Effective Use of Combination of Lung Malignity Markers

Jan Mocak[1,2]

[1] Department of Analytical Chemistry, Faculty of Chemical and Food Technology, Slovak University of Technology, Bratislava, Slovakia

[2] Department of Chemistry, Faculty of Natural Sciences, University of SS. Cyril & Methodius, Trnava, Slovakia

e-mail: jan.mocak@stuba.sk, jan.mocak@ucm.sk

University of SS. Cyril and Methodius
Trnava, Slovakia

S T U

Slovak University of Technology
Faculty of Chemical and Food Technology
Bratislava, Slovakia

Contents

Introduction – advantage of multidimensional methods of data analysis (MDA)

Tumor markers - diagnostic tools for lung malignity

Characterization of tumor markers by various MDA techniques (PCA, CA, LDA, logistic regression, KNN, artificial neural networks)

Assessment of diagnostic effectiveness of tumor markers

Optimal combination of tumor markers and its implementation in practice

Conclusions

Literature

Introduction

advantage of chemometrics methods: ***versatility - high level of abstraction*** when processing of data of everydays life (food, medicinal, pharmaceutical chemistry, environmental studies, etc.)

from chemometrical standpoint all data are defined by:
- ***objects*** / ***cases*** (e.g. blood samples) arranged by ***rows*** in the data table
- ***input variables***/***properties***/***descriptors*** (spectral, chromatographic, electrochemical signals, molecular properties, etc.) arranged by ***columns***

used variables (***input*** and ***output***/***target***) are: ***continuous*** (very many levels), ***categorical*** (low number of levels) - ordinal or nominal

implementation of ***multivariate (multidimensional) methods*** may enable: 1) design of new potential drugs, 2) prediction of the selected properties of the investigated compounds/species, 3) better monitoring or prediction of diagnosis (e.g. simultaneous use of several laboratory tests – orig. variables), 4) selection of optimal measurement method, 5) …

Tumor markers (TM)

TM are important indicators utilized for ***monitoring the course of malignant disease*** and ***therapy effect, assessment of prognosis*** and ***health risk***

TM are compounds ***present in human body*** and analysed in clinical laboratories

enhanced TM concentration usually indicates malignity, but TM concn. may be enhanced also in case of benign tumors, tbc, etc.

TM concentration is ***quickly available by a non-invasive procedure,*** but TM are ***less reliable*** compared to histology; nevertheless, fast start of therapy may save the life

at present various TM are used, which are more or less specific; **CEA** (carcinoembryonic antigen), **CYFRA** (cytokeratin fragment) – used for dg. of lung malignity by their determination in ***serum*** or ***exudation***

Processing of TM data

markers CEA and CYFRA 21-1 were determined in Institute of tuberculosis and respiratory diseases (ÚTaRCH) in Poprad

data set - 182 patients, 74 women and 108 men

86 patients – with ***malignant*** disease

96 patients – ***benign*** tumors or tuberculosis

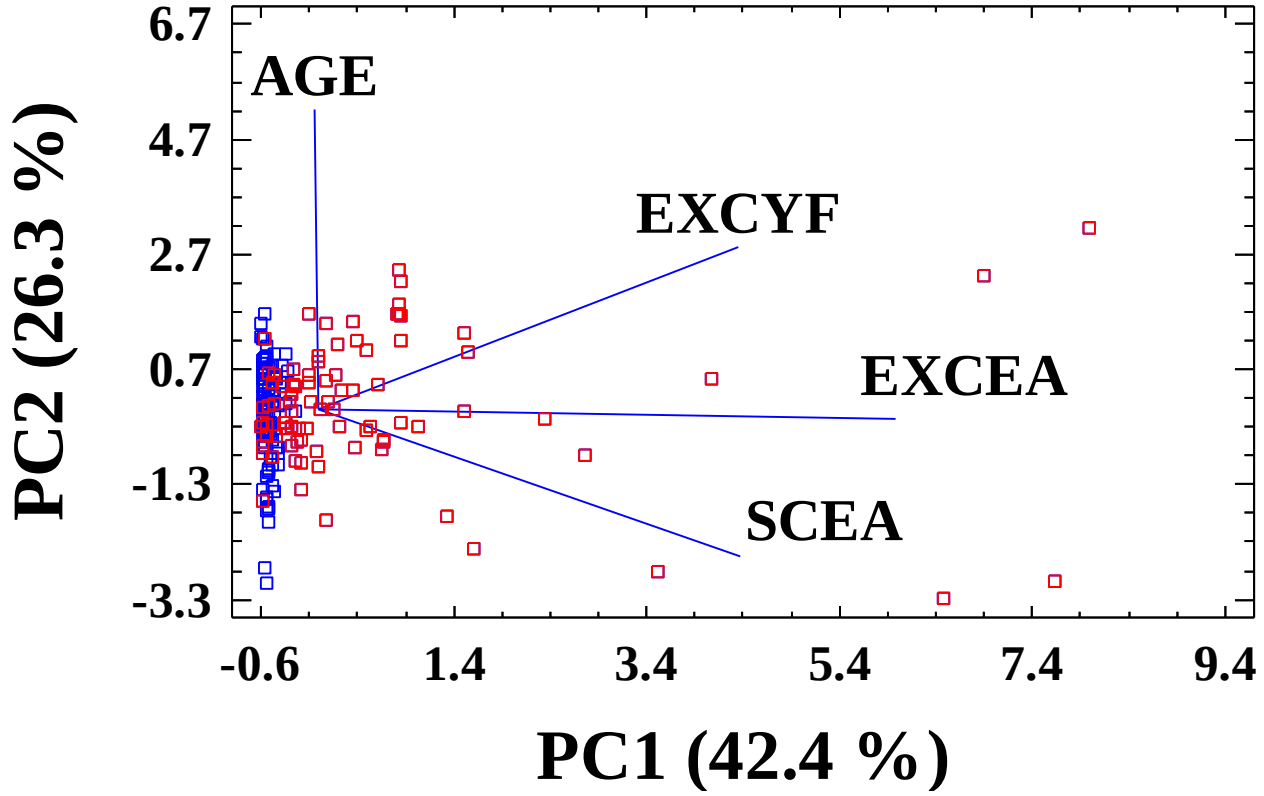
descriptors (variables): ***EXCEA*** & ***EXCYF*** in pleural exudation, ***SCEA*** & ***SCYF*** in serum, age (***AGE***), and patient's gender (***SEXN***) – categor. var.

during preliminary studies ***SCYF*** was found the least effective therefore it was eliminated (econom. reasons)

target categorical variable diagnosis (*DG*): 0 – ***benign***, 1 - ***malignant***

Mrazova V., Mocak J., Varmusova E., Kavkova D., Bednarova A.: Use of multidimensional data analysis for prediction of lung malignity. *J. Pharm. Biomed. Analysis* 50 (2009) 210-215.

