10.2478/v10048-008-0026-7

MEASUREMENT SCIENCE REVIEW, Volume 8, Section 1, No. 5, 2008

Statistical Method Based on Confidence and Prediction Regions for Analysis of Volatile Organic Compounds in Human Breath Gas

G. Wimmer jr.

Institute of Measurement Science, Slovak Academy of Sciences,
Dúbravská cesta 9, 84104 Bratislava, Slovakia
Mathematical Institute, Slovak Academy of Sciences
Štefánikova 49, 814 73 Bratislava
E-mail: wimmerg@mat.savba.sk

In this paper we introduce two confidence and two prediction regions for statistical characterization of concentration measurements of product ions in order to discriminate various groups of persons for prospective better detection of primary lung cancer. Two MATLAB algorithms have been created for more adequate description of concentration measurements of volatile organic compounds in human breath gas for potential detection of primary lung cancer and for evaluation of the appropriate confidence and prediction regions.

Keywords: breath gas concentration, lung cancer detection, linear regression model, confidence interval, prediction interval.

1. Introduction

IN THIS PAPER we present our contribution to the test system for statistical analysis of concentration measurements of volatile organic compounds (VOCs) in human breath gas for potential detection of primary lung cancer (PLC). In particular, we suggest a method to determine two types of confidence regions and two types of prediction regions for the VOC concentrations. The test system is created in the frame of the EU-Project "Breath-gas analysis for molecular-oriented detection of minimal diseases" (BAMOD) and is applicable also for the statistical analysis of VOC-measurements on cell and bacterial cultures by proton-transfer-reaction mass spectrometry (PTR-MS), as well as for the analysis of human breath gas concentration measurements based on selected-ion-flow-tube mass spectrometry (SIFT-MS). The database with data from breath gas samples of primary lung cancer patients and healthy individuals measured by PTR-MS consists of concentration measurements of product ions at 208 selected m/z values (mass-to-charge ratios), namely, m/z from 22 to 230 excluding 37. The MATLAB algorithm "Test_Of_Degree" determines the optimal statistical model for measured data. The MATLAB algorithm "Intervals" computes the proposed confidence and prediction regions.

2. Statistical model for repeated concentration measurements

Let the random variable $Y_{i,j}$, with its realization $y_{i,j}$, be the logarithm of the j-th independent concentration measurement of the product ions at the i-th mass-to-charge ratio m/z. For the J independent measurements of concentration of product ions at selected m/z values we consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y} = (Y_{1,1},\ Y_{1,2},\ \ldots, Y_{1,J},\ Y_{2,1},\ Y_{2,2},\ \ldots,\ Y_{2,J},\ \ldots,\ Y_{I,1}, Y_{I,2}, \ldots, Y_{I,J})'$, $\mathbf{X} = \mathbf{I}_{I,I} \otimes \mathbf{1}_{J,1}$ is a $IJ \times I$ matrix, the symbol $\otimes$ denotes the Kronecker product, $\boldsymbol{\alpha} = (\mu_1, \mu_2, \ldots, \mu_I)'$ is a vector of unknown parameters, $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \varepsilon_{1,2}, \ldots, \varepsilon_{I,J})'$ is a vector of normally distributed errors, i.e. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{W})$, $\mathbf{W} = diag(1/w_1, \ldots, 1/w_1, 1/w_2, \ldots, 1/w_2 \ldots, 1/w_I, \ldots, 1/w_I)$, σ^2 is an unknown scalar factor of the covariance matrix and w_i are the weights of the measurements $Y_{i,1}, \ldots, Y_{i,J}$ that are inversely proportional to the variances $var(Y_{i,1}), \ldots, var(Y_{i,J})$. As a reasonable approximation of the true values of the weights we use $w_i \approx 1/s_i^2$ $s_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(y_{i,j} - \bar{y}_i)^2$, where $\bar{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{i,j}$.

Further, we will assume that the valid model for the measurements $\mathbf{Y}$ could be of the following form

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where

$$\mathbf{U} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ & & & \vdots & \\ 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ & & & \vdots & \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ & \ddots & & & \\ & & \ddots & & \\ 1 & x_I & x_I^2 & \ldots & x_I^k \\ & & & \vdots & \\ 1 & x_I & x_I^2 & \ldots & x_I^k \end{pmatrix}$$

is a given $IJ \times (k+1)$ matrix, x_j, $j = 1, 2, \ldots, I$ are points, in which are measurements done (in this case they are m/z values), $k \in \{0, 1, \ldots, J-1\}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is a vector of unknown parameters, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{W})$. Model (2) supposes, that the mean values of $\mathbf{Y}$ are lying on a polynomial of degree k. It is easy to see, that model $(\mathbf{Y}, \mathbf{U}\boldsymbol{\beta}, \sigma^2\mathbf{W})$ is a submodel (see [1]) of the linear regression model $(\mathbf{Y}, \mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{W})$. If $\mathbf{Y}$ satisfies the submodel $(\mathbf{Y}, \mathbf{U}\boldsymbol{\beta}, \sigma^2\mathbf{W})$ then (see [1], p. 143)

