Lithuanian Journal of Physics, Vol. **48**, No. 3, pp. 209–217 (2008) doi:10.3952/lithjphys.48310

QUANTIFICATION OF PATTERN RECOGNITION QUALITY BY MULTIVARIATE NORMAL DISTRIBUTION FUNCTIONS

P. Serapinas, Ž. Ežerinskis, A. Acus

Institute of Theoretical Physics and Astronomy of Vilnius University, A. Goštauto 12, LT-01108 Vilnius, Lithuania
E-mail: ezerinskis@pfi.lt

Received 9 May 2008; revised 18 September 2008; accepted 18 September 2008

Analysis of the multivariate data distributions can be helpful or directly applicable in pattern recognition tests. Estimate of the volume of the critical region of overlapping distributions is essential in determination of the confidence level of classification. Mathematical tools for analysis of the multivariate distributions (included probability, false positives and false negatives, means for calculation of the critical region) are developed. Sum of the false negative and the false positive is found as a very approximate characteristic of the total uncertainty of classification. The false negative probability is extremely distribution coordinate dependent and analysis of the details of the overlapping distributions is needed to evaluate the real risk of misclassification of samples. Application of the multivariate distributions to the regional classification of wine samples according to the data of multielement analysis is presented as an example.

Keywords: multivariate distribution, uncertainty, pattern recognition, confidence test, food analysis

PACS: 02.50.Sk, 07.05.Ka, 82.80.Ms, 89.75.Kd

1. Introduction

Variety of information presented as numerous data form the basis of the modern decision making in deciding or rejecting suggested hypothesis or selecting between few of them. It is because and due to possibility of generation of the large data sets by the modern measurement techniques. As an example, data on composition of a lot of samples including somewhere up to sixty chemical elements is usual in recent environmental and geochemistry research, food chemistry and food authenticity studies, clinical and forensic toxicology. Even larger data sets are characteristic of the data for multivariate calibration, ultraviolet, visible, infrared, and mass spectrometry, gene and time series studies. Various techniques, including dispersion and correlation analyses, discriminant, factor, principal component, cluster, neural networks, and others [1] are used for more concise presentation, analysis, and interpretation of such data sets. Many aspects of the methodology of the use of those and related techniques were discussed in recent publications [1–6]

Naturally, in the classification or pattern recognition matters, evaluation of the quality of classification is an essential issue. In the present paper we show that analysis of the multivariate distributions of data or some their derivatives, as principal components, for example, can be helpful, or directly applicable for the pattern recognition studies. A lot of tables and mathematical expressions for analysis of the bivariate distributions can be found, but very few data concern multivariate distributions. Tables or convenient means for calculation of the probability density functions, α and β type errors, evaluation of the critical region of the overlapping distributions would be of interest. The aim of the present paper is to aid development of such means and application of the multivariate distributions in data analysis. In some cases the results were found to be extraordinary simple. Application of the results to classification of wine samples according to their country of origin is presented as an illustration.

2. Multivariate normal distributions. Theoretical treatment

The normal distribution characterizes the data where many small, independent effects additively contribute to each observation. The distribution is described by two parameters: location (typically mean or "average", μ) and scale (standard deviation or "variability", σ).

ISSN 1648-8504

The continuous probability density function (PDF) of the normal distribution is the Gaussian

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad (1)$$

with $\sigma > 0$. About 68% of values of the normal distribution are contained in the range of one standard deviation from the peak. About 95% of the values are within two standard deviations and about 99.7% are located within three standard deviations from the mean. Precisely, the area under the curve between $-n\sigma$ and $n\sigma$ is geometrical definition of the standard error function $\mathrm{erf}(x)$ of real argument

