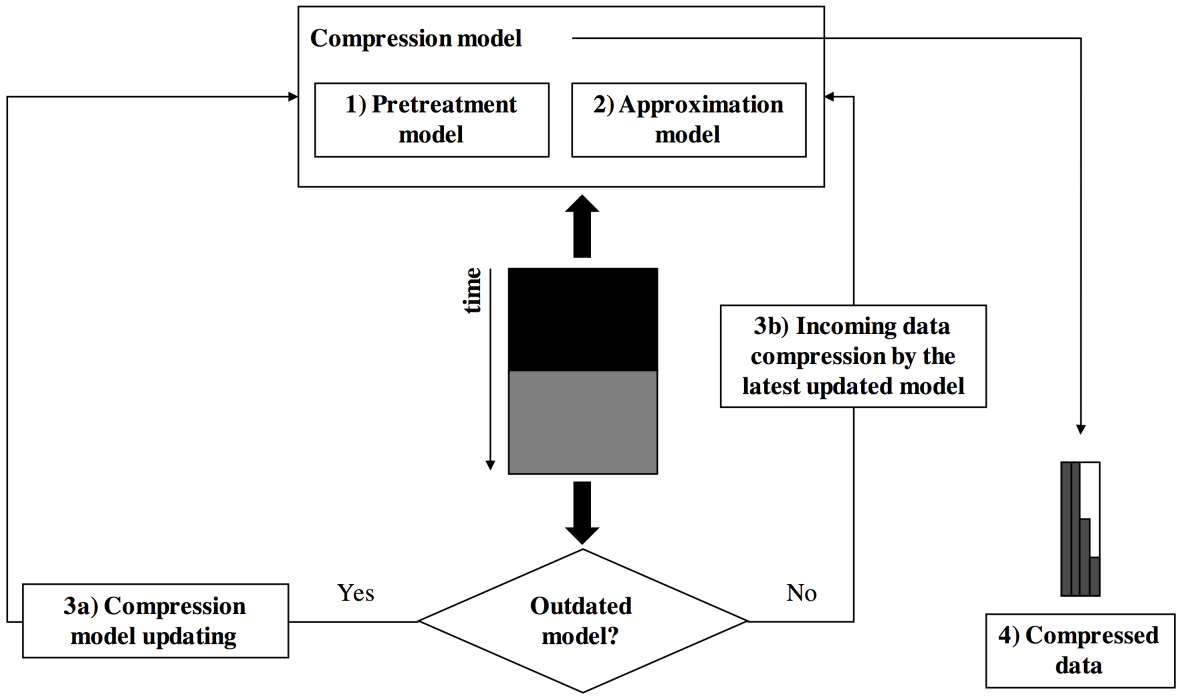

Figure 1 - Schematic representation of the OTFP algorithm: a first set of data (black block) is input to 1) a pretreatment and 2) a PCA-based dimensionality reduction stage. As new measurements are recorded (grey block), they can be either 3a) exploited for the reparametrisation of the compression model, if it is found to be outdated, or 3b) just approximated by its latest version. 4) Bilinear approximation loadings and preprocessing parameters are saved by keeping track of how they have been initially defined and/or changed during model updating. The time series of bilinear approximation scores are more or less continuously stored and deleted from memory to subsequently process new input data

process. Alternatively, it may deliver a sequence of input data blocks, each containing N_g records ($g = 1, 2, \ldots, G$) and the same set of J channels, e.g. N_g spectral profiles, constituted of absorbance values measured at J wavelengths and associated to individual pixels of an hyperspectral image. Such records are then treated by the following procedure:

1. The J-dimensional data are (optionally) submitted to a lossless preprocessing, linearising the responses and balancing the variable variances to ease the subsequent bilinear modelling. This step is domain-specific and the way it is executed has to be set *a priori*. For this reason, the best pretreatment strategy should always be selected based on both the nature of the handled instrumental equipment and technical knowledge;
2. The preprocessed data are projected onto the subspace defined by the bilinear loadings, $\mathbf{P}$, already established at this point in time, to estimate the scores for the respective components;
3. The residuals left in the data after the projection on known components are input to a bilinear (here PCA-like) modelling stage to detect new unknown components and isolate outliers. If new components are found, they are quantified in terms of new scores and loadings. Thus, the statistically redundant J original variables are replaced by a smaller number (A) of principal components (PCs). The number of such components determines the degree of

fidelity initially specified by the user. The algorithm automatically learns to identify and quantify all the systematic types of covariation in the data stream as it flows, while most of the random measurement errors and individual or irrelevant outliers are removed, provided the latter do not constitute a new pattern of variation. This compressed representation is suitable for graphical interpretation and quantitative use, and from it the pretreated data can be reconstructed at any time;

4. At regular intervals, the OTFP model may be refined and reorthogonalised in a linear updating stage;
5. The pretreatment information associated to the different blocks is stored as output together with the approximation model scores and loadings.

As specified before, the OTFP algorithm detects all the systematic types of covariation in the data stream - be it from the flow of observed objects (expected information) or from the measuring process itself (unexpected information, anyway needed for reliable interpretation and quantitative use of the data). Phenomena considered irrelevant during preprocessing, as well as individual outliers discovered by the OTFP algorithm, are noted and then excluded from the self-modelling process. So is much of the random, independent measurement error, since it does not represent a systematic pattern of covariation.
At any time, the systematic part of the data stream can be reconstructed from the data model, e.g. for visualisation. But this reconstruction is not mandatory; the compressed data model parameters, representing the known and/or unknown types of systematic phenomena in the data stream, are themselves suitable for efficient storage and transmission, human graphical interpretation and applied quantitative usage.
These steps will now be described. For further details, the reader may contact either the corresponding or the last author.

2.1. Input

The *ever-lasting* raw data stream, $\mathbf{X}$, divided into a sequence of blocks, $\mathbf{X}_g$ ($N_g \times J$, $g = 1, 2, \ldots, G$), is submitted to the optional preprocessing stage, which includes a linearisation and a signal-conditioning step, and then to the OTFP self-modelling. The number of observations encompassed by these blocks can be freely set by the user. Unless the preprocessing parameters and the OTFP centre and scaling vectors ($\mathbf{m}$ and $\mathbf{c}$) are established *a priori*, the start of the modelling process (i.e. for $\mathbf{X}_1$) requires sufficient observations to enable a precise and relevant initialisation of them.

2.1.1. Linearisation

The linearisation of the input data in $\mathbf{X}_g$ is domain-specific. For instance, non-linearities in light spectroscopy data may be reduced by transformation of the recorded light intensity, I, at each wavelength, first to transmittance, $T = \frac{I}{I_0}$, where I_0 represent the blank signal, and then to absorbance, $A = \log \frac{1}{T}$, to better conform to Beer's law of linear chemical responses.
Another aspect of the linearisation is to convert non-additive variation types (e.g. multiplicative light scattering in absorbance spectra, motions in RGB or hyperspectral videos, *etc.*) into additive signal contributions or preprocessing parameters. For instance, multi-channel pretreatments such

as Standard Normal Variate (SNV) [17], Multiplicative Scatter Correction (MSC) [18, 19] and Extended Multiplicative Signal Correction (EMSC) [20] can reparametrise multiplicative effects. Two-domain IDLE modelling [5] can convert confusing motion effects into nicely additive motion flow fields. Domain transforms, like Fast Fourier Transform (FFT) and wavelet analysis can change data locally from time to frequency domain. Such a more or less lossless, model-based preprocessing may produce additional parameters, which may be highly informative and must be stored for later data reconstruction.