$$F_k = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'\mathbf{W}^{-1}(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})/(I - (k+1))}{(\mathbf{Y} - \hat{\boldsymbol{\mu}})'\mathbf{W}^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}})/(IJ - I)} \sim F_{(I-(k+1)),(IJ-I)}, \tag{3}$$

where $\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}$ is the best linear unbiased estimator of $E(\mathbf{Y})$ in the model (1) and $\hat{\boldsymbol{\nu}} = \mathbf{U}(\mathbf{U}'\mathbf{W}^{-1}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}^{-1}\mathbf{Y}$ is the best linear unbiased estimator of $E(\mathbf{Y})$ in the submodel (2). The MATLAB algorithm "Test_Of_Degree" is searching for the lowest k, such that the hypotheses H_0:"Measured data are not inconsistent with the statement, that $E(\mathbf{Y})$ is lying on the polynomial of k-th degree", is not rejected.

3. Confidence and prediction regions

Let $\mathbf{Y} \sim N(\mathbf{U}\boldsymbol{\beta}; \sigma^2\mathbf{W})$ and $\{(m/z)_1, (m/z)_2, \ldots, (m/z)_I\} = \mathcal{A}$ where $(m/z)_i = x_i$, $i = \{1, 2, \ldots, I\}$ are the mass-to-charge ratios for which the product ions concentrations were measured. Further, let $\mathbf{x}_i = (1, x_i, x_i^2, \ldots, x_i^k)'$ and let $d_{x_i} = \sqrt{\mathbf{x}_i'(\mathbf{U}'\mathbf{W}^{-1}\mathbf{U})^{-1}\mathbf{x}_i}$.

For chosen $x_i \in \mathcal{A}$ the $(1-\alpha)100\%$-confidence interval for $\mathbf{x}_i'\boldsymbol{\beta}$ is given as

$$\left\langle \mathbf{x}_i'\hat{\boldsymbol{\beta}} - s d_{x_i} t_{(IJ-(k+1))}(1-\alpha/2), \; \mathbf{x}_i'\hat{\boldsymbol{\beta}} + s d_{x_i} t_{(IJ-(k+1))}(1-\alpha/2) \right\rangle, \tag{4}$$

where $s^2 = [(\mathbf{Y}-\hat{\mathbf{Y}})'\mathbf{W}^{-1}(\mathbf{Y}-\hat{\mathbf{Y}})]/(IJ-(k+1))$ is the unbiased estimator of σ^2 and $\hat{\boldsymbol{\beta}} = (\mathbf{U}'\mathbf{W}^{-1}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}^{-1}\mathbf{Y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, $\hat{\mathbf{Y}} = \mathbf{U}\hat{\boldsymbol{\beta}}$, and further, $t_{(IJ-(k+1))}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the t-distribution with $(IJ-(k+1))$ degrees of freedom.

At least $(1-\alpha)100\%$-confidence region for $\mathbf{x}_i'\boldsymbol{\beta}$ for all $x_i \in \mathcal{A}$ is given as

$$\left\langle \mathbf{x}_i'\hat{\boldsymbol{\beta}} - s d_{x_i} \sqrt{(k+1)F_{(k+1,(IJ-(k+1)))}(1-\alpha)}, \; \mathbf{x}_i'\hat{\boldsymbol{\beta}} + s d_{x_i} \sqrt{(k+1)F_{(k+1,(IJ-(k+1)))}(1-\alpha)} \right\rangle, \tag{5}$$

where $F_{(k+1,(IJ-(k+1)))}(1-\alpha)$ is the $(1-\alpha)$ quantile of the F-distribution with $k+1, (IJ-k-1)$ degrees of freedom.

Interval, which covers the next realization of the random variable $Y_{x_i} = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{x_i}, \varepsilon_{x_i} \sim N(0, \frac{\sigma^2}{w_i}), x_i \in \mathcal{A}$ with probability $(1-\alpha)$ is

$$\left\langle \mathbf{x}_i'\hat{\boldsymbol{\beta}} - s\sqrt{\frac{1}{w_i} + d_{x_i}^2}\, t_{(IJ-(k+1))}(1-\alpha/2), \; \mathbf{x}_i'\hat{\boldsymbol{\beta}} + s\sqrt{\frac{1}{w_i} + d_{x_i}^2}\, t_{(IJ-(k+1))}(1-\alpha/2) \right\rangle. \tag{6}$$

This random interval is called the $(1-\alpha)100\%$-prediction interval for a single future observation Y_{x_i}.