Chemometric characterization of TM by PCA biplot



182 patient samples, 3 TMs + Age are shown as the rays. Blue marks – non-malignant, red marks – malignant. PC1 expresses the extent of malignity. Software Statgraphics Centurion XV.

Principal Component Analysis

provides such a ***linear combination*** of original variables, which explains the maximum of data variability (keeps maximally the variance of the original data).

principal components, PC_1 … PC_r, are new (latent) ***uncorrelated*** **variables,** ***hierarchically ordered***

importance of each *PC* is given by the corresponding ***eigenvalue*** λ of the ***covariance matrix***, $PC_1 = \lambda_1 / \Sigma\lambda_i$

the number of the utilized *PC*'s, *r,* is usually smaller than the number of original variables *p*: $r \leq p$

PCA is used for ***variable reduction purposes***

common plots are: ***scatterplots*** (showing objects), ***loadings (components) plots*** (showing variables), and ***biplots*** (objects + variables)

Cluster Analysis, CA

CA - technique, which groups ***similar objects*** into ***clusters*** using some ***measure of similarity***

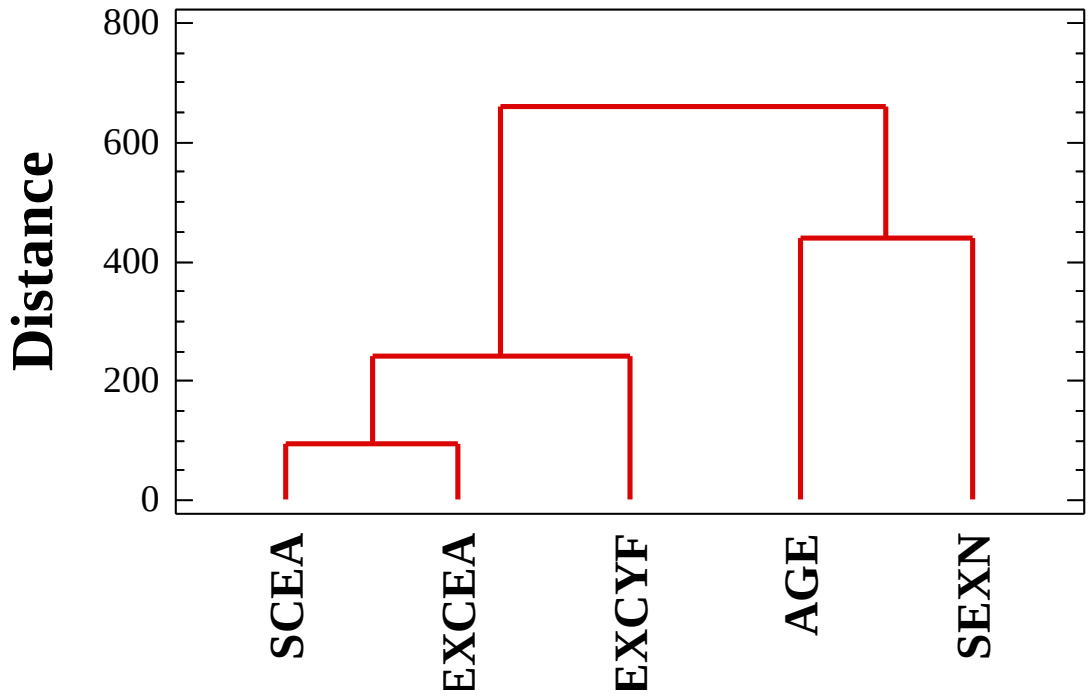
measure of similarity is defined by means of various ***distances*** like squared Euclidean distance, Euclidean distance, and Hamming d. (city-block d., Manhattan d.)

the smaller is the distance among objects the more similar they are - the calculated ***similarity matrix*** contains distances of all possible pairs objects

also ***variables can be clustered*** (instead of objects) to see their similarities

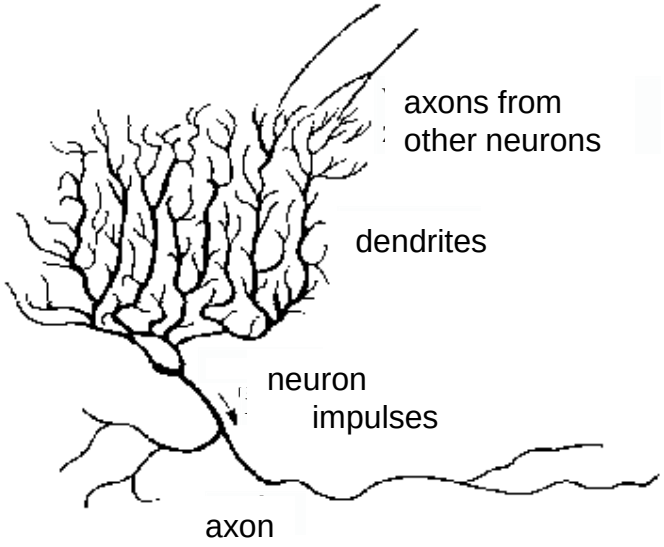
final output of ***hierarchical CA algorithms*** is ***dendrogram***, where stepwise clustering of objects/variables is demonstrated – starting with most related up to most diverse

