Bulletin of Statistics & Economics,
Autumn 2008; Volume 2; Number A08; Bull. Stat. Econ.
ISSN 0973-7022;

Test of the Hypothesis That One Group of Dependences is Consistent with Another Group of Dependences

Jiri Knizek[1], Jan Sindelar[2], Zdenek Pulpan[3], Borivoj Vojtesek[4], Rudolf Nenutil[4], Kristyna Brozkova[4], Viktor Drazan[5], Martin Hubalek[6], and **Ladislav Beranek[7]**

[1]Charles University in Prague, Faculty of Medicine in Hradec Kralove:
Department of Medical Biophysics, Simkova 870, 500 38 Hradec Kralove, Czech Republic
Email: *knizekj@lfhk.cuni.cz*

[2]Academy of Sciences of the Czech Republic, Institute of Information Theory and Automation:
Department of Stochastic Informatics, Czech Republic

[3] University of Hradec Kralove, Faculty of Education, Department of Mathematics

[4]Masaryk Memorial Cancer Institute in Brno: Experimental oncology division, Czech Republic

[5]Academy of Sciences of the Czech Republic: Institute of Biophysics in Brno, Czech Republic

[6]University of Defense in Brno, Faculty of Military Health Sciences in Hradec Kralove: Institute of Molecular Pathology, Czech Republic

[7] University of South Bohemia in Ceske Budejovice, Faculty of Education:
Department of Computer Science, Czech Republic

ABSTRACT

In this paper we describe the basic algorithmic characteristics of the test of the hypothesis that one group of dependences is consistent with another group of dependences for a case when the error disturbances of this data have normal distribution. Unpaired and paired versions are presented. The aim of our work was to find effective algorithms for biomarker identification in mass (MS) or Raman spectra (RS). This approach has been applied to mass spectral data and measured for identification of kidney carcinoma biomarker. Further more, the algorithm has been applied to Raman spectral data with the aim of recognizing the efficiency of heat denaturation of calf thymus DNA. The results of the applications show the need to find new regression algorithms, for example in non-parametric statistics.

Key words: Decision making; Test of hypothesis; Biomarkers; Mass spectra; Raman spectra

Mathematics Subject Classification: 62F03, 62J99

JEL Classifications: C12

1. INTRODUCTION

Rapid development of genomic and proteomic methods have 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. The application of advanced methodologies is necessary to give us more detailed structured information.

In the title of this paper and also in many of its paragraphs, the term *dependence* is used. The term *dependence* means the so-called *functional dependence*. Consequently, a certain value of the independent variable x always corresponds to only one certain value of the dependent variable y. It is assumed that *functional dependence* is identical with the studied *physical dependence*, and

eventually *biophysical dependence*, whose exact mathematically analytical form is in the majority of cases unknown.

The text also deals with the so-called *regression dependence*. In that case, for a certain value of the independent variable x, there always exists a certain random distribution of the dependent variable y (random quantity). An appropriate *regression function* is then considered as a mean value of *regression dependence*. We denote *regression function* with the symbol $\eta(x)$. In this article, the mentioned tests of hypotheses about mutual relations of dependences are, in fact, presented as *tests of hypotheses about mutual relations of regression functions*. At the same time, it is assumed that *the regression functions* are sufficient approximations of the appropriate so-called *functional* (i.e. *studied biophysical*) *dependences*.

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* $\eta_1(x)$, $\eta_2(x)$, ... , $\eta_M(x)$.

Apart from the above mentioned information, further in the text the so-called *experimental dependences are also discussed.* These *dependences* are created by the appropriate experimentally measured or observed data items corresponding with the independent variable x and dependent variable y .

Particularly often, however, the following problem occurs in medical and biological research.

It is necessary to compare the set of dependences of the type "*group of sick patients*" with the set of dependences of the type "*group of healthy patients*". In that case, what is effectively being dealt with are *experimental dependences* measured, e.g., on population samples of people, experimental animals, micro-organisms, etc. Or it is necessary to compare the set of *experimental dependences* of the type "*the given noxious substance was applied*" with the set of experimental dependences of the type "*the given noxious substance was not applied*". Subsequently, statistical data evaluation solves this corresponding decision-making problem. That means *the test of the hypothesis that one group of dependences is consistent with another group of dependences* is conducted (null hypothesis H_0). Generally, *the aim of the research is* to demonstrate that a given noxious substance has ascertainable influence, i.e., *to reject the null hypothesis* H_0 .

Very often the problem is specified in such a way that the first group of experimental dependences models the data measured on *the group of sick patients, while* the second group of experimental dependences models the data measured on *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 models the data of the type *"group of sick patients"* by the regression functions ${}_{\text{sick}}\eta_1(x)$, ${}_{\text{sick}}\eta_2(x)$, ... , ${}_{\text{sick}}\eta_{M_{\text{sick}}}(x)$, and the group of spectral dependences which models the data of the type *"group of healthy patients"* by the regression functions ${}_{\text{healthy}}\eta_1(x)$, ${}_{\text{healthy}}\eta_2(x)$, ... , ${}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x)$. The quantity x is a real independent variable that may represent time, effective mass in the case of MS, wave number in the case of RS, etc. The total number of dependences or the total number of regression functions then is $M = M_{\text{sick}} + M_{\text{healthy}}$.

In this paper, the algorithm for *the testing of the hypothesis that one group of dependences is consistent with another group of dependences* is described. There is a regression model presented in section *2* which serves as a basis for all *tests*. Section *3* deals with particular features of *the testing algorithm*. There is a general form of *null hypothesis* H_0 presented in subsection *3.1*. In subsection *3.2*, the so-called *definition matrix* is established and a special form of *null hypothesis* H_0 *about mutual relations of dependences* is presented. In subsection *3.3*, the relation for enumeration of *p*-value for *the test of null hypothesis* H_0 is introduced for cases where the vector of error disturbances has a normal distribution. An integral part of *the testing algorithm* is *a statistically optimized orthogonal polynomial regression of the set of all experimental dependences*. In subsection *3.4*, computation of optimized degrees of polynomials in this regression is presented. In subsection *3.5*, there is a description of *the test* itself, which forms the main part of this paper. In subsection *3.6*, the so-called *pair variant of the test* is presented, which is applied when measurement is performed on the same subject (man, experimental animal, micro-organism, etc.) before and after application of a given noxious substance. Section *4* treats some applications of *the test* in medical and biological research. Subsection *4.1* treats general features of mass (MS) and Raman (RS) spectra from the point of view of possible applications of *the test*. In subsection *4.2*, the way of spectra segmentation preprocessing is established. In subsection *4.3* and *4.4*, respectively, the concrete MS- and RS-application of *the test* is introduced.

