Investigation of metabonomics technique by analyze of NMR data, which method is better? Mean center or auto scale?

Seyed AbdolReza Mortazavi-Tabatabaei [1],Fariba Fathi[2], Fatemeh Ektefa[3], Mohsen Tafazzoli[2], Afsaneh Arefi Oskouie [4*], Mostafa Rezaie-Tavirani[1], Mohamad Reza Zali[5],Mohamad Rostami Nejad[5],Kamran Rostami[6]

[1]Reseach Student Committee, Proteomics Research Center, Faculty of Paramedical Sciences ,Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Department of Chemistry, Sharif University of Technology, Tehran, Iran
[3]Department of Chemistry, Tarbiat Modares University, Tehran, Iran
[4]Department of Basic Science, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Research Center for Gastroenterology and Liver Disease, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[6]Acute Medicine, Dudley Group of Hospital, Dudley, UK

*Corresponding Authors Email address: a_arefioskouie@yahoo.com (A. Arefi Oskouie)

ABSTRACT

The factors such as disease can disrupt homeostasis, resulting in perturbations of endogenous biochemicals that are involved in key metabolic profiles. Metabonomics is useful technique to quantitative description of endogenous metabolites present in a biological sample such as urine, plasma and tissue. High resolution ^{1}H nuclear magnetic resonance (NMR)-based metabonomics is a technique used to analyze and interpret multivariate metabolic data that correlate with changes of physiological conditions. Before any explanation for metabolite data, preprocessing the spectroscopic data is essential. In this paper, we show scaling effects in metabonomics investigation of patients diagnosed with Crohn's and Celiac disease. two techniques of scaling were applied as follows: mean centering and auto scaling. Results reveal that the mean centering is more useful to segregate patients from healthy subjects in the data set of Crohn's and Celiac disease.

Keywords: ^{1}H nuclear magnetic resonance; Crohn's disease; Celiac disease; Auto scale; Mean center; Principal component analysis

INTRODUCTION

Crohn's disease is a type of inflammatory bowel disease. This disease may influence any part of the gastrointestinal tract from mouth to anus causing a wide range of symptoms. preliminary symptoms are abdominal pain, diarrhea which may be bloody if inflammation is severe, vomiting or weight loss, as well as outside gastrointestinal tract such as skin rashes, inflammation of the eye, tiredness, arthritis and lack of concentration[1-3].

The diagnosis of Crohn's disease can sometimes be challenging[4] hence a number of tests are necessary to assist the physician in diagnosing of disease. The diagnosis of Crohn's disease with absolute certainty may not be possible even with a complete series of tests. A colonoscopy is approximately 70% effective in diagnosing of disease, and another tests being less effective[5].

Another gastrointestinal disease is Celiac disease resulting in chronic digestive disorder hurting to the lining of the small intestine causing malabsorption of minerals and nutrients[6]. The lining of the intestines contains parts named villi assisting absorb nutrients. The exact cause of celiac disease is unknown[7]. The immune system of people with celiac disease reacts by damaging these villi, when they use products that contain gluten. This damage affects the ability to absorb nutrients properly leading to a person becomes

malnourished. The disease can extend from infancy to late adulthood. Women are affected more often than men[8].

The factors such as disease, drug toxicity, genetic alteration and distorted physiological status can disrupt homeostasis, resulting in perturbations of endogenous biochemicals that are involved in key metabolic profiles in the cells and tissues of an organism. It is worthy of mentioning that monitoring perturbations in biofluid composition may yield expensive information concerning molecular mechanisms. These information can be used to diagnosis, prognosis and drug design[9].

Hundreds of low-molecular endogenous metabolites exist in a biological sample such as urine, plasma and tissue. The global quantitative descriptions of these metabolites are required. In this respect, metabonomics is useful to application in the medical which provides quantitative description [10, 11] . This technique is defined as "the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification" [12].

The metabonomics approach was pioneered firstly evolved by Nicholson and co-workers. Their initiating work led to a novel analytical technique for rapid discovery of biological dysfunctions in pharmaceutical and medical applications. High resolution ^{1}H nuclear magnetic resonance (NMR) spectroscopy is applied in metabolic profiling of biological fluids combined with multivariate analysis. The main purpose of metabonomics is identification the metabolites that correlate with changes of physiological conditions [13].

^{1}HNMR has several features that make it very useful technique. It is rapid, stable in time, basic sample preparation, any chemical species that contains protons gives rise to signal, to be able to detect broad range of metabolites in a non-targeted way and minor components in the presence of much larger signals.