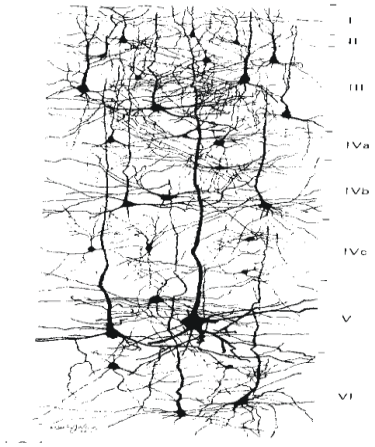
Chemometric characterization of TM by cluster analysis of variables

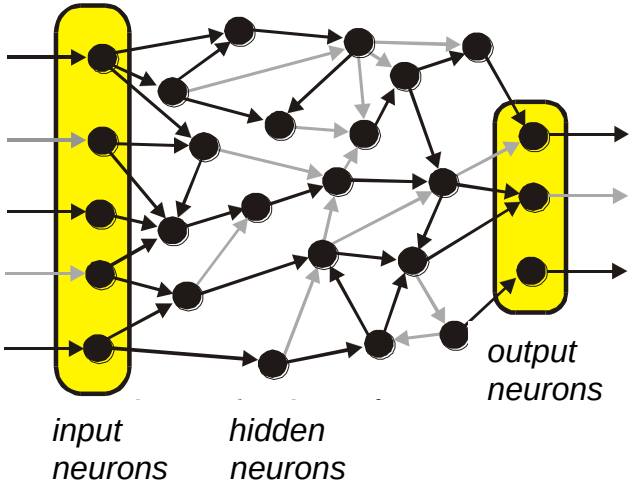


Clustering of variables. Ward method, squared Euclidean distance. Software Statgraphics Centurion XV.

Neural networks







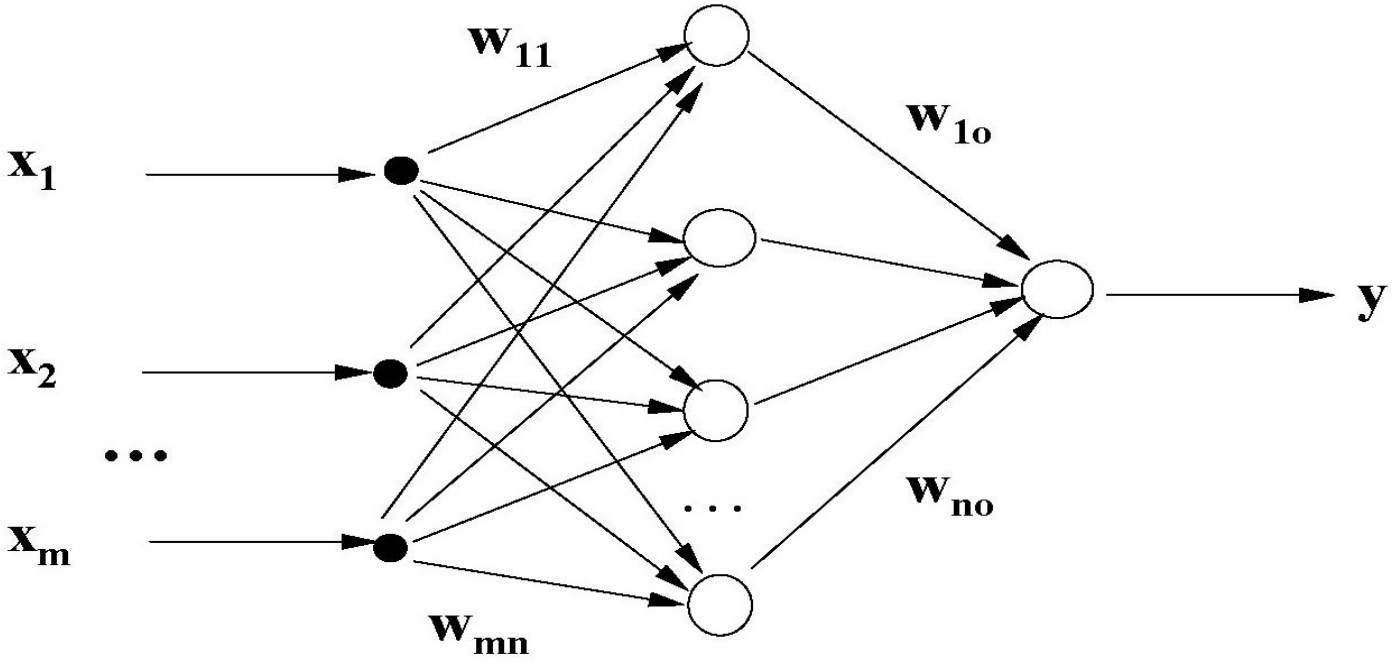
A B C

A: Typical neural cell with extensive dendrite system (at input) and a long branching axon (at output)

B: Mutual interconnections of the neurons by means of connections between dendrites and axons, which is observable in microanatomy of human brain

C: Architecture of a neural network with two types of links between neurons, dark/light lines represent exitation /inhibition links

Artificial neural networks (ANN)

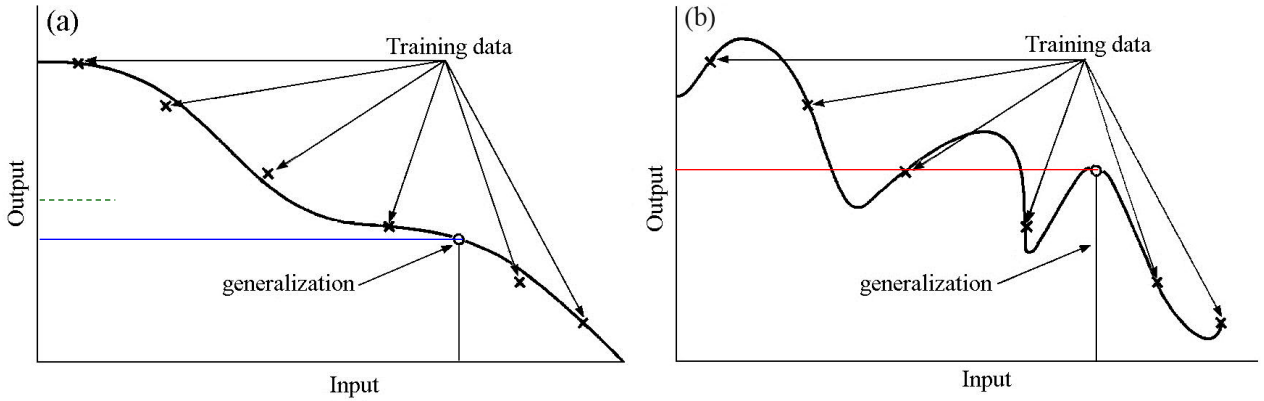


layer: input hidden output

three layer perceptron containing m input neurons (descriptors x_i), n hidden neurons and o output neurons (predictions y_j) with the corresponding weights w_{ij}