2. STATISTICAL MODEL

We consider hereafter orthogonal polynomials $\psi_0(x), \psi_1(x), \psi_2(x), ...$ of the independent variable x, where $\psi_\kappa(x)$ is a polynomial of κth degree, $\kappa = 0,1,2,\ldots$ (Forsythe, 1957; Ralston, 1973; Golub and Van Loan, 1996). We further consider that $M_{\text{sick}} + M_{\text{healthy}}$ of regression functions ${}_{\text{sick}}\eta_1(x)$, ${}_{\text{sick}}\eta_2(x)$, ... , ${}_{\text{sick}}\eta_{M_{\text{sick}}}(x)$ and ${}_{\text{healthy}}\eta_1(x)$, ${}_{\text{healthy}}\eta_2(x)$, ... , ${}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x)$ are in fact M regression functions $\eta_1(x) = {}_{\text{sick}}\eta_1(x)$, $\eta_2(x) = {}_{\text{sick}}\eta_2(x)$, ... , $\eta_{M_{\text{sick}}}(x) = {}_{\text{sick}}\eta_{M_{\text{sick}}}(x)$, $\eta_{M_{\text{sick}}+1}(x) = {}_{\text{healthy}}\eta_1(x)$, $\eta_{M_{\text{sick}}+2}(x) = {}_{\text{healthy}}\eta_2(x)$, ... , $\eta_M(x) = {}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x)$. Regression functions

$$\eta_i(x) = \sum_{\kappa=0}^{K_i} \beta_{i,\kappa}\, \psi_\kappa(x) = \boldsymbol{\beta}_i \boldsymbol{\psi}_i'(x), \quad i = 1,2,\ldots,M,$$

are linear combinations of our orthogonal polynomials. Here $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,K_i})$ are the vectors of regression coefficients and

$$\boldsymbol{\psi}_i(x) = (\psi_0(x), \psi_1(x), \ldots, \psi_{K_i}(x)) \tag{1}$$

are vectors of orthogonal polynomials, where $i = 1,2,\ldots,M$. The number K_i is the degree of the highest orthogonal polynomial used at the construction of regression function $\eta_i(x)$, $i = 1,2,\ldots,M$.

We consider for the fixed vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)'$ of the independent variable x value vectors of the dependent variables $\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})'$, $i = 1,2,\ldots,M$, which are random quantities.

We have $(T \times (K_i + 1))$ dimensional matrices

$$\boldsymbol{\Psi}_i = \begin{pmatrix} \psi_0(x_1) & \psi_1(x_1) & \dots & \psi_{K_i}(x_1) \\ \psi_0(x_2) & \psi_1(x_2) & \dots & \psi_{K_i}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_0(x_T) & \psi_1(x_T) & \dots & \psi_{K_i}(x_T) \end{pmatrix}, \quad i = 1, 2, \dots, M,$$

of the values of our orthogonal polynomials.

The statistical model for *"the test of the hypothesis that one group of dependences is consistent with another group of dependences"* is *"the disturbance-related sets of regression equations, the case with contemporaneously correlated disturbances"* (Judge et al., 1985). We can express this model in the form

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Psi}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Psi}_M \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_M \end{pmatrix}$$

or alternatively in the brief form

$$\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{\beta} + \boldsymbol{e}. \tag{2}$$

Herein, the vector $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_M)'$ is a vector of dependent variables, the vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_M)'$ is a vector of regression coefficients and the vector $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_M)'$ is a vector of error disturbances. The matrix $\boldsymbol{\Psi}$ is a block diagonal matrix with dimensions $(MT \times K)$ with sub-matrices $\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, ..., \boldsymbol{\Psi}_M$ on the diagonal, whereby the dimension K of the regression vector $\boldsymbol{\beta}$ is the sum of dimensions of single regression vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, ... , $\boldsymbol{\beta}_M$, i.e. $K = \Sigma_{i=1}^{M}(K_i + 1)$.

We suppose that the mean values of error disturbances satisfy $E[\boldsymbol{e}_i] = \mathbf{0}$ and $E(\boldsymbol{e}_i \boldsymbol{e}_{i'}') = \sigma_{ii'} \boldsymbol{I}_T$ for any positive σ_{ij}, where $i, i' = 1, 2, \dots, M$. It is assumed that variances $\sigma_{ii'}$, where $i, i' = 1, 2, \dots, M$, are practically constant. From here the covariance matrix $\boldsymbol{\Omega}$ of the vector $\boldsymbol{e}$ of error disturbance is given by the relation

$$\boldsymbol{\Omega} = E[\boldsymbol{e}\boldsymbol{e}'] = \boldsymbol{\Sigma} \otimes \boldsymbol{I},$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \dots & \sigma_{MM} \end{pmatrix}.$$

Estimation retrieval of the estimate of the regression coefficient vector $\boldsymbol{\beta}$ proceeds in two stages. First, the estimation $\hat{\boldsymbol{\beta}}$ regression coefficient vector is found. It is assumed that the system of normal equations

$$\boldsymbol{\Psi}' \boldsymbol{\Omega}^{-1} \boldsymbol{\Psi} \hat{\boldsymbol{\beta}} = \boldsymbol{\Psi}' \boldsymbol{\Omega}^{-1} \boldsymbol{y}$$

has a regular solution (Forsythe, 1957; Ralston, 1973; Golub and Van Loan, 1996), where we put $\boldsymbol{\Omega} = \boldsymbol{I}$ in the first stage of the solution totaling two stages.

Elements of the estimate $\hat{\boldsymbol{\Sigma}}_{(M \times M)}$ of the matrix of mixed variances are then calculated with the help of the relations

$$\hat{\sigma}_{ii'} = T^{-1} \hat{\boldsymbol{e}}_i' \hat{\boldsymbol{e}}_{i'}, \quad i, i' = 1, 2, \ldots, M,$$

where LS-residuals $\hat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{\Psi}_i \hat{\boldsymbol{\beta}}_i$ and where $i = 1, 2, \ldots, M,$ are sub-vectors of the estimates of error disturbances.