2.1.2. Weighing the variables for better signal conditioning

In general, for an optimal data approximation, the J originally measured descriptors in $\mathbf{X}_g$ are approximately centred, e.g. by subtraction of their mean values estimated from the data flowed up to the current step. They may then be weighed to ensure a better balance among their variances so that:

$$\mathbf{X}_{g,p} = (\mathbf{X}_g - \mathbf{1}\mathbf{m}^{\mathrm{T}}) \circ \mathbf{1}\mathbf{c}^{\mathrm{T}} \tag{4}$$

where $\mathbf{1}$ ($N_g \times 1$) is a vector of ones, $\mathbf{m}$ ($J \times 1$) and $\mathbf{c}$ ($J \times 1$) contain the model centre and the input weighing factors (these weighing factors could e.g. be defined as the inverse of the standard deviation values of the J recorded variables at the current step), respectively, while $\circ$ identifies the element-wise (Hadamard) product. The same pretreatment is applied to all consecutive data blocks until $\mathbf{m}$ and/or $\mathbf{c}$ are readjusted as part of the model updating operation (see below).

2.2. Fit to already established model subspace

The linearised, centred and weighed records in $\mathbf{X}_{g,p}$ are now projected onto the already established loadings $\mathbf{P}$ (if they exist at the current step), according to the linear structure model:

$$\mathbf{X}_{g,p} = \mathbf{T}_{g,p}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{g,p} \tag{5}$$

Clearly, the frequency at which such a projection step is carried out depends on the number of observations in $\mathbf{X}_{g,p}$, that is, as aforementioned, a user-defined parameter[iii].

2.3. Bilinear model expansion

In the present implementation of the OTFP, once calculated, the residual vectors in $\mathbf{E}_{g,p}$ are examined: if they are deemed small enough to be considered uninteresting noise, the respective original records are simply discarded and their scores gathered in $\mathbf{T}_{g,p}$. If this is not the case such residual vectors are introduced into a temporary repository to check whether they represent a new systematic trend in the data stream or not. At regular intervals or when its size or variance exceeds a specific user-defined threshold, this temporary repository is used for the estimation of a new set of loadings and scores. If their respective factors are found to explain a sufficiently high amount of the repository variation[iv], these new scores and loadings are appended to those of the already

[iii] In the case-studies described in Section 4, the projection frequency was found to affect only the computational time of the algorithmic procedure (as it increases, the number of data blocks the OTFP has to consecutively handle becomes larger) but not its final outcomes.

[iv] The scores for these new PCs are - implicitly - defined to be zero for all the previous observations.

established PCs in $\mathbf{P}$ and $\mathbf{T}_{g,p}$, respectively. Otherwise, if leverage analysis of the new scores points out that only scattered objects have contributed to them, these are dismissed as incidental outliers, their scores are stored, and the original model is retained.
Since the size of the entire scores matrix can become very large as the information flows, the scores are saved to the local disk at regular intervals and then deleted from memory along with $\mathbf{X}_{g,p}$ and $\mathbf{E}_{g,p}$[v].

2.4. Model updating

Whenever necessary (e.g. if the model is characterised by a relatively high bias), preprocessing parameters, loadings and scores for both old and new observations are readjusted to ensure PCA-like orthogonality and thus a more efficient compression of the data. For such an updating, the OTFP does not need to recall the whole array of scores stored on the local disk, but directly operates on two summary indices of such an array, which are kept in memory in place of it (namely its column-wise cumulative sum vector and its cross-product matrix). The dimensionality of the reestimated model is automatically established according to the user's desired optimisation criterion. Here, for simplicity, the percentage of data variance that has to be captured is used. This allows the original information stream to be retrieved with a predetermined reconstruction accuracy. Other criteria, based on e.g. the statistical significance of the eigenvalues associated to the single components [21, 22], may also be exploited.

3. Datasets

To evaluate the potential of the proposed method, 4 different sets of time series data were compressed and modelled as detailed before and reconstructed afterwards:

- High-speed multi-channel monitoring of a chemical reaction: 4329 multi-channel Vis-NIR spectra were measured in-line between 400 and 1098 nm (350 wavelengths) via a NIRS 6500 spectrophotometer, equipped with a fibre-optic bundle, during several replicates of the self-oscillating Belousov-Zabhotinsky (B-Z) reaction [23]. The final matrix had dimensions 4329×350. This example is intended to illustrate a new way to handle more or less continuous, high-dimensional measurements of a complex dynamic system not yet fully understood from a scientific point of view;

- Detailed remote characterisation of a set of related, complex objects: three 245×210-sized hyperspectral NIR images of three oranges were registered within the near-infrared spectral range 898-1690 nm (247 wavelengths) by a XEVA-FPA-1.7-320 line-scanner camera (Xenics, Belgium). To enable their handling, such three-way arrays needed to be unfolded into a unique matrix, so that a single pixel spectrum was contained in each one of its rows. After background removal, its dimensions were 72365×247. This example was chosen to illustrate how non-invasive bio-spectroscopy can reveal hidden aspects of related complex biological samples;

[v] In the case-studies described in Section 4, the storage of the scores on the local disk proved not to constitute a limiting step for the execution of the OTFP algorithm.

- Airborne environmental surveillance: an hyperspectral image was recorded by a push-broom device installed on an Unmanned Aerial Vehicle (DroneSpex, Norut AS - University Centre in Svalbard - Norwegian University of Science and Technology, Norway [24]), flying over Faial (Azorean Islands, Portugal). At each accumulation step, the optical sensor collected the absorbance values at 450 wavelengths in the visible light range between 420 and 640 nm for a strip of 245 pixels. A total number of 1000 consecutive snapshots were captured, which led to a three-way array of dimensions $1000 \times 245 \times 450$. Also in this case, it was unfolded into a 245000×450 matrix. This example is intended to show how data from a modern environmental monitoring instrument, a drone, can be automatically compressed for efficient storage and transmission and interpreted in the compressed state;

- Traditional industrial process analysis: 76 engineering variables, mainly including temperatures, pressures and flow rates, were recorded at hourly intervals to follow the evolution of a continuous industrial process. The complete data structure had dimensions 14561×76. This example illustrates the application of the OTFP to records measured over time by a relatively small set of conventional sensors.

4. Results and discussion

The power of the OTFP approach and the quality of the initial data retrieval were assessed in all the case-studies at hand according to the following indices:

- A: number of extracted PCs;
- EV_{raw}: percentage of explained raw data variance;
- EV_p: percentage of explained preprocessed data variance;
- $RMSRE$: Root Mean Square Reconstruction Error defined as $\sqrt{\frac{\sum_{n=1}^{N}\sum_{j=1}^{J}(x_{n,j}-\hat{x}_{n,j})^2}{NJ}}$, where $x_{n,j}$ is the (n, j)-th element of $\mathbf{X}$ and $\hat{x}_{n,j}$ refers to its respective reconstructed value;
- t_c: compression time expressed in seconds[vi];
- CR: compression ratio[vii].

EV_{raw}, EV_p and $RMSRE$ are strictly related to the OTFP approximation accuracy degree, while A, t_c and CR can be considered measures of computational speed and efficiency.
Calculations were executed by using Idletechs' prototype software in a Matlab R2012b environment (The MathWorks, Inc., Natick, Massachusetts, United States), set up on a MacBook Pro equipped with a 2.3 GHz Intel Core i7 and 8 GB 1600 MHz DDR3 RAM.

[vi] t_c is computed as the time needed to compress the entire concerned dataset.

[vii] CR is computed as the ratio between the memory usage of the uncompressed and compressed (preprocessing parameters, scores and loadings matrices) data structures, both saved as double precision .mat files.

4.1. High-speed multi-channel monitoring of the Belousov-Zhabotinsky reaction

Table 1 lists the values of the aforementioned parameters related to the Vis-NIR data compression. The initialisation measurements were centred and weighed ($\mathbf{c} = \frac{1}{\mathbf{m}+0.05}$) after baseline correction[viii]. The model centre vector, $\mathbf{m}$, was updated at regular intervals as new spectroscopic details were encountered in the process, while the variable weighing vector, $\mathbf{c}$, was kept constant for simplicity.