At least $(1-\alpha)100\%$-prediction region for Y_{x_i} for all $x_i \in \mathcal{A}$ is

$$\left\langle \mathbf{x}_i'\hat{\boldsymbol{\beta}} - s\sqrt{\frac{1}{w_i} + d_{x_i}^2}\sqrt{IF_{(I,(IJ-(k+1)))}(1-\alpha)}, \; \mathbf{x}_i'\hat{\boldsymbol{\beta}} + s\sqrt{\frac{1}{w_i} + d_{x_i}^2}\sqrt{IF_{(I,(IJ-(k+1)))}(1-\alpha)} \right\rangle. \tag{7}$$

In the situation, when $k << I$, the tolerance region could be more suitable than the prediction region (7), for more details see [2], [3].

The MATLAB algorithm "Intervals" computes the confidence interval (4) and prediction interval (6) for chosen k and $x_i \in \mathcal{A}$ and also the confidence region (5) and prediction region (7), respectively.

4. Some illustrative examples

Let us demonstrate the above mentioned confidence and prediction regions on the data, where the concentrations of selected VOCs were determined in exhaled breath samples of healthy subjects and of subjects with primary lung cancer by proton-transfer-reaction mass spectrometry (PTR-MS) in ppb (particles-per-billion) levels, sampled and measured at the Medical University of Innsbruck, Austria, during years 2006 and 2007. The measured counts were transformed using the knowledge of chemistry kinetics and reaction constants to concentrations of volatile organic compounds in ppb levels. The medians from at least 3 repeated concentration measurements of the selected compounds (m/z values) were taken per each breath sample and were used for analysis (in all cases is $\alpha = 0.05$).