The resulting estimate $\hat{\hat{\boldsymbol{\beta}}}$ is an EGLS-estimate of the regression coefficients' vector (Judge et al., 1985). It is arrived at by solution of the system of normal equations

$$\boldsymbol{\Psi}' \hat{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Psi} \hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{\Psi}' \hat{\boldsymbol{\Omega}}^{-1} \boldsymbol{y}, \quad \text{where} \quad \hat{\boldsymbol{\Omega}}^{-1} = \hat{\boldsymbol{\Sigma}}^{-1} \otimes \boldsymbol{I},$$

for which a regular solution is also supposed (Forsythe, 1957; Ralston, 1973; Golub and Van Loan, 1996).

This regression model has proven useful in the past at solving related problems of proteomics (Knizek et al., 2004b; Knizek, 2004a; Knizek, 2007a-c; Knizek, 2008).

3. HYPOTHESES TESTING

3.1 GENERAL FORM OF NULL HYPOTHESIS

Within the regression model (2), we can carry out *the test of the null hypothesis*

$$\left.\begin{array}{lcccc} H_0: & \boldsymbol{c} & \boldsymbol{\beta} & = & \boldsymbol{r} \\ & (J \times K) & (K \times 1) & & (J \times 1) \end{array}\right\} \tag{3}$$

for J linear relations among regression coefficients from the vector $\boldsymbol{\beta}$. An alternative hypothesis is the double-sided alternative, which, from J relations in (3) at least one, is not valid. The form of the $\boldsymbol{c}$ matrix of constants and the form of the $\boldsymbol{r}$ vector of constants in the relation (3) concretize *the null hypothesis* H_0. The rows of the matrix $\boldsymbol{c}$ have to be linearly independent.

3.2 *The definition matrix and the null hypothesis about mutual relations of dependences*

Hereafter, both the matrix $\boldsymbol{c}$ and the vector $\boldsymbol{r}$ depend on an abscissa x. That is why we replace the symbol $\boldsymbol{c}$ in relation (3) by the symbol $\boldsymbol{c}(x)$ and the symbol $\boldsymbol{r}$ with the symbol $\boldsymbol{r}(x)$.

We will test mutual linear relations among our regression functions. These relations will have the form

$$\sum_{i=1}^{M} k_{j,i}\,\eta_i(x)\,, \quad j=1,2,\ldots,J\,,$$

where the coefficients $k_{j,i}$, $j=1,2,\ldots,J$, $i=1,2,\ldots,M$ are real numbers. We summarize the coefficients from our linear relations into the matrix

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

which we denominate *the definition matrix.* We summarize the values of regression functions into the vector

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

The null hypothesis about linear relations among our regression functions then has the form

$$H_0:\boldsymbol{k}\boldsymbol{\eta}(x)=\boldsymbol{r}(x)\ . \tag{4}$$

We demonstrate that it is a special case of the null hypothesis (3). We summarize the vectors $\boldsymbol{\psi}_i(x)$ from (1), $i=1,2,\ldots,M$, into the block matrix

$$\boldsymbol{X}_{(M\times K)}(x)=\begin{pmatrix} \boldsymbol{\psi}_1(x) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\psi}_2(x) & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\psi}_M(x) \end{pmatrix}.$$

The equation $\boldsymbol{k}\boldsymbol{\eta}(x)=(\boldsymbol{k}\boldsymbol{X}(x))\,\boldsymbol{\beta}$ holds. If we then put in (3) that $\boldsymbol{c}(x)=\boldsymbol{k}\boldsymbol{X}(x)$, (3) assumes the form of (4).

Let's note that the rows of the matrix $\boldsymbol{c}(x)$ must be linearly independent and, consequently, also the rows of matrix $\boldsymbol{k}$ must be linearly independent.

3.3 p-Value

We calculate the *p-value* for the test of the hypothesis H_0: $\boldsymbol{c}(x)\boldsymbol{\beta}=\boldsymbol{r}(x)$, provided that the vector $\boldsymbol{e}$ of error disturbances has normal distribution, by means of the relation

$$p(x)=1-F_{J,MT-K}(\hat{\lambda})\,. \tag{5}$$

In relation (5), $F_{J,MT-K}(\hat{\lambda})$ is the distribution function of *F*-distribution with J and $MT-K$ degrees of freedom and test characteristic

$$\hat{\lambda} = \frac{(\boldsymbol{r}(x) - \boldsymbol{c}(x)\hat{\hat{\boldsymbol{\beta}}})'(\boldsymbol{c}(x)\,\hat{\boldsymbol{B}}\,\boldsymbol{c}'(x))^{-1}(\boldsymbol{r}(x) - \boldsymbol{c}(x)\hat{\hat{\boldsymbol{\beta}}}) \,/\, J}{(\boldsymbol{y} - \boldsymbol{\Psi}\hat{\hat{\boldsymbol{\beta}}})'(\hat{\boldsymbol{\Sigma}}^{-1}_{(M\times M)} \otimes \boldsymbol{I}_{(T\times T)})(\boldsymbol{y} - \boldsymbol{\Psi}\hat{\hat{\boldsymbol{\beta}}}) \,/\, (MT - K)},$$

where $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_{(K\times K)} = (\boldsymbol{\Psi}'_{(K\times MT)}(\hat{\boldsymbol{\Sigma}}^{-1}_{(M\times M)} \otimes \boldsymbol{I}_{(T\times T)})\boldsymbol{\Psi}_{(MT\times K)})^{-1}$ (Judge et al., 1985).

3.4 Statistically optimized orthogonal polynomial regression of the experimental dependences group

Before we begin to test the miscellaneous mutual linear relations among regression functions, it is necessary to determine the optimum values $K_1, K_2, \ldots, K_M$ of the degrees of orthogonal polynomials. A strategy is chosen for this process when the degree values of orthogonal polynomials are always identical, i.e., $K_1 = K_2 = \ldots = K_M$. We establish the artificial variable $\mathcal{K}$ for preservation of their values.

Estimations of the orthogonal polynomials degrees are determined with the help of a series of null hypothesis tests (3). The process is as follows.

First, the maximum possible degrees of orthogonal polynomials are selected. It was empirically ascertained that degrees higher than 15 have no practical meaning. Therefore, we initially choose $\mathcal{K} = \min(T - 2\,,\, 15)$.