Table 1 - Vis-NIR light absorbance spectra from the B-Z reaction: values of the compression quality indices. The number of original measured variables is reported in the first column

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
350	10	99.93	99.61	0.0019	12.5	26.81 $(\frac{9768173 \text{ bytes}}{364370 \text{ bytes}})$

In order to more clearly appreciate the performance of the OTFP, 3 uncompressed and reconstructed spectra associated to different reaction stages are displayed in Figure 2. The full approximation model is sketched in Figure 3, in terms of final model mean (Figure 3a), chosen weighing factors (Figure 3b), de-weighed and scaled loadings (Figure 3c) and lack-of-fit residuals (Figure 3d). This example has shown that the OTFP automatically discovered and quantified various systematic variation patterns in the complex, ill understood B-Z reaction. At our chosen fidelity

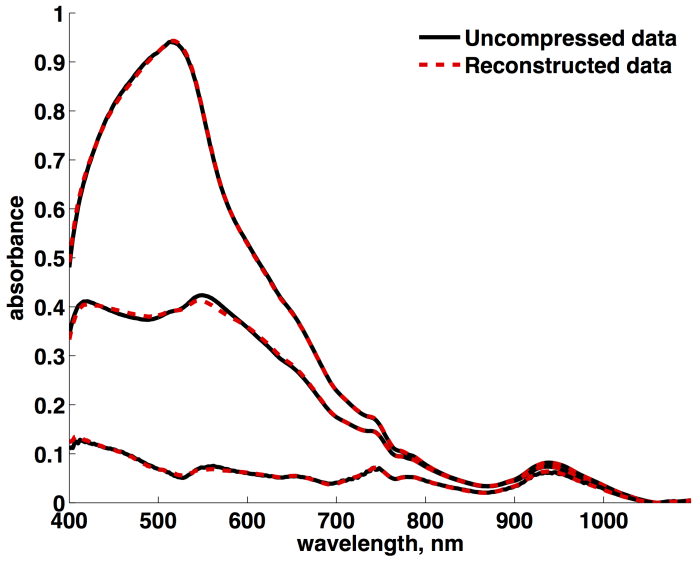

Figure 2 - Vis-NIR light absorbance spectra from the B-Z reaction at three different points in time: input (black solid lines) and OTFP modelled and reconstructed (red dotted lines) spectra

[viii]The reported results refer to the baseline-subtracted spectra for better illustration.

fraction (relative reconstruction error variance < 0.01%, resulting in 10 PCs), only very slight differences between the original and reconstructed profiles are detectable to the naked eye. Had we demanded higher fidelity fractions, more PCs would have been included. Conversely, had we demanded fewer PCs, that would have given higher reconstruction error variance. When submitting this high-dimensional data stream to the automatic model-based data compression, the main patterns of systematic variability in the data were automatically found and extracted. In this example, each high-dimensional spectrum was measured at a single space point only. The next example will show how an overwhelming data stream that arises when thousands of such high-dimensional spectra are measured in parallel by a hyperspectral camera can be dealt with by the OTFP.

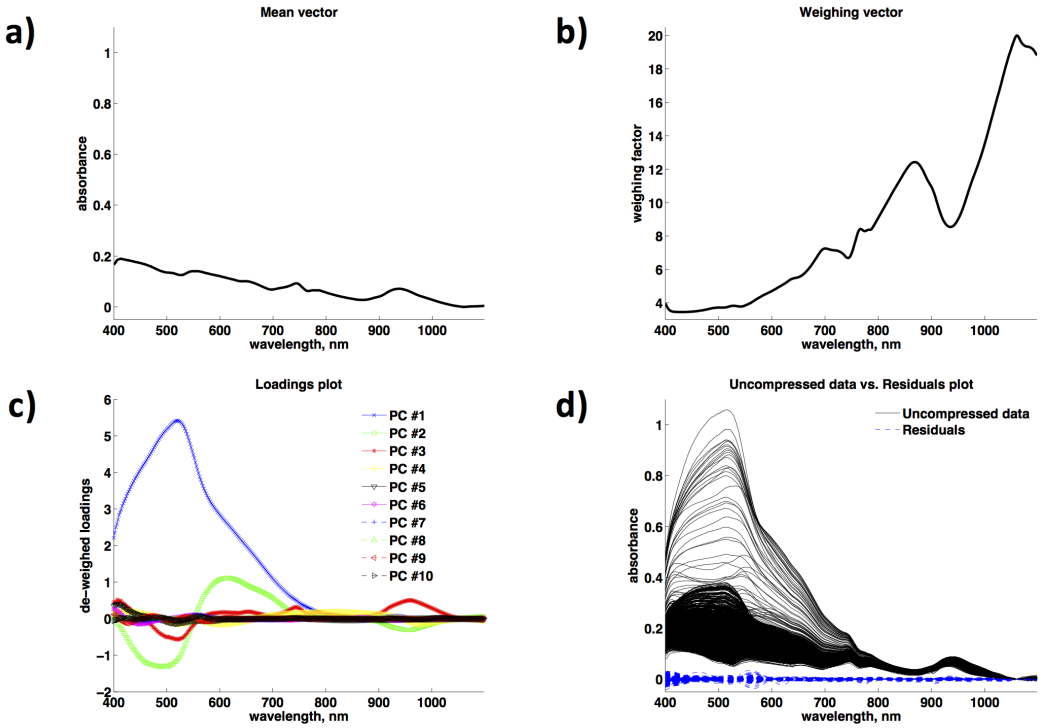

Figure 3 - Vis-NIR light absorbance spectra from the B-Z reaction: representation of the full compression model. a) Final mean vector[ix], b) variable weighing factors (kept constant throughout the algorithmic procedure), c) loadings profiles (divided by the channel weights, **c**, and scaled by their respective singular values) and d) input absorbance spectra (black solid lines) and lack-of-fit residuals (blue dotted lines)

4.2. Detailed remote characterisation of orange samples

This example concerns efficient quality control of physical objects - in this case oranges. The individual pixel NIR spectra were submitted to a model-based pretreatment, MSC, to remove the undesired light scattering effects and prevent actual chemical signals, often of lesser magnitude [25], from being overlooked. They were subsequently centred and weighed to down-scale noisy

[ix]The mean vector closely resembles the lower-absorbance spectral profiles, due to their high abundance in the Vis-NIR dataset.

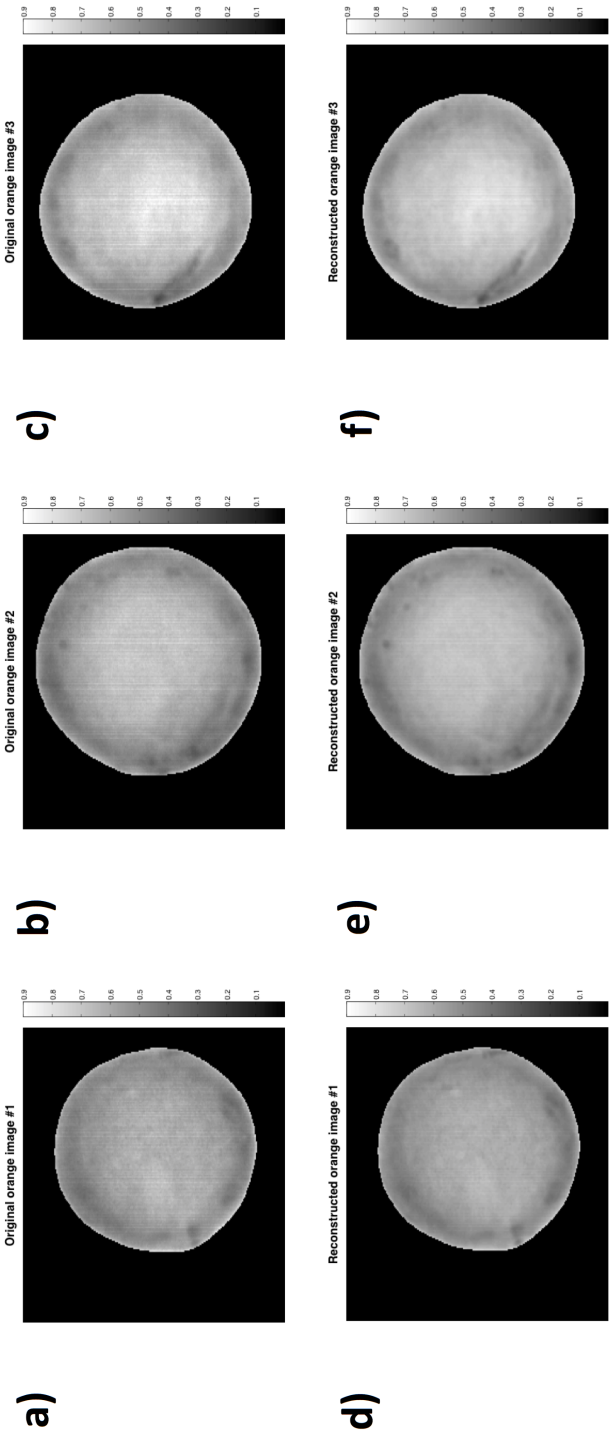

Figure 4 - Hyperspectral NIR images: a-c) Uncompressed and d-f) OTFP modelled and reconstructed grey-scale orange image #1, #2 and #3 at 1675 nm

