# Identification of Markers from a Set of Spectral Courses

JIRI KNIZEK
Charles University in Prague
Faculty of Medicine
Simkova 870
500 38 Hradec Kralove
CZECH REPUBLUC
knizekj@centrum.cz

LADISLAV BERANEK
University of South Bohemia
Institute of Applied Informatics
Branisovska 31a
370 05 Ceske Budejovice
CZECH REPUBLIC
beranek@ef.jcu.cz

PAVEL BOUCHAL
Masaryk University in Brno
Faculty of Science
Kamenice 753/5
625 00 Brno
CZECH REPUBLUC
bouchal@chemi.muni.cz

BORIVOJ VOJTESEK
Masaryk Memorial Cancer
Institute
Zluty kopec 543/7
602 00 Brno
CZECH REPUBLUC
nenutil@mou.cz

RUDOLF NENUTIL
Masaryk Memorial Cancer
Institute
Zluty kopec 543/7
602 00 Brno
CZECH REPUBLUC
nenutil@mou.cz

PAVEL TOMSIK
Charles University in Prague
Faculty of Medicine
Simkova 870
500 38 Hradec Kralove
CZECH REPUBLUC
tomsikpj@lfhk.cuni.cz

*Abstract:* A brief introduction of algorithms for the statistical identification of markers from a set of spectral courses is the topic of our paper. Partial results, demonstrated by pictures, are very promising. The proposed algorithm is generally applicable for an arbitrary problem of marker identification by tests in a set of quantifying dependences.

*Key–Words:* Marker, biomarker, regression, tests of hypotheses, software

## 1 Introduction

The rapid development of genomic and proteomic methods led to an enormous increase in experimental data. To be able to extract answers to important questions from these data, it is necessary to find an effective bio-statistical method for their processing. Application of advanced methodologies is necessary to give us more detailed, structured information.

A (dependence) biomarker, or (dependence) biological marker, is a (dependence) indicator of a biological state. It is a characteristic that is objectively measured and evaluated as a (dependence) indicator of normal biological processes, pathogenic processes, or pharmacologic (dependence) responses to a therapeutic intervention. It is used in many scientific fields. The presence and concentration of certain biomarker molecules is then identified and measured. That is why our proposed algorithm based exclusively on classical statistical decision making with the help hypotheses testing may facilitate prediction of certain clinical aspects of diseased patients.

Medical and biological research often deals with miscellaneous dependences. A particular experimental problem in which dependences are dealt with is then, in terms of the previous paragraph, described by *the means of regression functions* $\eta_1(x)$, $\eta_2(x)$, $\ldots$, $\eta_M(x)$ [13], [14].

Very often, the problem is specified in such a way that the first group of experimental dependences models the data measured in *the group of diseased patients* and the second group of experimental dependences models the data measured in *the group of healthy patients*. *Spectral methods* represent a large class of physical methods which are based on *two dimensional dependences*. We denote the group of spectral dependences which model the data of the type *“group of diseased patients“* by the regression functions ${}_{\text{diseased}}\eta_1(x)$, ${}_{\text{diseased}}\eta_2(x)$, $\ldots$, ${}_{\text{diseased}}\eta_{M_{\text{diseased}}}(x)$ and the group of spectral dependences which model the data of the type *“group of healthy patients“* by the regression functions ${}_{\text{healthy}}\eta_1(x)$, ${}_{\text{healhy}}\eta_2(x)$, $\ldots$, ${}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x)$. The quantity $x$ is a real independent variable that may represent time, effective mass in the case of MS[1], etc. Then the total number of de-