(*) We put $K_1 = K_2 = \ldots = K_M = \mathcal{K}$. Then we test the null hypothesis H_0, that all regression coefficients β_{1,K_1}, β_{2,K_2}, ... , β_{M,K_M} in orthogonal polynomials of the highest degrees are zero, against the two-sided alternative that it is not so in at least one of the M cases. The matrix $\boldsymbol{c}$ includes zero elements here. The only exceptions are the elements

$$c_{i,\,\xi_i} = \sum_{t=1}^{T} \psi_{K_i}(x_t)\,, \qquad \xi_i = \sum_{l=1}^{i}(1 + K_l)\,, \qquad i = 1, 2, \ldots, M\,.$$

The vector of constants $\boldsymbol{r} = \boldsymbol{0}$.

If the null hypothesis H_0 is rejected, the resulting estimations $\hat{K}_1, \hat{K}_2, \ldots, \hat{K}_M$ of the orthogonal polynomials degrees are equal to $\mathcal{K}$. If the null hypothesis is not rejected, we decrease the value of $\mathcal{K}$ by one. If $\mathcal{K} = 0$, we put $\hat{K}_1 = \hat{K}_2 = \ldots = \hat{K}_M = 0$. If $\mathcal{K} \neq 0$, we come back to the clause (*) and repeat the process from this clause on.

Very good efficiency of this procedure of searching for *estimation of the degrees of regression orthogonal polynomials* was verified using simulated data. See graphic demonstration Figure 1.

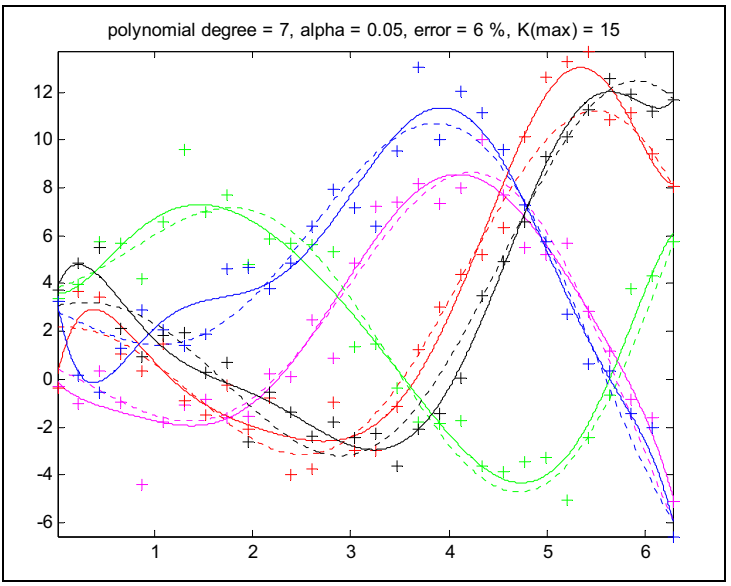


Figure 1 Demonstration of effectiveness of the process of searching for *estimations of regression orthogonal polynomials* on simulated data. Dashed line - simulated courses of *functional dependences*; crosses – randomized courses of *functional dependences*; solid line - statistical estimations of courses of *functional dependence* based on randomized courses of *functional dependences*.

3.5 Testing the hypothesis that one group of dependences is consistent with another group of dependences

As it was noted previously in the introduction, in medical and biological research the following problem often occurs. It is necessary to compare the set of *experimental (e.g. spectral) dependences of the type "group of sick patients"* with the set of *experimental dependences of the type "group of healthy patients*". What is being dealt with in these cases is, as a rule, experiments measured, for example, on population samples of people, experimental animals, micro-organisms, etc. In some cases it is necessary to compare the set of *experimental dependences of the type "the given noxious substance was applied"* with the set of *experimental dependences of the type "the given noxious substance was not applied*".

Our effort is to test the null hypothesis that, in the point x, every regression function ${}_{\text{sick}}\eta_1(x)$, ${}_{\text{sick}}\eta_2(x)$, … , ${}_{\text{sick}}\eta_{M_{\text{sick}}}(x)$ is equal to every regression function ${}_{\text{healthy}}\eta_1(x)$, ${}_{\text{healthy}}\eta_2(x)$, … , ${}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x)$, thus

$$H_0\colon\ {}_{\text{sick}}\eta_i(x) = {}_{\text{healthy}}\eta_{i'}(x),\quad i = 1,2,\ldots,M_{\text{sick}},\quad i' = 1,2,\ldots,M_{\text{healthy}}. \tag{6}$$

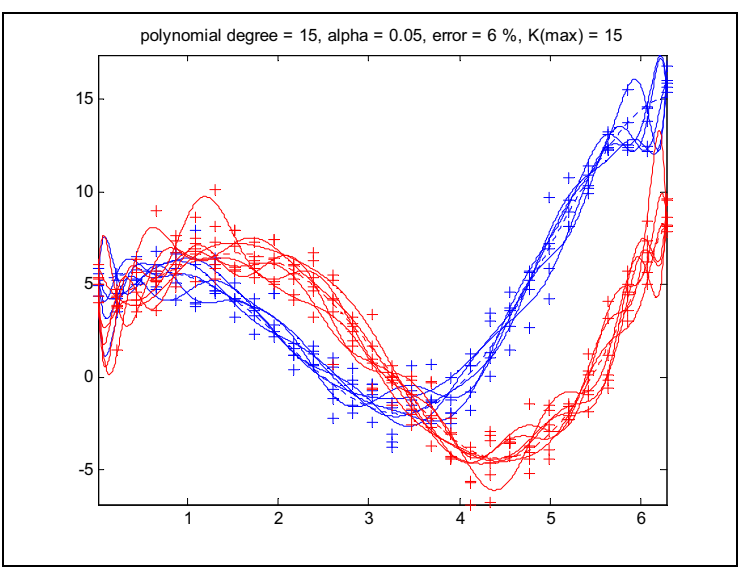
Matrix $\boldsymbol{c}(x)$, corresponding to (6), has $M_{\text{sick}} \cdot M_{\text{healthy}}$ rows, which are linearly dependent. Thus, it contains at the most $(M_{\text{sick}} + M_{\text{healthy}} - 1)$ linearly independent rows. Therefore, it is necessary to substitute (6) using a lesser number of equations. This can be done in various ways. For example, the following form of the null hypothesis has proven well:

$$H_0: \begin{cases} {}_{\text{sick}}\eta_1(x) = {}_{\text{sick}}\eta_2(x) = \ldots = {}_{\text{sick}}\eta_{M_{\text{sick}}}(x), \\ {}_{\text{sick}}\overline{\eta}(x) = {}_{\text{healthy}}\overline{\eta}(x), \\ {}_{\text{healthy}}\eta_1(x) = {}_{\text{healthy}}\eta_2(x) = \ldots = {}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x), \end{cases} \tag{7}$$

containing $(M_{\text{sick}} + M_{\text{healthy}} - 1)$ equations. Here, the means are ${}_{\text{sick}}\overline{\eta}(x) = M_{\text{sick}}^{-1} \Sigma_{i=1}^{M_{\text{sick}}} {}_{\text{sick}}\eta_i(x)$ and ${}_{\text{healthy}}\overline{\eta}(x) = M_{\text{healthy}}^{-1} \Sigma_{i'=1}^{M_{\text{healthy}}} {}_{\text{healthy}}\eta_{i'}(x)$. The corresponding *definition matrix* has the form

$$\boldsymbol{k} = \begin{pmatrix} \overbrace{\begin{matrix} 1 & -1 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 1 & -1 & 0 & 0 & 0 & \ldots & 0 & 0 \end{matrix}}^{(M_{\text{sick}}\text{-1})\text{ - times}} \\ \begin{matrix} M_{\text{sick}}^{-1} & M_{\text{sick}}^{-1} & M_{\text{sick}}^{-1} & \ldots & M_{\text{sick}}^{-1} & M_{\text{sick}}^{-1} & -M_{\text{healthy}}^{-1} & -M_{\text{healthy}}^{-1} & -M_{\text{healthy}}^{-1} & \ldots & -M_{\text{healthy}}^{-1} & -M_{\text{healthy}}^{-1} \end{matrix} \\ \overbrace{\begin{matrix} 0 & 0 & 0 & \ldots & 0 & 0 & 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 & -1 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & -1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 1 & -1 \end{matrix}}^{(M_{\text{healthy}}\text{-1})\text{ - times}} \\ \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{M_{\text{sick}}-\text{ times}} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{M_{\text{healthy}}\text{- times}} \end{pmatrix}.$$

A very high efficiency of the procedure for *testing of the null hypothesis* (7) has been verified on simulated data. See graphic demonstration in figure 2a and 2b. In the graphic demonstration in figure 2b, *p-values* $p(x)$ *of tests of hypotheses that one group of dependences is consistent with another group of dependences* are displayed. As many as 1000 equidistant abscissas were selected. The test has been performed in each of them; see figure 2a. The calculated *p-values* $p(x)$ form in Figure 2b the graph of dependence on the independent variable x.



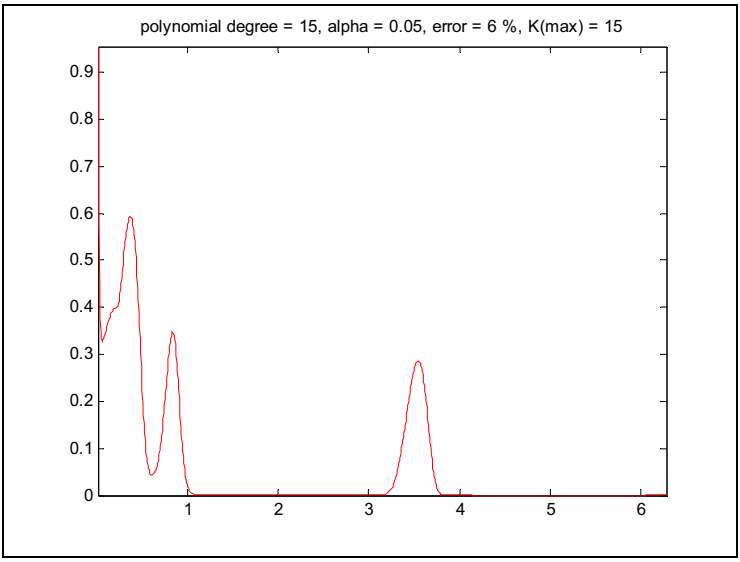
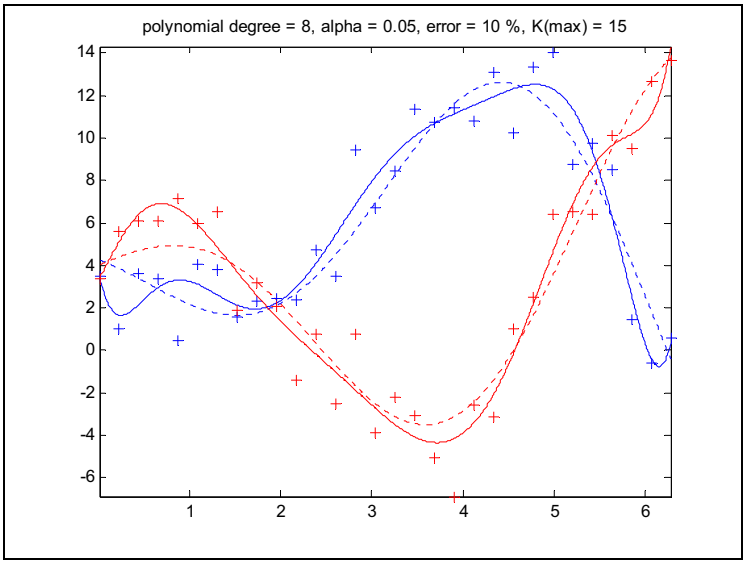


Figure 2a (upper) and 2b (lower) Demonstration of efficiency of *the test of the hypothesis that one (red) group of dependences is consistent with another (blue) group of dependences*. Upper graph: two simulated groups (red and blue) of stochastic *functional dependences*. Dashed line - simulated courses of *functional dependences*; crosses – randomized courses of *functional dependences*; solid line - statistical estimation of courses of *functional dependences* based on randomized courses of *functional dependences*. Lower graph: Corresponding dependence of *p-value* $p(x)$ on the appropriate independent variable x.