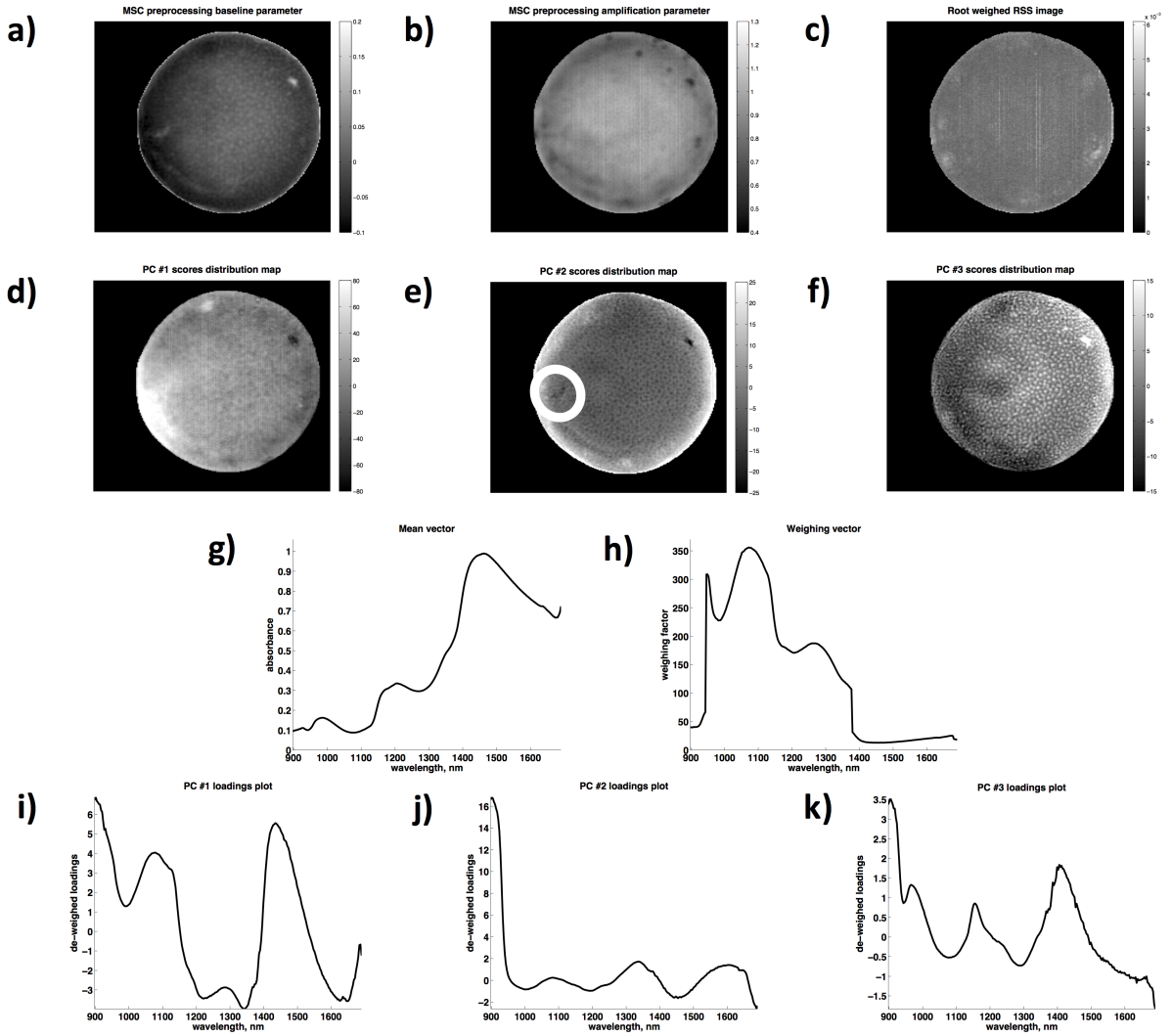

Figure 5 - Hyperspectral NIR images - Modelling of orange image #2: a) baseline variations and b) amplification variations, estimated by MSC preprocessing and used to correct the spectra of the individual pixels, c) summary of the unmodelled residuals (root Residuals Sum-of-Squares, RSS, of the weighed wavelength channels after the extraction of 5 OTFP PCs), d) PC #1, e) PC #2 and f) PC #3 grey-scale scores distribution maps, g) final wavelength mean vector and h) wavelength weighing factors (kept constant throughout the algorithmic procedure) i) PC #1, j) PC #2 and k) PC #3 loadings profiles (divided by the channel weights, **c**, and scaled by their respective singular values). The white circle in e) highlights a particular defect on the surface of the orange sample

spectral regions. Here, the model centre was continuously updated, the variable weighing factors kept constant all over the processing and the MSC parameters additionally stored along with all the other retained information.

As indicated in Table 2, the compression of the orange hyperspectral images also yielded satisfactory outcomes. In addition to a very precise data retrieval, since noise is partly filtered out, various imperfections, probably due to instrumental problems, are apparently removed (see Figure 4).

As an example of the added value the bilinear modelling offers unlike conventional compression

Table 2 - Hyperspectral NIR images: values of the compression quality indices. The number of original measured variables is reported in the first column

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
247	5	99.93	93.27	0.0096	43.8	33.29 $(\frac{129235545 \text{ bytes}}{3882254 \text{ bytes}})$

methods in terms of understanding and interpretability, the scores distribution maps[x] (or scores plots) of image #2 related to the first three extracted PCs and their corresponding loadings profiles are displayed in Figure 5 along with the MSC preprocessing parameters used to correct the spectra of the individual pixels, the corresponding root weighed Residuals Sum-of-Squares (RSS) image (after the extraction of five PCs), the final mean vector and the variable weighing factors resulting from the OTFP. PC #1 seems to reflect an overall lighting variation on the 3D orange. The texture of the orange peel is partly captured by PC #2, along with a particular defect located on the bottom-left area of its surface and a 3D illumination and/or penetration effect generating a gradual decrease in the scores values from the border to the centre of the sample. PC #3 seems to represent a purely textural component.
This example has shown that the self-modelling process simplified the interpretation and usage of the enormous amounts of data from a hyperspectral camera recording a series of similar objects. The model parameters gave high compression as well as interesting graphical insights. The next case-study will illustrate an even more overwhelming data stream from a continuously measuring hyperspectral camera installed on a flying drone.

4.3. Environmental surveillance by airborne hyperspectral imaging

The high compression of the hyperspectral push-broom image is proven by both Table 3 and Figures 6a and 6b. In this case the spectrum of each pixel at each point in time was just centred. Specifically, the model centre was continuously updated, while the variable weighing factors were set to 1 and kept constant all over the processing.
Despite the notable reduction in the memory usage, the uncompressed and reconstructed pseudo-RGB pictures, constructed by selecting only three of the available wavelength channels[xi], exhibit barely perceptible discrepancies.
It is well known that while bilinear models from orthogonal subspace estimation methods (including PCA and the present OTFP) capture the essential variation information in data, the individual components are not intended to be meaningful from a physicochemical perspective, due to their mutual orthogonality (see Equations 2 and 3). Relaxing these orthogonality constraints and possibly adding other criteria, such as non-negativity in loadings and scores, may give more meaningful individual component plots. For example, Figure 6 also includes three different component scores distribution maps and loadings profiles (Figures 6c, 6d, 6e, 6f, 6g and 6h), obtained by a Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [26] transformation of the global

[x] The darkness of the pixels is proportional to the value of their scores on the respective components.