ANN – 3 sets of investigated objects

overfitting - danger in all nonlinear calculation algorithms



3 sets of data (object distribution by random way):

training set - for calculation of mathematical model

selection (validation) set - for selecting the best ANNs

test set - independent prediction (category in ***classification mode*** or the output value in ***regression mode***)

Results of classification by *DG* in % success for various MDA methods

Classification method	Training set, success, %	Leave-one-out, success, %	Validation set, success, %
LDA	74.6	75.3	86.7
QDA	86.3	**85.2**	**86.7**
KNN, *k*=3	89.0	78.6	83.3
KNN, *k*=5	86.8	**82.4**	**86.7**
KNN, *k*=7	85.2	79.7	83.3
KNN, *k*=9	81.9	78.0	86.7
KNN, *k*=11	80.2	76.4	86.7
LR	88.5	88.5	**90.0**
LR (+SEXN)	89.6	**89.6**	**90.0**
ANN	88.5	80.0	96.7
ANN (+SEXN)	86.9	83.3	96.7
ANN RBF	97.8	**95.6**	**97.8**

4 variables (***EXCEA, SCEA, EXCYF, AGE***) were used except LR and ANN where also ***SEXN*** was used as the 5th categ. variable. SAS 9.1.3, Trajan 6.

Classification techniques of MDA

classification techniques of MDA make use of *a priori known categorization of the objects* belonging to the *training set* for *classification of other objects not yet categorized* (belonging to the *test set*)

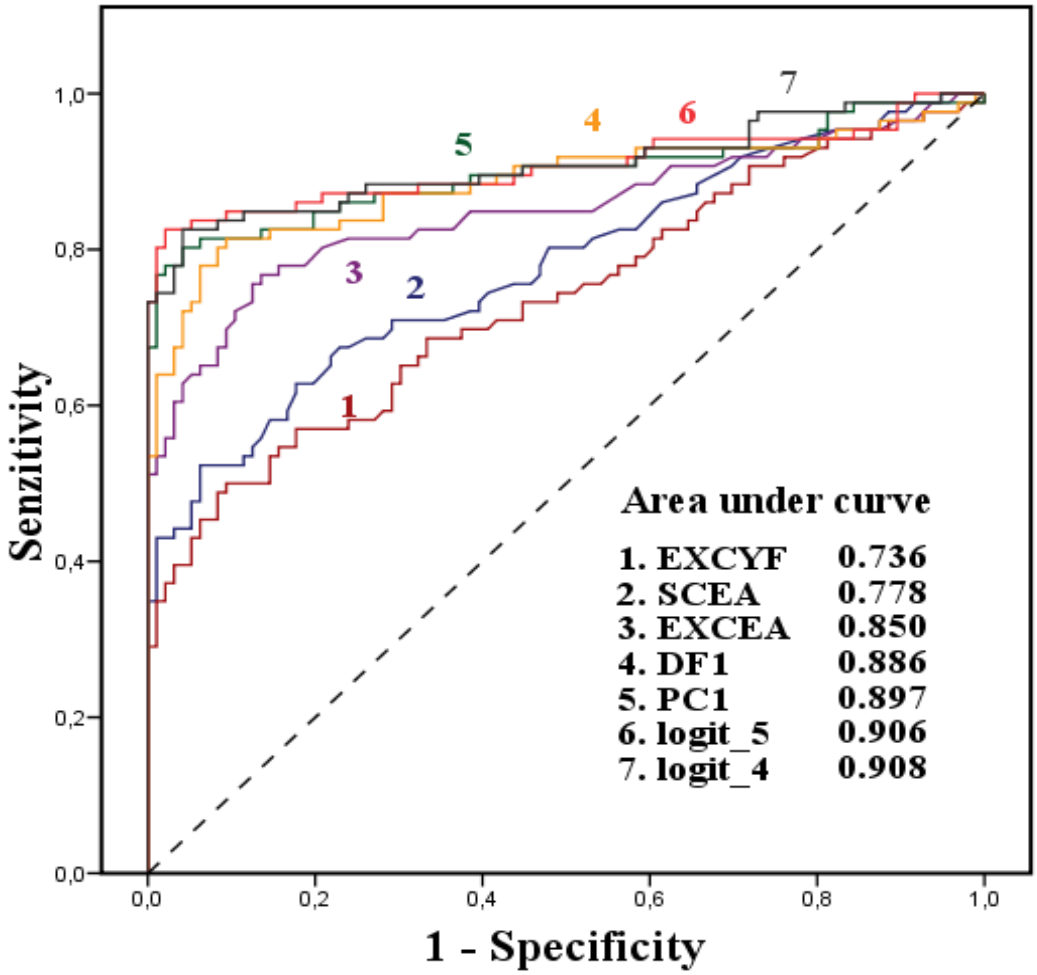
basic classification technique is *linear discriminant analysis, LDA,* which enables the *maximum discrimination* among *m* categories of objects by calculating one or more (*k*) *discriminant functions, DF,* and creating the decision boundaries between the categories ($k = m - 1$); *DFs* are new variables obtained by *linear combinations of original variables* (different compared to PCA); important step in classification is the *selection of variables* allowing *best* discrimination among the categories

logistic regression expresses the dependence of *probability ratios* of the states of *categorical variable* (e.g. dichotomous) and independent variables in the form $\ln [P/(1-P)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$

KNN uses categorization of *K nearest neighbour objects* (majority vote)

Vandenginste B.M.G., Massart D.L., et al: *Handbook of Chemometrics and Qualimetrics. Part B*. Elsevier, Amsterdam, 1998.