It can be seen from figures 2a and 2b that the efficiency of *the test of the hypothesis that one (red) group of dependences is consistent with another (blue) group of dependences* is very much exceeds expectations. Where simulated courses cross each other, the *p*-value $p(x)$ sharply increases and immediately decreases again. There, the null hypothesis H_0: $\boldsymbol{k}\,\boldsymbol{\eta}(x) = \boldsymbol{r}(x)$ is not rejected and *the difference between the two groups of regression function is not statistically confirmed*. On the contrary, in those intervals of the independent variable x where $p(x) < \alpha$, the null hypothesis is rejected, and as a result a *statistically justified suspicion exists that from* J *equations* $\boldsymbol{k}\,\boldsymbol{\eta}(x) = \boldsymbol{r}(x)$ *of the null hypothesis* H_0 *at least one is not valid*.

The algorithm also works very well in the case where only 1 red and 1 blue experimental *functional dependence* disturbed by normally distributed randomization were simulated. See figures 3a and 3b.



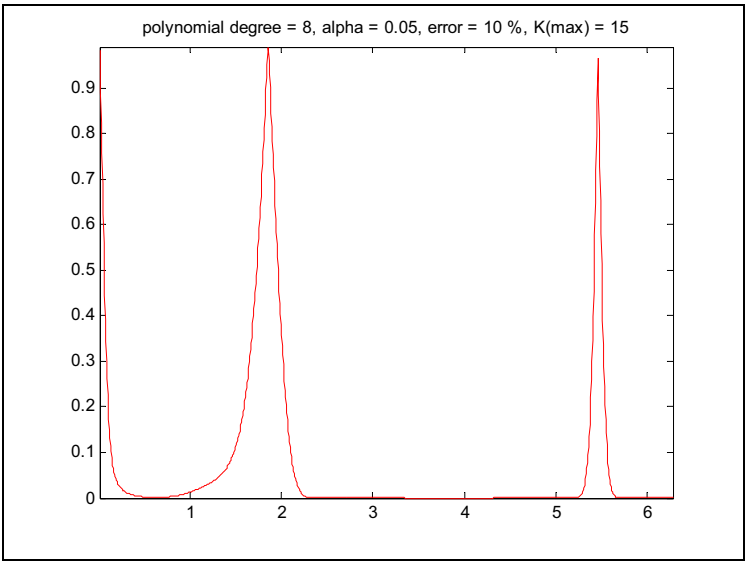


Figure 3a (up) and 3b (down) Demonstrating the efficiency of *the test of the hypothesis that one one-member group of dependences is consistent with another one-member group of dependences*. Upper graph: simulated two one-member groups (red and blue) of stochastic *functional dependences*. Dashed line - simulated courses of *functional dependences*; crosses - randomized courses of *functional dependences*; solid line - statistical estimations of courses of *functional dependences* based on randomized courses of *functional dependences*. Lower graph: Corresponding dependence of *p-value* $p(x)$ on the appropriate independent variable x.

3.6 The paired version of the test of the hypothesis that one group of dependences is consistent with another group of dependences

The paired version of the test is applied in the case where we have two equally large groups of regression functions, each with $M_{\text{sick}} = M_{\text{healthy}}$ members. We compare the pairs ${}_{\text{sick}}\eta_i(x)$ and ${}_{\text{healthy}}\eta_i(x)$ of regression functions, $i = 1, 2, \ldots, M_{\text{sick}}$. An example of this is the measurements carried out on the same patient before and after application of a medicine. Expressed more generally, it is measurement carried out on the same subject (human being, experimental animal, micro-organism, etc.) before and after application of a given noxious substance.

The task is to test the null hypothesis

$$H_0: \begin{cases} {}_{\text{sick}}\eta_1(x) = {}_{\text{healthy}}\eta_1(x), \\ {}_{\text{sick}}\eta_2(x) = {}_{\text{healthy}}\eta_2(x), \\ \quad\vdots \\ {}_{\text{sick}}\eta_{M_{\text{sick}}}(x) = {}_{\text{healthy}}\eta_{M_{\text{healthy}}}(x), \end{cases}$$

which consists of $M_{\text{sick}} = M_{\text{healthy}}$ equations. The corresponding *definition matrix* has the form

$$\boldsymbol{k} = \begin{pmatrix} M_{\text{sick}}\text{-times}\left\{ \begin{array}{cccccccc} 1 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \underbrace{0}_{M\text{-times}} & 0 & \ldots & 1 & 0 & 0 & \ldots & -1 \end{array} \right. \end{pmatrix},$$

where $M = 2M_{\text{sick}}$. Vector $\boldsymbol{r}(x) = \boldsymbol{0}$.

4 Results of applications on real data

This section discusses details of some possible *applications of the test* in medical and biological research.

4.1 Identification of biomarker areas in mass (MS) and Raman spectra (RS)

In this subsection, *physical properties* of mass and Raman spectra from the point of view of possible uses of the above described *test* are handled. At the same time, emphasis is placed on *physical characteristics of mass and Raman spectra of biomarkers* with regard to their *possible identification*.

We define *a biomarker* as a biochemical substance (often a protein) or its property (often a conformation or a post-translational modification) amount of which changes with intensity or degree of a studied biological process, e.g. disease. *The possibility (potential) of biomarker identification* using MS and/or RS assumes that changes in MS- and/or RS-peak size and shape are corresponding to changes in biomarker quantity. If, proportionally to the quantity, the size of the given biomarker MS-peak and/or RS-peak decreases or increases with progressive biological process, what is being dealt

with is *the* (our terminology) *decreasing or increasing*, respectively, *biomarker*. The *x*-coordinate, i.e., the effective mass and/or wave number of the MS-peak's and/or RS-peak's maximum, is *a characteristic constant of the appropriate biomarker*.

It is very important to understand the way mass and Raman spectra are measured and physically created in order to properly use *the test*. It is necessary to realize that the resulting MS and/or RS spectrum (e.g. of proteins from a tumor tissue) is a superposition of hundreds to thousands of spectra of particular biochemical substances (proteins, DNA and the like). It is assumed that only a small part of biochemical substances, approx. 5% of them, represent biomarkers.

By the so-called biomarker area, it is necessary to understand a section of the spectrum, in which one or more biomarkers express themselves in the same way, i.e., one or more biomarkers are increasing or one or more biomarkers are decreasing. At the same time however, this biomarker, or a group of them, may be "contaminated" by tens or hundreds of non-biomarkers or biomarkers having a contrary effect. For example in spectrum terminology, one biomarker MS-peak (and/or RS-peak), or a group of them, can be distorted by tens or hundreds of neighboring non-biomarker MS-peaks (and/or RS-peaks) or biomarker MS-peaks (and/or RS-peaks) that have a contrary effect.

