



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Big Data Approach to Analytical Chemistry

J. Dubrovkin

Computer Department, Western Galilee College
2421 Acre, Israel

Abstract: The Big Data approach to the Process Analytical Technology was proposed and discussed. The main stages of Analytical Big Data processing have been described. The Partial Least Squares regression and Principal Component Analysis of huge data matrices were performed using Fourier compression.

Keywords: Big Data, Process Analytical Technology, Partial Least Squares, Principal Component Analysis.

I. Introduction

Industrial processes represent analytical objects of continuously increasing complexity. The tremendous advance observed now in the Process Analytical Technology is related to the rapid development of both analytical instrumentation and mathematical methods of multivariate data analysis [1-3]. This success is associated with the rapid growth of the number of new analytical methods and the continuous expansion of the area of their application.

Consequently, the information to be analyzed comes from different analytical instruments (and other sources) in different formats into the data processing system rapidly. This process required data fusion into a single data base which must be build according to the new organizational principles [4, 5]. The corresponding informational models should simulate extremely complex problems by fitting to the massive amount of empirical semi-structured and unstructured data [6]. The algorithmic support and the interface of a computerized analytical system (often with limited computer resources) should be adjustable to systems with features of new types. One solution of these problems is the *cloud computing concept*, which opens new perspectives in analytical information management.

All of the above developments constitute the *Big Data Approach* (BDA) to the information technology, which stemmed from macroeconomics [7]. The BDA already has a revolutionary impact on all aspects of human life [7-16], its main applications being web marketing and advertising, telecommunication and network services, retail and financial services, government activities, social media, and health care [9]. The possibility of BDA application in chemistry was firstly suggested in [17, 18]. The perspectives of new searching techniques in the field of general chemistry that employ NoSQL Big Data bases belonging to the Royal Society of Chemistry were discussed in [19].

In the present study we consider the main principles of using BDA in analytical chemistry, including the algorithmic and instrumental aspects of computerized analytical systems, as well as possible BDA applications. The perspectives of analytical control in industry based on predictive analytics strategy are also considered. Standard algebraic notations are used in the mathematical part of the paper. Bold upper-case and lower-case letters denote matrices and column vectors, respectively. Upper-case and lower-case italicized letters represent scalars. All calculations were performed and the plots were built using the MATLAB program.

II. What is Big Data?

Below we present a brief review of the main features of Big Data, which are based on technical information [7-16]:

1. *Volume*. The volume of the largest commercial data bases may reach many petabytes (PB, 1 PB = 10^{15} Byte). For example, the total capacity of all US academic research libraries is about 2 PB [13].

2. *Variety*. Big Data is in no way a sum of unrelated data sets. In contrast to a “Big Band”, which represents the total of the data bases of a particular problem, Big Data is a huge set aligned with existing reference data [12]. Since all the relevant data are compiled in one set, the sampling errors decrease and, consequently, more measurement errors are acceptable. The loss in accuracy at the micro-level is compensated by deeper insight at the macro-level [8].

In this new situation, instead of the old approach “garbage in, garbage out”, information must be fitted to an appropriate data model in the process of dynamic recalibration [16]), where non-relevant data are rejected.

The use of Big Data creates the illusion of causality lack since mathematical methods allow correlating data to almost any target values [2]. Therefore, there exists the opinion that “Big Data is about *what*, not *why*” [8].

3. *Velocity*. New data are arriving faster than we can consume it, so it must be stored at extremely high rate (high velocity “event storms”) [15]. For example, about 90% of all the data in the world were generated in the period of 2011-2012.