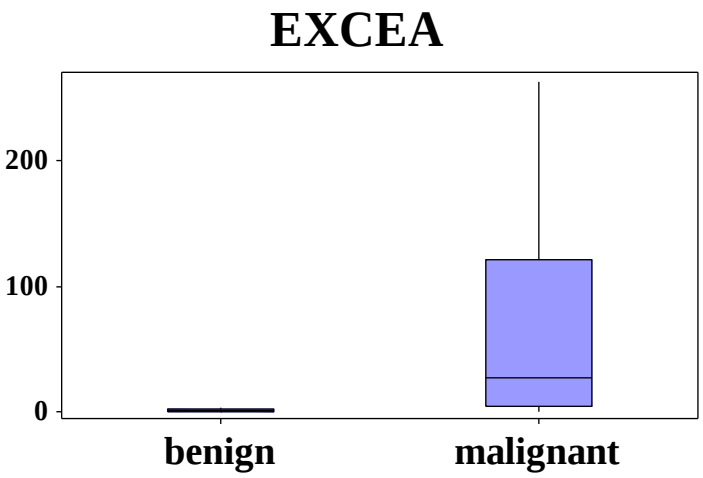
ROC curves for individual and composed TM

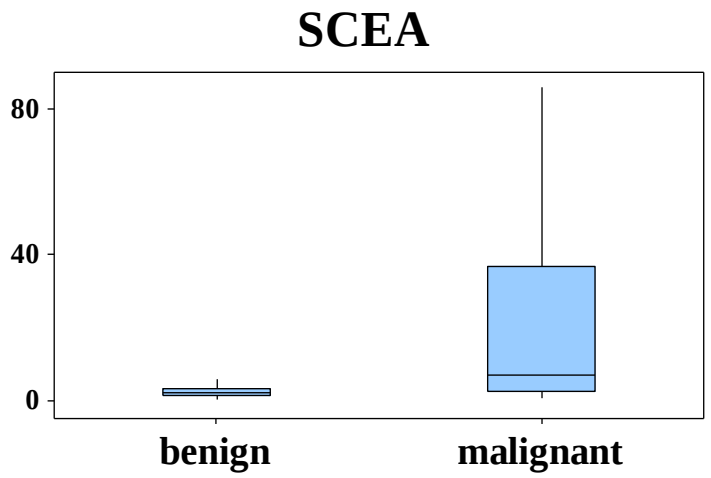


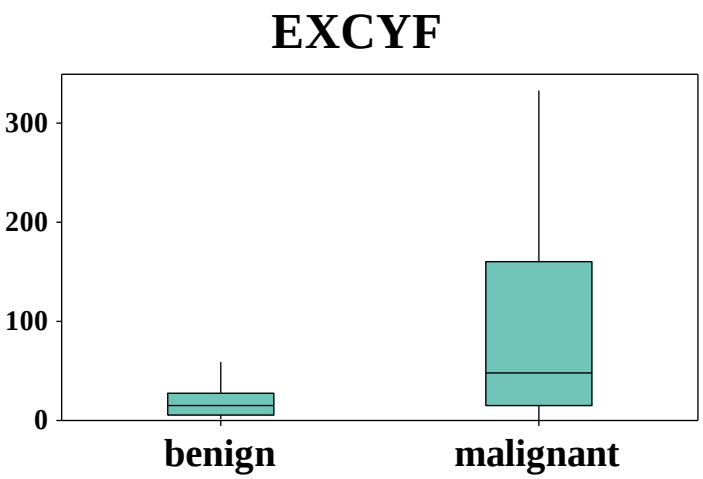
3 TM and 4 composed TM obtained by PCA (PC1 - 1st principal component), LDA (DF1 - 1st discriminant function) and logistic regression (logit_4 , logit_5 represent here dependent variables composed of 4 and 5 original variables, resp.). Software SPSS 15.0.

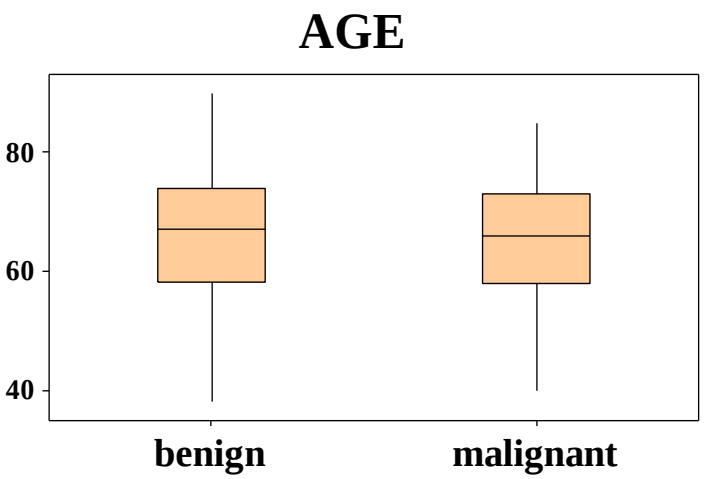
Balla B., Mocák J. Varmusová E., Kavková D., Tudík I.: Assessment of effectiveness of laboratory methods. *Chem Listy* 97 (2004) 333-338.