The relatively simplest is the mass spectrum SELDI TOF. One biochemical substance is represented here by one peak. In the case of mass spectrum MALDI TOF, one substance is represented by a group of about 8 peaks due to isotopic distributive splitting. Now, let us imagine that the resulting mass spectrum is a superposition of 500 to 3500 spectra of single substances defined in this way. *Biomarkers increasing or decreasing* present approximately 5% of these substances. It is clear that, with a very high probability, individual peaks mutually overlap each other with varying intensity. In other words, mass spectrum of a real sample is a very complicated formation.

RS is relatively the most complicated, as one substance is represented by approx. 60 peaks. The interpretation here is complicated (Schrader and Bougeard, 1995; Smith and Dent, 2004) because the size and position (of single signals) reflect not only the chemical composition of the studied molecule but also its secondary structure from the viewpoint of the stereochemistry and the interaction with other components in a sample (Benevides and Overman, 2005). Therefore, when processing Raman spectra, it is necessary to include in the analysis not only the change in signal heights at a given wave number but also the possibility of coincident changes in the peak width and position. It is possible to observe --- from the whole number of $3N-6$ vibrations of *N*-atomic molecule --- only those vibrations which are permitted by selection principles. In the case of biopolymers, for example DNA, molecule structure contributes to a decrease in the number of signals. A biopolymer is formed by linkage of only four kinds of nucleotides. The reason is that most measurable signals in RS are sensitive to the state of a group of a few atoms within the nucleotide. It is then possible to distinguish approximately 60 signals (peaks) in the Raman spectrum of DNA, but only some of them are characteristic of the given kind of nucleotides. Any change in the state of the molecule, from chemical disintegration to bending, changes the spectrum of the molecule at a number of places at once. RS of proteins, i.e., of tumor tissue, also gives a true picture of the stereo-chemical configuration of the substance. RS of a sample reflects not only its chemical composition, but also the physical chemical state of single components and the interaction among them. Thus there are temporary chemical aggregates which exist only on the basis of complex hydrogen structures and van der Waalse forces, etc. In RS, it is also necessary to view optical antipodes, both left-handed and right-handed, i.e., stereo-chemical configuration isomers, as different matters.

It follows from the given discourse on the principles of mass and Raman spectrum generation that by means of the proposed algorithm consisting in *testing the hypothesis that one group of dependences is identical to another group of dependences*, a biomarker area can be identified on the independent variable x axis rather than a biomarker as such. The reason is that *the possible mutual statistical distinctness of groups of dependences corresponding first to sick patients and then to healthy patients is, with high probability, caused not by one sole biomarker but by a group of biomarkers instead.*

4.2 Segmentation of mass and Raman spectra

Mass or Raman spectrum is a very extensive formation. Approximately 50 thousand points are scanned for graphic illustration and description of the mass or Raman spectrum. In practice, mass or Raman spectrum is conventionally displayed in such a format that the height-width ratio of the graphic picture approximately equals A4 format size. This way of displaying the mass spectrum often leads to the spectrum being displayed as a highly vibrating dependence with a number of very narrow peaks. The statistical procedures used (Tibshirani et al., 2004; Wu et al., 2003) are conceived in the corresponding sense.

For the approach based on the regression model (2), it is necessary to process a spectrum step by step, segment after segment, according to the possible effective masses or wave numbers. In practice, the segmentation into 35 to 2000 segments has proven; the number of their effective masses or wave numbers oscillates between 50 and 300. One of these segments is fixed. It is assumed that just the T effective masses or wave numbers, denoted $x_1, x_2 \ldots x_T$, fell within the segment set that we set.

4.3 Results of applications in an attempt to identify biomarker areas in mass spectrum

A large set of mass spectral data, measured with the aim of *identifying kidney carcinoma biomarkers*, was subjected to the above described *test of the hypothesis that one group of dependences is identical to another group of dependences*. This set of mass spectral data is represented partly by the set of dependences of the type *"sick"*, i.e., of kidney carcinoma, and partly by the set of dependences of the type *"healthy"*, i.e., patients without kidney carcinoma. The *algorithm of the test* "identified," with the help of the *p-value* $p(x)$, the presence of *biomarker areas* across practically the whole spectral span, i.e., in all equidistant abscissas in all segments. *The null hypothesis that one group of dependences is identical to another group of dependences* was rejected for the whole spectral span. Such a result does not correspond with reality and shows that the algorithm used is not appropriate for the given spectral data.

4.4 Results from applications in determining the efficiency of heat denaturation of calf thymus DNA with the help of Raman spectra

The test hypothesis that one group of dependences is identical to another group of dependences has been applied to data demonstrating the efficiency of heat denaturation of calf thymus DNA with the help of Raman spectra. (See Figure 4a and 4b). *The null hypothesis that one group of dependences is identical to another group of dependences* was, with the exception of the intersection

points, rejected for the whole span. This shows that it was statistically proved that by warming to 95°C *effective denaturation of calf thymus DNA will take place.*