[1] mass spectroscopy; especially by the use of spectral methods, is a monitored process that can take place in some relatively very narrow spectral region (in relation to the whole possible magnitude)

pendences or the total number of regression functions is

$$M = M_{\text{diseased}} + M_{\text{healthy}}.$$

In this paper, *a new regression algorithm for statistical identification of markers from a set of spectral courses* is described. There are *definition matters* presented in section **2**. There is *a model "set of multiple linear regressions"* presented in section **2.1** which serves as an initial basis of whole algorithm. Section **2.2** deals with a more narrow model for our purpose - model *"a set of orthogonal polynomial regressions"*, i.e., practically with model *"a set of spectral courses"*. Section **2.3** deals with the newly well-established statistical tool *"the definition matrix"* which simplifies the definition of various statistical tests of dependences. For the solution of the problem, key principals are presented in section **3**. Some numerical-mathematical aspects of used algorithms (and their solution with the help of *"highly effective algorithm for orthogonalization"*) are referred to in section **3.1**. Section **3.2**, named *"identification of markers by simultaneous tests in a set of quantifying dependences"*, deals with compliance to the fundamental *biophysical principles* at the algorithm application. There are *experimental results* presented in section **4**. Description of *"biomarker pictorial exemplifications on real data"* is presented in section **4.1**. Section **4.2** deals with very promising results of *identifying biomarker areas in SELDI-TOF mass spectra* of data which has been obtained from 10 patients suffering from renal cell carcinoma.

## 2 Definitions

### 2.1 The model "A set of multiple linear regressions"

The beginning of the algorithmic study described below is based on following test criterion:

$$\lambda_F = \frac{(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{C}\boldsymbol{R}')^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})/J}{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})\,(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/\,(MT-K)} \sim F_{(J,\,MT-K)},$$
$$\boldsymbol{C} = [\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})\boldsymbol{X}]^{-1}, \tag{1}$$

for the standard statistical model called the "Disturbance-Related Sets of Regression Equations"

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & & & \\ & \boldsymbol{X}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{X}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_M \end{bmatrix} \tag{2}$$

(or briefly $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$) for the null hypothesis

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}, \tag{3}$$

where the form of the $\boldsymbol{R}_{(J\times K)}$matrix of constants and the form of the $\boldsymbol{r}_{(J\times 1)}$ vector of constants in relation (3) concretize *the null hypothesis* $H_0$. Dimension $K$ of regression vector $\boldsymbol{\beta}$ is given as a sum of single regression vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_M$, i.e. $K = \Sigma_{i=1}^{M}(K_i + 1)$. The covariance matrix $\boldsymbol{\Omega}$ of the joint disturbance vector $\boldsymbol{e}$ is given by $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$ and so $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}$ [6], [12].

It is important that it is possible to test arbitrary linear mutual relations among particular multiple linear regressions in (2) with the help of the test criterion (1).

### 2.2 The model known as "A set of orthogonal polynomial regressions"

It is necessary to approach model (2) more narrowly for our purpose (namely) *"the statistical identification of markers from a set of spectral courses"*. Every multiple linear regression in model (2) is interpreted as an orthogonal polynomial regression describing one appropriate spectral course. Thus, we can test (with the help of (3) appropriately modified) arbitrary linear mutual relations among particular spectral courses [13], [14].

### 2.3 The definition matrix

When we summarize the values of regression functions (polynomial regressions) into the vector

$$\boldsymbol{\eta}\,(x){=}(\eta_1(x), \eta_2(x), \ldots, \eta_M(x))', \tag{4}$$

we can formally transcribe a null hypothesis (3) into the form

$$H_0 : \boldsymbol{k}\boldsymbol{\eta}\,(x) = \boldsymbol{r}\,(x)\,,$$

where an abscissa $x$ is the arbitrary value of used spectral independent variable and

$$\boldsymbol{k} = \begin{pmatrix} k_{1,\,1} & k_{1,\,2} & \ldots & k_{1,\,M} \\ k_{2,\,1} & k_{2,\,2} & \ldots & k_{2,\,M} \\ \vdots & \vdots & \ddots & \vdots \\ k_{J,\,1} & k_{J,\,2} & \ldots & k_{J,\,M} \end{pmatrix}$$

is the so called *definition matrix* [14]. Definition matrix $\boldsymbol{k}$ expresses generally all conceivable linear mutual relations among regression functions (4).