[xi] Around 445 nm, 535 nm and 575 nm, where the eye cones have their maximum sensitivity to blue, green and red light, respectively.

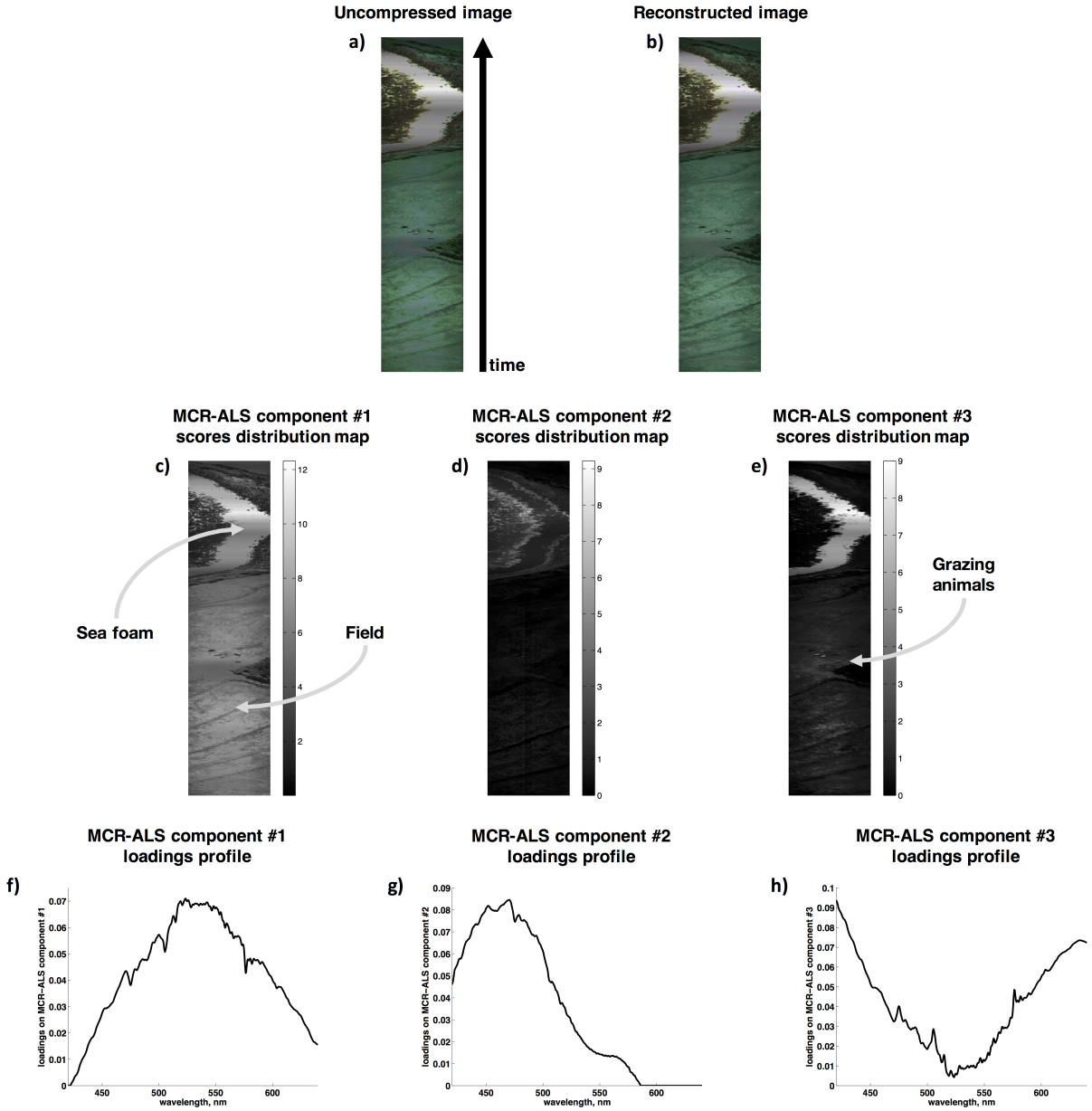

Figure 6 - Hyperspectral image from a push-broom camera installed on a flying drone: a) uncompressed and b) OTFP modelled and reconstructed images in pseudo-RGB colours, c) MCR-ALS component #1, d) MCR-ALS component #2 and e) MCR-ALS component #3 grey-scale scores distribution maps, f) MCR-ALS component #1, g) MCR-ALS component #2 and h) MCR-ALS component #3 loadings profiles

OTFP model. MCR-ALS is a soft bilinear-modelling technique, analogous to PCA, originally conceived for the resolution of multicomponent evolving chemical systems into pure individual contributions, not necessarily completely uncorrelated. It is based on an iterative sequence of optimisation steps, but requires appropriate initial guesses of these contributions to achieve a reliable solution. Here, the scores and the loadings represented in Figures 6c, 6d, 6e, 6f, 6g and 6h were

Table 3 - Hyperspectral image from a push-broom camera installed on a flying drone: values of the compression quality indices. The number of original measured variables is reported in the first column

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
450	3	99.82	99.02	0.015	300.2	45.02 $(\frac{241451269 \text{ bytes}}{5363455 \text{ bytes}})$

reestimated by executing MCR-ALS on the OTFP reconstructed data, appealing to the final OTFP loadings as input.
Although MCR-ALS components #1 and #3 are seemingly dominated by the sea foam pixels (whose corresponding signal was found to be saturated in a large spectral range), three distinct patterns are visibly recognisable: the field pixels in the first scores distribution map, the pixels surrounding the sea foam in the second scores distribution map and those capturing several animals grazing at the centre of the image in the third scores distribution map. Therefore, *ça va sans dire*, the OTFP may be employed for preliminary image treatment before further handling or segmentation.
Independent Component Analysis (ICA) [27, 28] or Parallel Factor Analysis (PARAFAC) [29, 30] and extensions of these coupled with various pixel clustering methods also belong to the rich flora of post-processing methods that can be applied to bilinear models like those coming from the OTFP.
The three first illustrations have shown how broad data stream from multi-channel sensors can be handled by the OTFP. The last example concerns a very different kind of data - a more or less random collection of individual, single-channel sensors. Traditional process industry is often extensively equipped with temperature-and-pressure sensors. Often, each new sensor gets its own display screen with its own alarm limits. How can the burden for the process operators be reduced as well as the number of false alarms? Perhaps by finding common patterns of covariation among the many sensors?

4.4. Analysis of an industrial manufacturing process

This example illustrates how the OTFP may be used for more rational handling of traditional industrial process data.
According to the quality indices reported in Table 4, the general performance of the OTFP when applied to this rather low-dimensional stream of industrial process data was found to be slightly worse than in the previous case-studies. This is not unexpected, given the low number of variables under study and their widely varying nature, and is a consequence of the fact that their correlation structure is not so strong that just few PCs can practically summarise all their significant variation. Nevertheless, for most of them an acceptable reconstruction was achieved, as Figure 7 confirms.
Besides, examining both scores and loadings can provide remarkable insights into the process behaviour, particularly if meaning can be ascribed to the input records - or at least to some of them - by human expert characterisation. This is illustrated by the scores plot in Figure 8a. PC #1 separates two groups of observations: blue dots and red squares refer in fact to Normal Operating Conditions (NOC) and shut-down time samples, respectively. As the latter present negative