$$P(|x|<n\sigma)=\mathrm{erf}\left(\frac{n}{\sqrt{2}}\right). \qquad (2)$$

In practice the range in which the distribution of variables is being regarded is always limited. If, for example, in "deciding whether or not a particular sample may be judged as likely to have been randomly drawn from a certain population" [7] we restrict ourselves to the range $\pm 2\sigma$ (power of the test 0.95), the probability of rejecting the null hypothesis that is actually true (false positive, α or type I error, or p level of significance) is 0.05. If the distributions partially cross, false negative error, or acceptance of the null hypothesis while, in fact, the alternative hypothesis is true, is possible. Naturally, no ambiguity arises if the critical region, where the two distributions overlap, is small as compared to the selected level of significance α. If the critical region is comparable to α, then careful analysis of the probability distributions inside the critical region is necessary. Many tables and procedures can be found to help analysis of the univariate distributions. In contrast, for the multivariate distributions direct calculations usually have to be performed. The probability density function of the bivariate normal distribution is

$$f(x,y;\mu,\sigma)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\right.$$
$$\left.\times\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho xy}{\sigma_x\sigma_y}\right]\right\}, \qquad (3)$$

where σ_x and σ_y are the standard deviations of the x and y values, ρ is the correlation coefficient.

The d-dimensional Gaussian (multivariate probability distribution) function is defined as

$$f(\vec{x};\vec{\mu},\sigma)=\frac{1}{(2\pi)^{d/2}\sqrt{\det\sigma}}$$
$$\times\exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^{\top}\sigma^{-1}(\vec{x}-\vec{\mu})\right], \qquad (4)$$

where $^{\top}$ denotes transposition and $^{-1}$ inverse operation, correspondingly. σ is a covariance matrix, with

$$\sigma_{ij}=\langle(x_i-\mu_i)(x_j-\mu_j)\rangle. \qquad (5)$$

Here the angle brackets denote expectation value of the quantity inside:

$$\mu_i=\langle x_i\rangle. \qquad (6)$$

Using Eq. (4) the probability density function of the trivariate normal distribution takes the following explicit form:

$$f(x_1,x_2,x_3;\mu,\sigma)=\frac{1}{2\sqrt{2}\pi^{3/2}\sqrt{\rho_t}}\exp\left\{-\frac{1}{2\rho_t}\right.$$
$$\times\Big[(x_3-\mu_3)\Big((x_3-\mu_3)(\sigma_{11}\sigma_{22}-\sigma_{12}^2)$$
$$+(x_2-\mu_2)(\sigma_{12}\sigma_{13}-\sigma_{11}\sigma_{23})$$
$$+(x_1-\mu_1)(\sigma_{12}\sigma_{23}-\sigma_{13}\sigma_{22})\Big)$$
$$+(x_2-\mu_2)\Big((x_3-\mu_3)(\sigma_{12}\sigma_{13}-\sigma_{11}\sigma_{23})$$
$$+(x_2-\mu_2)(\sigma_{11}\sigma_{33}-\sigma_{13}^2)$$
$$+(x_1-\mu_1)(\sigma_{13}\sigma_{23}-\sigma_{12}\sigma_{3,3})\Big)$$
$$+(x_1-\mu_1)\Big((x_3-\mu_3)(\sigma_{12}\sigma_{23}-\sigma_{13}\sigma_{22})$$
$$+(x_2-\mu_2)(\sigma_{13}\sigma_{23}-\sigma_{12}\sigma_{33})$$
$$\left.+(x_1-\mu_1)(\sigma_{22}\sigma_{33}-\sigma_{23}^2)\Big)\Big]\right\}. \qquad (7)$$

Here

$$\rho_t=-\sigma_{33}\sigma_{12}^2+2\sigma_{13}\sigma_{23}\sigma_{12}-\sigma_{11}\sigma_{23}^2-\sigma_{13}^2\sigma_{22}$$
$$+\sigma_{11}\sigma_{22}\sigma_{33}. \qquad (8)$$