Considerable confusion appears to exist in the metabonomics literature real need for the role of preprocessing the acquired spectroscopic data. A number of studies have presented various data manipulation approaches, some suggesting an optimum methods [14]. The goal of the study was to discus scaling effects in NMR spectroscopic. Metabonomics data sets of patients diagnosed with Crohn's and Celiac diseases were applied in classification model.

MATERIALS AND METHODS

Sample collection

Twenty-six adult patients (11males and15females with mean age of 33.6 ± 11.3 years) diagnosed with Crohn´s disease as well as twenty-seven adult patients (11males and 16 females with mean age of 33.6 ± 11.3 years) diagnosed with Celiac disease at the Research Center for Gastroenterology and Liver Disease, Shahid Beheshti University of Medical Sciences, contributed in this study. Also for each diseases, the control group including twenty-nine healthy subjects (HS) (15 males and 14 females with mean age 34.7±12.2 years) introduced by this center. The blood samples were collected in eppendorf tubes and kept at room temperature for 20 min. The samples were then centrifuged at 2500 rpm for 10 minutes and stored at −80 ˚C until NMR analysis.

^{1}HNMR spectroscopy

All Nuclear magnetic resonance experiments were carried out at a proton frequency of 500.13 MHz and at 300 K on a Bruker Avance 500 spectrometer equipped an inverse detection probe (5mm) with z-gradients. For making a field-frequency lock, serum samples (500 µl) were diluted with 100 µl D_2O. ^{1}H NMR spectra were achieved using the Carr-Purcell-Meiboom-Gill (CPMG) spin-echo sequence [15] with pre-saturation. The CPMG pulse program removes broad signals from high-molecular-weight molecules that would otherwise complicate interpretation of the spectra. These spectra were acquired using a spin-echo loop time ($2n\pi$) of 43.9 ms, with a 2.5 s relaxation delay between pulses. Typical parameters were: spectral width: 8389.26 Hz; time domain points: 32K ; number of scans: 154; spectrum size: 32 K and line broadening: 0.3 Hz. Earlier to Fourier transformation, an exponential line broadening function of 0.3 Hz was applied to the Free Induction Decay(FID).

Data pre-processing

Bining and normalization

In NMR-based metabonomics studies, binning is a rapid method to make data sets. The bin width was applied in order to modeling purposes and determination of the effects of concentration changing in biofluids such as urine, blood, plasma and serum [16, 17].

ProMetab software (version prometab_v3_3) in MATLAB (The MathWorks, Natick, MA) [18] was used to segment each NMR spectrum into 205 chemical shift bins between 0.2 and 10.0 ppm. The bin width is 0.04 ppm. Further, the area within each spectral bin was integrated using the ProMetab software and produced a 1 × 205 vector to describe the spectrum. Water resonances were removed.

Indeed, NMR spectra require another pre-processing step calling normalization before conducting statistical analysis to detect delicate variations from metabolic profiles. After phase/ baseline correction, normalizing was done [19].

Scaling

Before any explanation for metabolite data, they must be normalized, scaled and cleaned up if there is any removable noise. After binning and normalization, scaling techniques were applied. Many approaches can be used for scaling. Scaling to unit variance (auto scale), and mean centering are two methods of the most popular of them. Scaling is executed on the columns of data (i.e. on each spectral intensity across all samples). In auto scaling technique, each column of the table can be scaled so that it has unit variance by dividing each value in the column by the standard deviation of the column(14). In the other hand, in mean center scaling technique, each column of the table can be achieved a mean of zero by subtracting the column mean from each value in the column.

$$X_{MC} = X - \bar{X}$$

Scaling affects the results of classification analysis, since it determines what correlations are important.

Other scaling methods include: Pareto scaling and vast scaling. Pareto scaling is very similar to autoscaling but the square root of the standard deviation is used as the scaling factor instead of the standard deviation. Vast scaling is an extension of autoscaling.

Statistical analysis

Orthogonal signal correction (OSC)

Orthogonal signal correction (OSC) is a technique applying to biofluid NMR data to minimize the influence physical and biological of inter- and intra-spectrometer variation during data acquisition. This procedure also minimizes the effects of innate physiological variation in high resolution ^{1}HNMR spectra of biofluids. The removal of orthogonal variation exposed features of interest in the NMR data and facilitated interpretation of the derived multivariate models(9, 20).