# 3 Main idea

## 3.1 A highly effective algorithm for orthogonalization

Computational practice showed that the currently used Gram-Schmidt's polynomials [5], [19] are not able to provide satisfactory measure of orthogonality. For our purpose – the polynomial approximation of a set of spectral courses – we had to use special, outstandingly efficient algorithms [1], [5], [7] - [10], [16], [18].

## 3.2 Identification of markers using simultaneous tests in a set of quantifying dependences

As substantial limitation while using the *"Test of the Hypothesis That One Group of Dependences is Consistent with Another Group of Dependences"* [14] is that the null hypothesis $H_0 : \boldsymbol{k}_{(J \times M)} \boldsymbol{\eta}_{(M \times 1)}(x) = \boldsymbol{r}_{(J \times 1)}(x)$ can be rejected in favour of the double-sided alternative that at least one of the $J$ linear relations $\boldsymbol{k}_{(J \times M)} \boldsymbol{\eta}_{(M \times 1)}(x) = \boldsymbol{r}_{(J \times 1)}(x)$ is not valid. However, *the biophysical principles* of the problem force the experimenter to assume that changes in the concentration of a given biomarker are *natural*, i.e., *complete*. It means that the experimenter would need to reject the null hypothesis $H_0 : \boldsymbol{k}_{(J \times M)} \boldsymbol{\eta}_{(M \times 1)}(x) = \boldsymbol{r}_{(J \times 1)}(x)$ in favour of the double-sided alternative that all $J$ linear relations $\boldsymbol{k}_{(J \times M)} \boldsymbol{\eta}_{(M \times 1)}(x) = \boldsymbol{r}_{(J \times 1)}(x)$ together are not valid. Resulting from these necessities is the fact that mutual conformity is available *only and only* in the cases where the number of tested linear relations is $J = 1$.

It emerges from these reasons that instead of testing one null hypothesis $H_0 : \boldsymbol{k}_{(J \times M)} \boldsymbol{\eta}_{(M \times 1)}(x) = \boldsymbol{r}_{(J \times 1)}(x)$, we must test $\kappa$ simultaneous null hypotheses $H_0^j : \boldsymbol{k}^j_{(1 \times M)} \boldsymbol{\eta}_{(M \times 1)}(x) = \boldsymbol{r}^j_{(1 \times 1)}(x) = r^j(x)$, where the index for the $j^{\text{th}}$ simultaneous null hypothesis is $j = 1, 2, \ldots, \kappa$. The size of the number $\kappa$, the concrete form of *definition row vectors* $\boldsymbol{k}^j_{(1 \times M)}$ and elements $r^j(x)$ is then dependent on whether our data are paired, unpaired or combined. This means that (for a given abscissa $x$) the appropriate *simultaneous* null hypotheses are rejected when un-equalities

$$\begin{aligned} &p_j(x) < \alpha/\kappa, \quad j = 1, 2, \ldots, \kappa, \\ &p(x) = p_{j'}(x) = max_{j=1,2,\ldots,\kappa} \; p_j(x) < \alpha/\kappa, \end{aligned} \tag{5}$$

are simultaneously valid. Along with this condition, the appropriate *power analysis-un-equalities*

$$\begin{aligned} &1 - \beta_j(x) \geq convention \; limit, \quad j = 1, 2, \ldots, \kappa, \\ &1 - \beta(x) = 1 - \beta_{j'}(x), \end{aligned} \tag{6}$$

must be fulfilled.

The requested power of the test (in other words *the convention limit*) depends on the test significance level $\alpha$: $1 - \beta_{\text{req}}(\alpha = 0.05) = 0.8$ and $1 - \beta_{\text{req}}(\alpha \leq 0.01) = 0.95$ [3], [4].

For test significance levels $\alpha$ greater than $\alpha = 0.05$, the requested powers of the test are $1 - \beta_{\text{req}}(\alpha = 0.1) = 0.6125$, possibly $1 - \beta_{\text{req}}(\alpha = 0.2) = 0.2375$.