Note that for bivariate distribution $\sigma_{11}=\sigma_1^2, \sigma_{22}=\sigma_2^2$, and $\sigma_{12}=\rho\sigma_1\sigma_2$. Some data for comparison of the normal, bivariate, and trivariate probability distributions are presented in Table 1. Because volume of ellipsoid is always less than volume of surrounding cuboid, it follows that probability included in the ellipsoidal volume within some standard deviation from the mean is also always less than the normal distribution probability of the same interval raised to power of dimension of the space. In particular, the normal probability

Table 1. Comparison of the included ellipsoidal probability of the univariate (normal), bivariate, and trivariate distributions. The second and the third entries of the univariate distribution column represent the first entry raised to the power of 2 and 3 correspondingly.

Range	Normal,	squared,	cubed	Bivariate	Trivariate
0.5σ	0.3829,	0.1466,	0.0561	0.11750	0.03086
1.0σ	0.6827,	0.4661,	0.3182	0.39347	0.19875
1.5σ	0.8664,	0.7506,	0.6503	0.67534	0.47783
2.0σ	0.9545,	0.9111,	0.8696	0.86467	0.73854
2.5σ	0.9876,	0.9753,	0.9632	0.95607	0.89994
3.0σ	0.9973,	0.9946,	0.9919	0.98889	0.97071

squared is always slightly larger than elliptical bivariate probability. The same is true for trivariate probability (Table 1) when compared with the normal probability raised to power of 3. The presented data for these distributions depend neither on the variables ratio nor on the covariance matrix.

Included ellipsoidal probabilities of higher dimensional multivariate distributions can be described by the following simple formula:

$$P(d, |x| < n\sigma) = 1 - \frac{\Gamma(\frac{d}{2}, \frac{n^2}{2})}{\Gamma(\frac{d}{2})}\,. \qquad (9)$$

Here d is the space dimension and Γ denotes the usual gamma function:

$$\Gamma(z) = \int\limits_0^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t\,, \quad \mathrm{Re}(z) > 0\,, \qquad (10)$$

$$\Gamma(z, z_1) = \int\limits_{z_1}^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t\,. \qquad (11)$$

From (9) it comes as a little surprise that ellipsoidal probabilities in even-dimensional space can be expressed using the elementary functions only. For example, for bivariate distribution (3) the included probability within 2σ variance interval is $P(2, |x| < 2\sigma) = 1 - 1/\mathrm{e}^2$. In odd-dimensional spaces this probability includes single one-dimensional $\mathrm{erf}(z)$ function. For example, for trivariate distribution the included probability for the same deviation is $P(3, |x| < 2\sigma) = -4/(\sqrt{2\pi}\mathrm{e}^2) + \mathrm{erf}(\sqrt{2})$. These can be easily checked to have the same numerical values as presented in Table 1. More results on exact expansion of (9) are presented in Appendix.

If two distributions overlap, the volume of the critical region can be found as an integral common to both distributions. If some level of significance is accepted, the range of the distribution is restricted by the corresponding ellipsis and the critical region is part of the overlapping distribution inside the ellipsis. Risk for the corresponding false negative error must be accounted for when the data inside the ellipsis are being regarded.

3. Analysis of the real distributions. Discussion

Characteristic examples of the bivariate and 3-dimensional data distributions are presented in Figs. 1 and 2 below. The data represent the problem of classification of wine samples measured in [8]. The absolute concentrations of 19 elements, namely Li, B, Na, Mg, Al, K, Ca, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Rb, Sr, Ba in 102 wines from Bulgaria (5), Chile (25), France (26), Hungary (7), Italy (6), Spain (28), California (5), measured by double focusing sector field mass spectrometer Element2 were taken for the analysis below. The number of the tested samples is given in parentheses. Tm was used as an internal standard. Relative measurement uncertainty usually did not exceed 10%. Step by step approach from all the samples to smaller classes (see Fig. 1) was used for classification. The Anova F-test was used to select the most informative elements at each classification step and enabled reduction of noise. More details on the measurement procedure can be found in [8]. In particular, elements Rb, Sr, Li, and Zn had large F-ratio values and were most useful for the current classification. Principal component analysis (PCA) for the selected elements was used to minimize correlations between measured data. Sometimes even one principal component was enough to distinguish between the two populations. Usually the first principal component explained 40–80% (44% in Fig. 1 and 60% in Fig. 2) of the dispersion of the data, second principal component 10–40% (33% in Fig. 1 and 25% in Fig. 2), but sometimes the higher principal components are important. For example, the 3rd principal component explained 22.3% of the variance relative to Fig. 1. The 3rd and 4th components (not shown) explain 11.0 and 3.7% of the variance relative to Fig. 2, correspondingly.