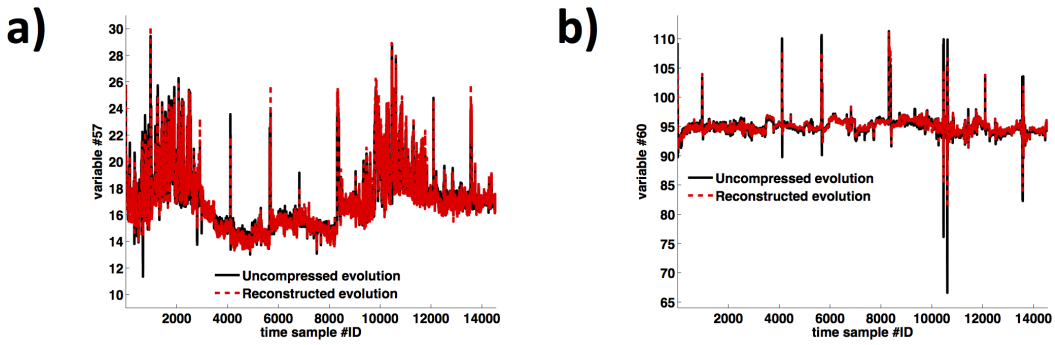

Figure 7 - Industrial process data: Uncompressed (black solid line) and OTFP modelled and reconstructed (red dotted line) temporal evolution of a) variable #57 and b) variable #60

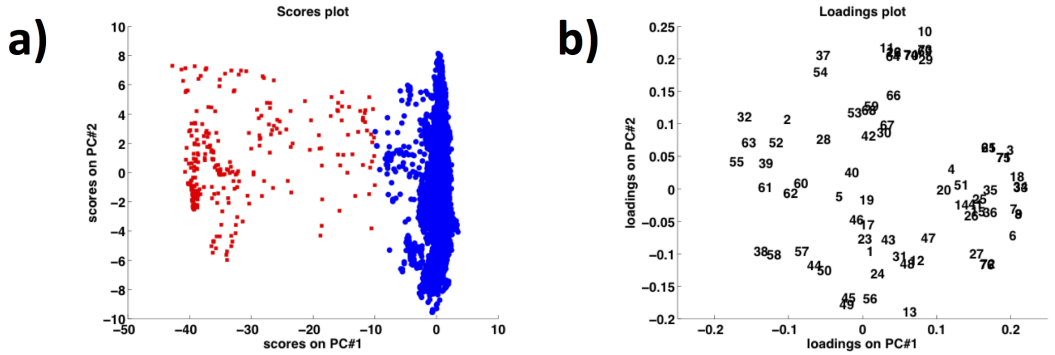

Figure 8 - Industrial process data: a) PC #1/PC #2 scores (blue dots and red squares refer to Normal Operating Conditions and shut-down time samples, respectively) and b) loadings plots (the numbers correspond to the #IDs of the original variables and are represented according to their respective PC #1/PC #2 loadings values)

projection coordinates on this component, they will be characterised by lower-than-average values of all the measured variables featuring a relatively large positive PC #1 loading (which actually assumed a nearly 0-level during shut-down periods) and *vice versa* (see Figure 8b). On the other hand, within-cluster differences seem to be mainly spotted by PC #2.

Table 4 - Industrial process data: values of the compression quality indices. The number of original measured variables is reported in the first column

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
76	13	99.47	81.33	0.4640[xii]	49.5	3.35 ($\frac{4895674 \text{ bytes}}{1459315 \text{ bytes}}$)

[xii] As the original variables were characterised by different units of measurements, the reported RMSRE value concerns the final centred and weighed data array.

5. Comparison with classical PCA

To what extent does the OTFP mimic the corresponding traditional data modelling strategy? The present implementation of the OTFP employs similar criteria in the model updating stage to those of standard PCA, so it is natural to compare both the approaches. While the OTFP needs to hold only a small part of the data in memory at a time, traditional PCA requires all the data to be held in memory at the same time, at least if both loadings and scores are to be assessed. The time span of the hyperspectral drone imaging example (Figure 6 and Table 3) was chosen short enough to allow conventional PCA to be run and its solution to be compared to the final OTFP model. Figure 9 permits to appraise the performance of the two methods for the same dataset. Figure

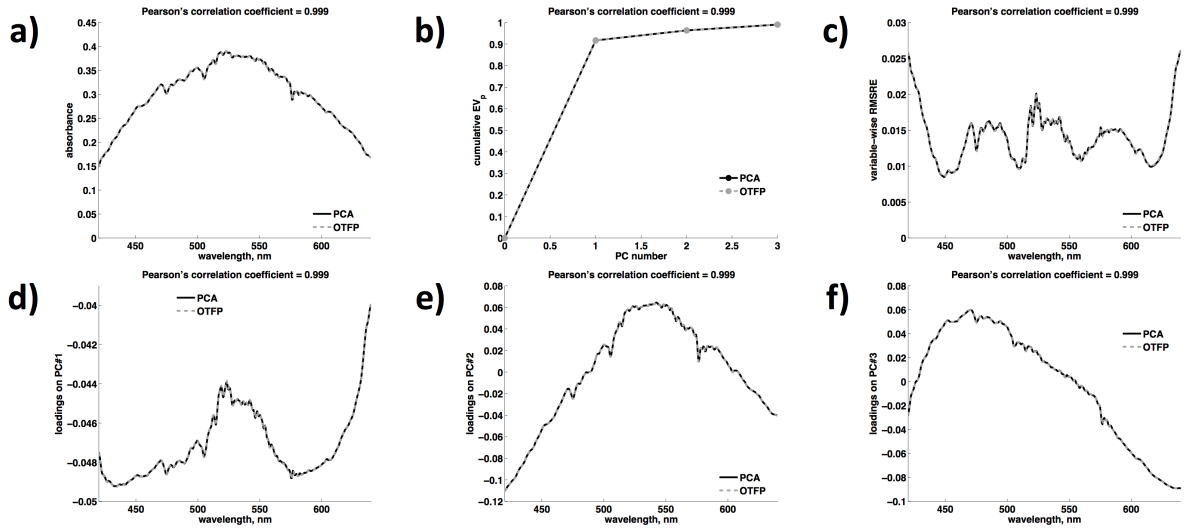

Figure 9 - Hyperspectral image from a push-broom camera installed on a flying drone - Classical PCA (black solid line) *vs.* OTFP (grey dotted line): a) Mean vectors, b) cumulative percentages of explained preprocessed data variance, c) lack-of-fit (root mean square error) for the individual variables after the extraction of 3 OTFP PCs (negligible if compared with the original signal magnitude) d) PC #1, e) PC #2 and f) PC #3 loadings. Variable weighing factors (not shown) were set to 1 for all the spectral wavelengths and kept constant all over the OTFP

9a shows that the mean spectrum used for model centring in PCA is more or less identical to the model centre vector, **m**, gradually developed by the OTFP. The outcomes of the two techniques are also virtually indistinguishable if looking at the plot of the cumulative percentage of explained preprocessed data variance (Figure 9b) and the variable-wise RMSRE (after the extraction of three PCs, Figure 9c) as well as at the loadings profiles of PC #1 (Figure 9d), PC #2 (Figure 9e) and PC #3 (Figure 9f). The corresponding spatiotemporal OTFP scores (not displayed due to the high number of data points) were also found to be very similar to the PCA ones. Consequently, both PCA and the OTFP led to practically identical values of the diagnostic indices listed at the beginning of Section 4 except for t_c. The decomposition was in fact achieved faster by PCA, which had simultaneous access to all the available information. On the other hand, the OTFP had to handle it by evolving its bilinear model on-the-fly as the data flowed. Nevertheless, the comparison highlighted the OTFP can be considered a feasible alternative to standard PCA, when

this latter is not applicable (e.g. when the measurements are collected in real time or the size of the analysed matrix is prohibitively large).

6. Discussion

The OTFP treats the incoming data one record or one batch of records at a time, and gradually develops a compact quantitative model of this data stream from the covariation patterns that it discovers. Still, Figure 9 illustrates the OTFP behaved quite similarly - at least for the first three components - to the corresponding conventional *global* multivariate data modelling method (in this case PCA), which analyses all the data simultaneously. Their results are almost identical even if the OTFP repeatedly has to make sense out of small chunks of data as they arrive. So it has to make many temporary decisions about what to throw away as random noise, while the global PCA has access to all the data at once. On the other hand, this is precisely the motivation behind the development of the OTFP tool, that is to always maintain an updated, compact summary of all the systematic changes, which have taken place in an otherwise overwhelming, *ever-lasting* high-dimensional data stream, with low computational or memory requirements.