Principal component analysis (PCA)

Studies were based on more than three genes. If we want to account for all of them in the analysis, we must reduce the multidimensional information in a lower dimensional space, such as 2 and 3 dimensions. The reason for reduction the multidimensional information is that there has no way to imagine more than three dimensions. One way of doing the reduction is by means of principal components (PC). PCs are orthogonal vector that describe a space of lower dimension accounting the maximum variation in the original space(21). The original space is made up of the gene expression profiles and has as high a variance as possible (that is, accounts for as much of the variability in the data as possible). It is possible to sort samples in scatter plots based on the principal components. The scatter plots will reflect most of the information in the original data set of higher dimension. The number of principal components is less than or equal to the number of original variables.

Principal component analysis (PCA) is a mathematical process that uses an orthogonal transformation to convert set of observations of possibly correlated variables into set of values of uncorrelated variables (principal components) [22]. PCA is sensitive to the relative scaling of the original variables.

PCA is applied abundantly in all forms of analysis - from neuroscience to computer graphics - because it is a simple. Also it is 770non-

parametric method of extracting appropriate information from confusing data sets.

DISCUSSION AND CONCLUSION

In metabonomics, data are usually presented as a table where each row corresponds relates to a given sample or analytical experiment. Each column relates to a single measurement in that experiment, typically individual spectral peak intensities or metabolite concentrations. All calculations were performed using Matlab 5.3.1 [23]. Two matrixes, predictor variables (X) and predicted variable(s) (Y), as input data are exported to MATLAB software. Matrix X is NMR data which rows are number of samples and the columns are variables. Also in matrix Y 1 and 0 correspond to patients and healthy subjects, respectively.

After binning and normalization, we apply OSC technique on biofluid NMR data. The number of OSC components to calculate are 2. Optional inputs are the maximum number of iterations used in attempting to maximize the variance captured by orthogonal component (iter) {default = 0}, and tolerance on percent of x variance to consider in formation of the final w vector (tol) {default = 99.9}. Therefore**,** PCA are performed with the PLS-Toolbox version 2.0 [24]. Both the graphical ('pcagui') and the command-line ('pca') versions of PCA have been used. The command-line version is only used to automate PCA. After this processes, we use mean center and auto scale methods for classifying data into patients and healthy subjects of both Celiac and Crohn's disease.

Figure 1 and 2 present results of auto scale and mean center**,** respectively, for Crohn's disease**.** The representative points of the serum samples of Crohn's disease are mapped in the space spanned by the first two principal components PC1 versus PC2.