Of course, it is not evident in advance that application of the PCA will aid regional classification. The main tendencies of data variation highlighted by the principal components cannot be necessarily due to the regional effects. Studies in [8], where comparison of the data classification capabilities of the raw element concentration data and the principal components was undertaken, revealed that application of the PC increased classification capabilities in comparison to the raw data in the case under study. In addition, the results of the PC distribution tests were in correspondence

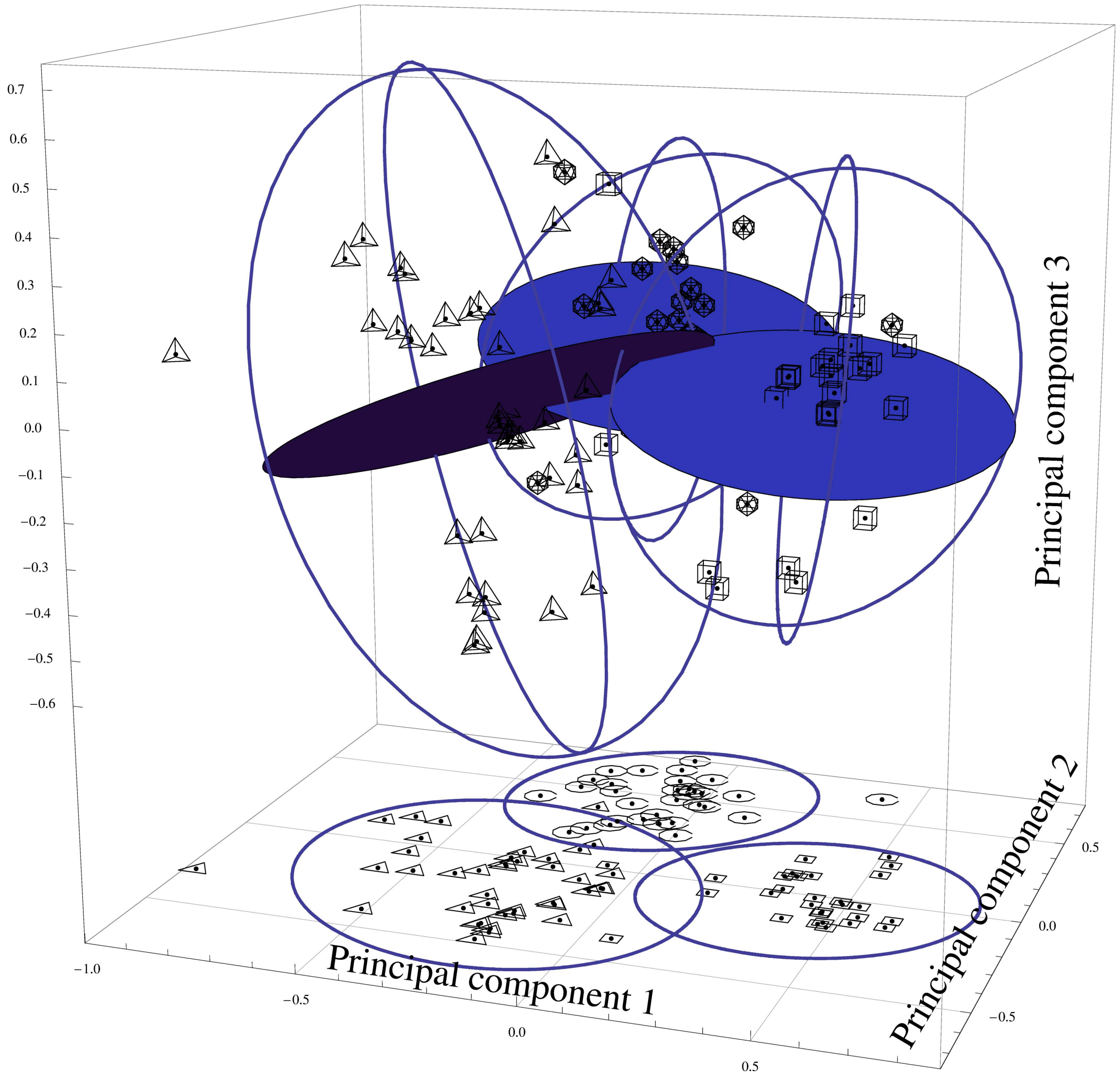

Fig. 1. Results of PCA analysis of concentrations of Sr, Rb, and Zn in wines from different countries. Square (cuboid) □ denotes samples from Chile and California, triangle (tetrahedron) △ is for wines from Spain, Bulgaria, and Hungary, and circle (sphere) ○ represents brands from France and Italy. 3D ellipsoids are drawn at 95% confidence level. 2D ellipses are calculated in the same way using $x-y$ projection points.

with the normal distribution, although the power of the test was not high because of the small data sets.

Two and three principal component scatters of all the measured data for Sr, Rb, and Zn are presented in Fig. 1. The points tend to group into 3 batches. The one consists of samples from Spain, Bulgaria, and Hungary, the other is formed by wines form Chile and California, and the last one is wines from France and Italy. The mean values of the principal components for each group and the corresponding covariance matrices of the trivariate distribution are listed in Table 2. One could notice