The OTFP uses a multivariate data driven approximation model (here, PCA-like) as a generic Taylor expansion around a chosen set point or model centre, to summarise whatever known or unknown phenomena has caused the systematic covariation patterns in the input data stream. The OTFP data model has a linear, additive structure and therefore gives the best approximation performance when non-additive and/or non-linear effects in the input data have been corrected for in the preprocessing step. Preprocessing is then helpful for reducing response curvature and other types of non-linearities[xiii]. Response linearity was improved in the first example (Figures 2 and 3) by converting the fibre-optic transflection data to absorbance values. Patterns known to be non-interesting may be removed during preprocessing, as illustrated by the simple baseline correction in the same case-study. In the second example (see Figure 4), unknown additive baseline variations and multiplicative amplification variations were estimated, parametrised and removed jointly by MSC. On the other hand, when the input variables are given in very different units, preprocessing should also scale them to balance their uncertainty levels - or, if that is unknown, to balance their total variances as shown for the industrial process data (Figures 7 and 8).

The OTFP components are mathematical basis vectors that characterise the data stream. When plotted in combination they give useful insights into the main patterns of data variation, as illustrated in Figure 8. But such orthogonal basis vectors are not meant to be interpreted individually. Therefore, the OTFP solution may be at any time readjusted for better visualisation and more causal interpretation. This was shown by the conversion of the orthogonal, PCA-like OTFP component profiles into more graphically distinct ones by requiring non-negative scores and loadings in an MCR-ALS-based post-processing.

[xiii] In case preprocessing is not of help, complex non-linearities and system heterogeneities may be handled by e.g. automatically splitting the data stream under study into two or more disjoint OTFP models (in a similar way as for the well-known static Soft Independent Modelling of Class Analogy - SIMCA - approach [31, 32]).

7. Conclusions and perspectives

In the near future, a drastic increase in the collection and use of high-dimensional, continuous measurements is expected. Rational use of such data streams requires generic data modelling tools that not only give good predictions and classifications as the information flow evolves, but that also reveal its essential structure for human interpretation and efficient compression. In this article the On-The-Fly Processing (OTFP) tool for the on-the-fly gradual modelling and compression of continuous quantitative data streams was proposed. It is based on an evolving implementation of PCA that updates on-line, when necessary, both preprocessing parameters and principal component structure (whose changes and possible expansion can be optionally monitored in real time through intuitive graphical displays). It combines the advantages of three different ways of attaining PCA or PCA-like bilinear decompositions, while avoiding their disadvantages:

- repeated use of conventional PCA, each time bringing increasing amounts of data into memory for simultaneous analysis, which yields bilinear models relevant for both past and present observations, but becomes prohibitively slow and memory-demanding for *everlasting* data streams;
- moving-window PCA, which repeatedly merges new observations with a bilinear subspace loading summarising past observations, ensuring that the bilinear model is up-to-date and thus relevant for the latest observations at any given moment, while losing relevance for older observations;
- eigen-analysis of the cumulative $J \times J$ cross-product matrix, a simple and fast computation as long as J is not too large, which is suitable for parallelisation and *out-of-core* estimation of the PCA loadings with relevance also for past observations, but without quantitative scores for them.

The OTFP discovers new systematic patterns of covariation in multi-channel data streams, and thereby extends its current bilinear model with new dimensions in a computationally efficient way. Over time, the observation scores are stored to disk in packets and then deleted from memory. The model is continuously updated to be as PCA-like as possible, but in such a way that past scores can always be recalled and compared to present ones.
The algorithmic procedure exhibited very satisfactory performance in terms of compression rate and time and quality of the input reconstruction, especially if measurement series underlain by strong correlation structures (e.g. Vis-NIR spectra or hyperspectral images) were dealt with. On the other hand, in the industrial process example, its power was not as prominent, probably due to a lower degree of intercorrelation in this data stream. Still, the retrieval of the temporal evolution of the original variables was reasonably precise. This could represent an important cross-road for manufacturing companies, whose modern information storage systems are commonly based on univariate calculations, not taking into account the possible interdependences among various instrumental responses, destroying their essential multivariate nature and eliminating much of their meaningful content [33]. Last but not least, the scores and loadings estimated through the PCA-based dimensionality reduction feature distinctive interpretability properties, extremely helpful for data understanding, utilisation and further exploration by complementary statistical approaches

(e.g. MCR-ALS). As far as the authors are concerned, no available compressor guarantees such a noteworthy added value.
In the near future, this strategy for continuous, automatic model development based on multi-channel measurements, may become useful also for processing a wider range of data stream types. For instance, Internet of Things (IoT) will result in an enormous increase of technical measurements in many fields of interest (medicine, industry, communications *etc*.). Many of these IoT sensors will be multi-channel (e.g. cameras, spectrometers). Others will be univariate, but even these will generate multi-channel data: the time series from one single, more or less continuous data source will lead to high-dimensional frequency spectra (spectrograms), after suitable domain transforms (e.g. by FFT or wavelet analysis). Since the methodology here relies solely on linear algebra, it is expected to work properly also within the more general BIG DATA context.

8. Acknowledgments

This research work was partially supported by the Spanish Ministry of Economy and Competitiveness under the project DPI2014-55276-C5-1R, Shell Global Solutions International B.V. (Amsterdam, The Netherlands), Idletechs AS (Trondheim, Norway), the Norwegian Research Council (Grant 223254) through the Centre of Autonomous Marine Operations and Systems (AMOS) at the Norwegian University of Science and Technology (NTNU, Trondheim, Norway) and the Ministry of Education, Youth and Sports of the Czech Republic (CENAKVA project CZ.1.05/2.1.00/01.0024 and CENAKVA II project LO1205 under the NPU I program). The authors want to acknowledge Prof. Bjørn Alsberg for providing the Vis-NIR equipment and the Laboratório de Sistemas e Tecnologia Subaquática (LSTS, University of Porto), the Hydrographic Institute of the Portuguese Navy and the University of the Azores for carrying out the REP15 exercise, during which the hyperspectral push-broom image was collected.

9. References

[1] L. Buydens, Towards tsunami-resistant chemometrics, Anal. Scien. 0813 (2013) 24–30.
[2] G. Moore, Cramming more components onto integrated circuits, Electronics 38 (1965) 114–117.
[3] A. Katal, M. Wazid, R. Goudar, Big Data: issues, challenges, tools and good practices, in: Sixth International Conference on Contemporary Computing (IC3), IEEE, 2013, pp. 404–409.
[4] D. Salomon, G. Motta, Handbook of Data Compression, 5th Edition, Springer-Verlag Inc., London, UK, 2010.
[5] H. Martens, Quantitative big data: where chemometrics can contribute, J. Chemometr. 29 (2015) 563–581.
[6] I. Joliffe, Principal Component Analysis, 2nd Edition, Springer-Verlag Inc., New York, USA, 2002.
[7] S. Wold, A theoretical foundation of extrathermodynamic relationships (Linear Free Energy relationships), Chem. Scripta 5 (1974) 97–106.
[8] S. Wold, M. Sjöström, Chemometrics and its roots in physical organic chemistry, Acta Chem. Scand. 52 (1998) 517–523.
[9] H. Martens, K. Tøndel, V. Tafintseva, A. Kohler, E. Plahte, J. Vik, A. Gjuvsland, S. Omholt, New perspectives in Partial Least Squares and related methods, 1st Edition, Vol. 56, Springer-Verlag Inc., New York, USA, 2013, Ch. PLS-based multivariate metamodeling of dynamic systems, pp. 3–30.
[10] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (1936) 211–218.
[11] A. Balsubramani, S. Dasgupta, Y. Freund, The fast convergence of incremental PCA, in: Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 3174–3182.