4. *Technologies* include image processing, computer vision, machine learning, speech recognition, and natural language processing.
5. *Sophisticated On-Line Analytical Processing (OLAP)* such as the use of 12 Godd's rules for Basic, Special and Reporting features, and Dimensions; cluster and topological data analysis; symbolic data analysis; intuitionistic fuzzy logic; data mining; Pareto analysis. It is expected that, in the future, OLAP will be based on prescriptive analytics [20].
6. *NoSQL databases*. In contrast to relational SQL databases, which use structured tables to store multiple related attributes, NoSQL databases store binary data objects (e.g., documents and multi-attribute tuples) along with the associated keys [20].
7. *New data visualization techniques*, e.g., high resolution maps, networks, and graphs.
8. *New-type "data scientists"*, who are able to extract information from large unstructured data sets (texts, images, sound and video) and then present the results to non-analytical experts [15].
9. *Extremely high requirements for hardware and software*. For example, Oracle Big Data Appliance [14] (the combination of hardware and special-purpose data-base management system software) has the following parameters: main memory of 864 GB and main storage of 432 TB, 18 servers that include a Sun server (2 CPUs: 6 core Intel processors), 48 GB memory per node, and 12 x 2 TB disks per node. Another example of the IBM Big Data Platform is presented in [15].
10. *Development, deployment, and ongoing operations cost millions of dollars*.

III. Analytical Big Data processing

According to modern concepts [1-3], analytical control of production steps is an integral part of the business processes in industry and both should be considered together. The Data Fusion Framework established by the Joint Directors of Laboratories (US Department of Defense) [5] separates the whole system of a technological process into its objects. The step-by-step breakdown of the process (object identity, event identity, and decision making) is used to convey the business process at each stage [2]. The analytical control based on the principles of the Process Analytical Technology/Quality by Design approach [2] may be considered in the framework of the Waterfall Model (WM) [5]. The main stages of this model are:

- Data collection, using sensors.
- Preprocessing.
- Feature extraction.
- Pattern processing.
- Situation assessment.
- Decision making.

Accordingly to WM we suggest the following simplified diagram of the analytical control (Fig. 1), which includes the treatment of sensor data, semi-structured and unstructured data:

1. On-line raw data collection: analytical data and instrumental parameters.
2. Data converting to the standard format (e.g., XML) [21].
3. Preprocessing [22], data compression [23], and data fusion [24].
4. Calibration and validation mode [25], which includes creating a new mathematical model or correcting the "old" model (by means of, e.g., step-wise regression [26]). Measurement mode: generating output data.
5. Qualitative analysis, in which the methods of business solutions, such as data mining [27], can be used.
6. The Process Trajectory Control (analyzing data in motion [15]): decision making using "output" analytical information [3] and industry specific transaction data [15].

IV. Computer simulation of Big Data processing

1. Data set modeling.

The analytical Big Data processing steps are simulated using a computer program written in MATLAB (Fig. 2). The input parameters of the program are: the number of the components in a mixture (n), the number of mixtures (p), the number of bands (t) in the spectrum of a mixture and their parameters, background and noise level. The i, j -element of the matrix $\mathbf{S}$ ($m \times n$) of the pure component spectra was represented as:

$$s_{ij}(v) = \sum_{i=1}^t A_{ij} \exp \left\{ -4 \ln 2 \left[\frac{v - (v_0)_{ij}}{w_{ij}} \right]^2 \right\}, \quad (1)$$

where v ($1, 2, \dots, m$) is the analytical point (e.g., wavelength), A_{ij} is the value of a band maximum: $A_{ij} = \eta$, $\eta = |1 + \text{randn}(1)|$, $\text{randn}(1)$ is a normal random variable with zero mean and unit variance; $(v_0)_{ij}$ is the band maximum location: $(v_0)_{ij} = (m/3)\eta$, and w_{ij} is the full band width at half maximum: $w_{ij} = (3m/t)\eta$.

The r -th component of the concentration vector ($\mathbf{c}_q$) of mixture q is

$$c_{rq}^{(norm)} = c_{rq} / \sum_{q=1}^n c_{rq}, \quad (2)$$

where $c_{rq} = |\text{rand}(1)|$.

Figure 1 Analytical big data processing steps.