# 4 Experiments

## 4.1 Pictorial exemplifications of real data, Figures 1-9

The *potential biomarker areas* were obtained by the proposed data-treatment of the mass spectral data, measured with the aim of *identifying renal cell carcinoma biomarkers*. Two experimental groups (diseased and healthy, i.e. red and blue) are demonstrated in figures. Pentagrams "$\star$" : discrete courses of the measured (renal cell carcinoma) spectrum; pentagrams "$\star$": discrete courses of the measured (not from renal cell carcinoma) spectrum; solid lines: statistical estimations of the courses of *function dependences* based on the experimental courses of "$\star$" and "$\star$".

Conventional decision making conditions (5) and (6) are satisfied in the whole measurement range at all the $1^{st}$-$9^{th}$ figures. The physical unit of the independent variable (effective mass) in all pictures is Dalton. The physical unit of the dependent variable (intensity of mass-spectrum) in all pictures is as a %. Appropriate potential biomarker areas are then located around the *x*-ordinates of appropriate dependent variable maximums.

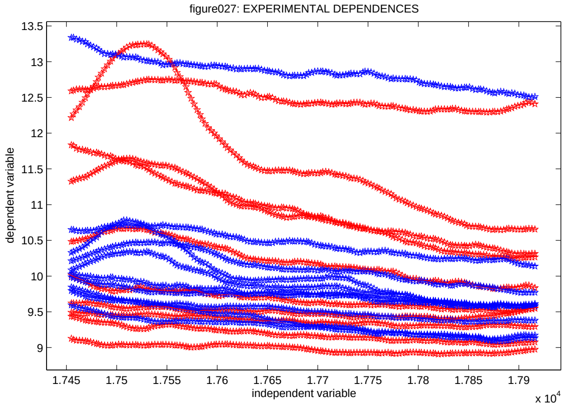

**Figure 1.** See very detail comments in the section "Pictorial exemplifications of real data, Figures 1-9".

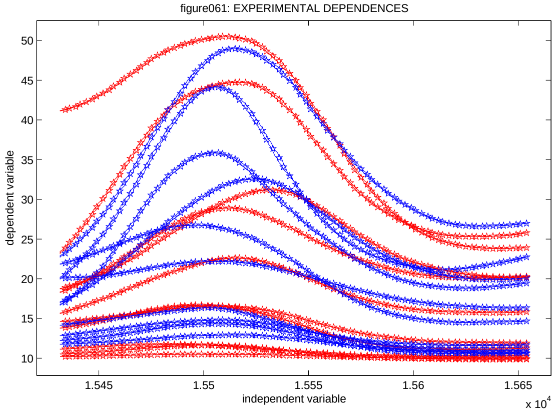

**Figure 2.** See very detail comments in the section “Pictorial exemplifications of real data, Figures 1-9”.

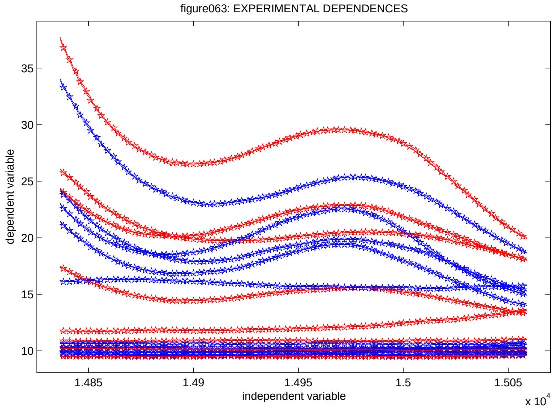

**Figure 3.** See very detail comments in the section “Pictorial exemplifications of real data, Figures 1-9”.

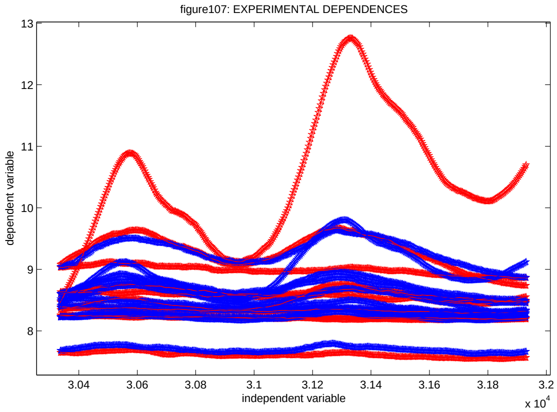

**Figure 4.** See very detail comments in the section “Pictorial exemplifications of real data, Figures 1-9”.

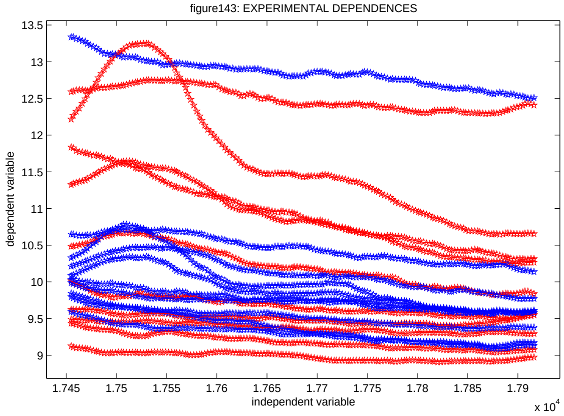

**Figure 5.** See very detail comments in the section “Pictorial exemplifications of real data, Figures 1-9”.

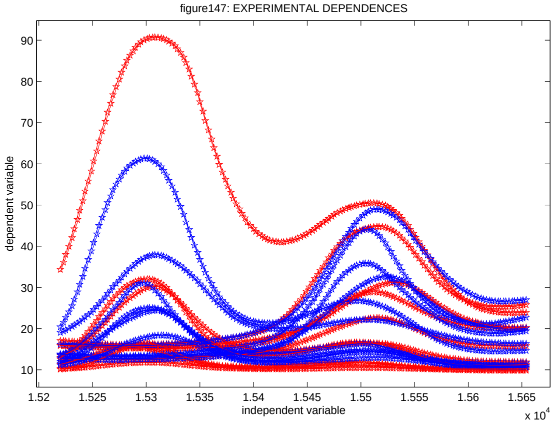