[12] N. Halko, P. Martinsson, Y. Shkolnisky, M. Tygert, An algorithm for the Principal Component Analysis of large data sets, SIAM J. Sci. Comput. 33 (2011) 2580–2594.
[13] N. Kettaneh, A. Berglund, S. Wold, PCA and PLS with very large data sets, Comput. Stat. Data An. 48 (2005) 69–85.
[14] E. Rabani, S. Toledo, Out-of-core SVD and QR decompositions, in: SIAM Proc. S., Society for Industrial and Applied Mathematics, 2001.
[15] F. Vogt, M. Tacke, Fast Principal Component Analysis of large data sets, Chemometr. Intell. Lab. 59 (2001) 1–18.
[16] J. Camacho, Visualizing Big Data with Compressed Score Plots: approach and research challenges, Chemometr. Intell. Lab. 135 (2014) 110–125.
[17] R. Barnes, M. Dhanoa, S. Lister, Standard Normal Variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[18] H. Martens, S. Jensen, P. Geladi, Multivariate linearity transformation for near-infrared reflectance spectrometry, in: O. Christie (Ed.), Proc. Nordic Symp. on Applied Statistics, Stokkand Forlag Publ., Stavanger, Norway, 1983, pp. 208–234.
[19] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491–500.
[20] H. Martens, J. Nielsen, S. Engelsen, Light scattering and light absorbance separated by Extended Multiplicative Signal Correction. Application to near-infrared transmission analysis of powder mixtures, Anal. Chem. 75 (2003) 394–404.
[21] I. Endrizzi, F. Gasperi, M. Rødbotten, T. Næs, Interpretation, validation and segmentation of preference mapping models, Food Qual. Prefer. 32 (2014) 198–209.
[22] R. Vitale, J. Westerhuis, T. Næs, A. Smilde, O. de Noord, A. Ferrer, Selecting the number of factors in Principal Component Analysis by permutation testing - Theoretical and practical aspects, Submitted.
[23] A. Zaikin, A. Zhabotinsky, Concentration wave propagation in two-dimensional liquid-phase self-oscillating system, Nature 225 (1970) 535–537.
[24] K. Nordkvist, Ocean color retrieval using DroneSpex - A miniature imaging spectrometer, Master's thesis, Department of Space Science, MSc Programmes in Engineering, Space Engineering, Luleå University of Technology (2007).
[25] K. Esbensen, Multivariate Data Analysis - in practice, 5th Edition, CAMO Process AS, Oslo, Norway, 2002.
[26] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemometr. Intell. Lab. 76 (2005) 101–110.
[27] P. Comon, Independent component analysis, a new concept?, Signal Process. 36 (1994) 287–314.
[28] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, 1st Edition, John Wiley & Sons Inc. New York, USA, 2001.
[29] F. Hitchcock, The expression of a tensor or a polyadic as a sum of products, J. Math. Phys. Camb. 6 (1927) 164–189.
[30] R. Bro, PARAFAC. tutorial and applications, Chemometr. Intell. Lab. 38 (1997) 149–171.
[31] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127–139.
[32] S. Wold, M. Sjöström, Chemometrics: Theory and Application, 1st Edition, Vol. 52, American Chemical Society, Washington D.C., USA, 1977, Ch. SIMCA: a method for analyzing chemical data in terms of similarity and analogy, pp. 243–282.
[33] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, Int. J. Adapt. Control Signal Process. 19 (2005) 213–246.

Figure captions

- Figure 1: Schematic representation of the OTFP algorithm: a first set of data (black block) is input to 1) a pretreatment and 2) a PCA-based dimensionality reduction stage. As new measurements are recorded (grey block), they can be either 3a) exploited for the reparametrisation of the compression model, if it is found to be outdated, or 3b) just approximated by its latest version. 4) Bilinear approximation loadings and preprocessing parameters are saved by keeping track of how they have been initially defined and/or changed during model updating. The time series of bilinear approximation scores are more or less continuously stored and deleted from memory to subsequently process new input data

- Figure 2: Vis-NIR light absorbance spectra from the B-Z reaction at three different points in time: input (black solid lines) and OTFP modelled and reconstructed (red dotted lines) spectra

- Figure 3: Vis-NIR light absorbance spectra from the B-Z reaction: representation of the full compression model. a) Final mean vector, b) variable weighing factors (kept constant throughout the algorithmic procedure), c) loadings profiles (divided by the channel weights, **c**, and scaled by their respective singular values) and d) input absorbance spectra (black solid lines) and lack-of-fit residuals (blue dotted lines)

- Figure 4: Hyperspectral NIR images: a-c) Uncompressed and d-f) OTFP modelled and reconstructed grey-scale orange image #1, #2 and #3 at 1675 nm

- Figure 5: Hyperspectral NIR images - Modelling of orange image #2: a) baseline variations and b) amplification variations, estimated by MSC preprocessing and used to correct the spectra of the individual pixels, c) summary of the unmodelled residuals (root Residuals Sum-of-Squares, RSS, of the weighed wavelength channels after the extraction of 5 OTFP PCs), d) PC #1, e) PC #2 and f) PC #3 grey-scale scores distribution maps, g) final wavelength mean vector and h) wavelength weighing factors (kept constant throughout the algorithmic procedure) i) PC #1, j) PC #2 and k) PC #3 loadings profiles (divided by the channel weights, **c**, and scaled by their respective singular values). The white circle in e) highlights a particular defect on the surface of the orange sample

- Figure 6: Hyperspectral image from a push-broom camera installed on a flying drone: a) uncompressed and b) OTFP modelled and reconstructed images in pseudo-RGB colours, c) MCR-ALS component #1, d) MCR-ALS component #2 and e) MCR-ALS component #3 grey-scale scores distribution maps, f) MCR-ALS component #1, g) MCR-ALS component #2 and h) MCR-ALS component #3 loadings profiles

- Figure 7: Industrial process data: Uncompressed (black solid line) and OTFP modelled and reconstructed (red dotted line) temporal evolution of a) variable #57 and b) variable #60

- Figure 8: Industrial process data: a) PC #1/PC #2 scores (blue dots and red squares refer to Normal Operating Conditions and shut-down time samples, respectively) and b) loadings plots (the numbers correspond to the #IDs of the original variables and are represented according to their respective PC #1/PC #2 loadings values)

- Figure 9: Hyperspectral image from a push-broom camera installed on a flying drone - Classical PCA (black solid line) *vs.* OTFP (grey dotted line): a) Mean vectors, b) cumulative percentages of explained preprocessed data variance, c) lack-of-fit (root mean square error) for the individual variables after the extraction of 3 OTFP PCs (negligible if compared with the original signal magnitude) d) PC #1, e) PC #2 and f) PC #3 loadings. Variable weighing factors (not shown) were set to 1 for all the spectral wavelengths and kept constant all over the OTFP

Table 1

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
350	10	99.93	99.61	0.0019	12.5	26.81 $(\frac{9768173 \text{ bytes}}{364370 \text{ bytes}})$

Table 2

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
247	5	99.93	93.27	0.0096	43.8	33.29 $(\frac{129235545 \text{ bytes}}{3882254 \text{ bytes}})$

Table 3

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
450	3	99.82	99.02	0.015	300.2	45.02 $(\frac{241451269 \text{ bytes}}{5363455 \text{ bytes}})$

Table 4

J	A	EV_{raw}	EV_p	$RMSRE$	t_c	CR
76	13	99.47	81.33	0.4640	49.5	3.35 $(\frac{4895674 \text{ bytes}}{1459315 \text{ bytes}})$

Table captions

- Table 1: Vis-NIR light absorbance spectra from the B-Z reaction: values of the compression quality indices. The number of original measured variables is reported in the first column

- Table 2: Hyperspectral NIR images: values of the compression quality indices. The number of original measured variables is reported in the first column

- Table 3: Hyperspectral image from a push-broom camera installed on a flying drone: values of the compression quality indices. The number of original measured variables is reported in the first column

- Table 4: Industrial process data: values of the compression quality indices. The number of original measured variables is reported in the first column