appreciable correlation between principal components within the groups while, from the very PC concept, no correlation between the principal components for the batch in general is possible. Specific rules of data classification is the source of this correlation.

If, for example, 95% confidence level for classification is accepted, the distributions partially overlap (Fig. 1). It is to be noted that if we draw confidence ellipsoids matching multivariate PDF parameters, then these ellipsoids will not generally coincide with the ellipsoids which include 95% of the data points. It

Table 2. Principal components analysis of Rb, Sr, and Zn element concentration data in samples of wines from Chile and California (□), Spain, Bulgaria, and Hungary (△), and France and Italy (○).

Data	△ group			□ group			○ group		
	Covariance matrix								
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
PC1	0.0400	−0.0066	−0.0128	0.0237	0.0014	0.0076	0.0223	0.0031	0.0057
PC2	−0.0066	0.0383	0.0069	0.0014	0.0188	−0.0147	0.0031	0.0210	−0.0069
PC3	−0.0128	0.0069	0.0690	0.0076	−0.0147	0.0419	0.0057	−0.0069	0.0317
	Mean								
	−0.264	−0.225	−0.0148	0.429	−0.107	−0.0008	−0.072	0.382	0.0193

is because ellipsoid constructed by descriptive statistic methods ensures that the prescribed fraction of data (in our case 95% of points) lie inside the locus of ellipsoid without any explicit hypothesis about data distribution probability function. Despite the fact that both ellipsoids are centred on group mean value and their orientations coincide (it is calculated using the same covariance matrix), lengths of major / semi-minor / minor axes generally differ.