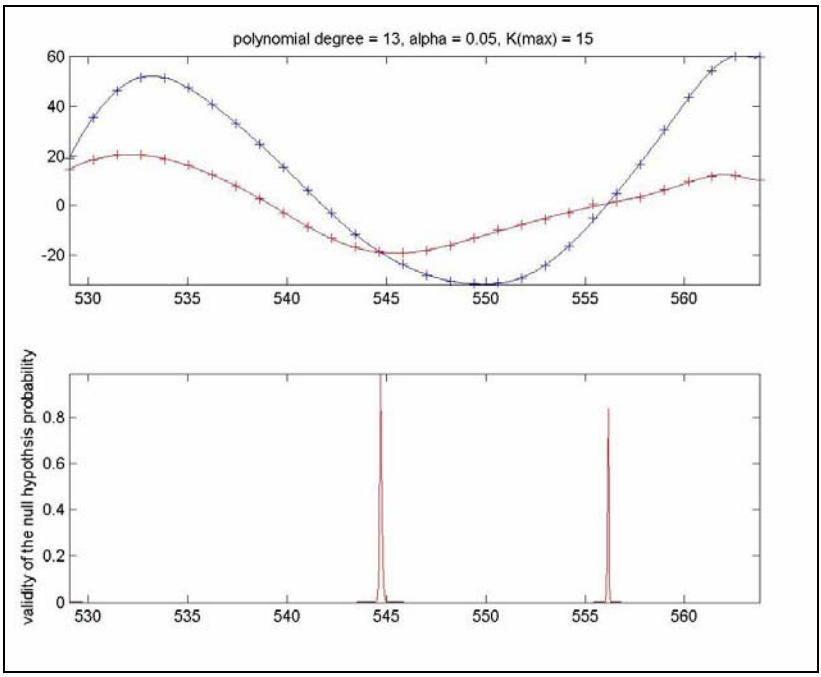


Figure 4a (upper) and 4b (lower) Demonstration of the efficiency of heat denaturation of calf thymus DNA with the help of RS: Upper graph: RS of 95°C denatured calf thymus DNA (red) and RS of 20°C native calf thymus DNA (blue). Crosses - measured spectral courses; solid line - statistical estimation courses. Lower graph: Corresponding dependence of *p-values* $p(x)$ on the independent variable x. As we can see, the corresponding *p-value* is nonzero only in those coordinates where red and blue courses intersect each other.

4. 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 test of the hypothesis that one group of dependences is identical to another group of dependences"* when data error disturbances have normal distribution.

Testing on simulated data has shown that in the event of fulfilment of stochastic presumptions about error uncertainties of the data, the algorithm of *the test* works very well. (See Figure 1-3).

Results of applications in an attempt to identify the biomarker area in mass spectra have shown that the *„the test of the hypothesis that one group of dependences is identical to another group of dependences„* is not fit for the given data from practice, i.e. a large set of mass spectral data measured with the aim of identifying carcinoma kidneys biomarker. The main reasons for this are:

- error disturbances of the data do not have normal distribution,
- biomarker behavior of the spectra has excessive character.

This finding is key to determining the direction of further bio-information science research in these issues. It is necessary to derive adequate relations within intentions of robust statistics (*M*-estimates, *L*-estimates, *R*-estimates) (Huber, 1981) and mathematical gnostics (Kovanic, 1986). *R-estimates especially* are procedures that are not directly based on concrete values, but *on their orders*. Hence, the request that the vector of error disturbances should have normal distribution is not valid here.

Mathematically gnostical processing does not work with a priori assumption about stochastic distribution of the error data uncertainties. An inherent feature of this method is, among others, the fact that it automatically searches for and finds clusters that cause excessive behaviour.

Acknowledgements

This work was supported by the grants IGA MZ CR NR / 8338 - 3/2005 and MZ0MOU2005.

The generation of orthogonal polynomials has been realized with the help of program procedures created by RNDr. Petr Tichy, PhD. and Prof. Dipl. Eng. Zdenek Strakos, DrSc. from the Computational Methods Department of the Computer Science Institute of the Academy of Sciences, Czech Republic.

References

1. Benevides, J. M., and Overman, S.A. ,2005, *Raman, polarized Raman and ultraviolet resonance Raman spectroscopy of nucleic acids and their complexes,* J. Raman Spect. 36:4, 279-299.
2. Forsythe G. E., 1957, *Generation and Use of Orthogonal Polynomials for Data-fitting on a Digital Computer,* J Soc Indust Appl Math. 5, 74-88.
3. Golub, G. H, and Van Loan, C. F., 1981, *Matrix Computations,* The Johns Hopkins University Press, Baltimore, 1996, 694.
4. Huber, P., 2004, *Robust Statistics,* J. Wiley, New York.
5. Judge, G. G, Griffiths, W. E, Hill, R. C, Lutkepohl, H., Tsoung-Chao, L., 1985, *The Theory and Practice of Econometrics,* J. Wiley, New York.
6. Knizek J., Pulpan Z., Hubalek M., Beranek L., Pokorny P., 2007a, *Mass spectrometry – some trends in contemporary biological research.* Cs. cas. fyz. **4**, 208.
7. Knizek J., Pulpan Z., Hubalek M., Beranek L., Pokorny P., 2007b, *Stochastic Model of Mass Spectrum Random Disturbances and its Simulation*, Summer School DATASTAT 06, Proceedings, Masaryk University, ISBN 978-80-210-4493-7.
8. Knizek J., Pulpan Z., Vojtesek B., Nenutil R., Brozkova K., Drazan V., Hubalek M., Beranek L., Kubacek L., 2007c, *The disturbance-related sets of regression equations – natural methodology for medical and biological research by means of spectra* (In: XXX. Days of medical biophysics), Publishing house of Charles University, Prague, p. 28, abstract, ISBN 978-80-239-9421-6.
9. Knizek J., Sindelar J., Beranek L., Vojtesek B., Nenutil R., Brozkova K., Drazan V., Hubalek M., Kubacek L., Spring 2008, *Power function for tests of null hypotheses on mutual linear regression functions' relations*, Bulletin of Statistics & Economics, Volume 2; Number S08; Bull. Stat. Econ.; ISSN 0973-7022.

10. Knizek, J., 2004a, *Theoretical basis of new methodology of mathematical-statistical decision making with the help of biomarkers from mass spectra*. Acta Medica (Hradec Kralove) 47:4, 291.

11. Knizek, J., Bergmann, M., Sindelar, J., Kovarova, H., 2004b, *MIAPS - programovy system pro odhadovani poradi vzajemne podobnosti/nepodobnosti chovani proteinu v case*, Acta Medica Supplementum 47:2,131-135.

12. Kovanic, P., 1986, *A New Theoretical and Algorithmical Basis for Estimation, Identification and Information*, Automatica 22:6, 657-674.

13. Ralston, A., 1973, *A First Course in Numerical Analysis*, McGraw Hill Book Company, New York.

14. Schrader, B., and Bougeard, D., 1995, *Infrared and Raman Spectroscopy*, Wiley-VCH Verlag GmbH, Weinheim.

15. Smith, E. and Dent, G., 2004, *Modern Raman Spectroscopy: A Practical Approach*, Wiley, New York.

16. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., Le, Q. T., 2004, *Sample classification from protein mass spectrometry by „peak probability contrasts”*, Bioinformatics - Bioinformatics Advance Access. Oxford University Press, 1-34.

17. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H., 2003, *Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data*, Bioinformatics 9:13,1636-43.