SUMMARY OF ALGORITHMIC FRAGMENTS FOR STATISTICAL IDENTIFICATION OF MARKERS FROM A SET OF SPECTRAL COURSES

Jiří Knížek[1], Ladislav Beránek[2], Petr Bouchal[3,4], Bořivoj Vojtěšek[4], Rudolf Nenutil[4] and Pavel Tomšík[5]

[1]Department of Medical Biophysics, Faculty of Medicine in Hradec Kralove,
Charles University in Prague, Czech Republic

[2]Department of Applied Mathematics,
University of South Bohemia, Czech Republic

[3]Department of Biochemistry, Faculty of Science,
Masaryk University, Czech Republic

[4]Regional Centre for Applied Molecular Oncology,
Masaryk Memorial Cancer Institute, Czech Republic

[5]Department of Medical Biochemistry, Faculty of Medicine in Hradec Kralove,
Charles University in Prague, Czech Republic

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. Therefore, our next effort will be directed at the construction of the 1st prototype of some semi-commercial software for the identification of markers from a set of spectral courses.

Keywords

marker, biomarker, regression, tests of hypotheses, software

Introduction

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. See e.g. Wikipedia.

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. The application of advanced methodologies is necessary to give us more detailed, structured information.

The model “A set of multiple linear regressions”

The beginning of the algorithmic study described below lay in finding a reality so that statisticians were able to derive a test criterion

$$\lambda_F = \frac{(\mathbf{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{RCR}')^{-1}(\mathbf{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})/J}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(MT-K)} : F_{(J,\,MT-K)} \qquad (1)$$

$$\boldsymbol{C} = [\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})\boldsymbol{X}]^{-1}$$

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

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

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

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \mathbf{r} \qquad (3)$$

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 $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}$ [12], [6].

Of importance is, 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).

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. So, we can test (with the help of (3) appropriately modified) arbitrary linear mutual relations among particular spectral courses [14], [17].

The definition matrix

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

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

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* [17]. Definition matrix $\boldsymbol{k}$ expresses generally all conceivable linear mutual relations among regression functions (4).

A highly effective algorithm for orthogonalization

Computational practice showed that the currently used Gram-Schmidt’s polynomials [5], [19], are not able to realize satisfactory measure of orthogonality. For our purpose – the polynomial approximation of a set of spectral courses – we had to use special, outstandingly efficient algorithms [18], [1], [10], [7], [8], [9].

Identification of markers by 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”* [17] 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 favor 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 favor 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 κ simultaneous null hypotheses $H_0^j\!: \boldsymbol{k}_{(1\times M)}^j\boldsymbol{\eta}_{(M\times 1)}(x) = \boldsymbol{r}_{(1\times 1)}^j(x) = r^j(x)$, where the index for the j^{th} simultaneous null hypothesis is $j = 1, 2, \ldots, \kappa$. The size of the number κ, 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

$$p_j(x) < \alpha/\kappa\,, \qquad j = 1,2,\ldots,\kappa\,, \qquad (5)$$

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

$$1-\beta_j(x) \geq convention\,limit\,, \qquad j = 1,2,\ldots,\kappa\,, \qquad (3)$$

must be fulfilled.

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

For test significance levels α 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.

Experimental results

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. gray and black) are demonstrated in figures. Gray pentagrams "☆": discrete courses of the measured (renal cell carcinoma) spectrum; black points "♦": 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 "☆" and "♦".

Conventional decision making conditions (5) and (6) are satisfied in the whole measurement range at all the 1st-9th 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.