**Figure 6.** See very detail comments in the section “Pictorial exemplifications of real data, Figures 1-9”.

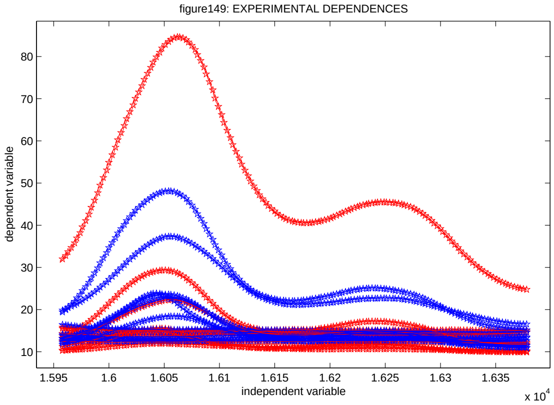

**Figure 7.** See very detail comments in the section “Pictorial exemplifications of real data, Figures 1-9”.

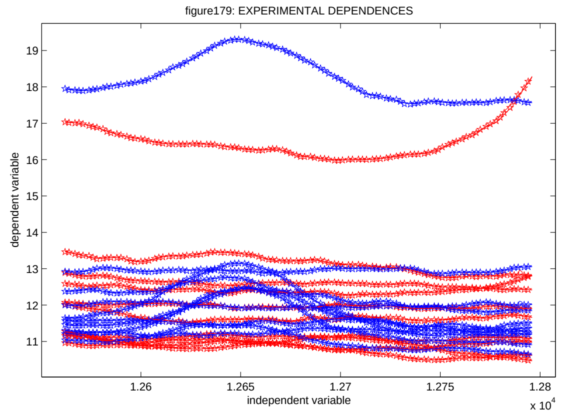

**Figure 8.** See very detail comments in the section "Pictorial exemplifications of real data, Figures 1-9".

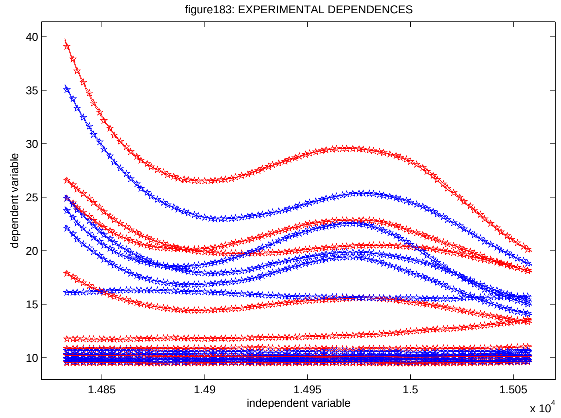

**Figure 9.** See very detail comments in the section "Pictorial exemplifications of real data, Figures 1-9".

## 4.2 Identifying biomarker areas in SELDI-TOF mass spectra

A large set of (normalized) mass spectral data, measured with the aim of *identifying renal cell carcinoma biomarkers*, was subjected to the algorithm described above. A group of data was obtained from 10 patients suffering from renal cell carcinoma. One group of data was obtained from renal cell carcinoma tissue, the second group of data was obtained from the same patients but from healthy (i.e. not renal cell carcinoma) tissue. Naturally, *the paired version* of the proposed algorithm was used here. Spectra were divided into segments containing 200 points. The findings of the already *discovered biomarker* "$\alpha$B-crystallin"[2] [11] by the proposed algorithm *was confirmed*. The proposed algorithm is *very sensitive*, because additional *potential biomarker areas* have been found. It managed to find at least 12 cases of other *biomarker areas*. See figures 1-9[3].

## 5 Discussion and conclusions

There is no doubt at present that computerized technologies in medicine and biological research, e.g. proteomics and genomics, need new approaches. This paper deals with *"The Regression Algorithm for Statistical Identification of Markers From a Set of Spectral Courses"* in cases where data error disturbances have a normal distribution.