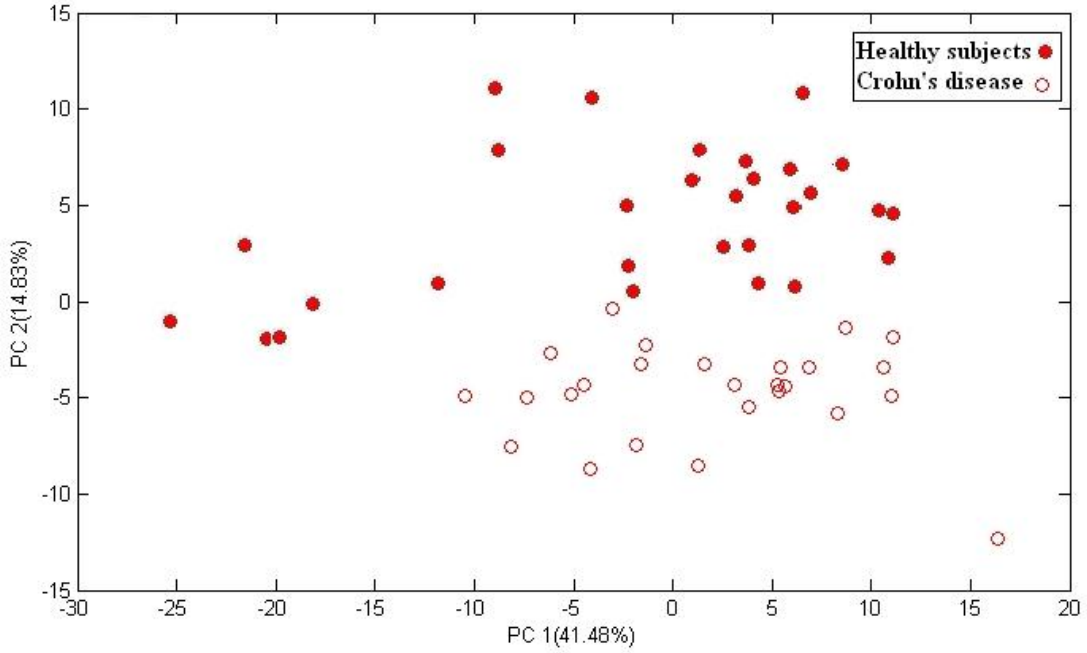

Figure1. Results of auto scale for Crohn's disease

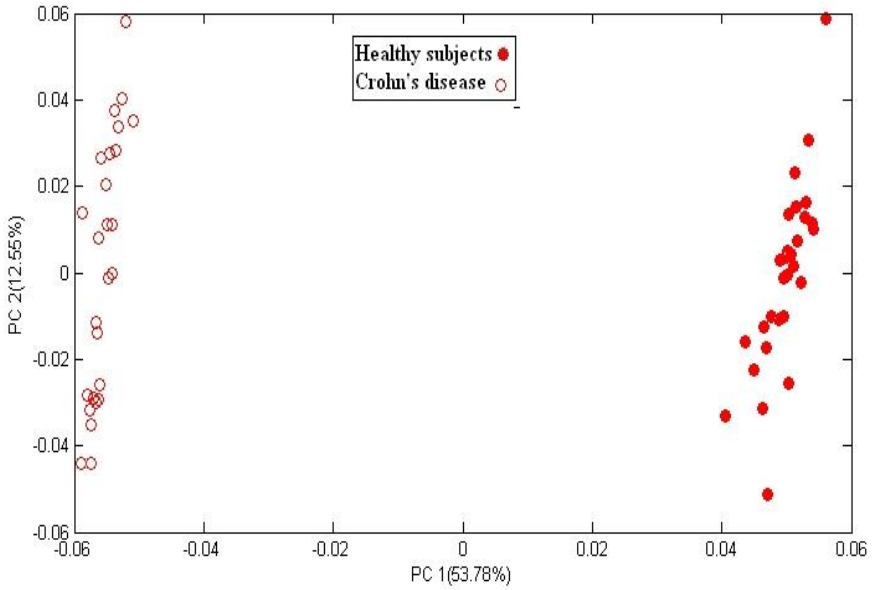

Figure2. Results of mean center for Crohn's disease

In Figure1 variability of 56.31% was expressed by two PCs with auto scale which 41.48% was described by PC1, 14.83% was described by PC2. In Figure 2 Variability of 66.33% was expressed by two PCs with mean center.

A comparison between Figure 1 and Figure 2 reveals that segregation patients and healthy subjects was not achieved with auto scaling but there is an absolute classification when we apply mean center scaling. This result was obvious in Celiac disease, too (see Figure3 and figure4).