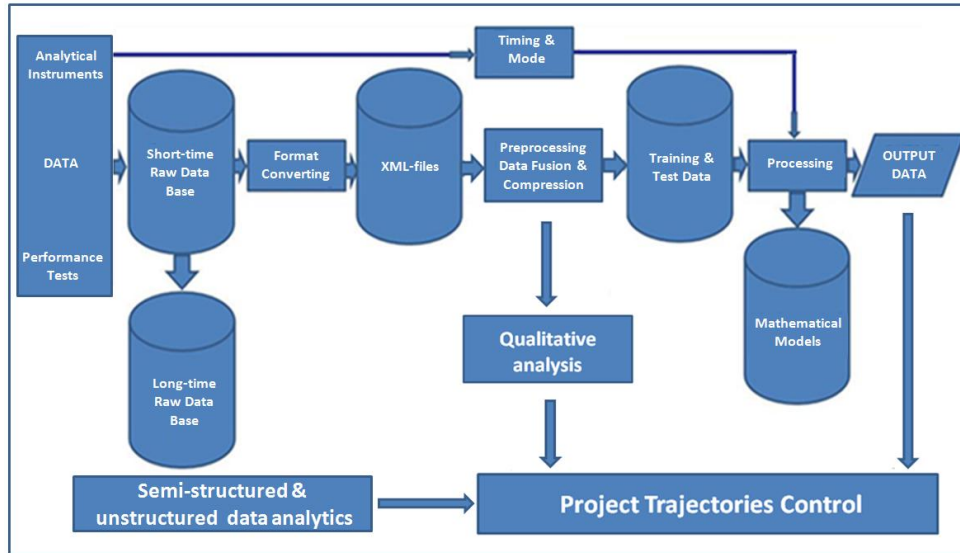
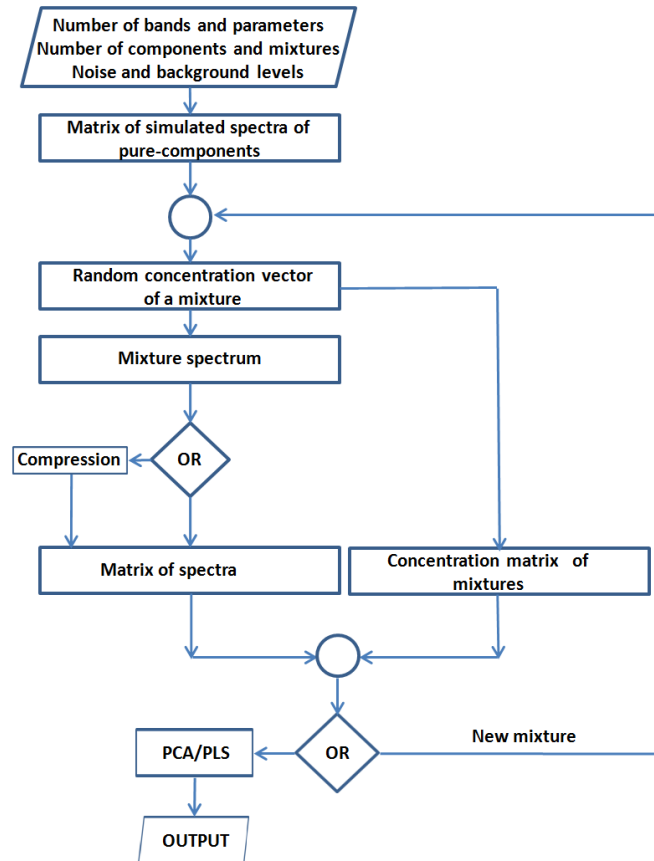


Figure 2 Computer simulation diagram.