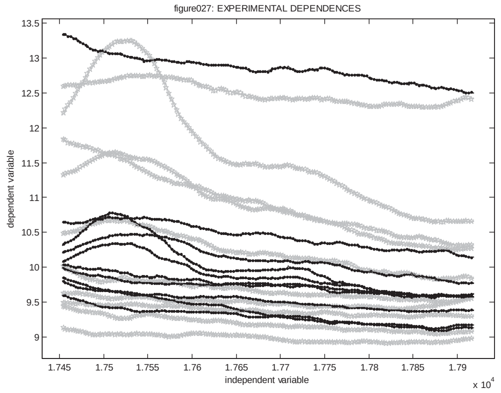

Fig. 1: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

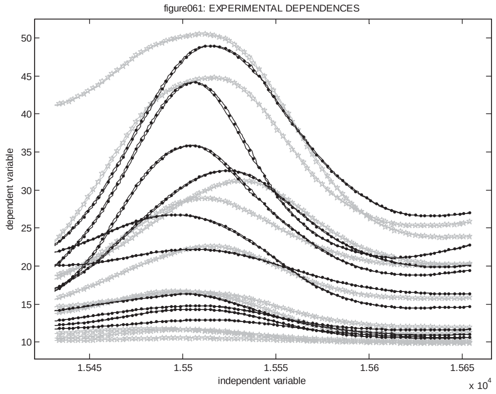

Fig. 2: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

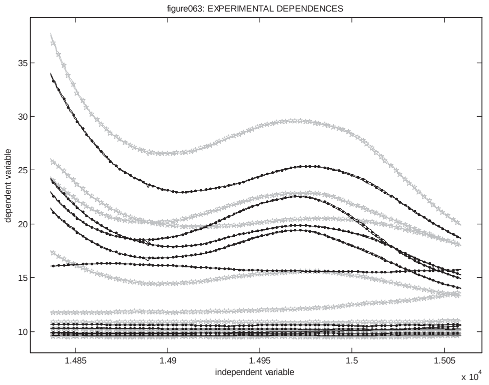

Fig. 3: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

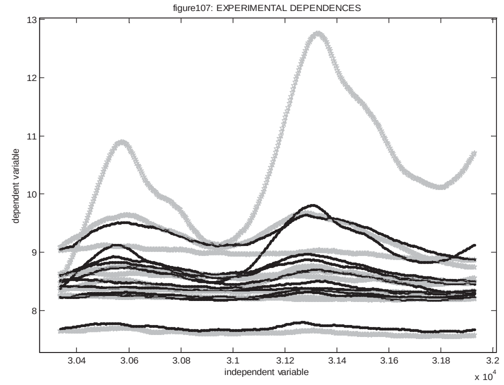

Fig. 4: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

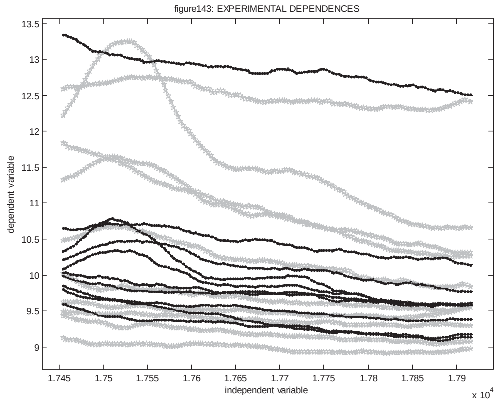

Fig. 5: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

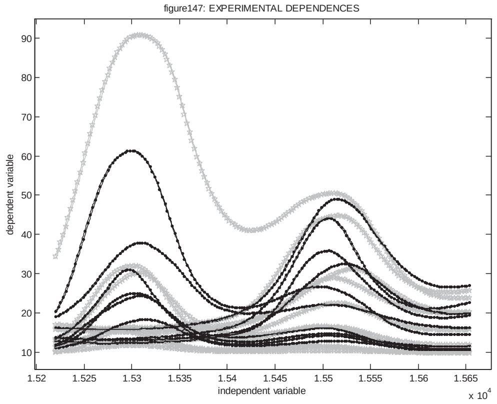

Fig. 6: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

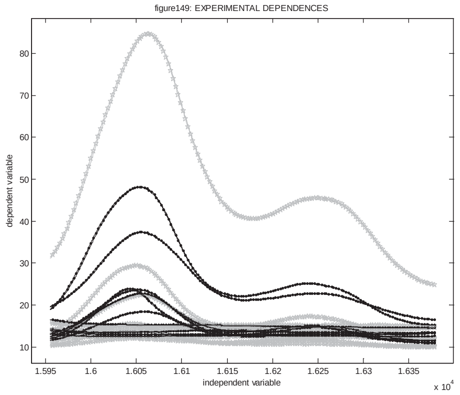

Fig. 7: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

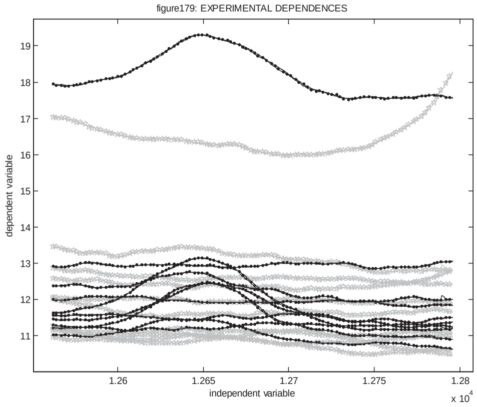

Fig. 8: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

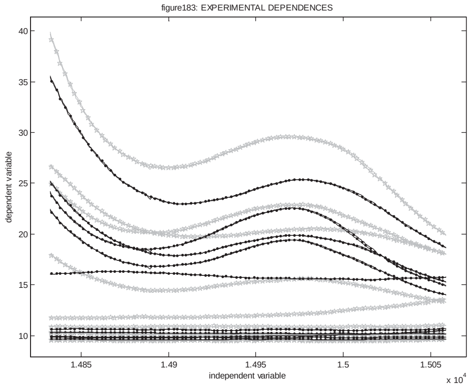

Fig. 9: See comments in the section "Pictorial exemplifications of real data, Figures 1-9".

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* "αB-crystallin"[1] [11] by the proposed algorithm *was confirmed*. The proposed algorithm is *very sensitive*, as yet other *potential biomarker areas* were found. It managed to find at least 12 cases of other *biomarker areas*. See figures 1-9[2].

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 Summary of Algorithmic Fragments 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* [16], [13].

Acknowledgement

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., 1951, vol. 9, p. 17-29.

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

[3] Cohen, J. *Statistical Power Analysis for the Behavioral Science*. Mahwah, New Persey: Lawrence Erlbaum, 1988, 2nd ed., p.567.

[4] Daly, L.E., Bourke, G.J. *Interpretation and Uses of Medical Statistics*. Oxford: Blackwell Science, 2000, 5th ed., p. 276-279.

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