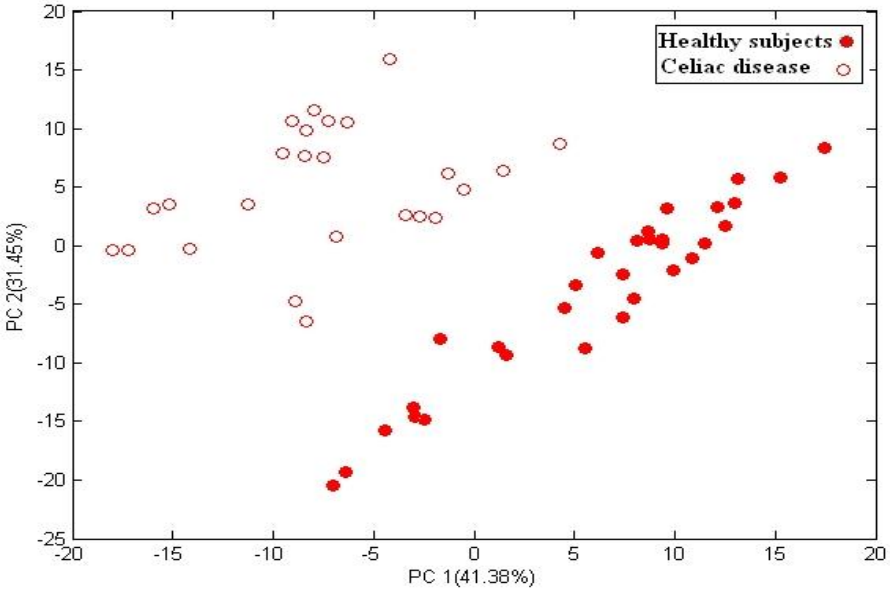

Figure3. Results of auto scale for Celiac's disease

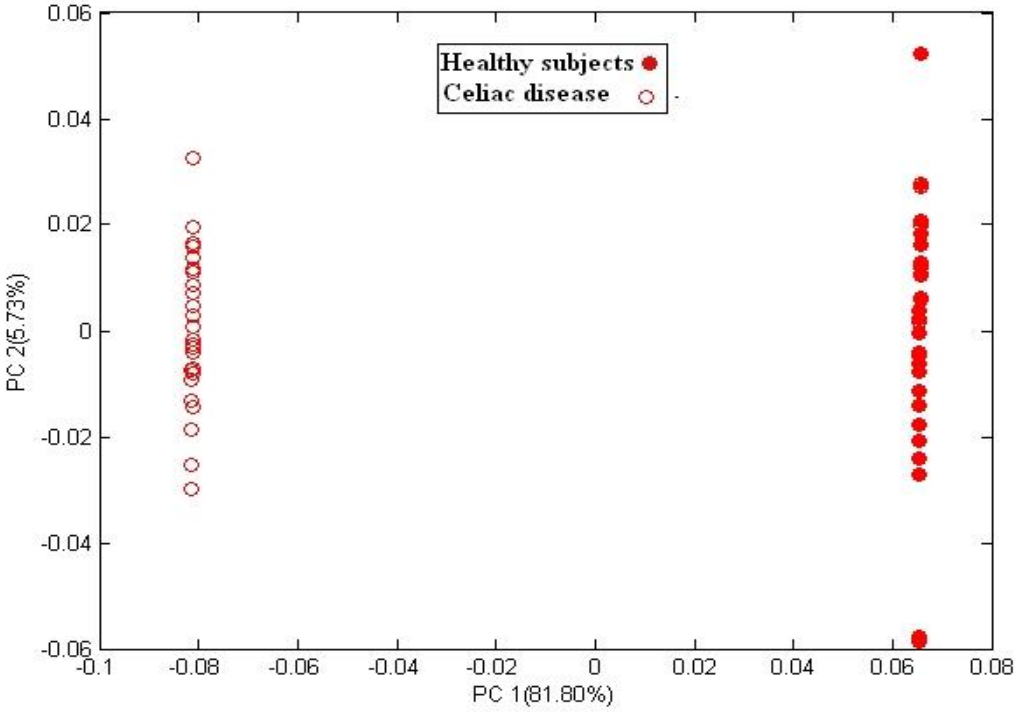

Figure 4. Results of mean center for Celiac's disease

The most chemical measurements are inherently multivariate. This means that more than one measurement can be made on a single sample. A clear example is NMR spectroscopy that can be recorded a spectrum at hundreds of chemical shifts on a single sample. In this study we intend to answer to matter such as: Which compounds behave similarly? Which people belong to a similar group? PCA is one of several multivariate statistics techniques helping classification data into specific groups and exploring patterns in these data. In explanation of PCA was expressed that the number of significant PCs is ideally equal to the number of significant components. In this study the number of PCs is equal to the number of samples. Each PC is characterized by two sets of information, the scores and the loadings. The scores in the case of NMR correlate to the number of samples (two groups), and the loadings correlate to the spectra. Scores plots often give useful information about the relationships between the samples (rows) in the data set. Plots can be done as the projections of the samples onto a single eigenvector versus sample number or onto the plane formed by two eigenvectors. A projection of the samples onto the two eigenvectors associated with the largest eigen values depicts the largest amount of information about the relationship between the samples that can be shown in two (linear) dimensions.

Figures 2 and 4, respectively, show the patient group of Crohn's and Celiac diseases behave very similarly whereas group of control behave in a diametrically opposite manner. We show that the mean centering is more useful to segregate patients and healthy subjects. We demonstrated that mean centering option is more effective at eliminating correlation from the PCA residuals than auto scaling in this situation. Imagine for the data set where the variables have widely different deterministic variances, but have added measurement noise of identical magnitudes. Furthermore, assume that the deterministic variation in the data set is confined to a subspace of the data space. In this case the auto scaling option increases the variance of variables that are

mostly noise relative to those that are mostly deterministic variation raising the effective noise level of the data set. Any PCA model that is formed from this data would be more likely to capture noise in the model and thus less likely to include all deterministic variation.

In this study, we present scaling effects in NMR spectroscopic metabonomics data sets from serum of patients diagnosed with Crohn's and Celiac diseases. We applied two scaling techniques, mean centering and auto scaling, for segregate patients diagnosed with Crohn's disease and healthy subjects, and patients diagnosed with Celiac disease and healthy subjects. Data sets of Crohn's and Celiac diseases both of them, mean centering is more useful to segregate patients and healthy subjects because of the mean centering option is more effective at eliminating correlation from the PCA residuals than the auto scaling in this situation.