Detailed probability distribution analysis is needed to decide how essential is the risk in classification and must or may not the accepted confidence level be reduced. The natural expectation is that disposition of more data should provide better separation of samples. This is completely confirmed by calculations of the bivariate probability function intersection integrals for different groups of data presented in Table 3. The integrals represent the part of the interfering distribution inside the 95% confidence level ellipse being regarded and can be interpreted as the hypothesis false negative error that the point in the intersection area belongs to this group, while in reality it belongs to another one. As follows from the table, classification of a sample found inside the parameter region characteristic of the △ group according to the two elements data (Sr, Rb) is not possible, because overlapping of the other distributions is very large. At the same time identification of the samples from □ and ○ groups at least at confidence level about 0.94 seems possible. Discrimination is much better for the three elements (Sr, Rb, Zn) data set (see Fig. 1 also). According to the integral characteristics presented in Table 3, classification of all the data at confidence level about 0.9 is possible. In reality it is evident that at the crossing line of the two overlapping distributions the probabilities to find a sample as originating from any of the two distributions is equal, while at the opposite side of the ellipse probability of an error in classification is negligible. Sum of the two false negatives □ and ○ relative to △ could seem as some integral characteristic of possible error, but it is clear from Fig. 1 that either one or another is possible, not both. In such a manner the integral characteristics evaluate only the mean probability of classification of a large number of samples. They account for the decreasing character of the characteristic distributions but do not correspond to the problem of classification of particular sample represented by particular data set.

As another example, it seems trivial that two batches can be classified as separate if the variation of the describing parameters within the batches is small as compared to the differences between the mean values. Nevertheless, even if the centres of the distributions coincide but dispersions are very different, classification can be possible. In Fig. 2 the PC analysis of concentrations of Sr, Rb, Zn, and Li in wines from Spain and France is presented. The overlap of the two batches at 95% confidence level is negligible. The data for Bordeaux wines (France) are marked by triangles (△) in the figure. Naturally, they are found inside the ellipsis characteristic of the wines from France. Nevertheless, it follows from calculation of the bivariate PDF cross-section integral (at 95% confidence level) that probability of the wines from other regions of France to have characteristics similar to those from Bordeaux is comparatively small. The integral of probability density function of the data of wines from France (with Bordeaux district excluded) over 95% confidence ellipse area (denoted by solid thick line in Fig. 2) of Bordeaux wines is 13.5%. This can be easily seen from very different shapes of PDF characteristic of wines from Bordeaux and France in general (Fig. 2). If Bordeaux district is included in PDF parameters estimation, the calculated integral value increases to 14.0%. If instead of two principal components as shown in Fig. 2 we take three largest principal components, the trivariate PDF integral (false negative from other regions of France relative to Bordeaux) reduces to 7.2%. If all four PCA components are taken into account, the integral further reduces to below 2%, illustrating the high potential of application of the higher dimensions.

Table 3. The bivariate probability function intersection integrals for different groups of data. False Negative (FN) is part of the interfering distribution inside the 2D 95% confidence level ellipse being regarded. The bivariate distributions were calculated for the PC of the two elements (Sr, Rb) and three elements (Sr, Rb, Zn) data sets for wines from Chile and California (□), Spain, Bulgaria, and Hungary (△), and France and Italy (◯). The third PC component of the three-element case is omitted as it has not improved the classification.

	Sr, Rb PCA; FN from			Sr, Rb, Zn PCA; FN from		
Null hypothesis	△	□	◯	△	□	◯
△	–	23.4%	59.2%	–	3.80%	8.15%
□	2.23%	–	0	2.01%	–	0
◯	6.24%	0	–	4.31%	0	–