Figure 1

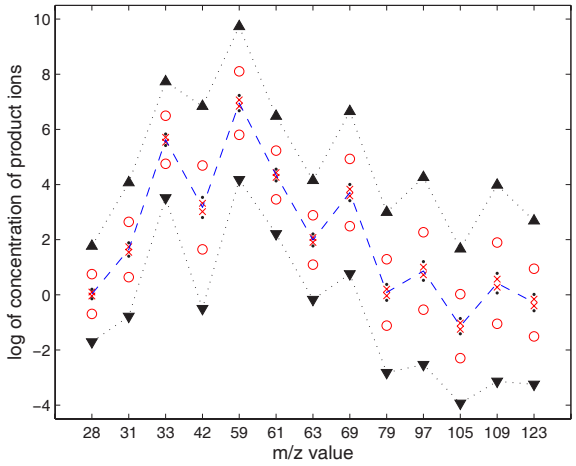

Figure 2

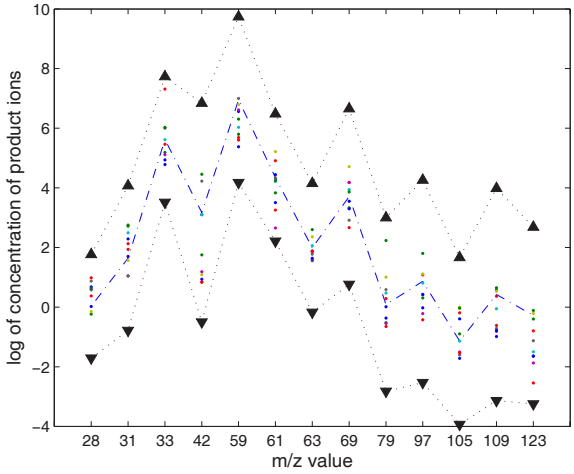

Figure 3

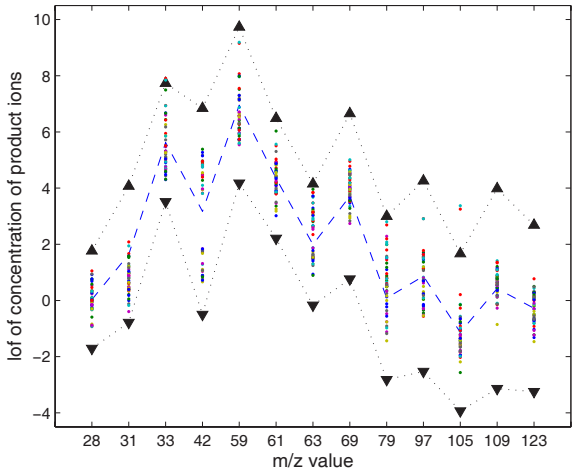

Figure 4

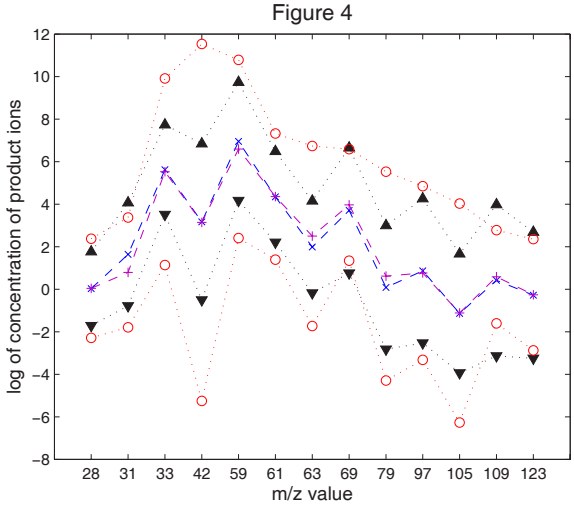

Figure 1 presents a plot of confidence intervals given by equation (4) (symbol $\times$), the confidence region given by (5) (symbol $\cdot$), the prediction intervals given by (6) (symbol $\circ$), and the prediction region given by the equation (7) (symbol "filled triangle") for mass-to-charge ratios 28,31,33,42,59,61,63,69,79,97,105,109,123 from $J = 300$ repeated measurements evaluated from healthy subjects. The estimated expected value $E(\mathbf{Y}) = \hat{\boldsymbol{\nu}} = \mathbf{U}(\mathbf{U}'\mathbf{W}^{-1}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}^{-1}\mathbf{Y}$ for the model (2) with lowest $k = 12$ is lying on the dashed line.

Figure 2 presents a plot of the prediction region given by the equation (7) for mass-to-charge ratios 28,31,33,42,59,61,63,69,79,97,105,109,123 from $J = 300$ repeated measurements evaluated from healthy subjects with further proper measurements done on health subjects (10 points). Estimated expected value $E(\mathbf{Y})$ for the model (2) with lowest $k = 12$ is lying on the dashed line.

Figure 3 presents a plot of the prediction region given by the equation (7) for mass-to-charge ratios 28,31,33,42,59,61,63,69,79,97,105,109,123 from $J = 300$ repeated measurements evaluated from healthy subjects with proper measurements done on subjects with primary lung cancer (50 points). Estimated expected value $E(\mathbf{Y})$ for the model (2) with lowest $k = 12$ is lying on the dashed line.

Figure 4 presents a plot of the prediction region given by the equation (7) for mass-to-charge ratios 28,31,33,42,59,61,63,69,79,97,105,109,123 from $J = 300$ repeated measurements evaluated from healthy subjects (symbol "filled triangle") with proper prediction region given by the equation (7) evaluated from $J = 300$ subjects with primary lung cancer (symbol $\circ$). Estimated expected value $E(\mathbf{Y})$ for the model (2) with lowest $k = 12$ from healthy subjects is located on the dashed line with (symbol $\times$) and estimated expected value $E(\mathbf{Y})$ for the model (2) with lowest $k = 12$ from subjects with primary lung cancer is located on the dashed line with (symbol $+$).

5. Discussion and conclusions

The repeated measurements on tested persons are fully characterized by the suggested confidence and prediction regions, given in (4), (5), (6), (7). According to these regions we can make judgments about the repeatability and the random error of the measuring method and/or measuring device. We can explicitly see whether other measurements differ from measurements on the tested person and/or to what extend they are different. It is expected that by using the above mentioned statistical tools it would be possible to better characterize also some groups of subjects in interest (hospital staff, cancer-smokers, control-smokers, cancer-nonsmokers, control-nonsmokers, etc.). Unfortunately, no differences have appeared in examples from section (4). In Figure 3, it is visible that almost all measurements done on subjects with primary lung cancer are lying in the prediction region given by the equation (7) for healthy subjects. Therefore we cannot see any differences between healthy subjects and subjects with primary lung cancer at these mass-to-charge ratios (m/z values). From the next example, Figure 4, we can deduce that estimated expected values $E(\mathbf{Y})$ from healthy subjects and from subjects with primary lung cancer are nearly the same. Further, we have not found differences between healthy subjects and subjects with primary lung cancer on other tested combinations either. Based on this finding we suspect that we have no evidence that it is possible to detect primary lung cancer from volatile organic compounds in human breath gas through this suggested method.

Acknowledgement

The research was supported by the EU project BAMOD – the FP6 project within the framework of the specific research and technological development programme "Integrating and strengthening the European Research Area, LSHC-CT-2005-019031 STREP, and by the grant of the Scientific Grant Agency of the Slovak Republic (VEGA) project No. 1/3016/06.

References

[1] Anděl, J. *Matematická statistika.* (Mathematical Statistics. In Czech). SNTL/ALFA, Praha 1978.

[2] NIST-SEMATECH. *e-Handbook of Statistical Methods*, National Institute of Standards and Technology, http://www.itl.nist.gov/div898/handbook/, 23-March-2007.

[3] Zvára, K. *Regresní analýza.* (Regression Analysis. In Czech). Academia, Praha 1989.