Box-plots of individual variables

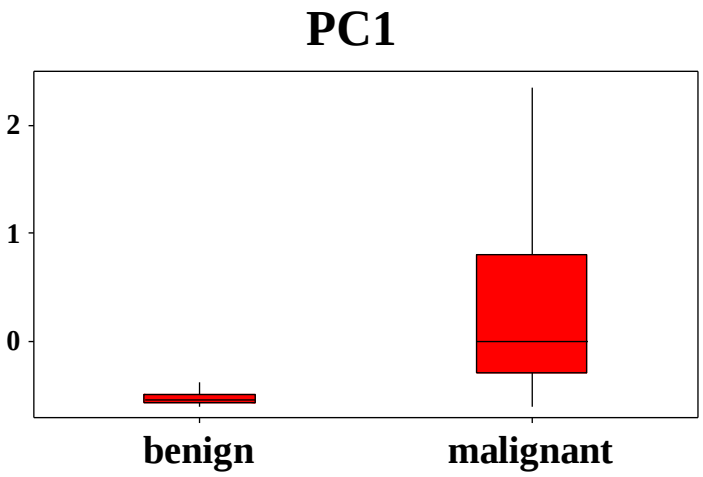


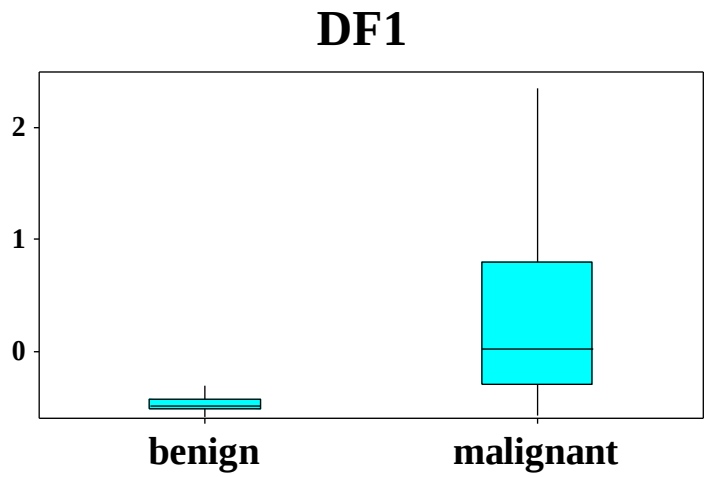


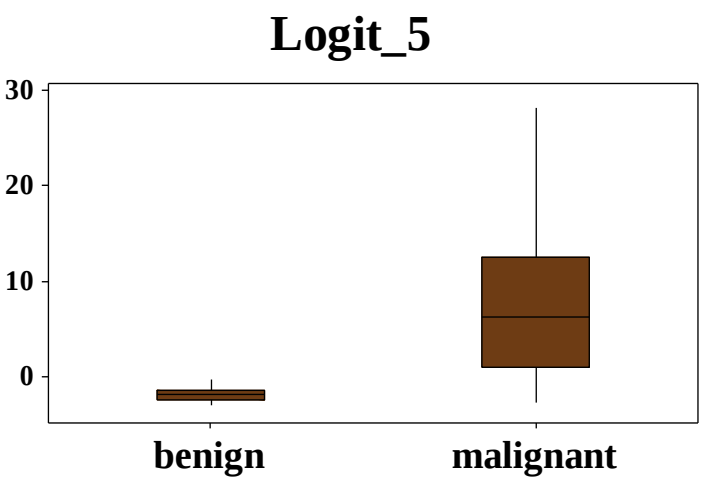


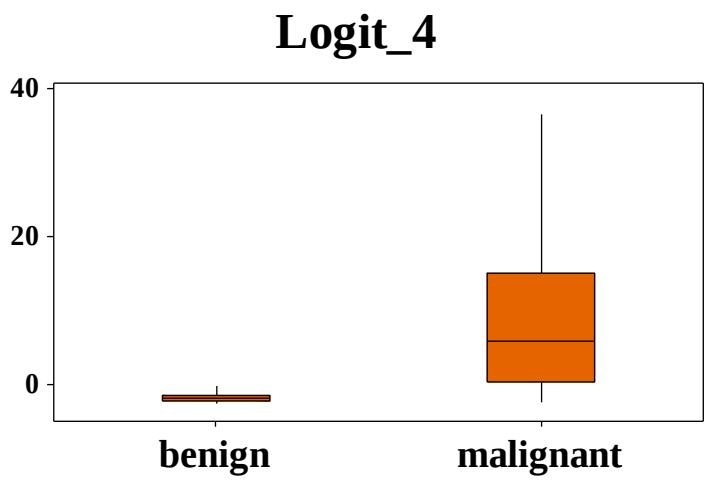


Box-plots of composed variables

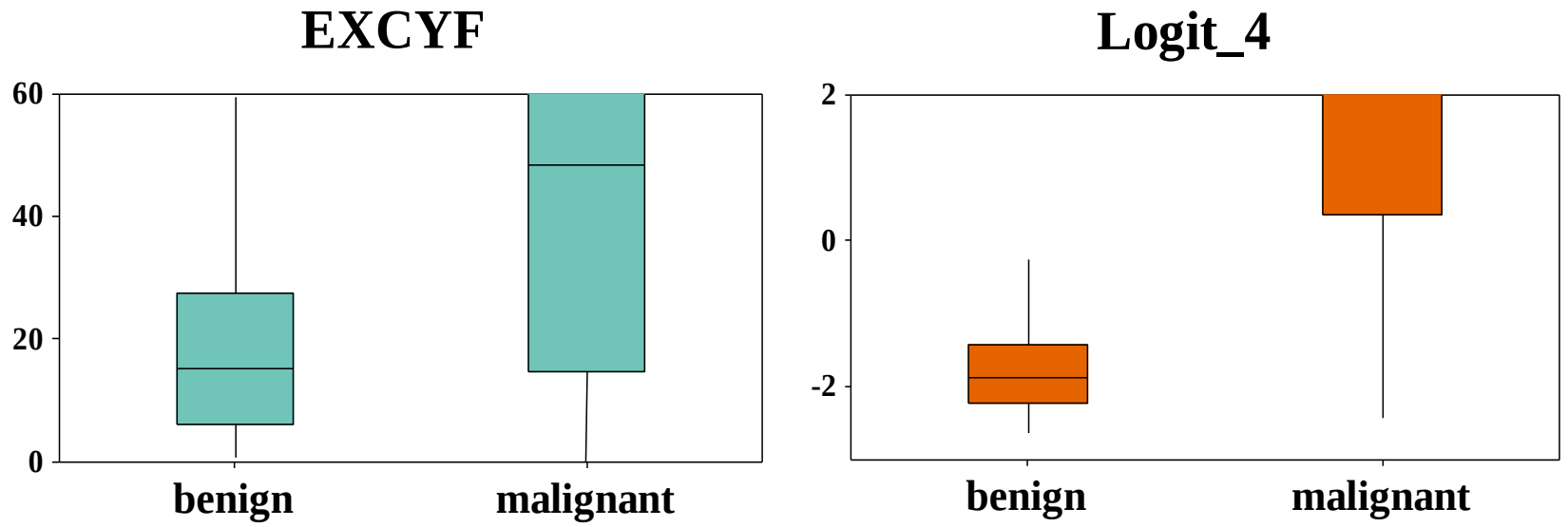








Comparison of the selected individual TM with composed TM

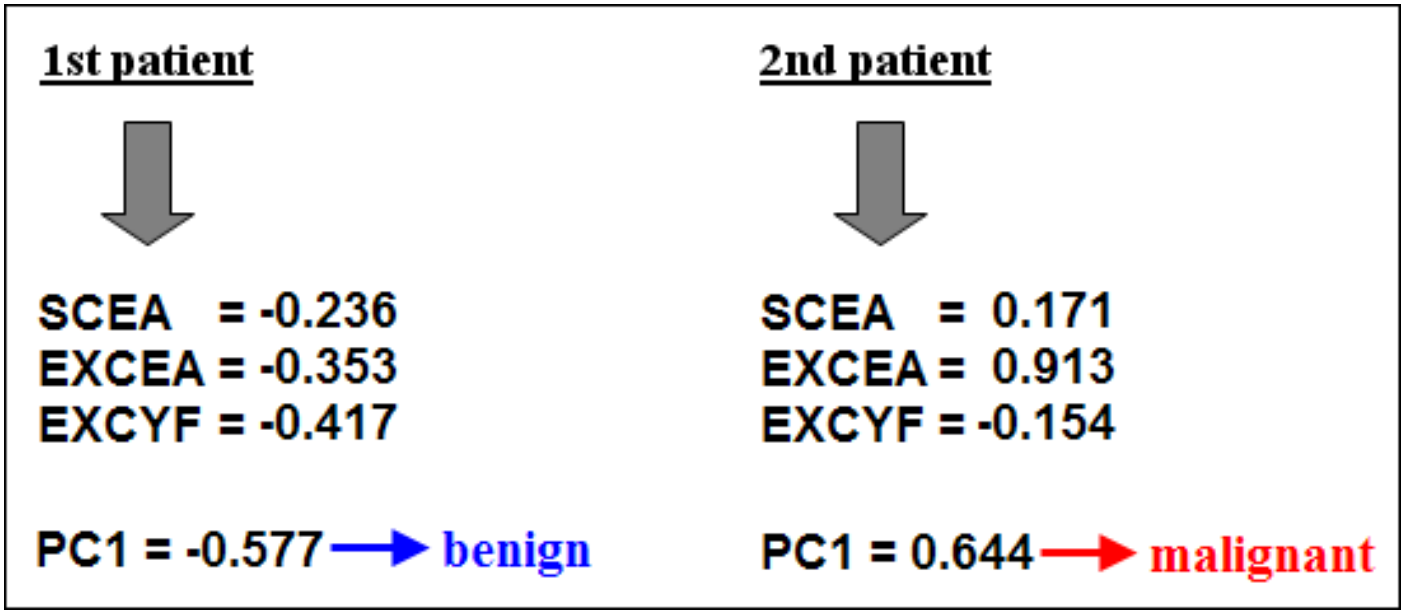


Magnified central parts of box-plots demonstrate the positions of lower quartile, median and upper quartile. The ends of the lower and upper straight-lines correspond to 5 % and 95 % percentiles, respectively.

Computer assisted prediction of diagnosis (lung tumour malignity)

Scheme for diagnosis prediction according to standardized values of tumor marker concn. using the **cut-off value -0.34** established at the max. of ***Efficiency*** (implemented into hospital information system)

$$PC1 = 0.6957\ EXCEA + 0.5074\ EXCYF + 0.5086\ SCEA$$



Conclusions

- Determination of concentration of tumor markers followed by the processing of the received data by an appropriate technique of multidimensional data analysis leads to a fast orientational diagnosis, which should be implemented into clinical practice
- Developed approach is a very promising perspective of the computer aided clinical diagnosis
- In comparison of the results of diagnosis prediction, the artificial neural networks were found as the best method for recognition of lung tumors

References – medicinal applications

Tumour & cardiovascular markers