The spectrum of mixture q has the form

$$\mathbf{x}_q = \mathbf{S}\mathbf{c}_q + \mathbf{b}_q + \boldsymbol{\xi}_q, \quad (3)$$

where $\mathbf{b}_q$ is the vector of the polynomial background. The i -th element of this vector equals

$$b_{iq} = l_B \left(\eta_1 + \frac{1}{m} \eta_1 i + \frac{1}{m^2} \eta_1 i^2 \right), \quad (4)$$

where l_B is the background level relative to the maximum measured value in each mixture spectrum, $\eta_1 = 1 + 0.3\text{randn}(1)$;

The i -th element of the noise vector ($\boldsymbol{\xi}_q$) is

$$\xi_{iq} = l_N \text{randn}(1), \quad (5)$$

where l_N is the noise level relative to the maximum measured value in each mixture spectrum. The matrix of the spectra of the mixtures under consideration (the response matrix):

$$\mathbf{X} = \mathbf{S}\mathbf{C} + \mathbf{B} + \mathbf{N}, \quad (6)$$

where $\mathbf{C}$ ($n \times p$) is the concentration matrix, $\mathbf{B}$ is the matrix of polynomials and $\mathbf{N}$ is the matrix of the random noise. Matrix $\mathbf{X}$ is filled with vectors (Eq. 3) in an iterative fashion. There is an option of compressing each $\mathbf{x}_q$ vector. Matrices $\mathbf{X}$ and $\mathbf{C}$ serve as analytical input for the Partial Least Squares (PLS) regression and Principal Component Analysis (PCA).

For the sake of simplicity, the preprocessing step is not shown in Fig. 2. This procedure may be included after creating vector $\mathbf{x}_q$ and/or matrix $\mathbf{X}$.

Different types of data sets that are build according to the diagram presented in Fig. 2 should be arranged for further qualitative and quantitative analysis.

2. Data Arrangement

A very simple example of arranging the data obtained from the pure-component NIR and UV-VIS spectra (Fig.3) is demonstrated in Fig. 4. The background and the noise levels were set at 1%. Both the calibration and the validation sets were composed of 4000 spectra. The relative standard error of prediction (RSEP) is

$$RSEP (\%) = 100 \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^n (\hat{c}_{ij} - c_{ij})^2}{\sum_{i=1}^p \sum_{j=1}^n c_{ij}^2}}, \quad (7)$$

where c_{ij} and $\hat{c}_{ij}$ are true and evaluated concentrations of j -th component of the i -th mixture in the validation set. Each RSEP value shown in Table 2 was calculated for 20 independent experiments with the confidence interval of 95%.

Figure 3 Spectra of the pure components used for data arrangement.

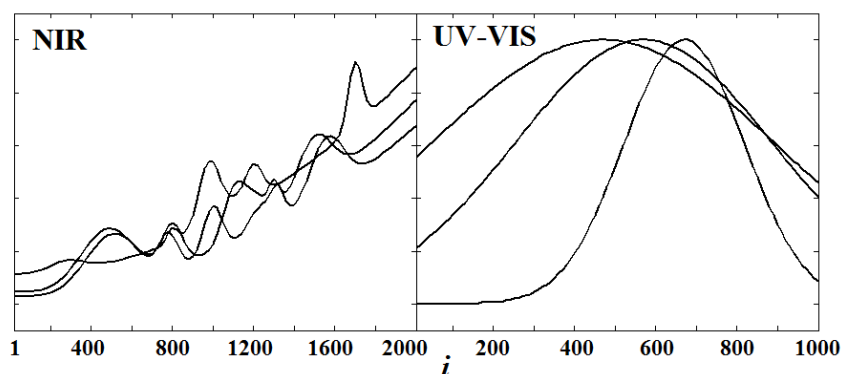


Figure 4 Diagram of data arrangement.

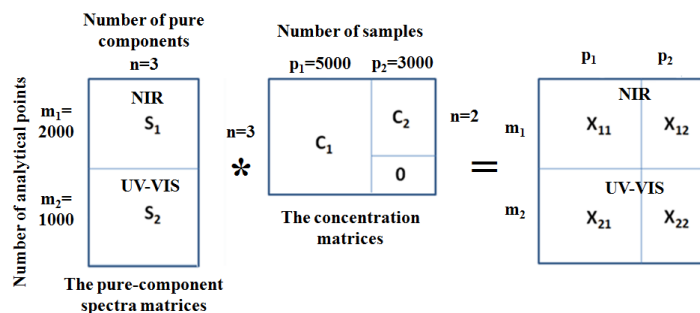


Table 2 Dependence of RSEP values on the number of PLS component.