ACKNOWLEDGMENT

We gratefully acknowledge financial support from Iran National Science Foundation (INSF), Sharif University of Technology and Shahid Beheshti University of Medical Sciences. This paper is resulted from PhD thesis of Seyed AbdolReza Mortazavi-Tabatabaei.

REFERENCES

1. Baumgart D, Carding S. Inflammatory bowel disease: cause and immunobiology. The Lancet 2007;369:1627–40.
2. Baumgart D, Sandborn W. Inflammatory bowel disease: clinical aspects and established and evolving therapies. The Lancet. 2007;369:1641–57.
3. Xavier R, Podolsky D. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007;448:427–34.
4. Mark P, Chang M, Chow EJ, Tabibzadeh S, Kirit-Kiriak V, Targan SR, et al. Identification of a prodromal period in Crohn's disease but not ulcerative colitis. American Journal of Gastroenterology 2000;95:3458–62.
5. Scheinfeld N, Teplitz E, McClain S. Crohn's disease and lichen nitidus: a case report and comparison of common histopathologic features. Inflammatory bowel diseases 2001;7:314–8.
6. Torres M, Casado M, Rios A. New aspects in celiac disease. World J Gastroenterol 2007;13:1156-61.
7. Holtmeier W, Caspary WF. Celiac disease. Orphanet Journal of Rare Diseases. 2006;1:3-11.
8. Goddard C, Gillett H. Complications of coeliac disease: are all patients at risk? Postgrad Med J 2006;82:705–12.
9. Beckwith-Hall BM, Brindle JT, Barton RH, Muireann Coen, Holmes E, Nicholson JK, et al. Application of orthogonal signal correction to minimise the effects of physical and biological variation in high resolution 1H NMR spectra of biofluids. Analyst. 2002;127:1283–8.
10. Holmes E, Nicholls A, Lindon J, Connor S, Connelly J, Haselden J, et al. Chemometric Models for Toxicity Classification Based on NMR Spectra of Biofluids. 2000;13:471-8.
11. Lindon J, Nicholson J, Holmes E, Everett J. Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. Concepts in Magnetic Resonance. 2000;12:289-320.
12. Nicholson J, Lindon J, Holmes E. Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999;29:1181-9.
13. Nicholson J, Wilson I. Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat Rev Drug Discov. 2003;2:668-76.
14. Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC. Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets. Anal Chem. 2006;78:2262-7.
15. Zhang G, Hirasaki G. CPMG relaxation by diffusion with constant magnetic field gradient in a restricted geometry: numerical simulation and application,. Journal of Magnetic Resonance. 2003;163:81–91.
16. Spraul M, Neidig P, Klauck U, Kessler P, Holmes E, Nicholson J, et al. Automatic reduction of NMR spectroscopic data for statistical and

pattern recognition classification of samples. J Pharm Biomed Anal. 1994;12:1215-25.
17. Holmes E, PJFoxall, Nicholson J, Neild G, Brown S, Beddell C, et al. Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic resonance spectra of human urine from normal and pathological states. Anal Biochem. 1994;220:284-96.
18. Huifeng W, Xiaoli L, Jianmin Z, Junbao Y. NMR-Based Metabolomic Investigations on the Differential Responses in Adductor Muscles from Two Pedigrees of Manila Clam Ruditapes philippinarum to Cadmium and Zinc. Mar Drugs. 2011;9:1566-79.
19. Teahan O, Gamble S, Holmes E, Waxman J, Nicholson JK, Bevan C, et al. Impact of Analytical Bias in Metabonomic Studies of Human Blood Serum and Plasma. Anal Chem. 2006;78:4307-18.
20. Niazi A, Azizi A. Orthogonal Signal Correction – Partial Least Squares Method for Simultaneous Spectrophotometric Determination of Nickel, Cobalt, and Zinc. Turk J Chem. 2008;32:217 – 28.
21. McVean G. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet. 2009;5:e1000686.
22. Vandeginste BGM, Massart DL, Buydens LMC, Jong Sd, Lewi PJ, Smeyers-Verbeke J. Handbook of Chemometrics and Qualimetrics. B P, editor. Amsterdam: Elsevier; 1998.
23. The Mathworks I, Matlab version 5.3.0.10183, Natick, MA. 1999.
24. Eigenvector Research Inc. P-Tc. 1998.