Mrazova V., Mocak J., Varmusova E., Kavkova D., Bednarova A.: Use of multidimensional data analysis for prediction of lung malignity. *J. Pharm. Biomed. Analysis* 50 (2009) 210-215.

Mrázová V., Mocák J., Varmusová E., Kavková D.: Computer-aided diagnosis of lung malignity using multidimensional analysis of tumour markers data. *Nova Biotechnol.* 8 (2008) 65-70.

Balla B., Mocák J. Varmusová E., Kavková D., Tudík I.: Evaluation of effectiveness of laboratory methods. *Chem Listy* 97 (2004) 333-338.

Balla B., Mocak J., Pivovarnikova H., Balla J.: Comparative study of cardiovascular markers data by various techniques of multivariate analysis. *Chemom. Intell. Lab. Systems* 72 (2004) 259-267.

Statines - effects of statine therapy

Ďurčeková T., Ján Mocák J., Balla J., Boronová K.: Effect of statin therapy on biochemical laboratory tests - A chemometrics study. *J. Pharm. Biomed. Analysis* 54 (2011) 141-147.

Ďurčeková T., Mocák J., Balla J., Gromanová G., Boronová K.: Comparison of ten biochemical laboratory tests before and after treatment by statins. *Nova Biotechnologica* 9 (2009) 133 -140.

Glycated haemoglobine – harmonisation of reference systems

Mrázova V., Mocak J., Bednárová A, Balla J.: Comparison of IFCC and NGSP methods for determination of glycated haemoglobin using advanced regression techniques. *Centr. Eur. J. Chem.* 8 (2010) 1216-1222, doi: 10.2478/s11532-010-0096-x.

Mrázova V., Mocak J., Balla J., Bednárová A.: Comparison of analytical methods using various types of regression. *Chem. Listy* 104 (2010) s676-s679.