Number of PLS components	2	3	4	5	6	7	8	9
RSEP, %	2.4±0.07	1.5±0.08	0.88±0.021	0.63±0.014	0.54±0.016	0.59±0.017	0.64±0.020	0.65±0.019

The results presented in Table 2 show that the optimal number of PLS components (which gives the minimal values of RSEP) is 6.

3. Data Compression

In the current situation of the ever-increasing number of the calibration mixtures to be employed, the size of matrix $\mathbf{X}$ composed of data obtained by the modern digital analytical instruments may be so large that it cannot be stored in a common data bases. Therefore, the data should be compressed. To do this, we used the discrete Fourier transform (DFT) [28] in the following way.

The k^{th} -harmonic of the direct DFT has the form

$$f_k = \frac{2}{m} \sum_{a=1}^{m-1} x(ah) \sin(k\alpha \pi/m), \quad (8)$$

where $x(ah)$ is the spectrum ordinate in the point ah and h is the sampling interval.

The original curve is restored by the inverse DFT:

$$x(v) = \sum_{k=1}^N f_k \sin(v \pi/mh), \quad (9)$$

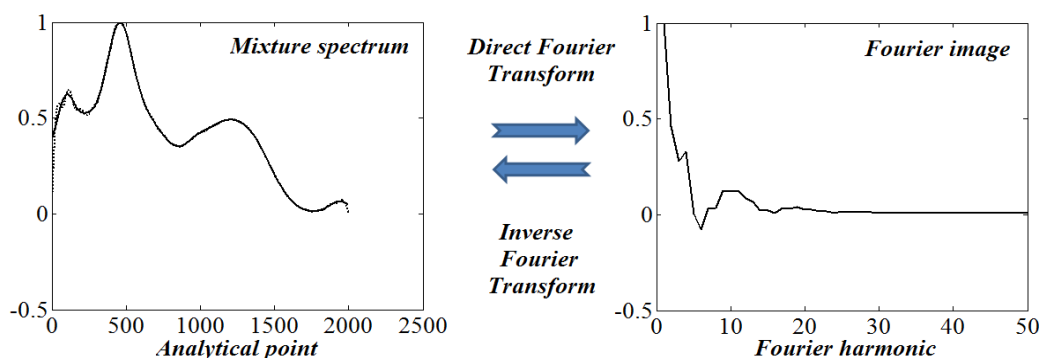
where N is the number of the highest Fourier harmonic, the choice of N being crucial here. However, since only direct transform is used for compression, the choice of the highest harmonic is a simple empirical procedure [20]. The curves presented in Fig. 5 suggest $N = 25$ as a good choice. Fourier harmonics can be readily computed using the well-known Fast Fourier transform algorithm.

4. PLS-regression of huge matrices of variables

The spectra of 30,000 10-component mixtures were simulated; the number analytical point was 2000; each spectrum included strongly overlapping 20 bands; the background and noise levels were set 1%. To perform Fourier compression 25 first harmonics were used. The calibration set was composed of 27,000 spectra, while the validation set contained 3,000 spectra. The “non-compression” and “compression” experiments were conducted independent. The values of RSEP were calculated based on 20 independent experiments, with the confidence of 95%. The following conclusions can be drawn from the obtained results (Table 3):

- The loss of information using DFT compression is negligible.
- Fourier data compression is highly effective computational tool. For example, my old PC with Intel® Core™ 2 Duo CPU 2.00 GHz, RAM 4.00 GB can hardly perform PLS-regression with 2000 x 50,000 (800 MB) matrix by MATLAB package. By means Fourier compression it can do this easily with matrices larger than 25 x 2,000,000 (400 MB).

Figure 5 Fourier treatment of a spectral curve.