[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 approxima-tion. 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, 2005, Vol. 50, p. 1069–1075.

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

[10] Higham, N. J. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2002, 2nd ed.

[11] Holcakova, J., Hernychova, L., Bouchal, P., Brozkova, K., Zaloudik, J., Valik, D., Nenutil, R., Vojtesek, B. *Identification of alphaB-crystallin, a biomarker of renal cell carcinoma by SELDI-TOF MS*, The International Journal of Biological Markers, Italy: Wichtig editore, 2008, Vol. 23, no. 1, p. 48-53. ISSN 0393-6155.

[12] Judge, G. G, Griffiths, W. E, Hill, R. C, Lutkepohl, H., Tsoung-Chao, L. *The Theory and Practice of Econometrics*, New York: J. Wiley, 2005.

[13] Knizek, J. *Marker Statistics I.: Regression analysis of dependences in medicine and molecular biology*, VDM Publishing House Ltd., Mauritius, 2011, ISBN-NR.: 978-3-639-33015-1.

[14] 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, 2008, Vol. 2, no. S08; p. 26-33.

[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, 2010, Vol. 5 [Special], no. A10, p. 12-20.

[16] Knizek, J, Sindelar J, Vojtesek B, Bouchal P, Nenutil R, Beranek L & Dedik O. *Using Markers to Aid Decision Making in*

[1] Ciphergen-software [2]

[2] 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.

Diagnostics, International Journal of Tomography & Statistics, 2011, Vol. 16, no. W11, p. 41-55.

[17] 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, 2008, Vol. 2, no. A08, p. 2-18.

[18] 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, 2010, Vol. 7, no. J10, p. 48-60.

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

Jiří Knížek, Ing., CSc.
Department of Medical Biophysics,
Faculty of Medicine in Hradec Kralove,
Charles University in Prague,
Šimkova 870,
CZ-500 38 Hradec Králové

E-mail: jknizek@lfhk.cuni.cz
Phone: +420 495 816 464