The proposed algorithm works in practice very well. At first sight, this property of the algorithm could appear rather unexpected, considering the very rigorous necessary requirements for the simultaneous testing (1) of the appropriate $p(x)$-values.

The discovered principles are generally usable in analogical spectroscopy studies, i.e., not only for treatment of MS for the purpose of biomarker identification. They are even generally applicable to the arbitrary problem of marker identification (used in miscellaneous branches of human activity) by simultaneous tests in a set of quantifying dependences.

With the help of an appropriate mass spectra database analysis, the proposed methodological approach will lead to the construction of *a clinic running system* which will allow *statistical decision making concerning suspicion of disease in patients* [17], [18].

**Acknowledgements:** This work was supported by the Czech Science Foundation (grant No. P304 /10/0868) and by the European Regional Development Fund (RECAMO; CZ 1.05/2.1.00/03.0101). This study was supported by the programs PRVOUK P37/01, PRVOUK P37/09 and GACR P303/12/P536.
.

*References:*

[1] Arnoldi, W.E., The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, **9**, 17-29 (1951).

[2] Ciphergen® Biosystems, Inc., ProteinChip software 3.1. Operation Manual (2002).

[2] Ciphergen-software [2]

[3] Numbers of *biomarker areas* in particular figures: f.1: 1, f.2: 1, f.3: 1, f.4: 2, f.5: 1, f.6: 2, f.7: 2, f.8: 1, f.9: 1. Note: Numbers of figures in headings of particular figures (e.g. "figure063" and the like) are order numbers of particular (200 points) original spectral segments.

[3] Cohen, J., *Statistical Power Analysis for the Behavioral Science.* Mahwah, New Persey: $2^{nd}$ edn Lawrence Erlbaum (1988).

[4] Daly, L.E., Bourke, G.J., *Interpretation and Uses of Medical Statistics.* Oxford: Blackwell Science $5^{th}$ edn; 276-279 (2000).