The dotted line is the restored spectrum.

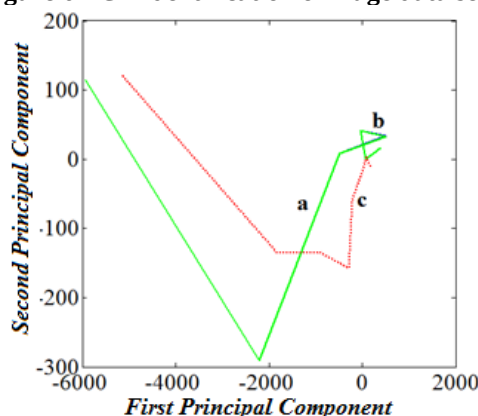
Table 3 Dependences of RSEP values on the number of PLS components obtained with and without data compression.

Number of PLS components	8	9	10	11	12	13	14	15
No compression RSEP, %	17.4± 1.0	5.9± 3.4	3.0± 1.1	2.9± 1.3	2.9± 1.3	2.9± 1.3	2.9± 1.3	2.9± 1.3
Compression RSEP, %	17.2± 0.5	5.4± 2.7	2.8± 1.3	2.8± 1.7	2.8± 1.0	2.7± 1.4	2.7± 1.2	2.8± 1.1

5. The Principal Component Analysis of huge matrices of variables

Two huge independent matrices $\mathbf{X}_1$ (2000 x 2,000,000) (data set 1) and $\mathbf{X}_2$ (2000 x 1,000,000) (data set 2) were simulated. The number of bands in each spectrum was 20, the number of mixture components was 10. The noise and the background levels were set at 0.5%. Two data-subset matrices, $\mathbf{X}_{11}$ and $\mathbf{X}_{12}$, (2000 x 1,000,000) each, were randomly selected from data set 1. The principal component analysis was performed using the matrices compressed to the size of 25 x 1,000,000. The results presented in Fig. 6 show that the plots corresponding to data sets 1 and 2 are clearly distinguishable, while the plots for subsets $\mathbf{X}_{11}$ and $\mathbf{X}_{12}$ practically coincide.

Figure 6 PCA identification of huge data sets.



a, b - two subsets of data set 1; c - independent data set 2. The range of scores is 1:8.