References – pharmaceutical applications

QSAR (quantitative structure – activity relationships) of *N*-benzylsalicylamides:

Nemeček P., Ďurčeková T., Mocak J., Waisser K.: Chemometrical analysis of new QSAR parameters and their exploitation for prediction of biological activity. *Chem. Papers* 63 (2009) 84-91.

Nemeček P., Ďurčeková T., Mocák J., Lehotay J., Waisser K.: Study of relationships between biological activity and physico-chemical properties of potential antituberculotics. *Nova Biotechnol.* 6 (2006) 37-47.

QSAR and QSRR (structure – chromatographic retention) of phenylbenzoxazinediones:

Nemecek P., Mocák J., Lehotay J., Waisser K.: Prediction of antimycobacterial activity of 3-phenyl-2*H*-1,3-benzoxazin-2,4(3*H*)-dione derivatives. *Anal. Sciences* (2010), in press.

Nemeček P., Mocák J., Lehotay J., Waisser K.: QSRR study of potential antituberculotic agents. *Chem. Listy* 104 (2010) s470-s474.

Nemeček P., Mocák J., Lehotay J., Waisser K.: Prediction of HPLC Retention Factor of Potential Antituberculotics by QSRR. *J. Liq. Chrom. & Rel. Tech.* (2010), in press.

QSAR and QSRR of esters of alkoxyphenylcarbamic acid:

Durcekova T., Mocak J., Lehotay J., Cizmarik J., Boronova K.: Study of anaesthetical activity of esters of alkoxyphenylcarbamic acid using chemometrical methods. *Pharmazie* 65 (2010) 169-174.

Ďurčeková T., Boronová K., Lehotay J., Mocák J., Čižmárik J.: Utilization of artificial neural networks for the study of the correlations between calculated and measured values of retention factors of esters of alkoxysubstituted phenylcarbamic acid. *Čes. slov. farmacie* 59 (2010) 205-209.

The castle of Bratislava

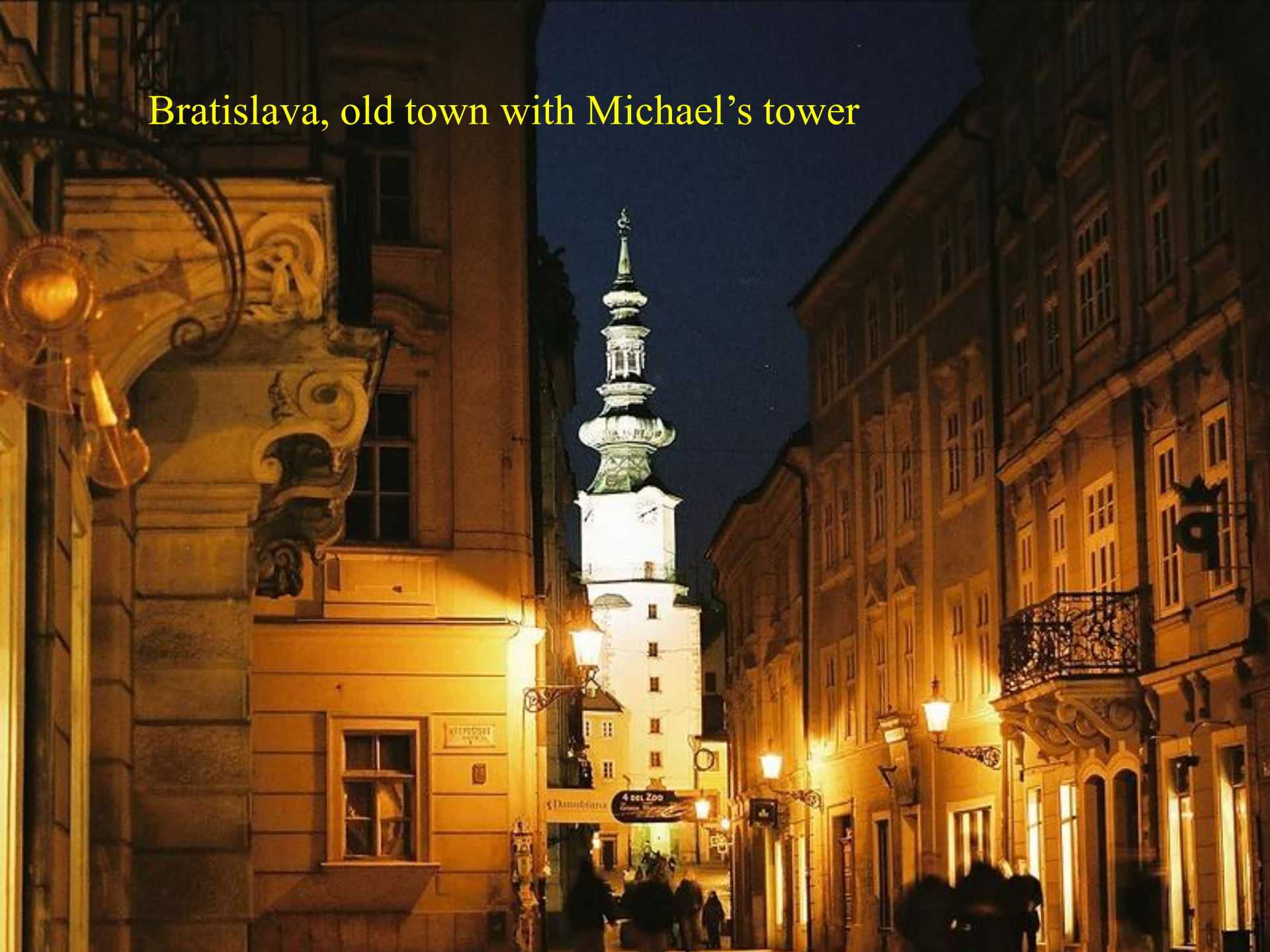
The castle of Bratislava


Bratislava, Grassalkovich palace, site of the president


Bratislava, Grassalkovich palace

Bratislava, Slovak National Theatre


Bratislava, Town Hall


Bratislava, old town with Michael's tower

Thank you!