[5] Forsythe, G.E., *Generation and Use of Orthogonal Polynomials for Data-fitting on a Digital Computer,* J Soc Indust Appl Math. **5**, 74-88 (1957).

[6] Gatignon, H., *Statistical Analysis of Management Data.* Kluwer Academic Publishers (New York, Boston, Dordrecht, London, Moscow) (2003).

[7] Gautschi, W., *Orthogonal polynomials: computation and approximation*. Numerical Mathematics and Scientific Computation. Oxford Science Publications. Oxford University Press, New York (2004).

[8] Giraud, L., Langou, J., Rozloznik, M., On the loss of orthogonality in the Gram-Schmidt orthogonalization process. *Computers & Mathematics with Applications*, **50**, 1069–1075 (2005).

[9] Giraud, L., Langou, J., Rozloznik, M., van den Eshof, J., Rounding error analysis of the classical Gram-Schmidt orthogonalization process. *Numer. Math.*, **101**, 87-100 (2005).

[10] Higham, N.J., *Accuracy and stability of numerical algorithms*. Second edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002).

[11] Holcakova, J., Hernychova, L., Bouchal, P., Brozkova, K., Zaloudik, J., Valik, D., Nenutil, R., Vojtesek, B., *Identification of $\alpha$B-crystallin, a biomarker of renal cell carcinoma by SELDI-TOF MS*, The International Journal of Biological Markers, Italy: Wichtig editore, 23, 1, 48-53 (2008).

[12] Judge, G.G, Griffiths, W.E, Hill, R.C, Lutkepohl, H., Tsoung-Chao, L., *The Theory and Practice of Econometrics*, J. Wiley, New York (1985).

[13] Knizek, J., Sindelar, J., Beranek, L., Vojtesek, B., Nenutil, R., Brozkova, K., Drazan, V., Hubalek, M. & Kubacek, L., *Power function for tests of null hypotheses on mutual linear regression functions' relations*, International Journal of Applied Mathematics & Statistics, Volume 2; Number S08; Bull. Stat. Econ., ISSN 0973-7022. pp. 26-33 (2008).

[14] Knizek, J., Sindelar, J., Pulpan, Z., Vojtesek, B., Nenutil, R., Brozkova, K., Drazan, V., Hubalek, M. & Beranek, L., *Test of the Hypothesis That One Group of Dependences is Consistent with Another Group of Dependences*, International Journal of Applied Mathematics & Statistics, Volume 2; Number A08; Bull. Stat. Econ., ISSN 0973-7022. pp. 2-18 (2008).

[15] Knizek, J., Sindelar, J., Vojtesek, B., Bouchal, P., Nenutil, R. & Beranek, L., *Identification of Markers by Simultaneous Tests in a Set of Quantifying Dependences*, International Journal of Statistics & Economics (formerly known as the “Bulletin of Statistics & Economics”), A10, Volume 5 [Special], Number A10. pp. 12-20 (2010).

[16] Knizek, J., Tichy, P., Beranek, L., Sindelar, J., Vojtesek, B., Bouchal, P., Nenutil, R. & Dedik, O., *Note on Generating Orthogonal Polynomials and Their Application in Solving Complicated Polynomial Regression Tasks*, International Journal of Mathematics and Computation, ISSN 0974-570X (Online), ISSN 0974-5718 (Print), Vol. 7; No. J10; June 2010. pp. 48-60 (2010).

[17] Knizek, J., Sindelar, J., Vojtesek, B., Bouchal, P., Nenutil, R., Beranek, L. & Dedik, O., *Using Markers to Aid Decision Making in Diagnostics*, International Journal of Tomography & Statistics, ISSN 0973-7294 (Online), ISSN 0972-9976 (Print), W11, Volume 16, Number W11. pp. 41-55 (2011).

[18] Knizek, J., 2011b, *Marker Statistics I.: Regression analysis of dependences in medicine and molecular biology*, VDM Publishing House Ltd., Mauritius (2011).

[19] Ralston, A., *A First Course in Numerical Analysis*, McGraw Hill Book Company, New York (1973).