References

- [1] A. L. Pomerantsev and O. Ye. Rodionova, "Process analytical technology: a critical view of the chemometricians", *J. Chemometrics*, vol. 26, 2012, pp. 299-310.
- [2] R. W. Kessler, "Perspectives in process analysis", *J. Chemometrics*, vol. 27, 2013, pp. 369-378.
- [3] A. Bogomolov, "Multivariate process trajectories: capture, resolution and analysis", *Chem. Intel. Syst.*, vol. 108, 2011, pp. 49-63.
- [4] A. Biancolillo, M. Bevilacqua, R. Bucci, A. Magri and F. Marini. "Data fusion protocols for food authentication", XVII euroANALYSIS, Warsaw, 2013 (www.euroanalysis2013.pl).
- [5] J. Esteban, A. Starr, R. Willets, P. Hannah and P. Bryanston-Cross, "A review of data fusion models and architectures: towards engineering guidelines", *Neural Computing and Applications*, vol. 14, 2005, pp. 273 – 281.
- [6] J. Isaeva, S. Sæbo, J. S. Wyller, S. Nhek and H. Martens, "Fast and comprehensive fitting of complex mathematical models to massive amounts of empirical data", *Chem. Intel. Syst.*, vol. 17, 2012, pp. 13-21.
- [7] I. Valova and M. Noirhomme-Fraiture, "Processing of large data sets: evolution, opportunities and challenges", *Proceedings of PCaPAC08*, Ljubljana, Slovenia, 2008, pp. 198-200.
- [8] V. M. Schonberger and K. Cukier, *Big Data: A revolution that will transform how we live, work, and think*. N. Y.: Houghton Mifflin Harcourt Publishing Company, 2013.
- [9] J. Vaughan (TechTarget Inc), 2013, "Putting Your Big Data Strategy on the Right Track", www.SearchBusinessAnalytics.com
- [10] "Big data. IBM's big data portfolio of products for the big data platform", <http://www-01.ibm.com/software/data/bigdata/platform/product.html>
- [11] "Big Data and Analytics", <http://www.idc.com/prodserv/FourPillars/bigData/index.jsp>
- [12] "Big Data without the Big Bang or big bucks (Gabriel Consulting group)" <http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=POL03133USEN>
- [13] M. Rouse, "How many bytes for...", <http://searchstorage.techtarget.com/definition/How-many-bytes-for>.
- [14] Oracle Big Data Appliance, <http://www.oracle.com/us/products/database/big-data-appliance/overview/index.html>.
- [15] M. Ferguson (Intelligent Business Strategies), "Architecting a big data platform for analytics ", <http://www.ndm.net/datawarehouse/pdf/Netezza%20-%20Architecting%20A%20Big%20Data%20Platform%20for%20Analytics.pdf>
- [16] L. Tucci, "Prescriptive analytics is coming to a future near you", [http://searchcio.techtarget.com/opinion/Prescriptive-analytics-is-coming-to-a-future-near-you?asrc=EM_ERU_25972454&utm_medium=EM&utm_source=ERU&utm_campaign=20140106_ERU%20Transmission%20for%2001/06/2014%20\(UserUniverse%20](http://searchcio.techtarget.com/opinion/Prescriptive-analytics-is-coming-to-a-future-near-you?asrc=EM_ERU_25972454&utm_medium=EM&utm_source=ERU&utm_campaign=20140106_ERU%20Transmission%20for%2001/06/2014%20(UserUniverse%20)
- [17] B. C. Gibb, "Big (chemistry) data", *Nature Chemistry*, vol. 5, 2013, pp. 248-249.
- [18] B. Bloom, "Toronto firm does big data chemistry in the cloud", <http://www.itworldcanada.com/news/toronto-firm-does-big-chemistry-in-the-cloud/145625>.
- [19] J. Leonard, "A positive reaction: big data technology at the Royal Society of Chemistry", <http://www.computing.co.uk/ctg/analysis/2305957/a-positive-reaction-big-data-technology-at-the-royal-society-of-chemistry>
- [20] M. Chapple, "Introduction to NoSQL. Building Databases to Support Big Data", <http://databases.about.com/od/otherdatabases/a/Introduction-To-Nosql.htm>.
- [21] Jianyong Zhu, "Automating laboratory operation by integrating laboratory information management systems (LIMS) with analytical instruments and scientific data management system (SDMS), scholarworks", URL: http://www.iupui.edu/bitstream/handle/1805/323/Thesis_JayZhu_Final.pdf.
- [22] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jaansen, G. Downey, L. Blanchet and L. M. C. Buydens, "Breaking with trends in pre-processing?", *TrAC*, vol. 50, 2013, pp. 96-106.
- [23] J. M. Dubrovkin, "Effectiveness of spectral coordinate transformation method in evaluation the unknown spectral parameters", *J. Appl. Spectr.*, vol. 38, 1983, pp. 191-194.
- [24] R. Tauler, M. Maeder and A. de Juan, "Multiset data analysis: extended multivariate curve resolution", In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, vol. 2, pp. 473-505. Elsevier, 2009.
- [25] H. Martens and T. Næs, *Multivariate Calibration*. New York: Wiley, 1992.
- [26] A. F. Sever and A. J. Lee, *Linear Regression Analysis*, 2nd ed., New Jersey: John Wiley and Sons, 2003.
- [27] Yu Tang, Yi-Zeng Liang and Kai-Tai Fang. "Data Mining in Chemometrics: Sub-structures. Learning via peak combinations. Searching in mass spectra", *J. Data Science*, vol. 1, 2003, pp. 481-496.
- [28] C. Lanczos, *Applied analysis*, New York: Dover Publication, Inc, 1988.