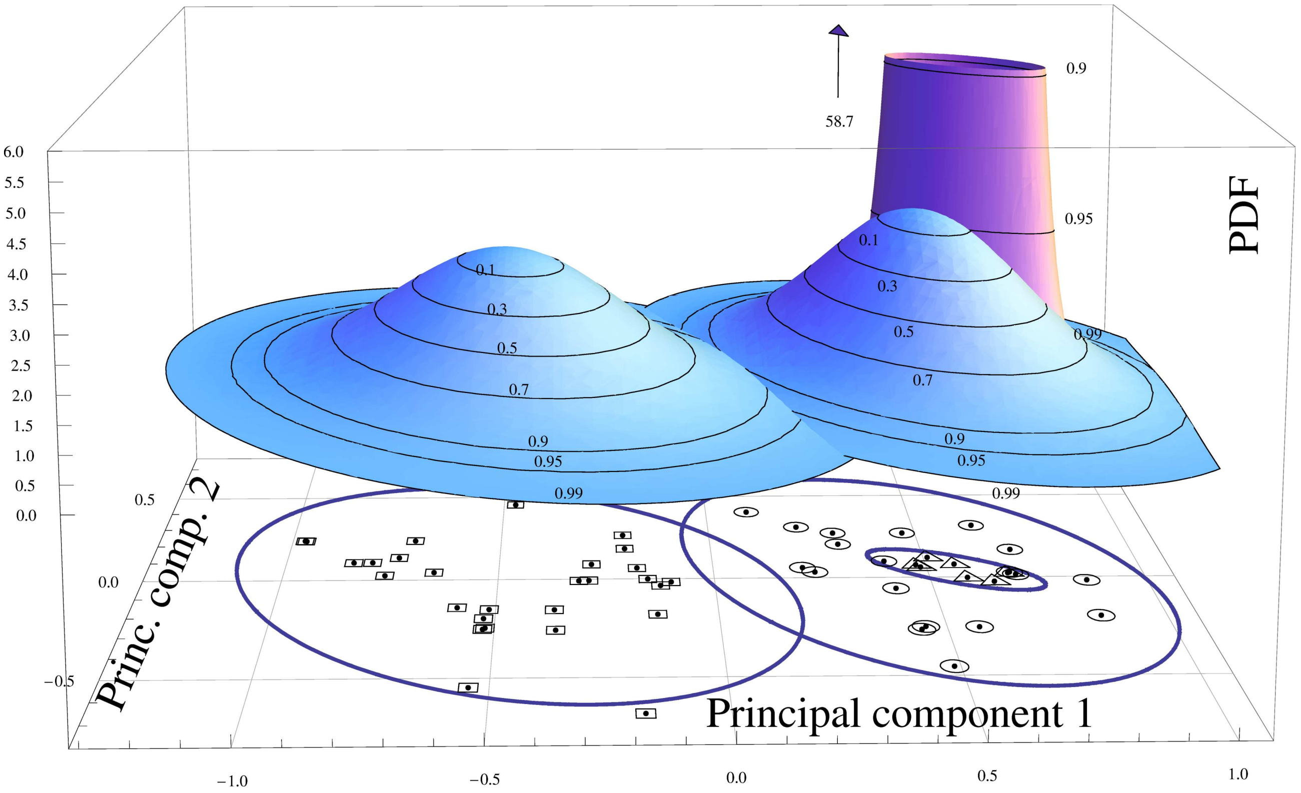

Fig. 2. Results of PCA analysis of concentrations of Sr, Rb, Zn, and Li chemical elements in wine from Spain (□) and France. Two largest PCA components, which explain correspondingly 60.3 and 25.0% of variance, are shown. The remaining PCA components explain 11.0 and 3.7% of data correspondingly. Bordeaux wines are marked by △ and brands from remaining France districts are represented by circles ◯. Ellipsoids in the $x-y$ plane (solid lines) are drawn at 95% confidence level. Isocontours of bivariate PDF surfaces denote ellipsoids of included probability.

As limited number of samples is tested, deviation of the determined standard deviation value from the true one can be significant. The Student's t coefficient can be included into calculation of the axes of the ellipses to account for the uncertainty. Multivariate Student's distributions ought to be used instead of the normal ones in the case (see [9] also). Naturally, generic principles of calculations discussed above and presented in the Appendix remain the same.

4. Conclusions

In such a manner we find that analysis of the multivariate distributions is an effective, transparent, and convenient tool to evaluate the accuracy of classification and the main sources of uncertainty. Mathematical means for application of the method are developed and presented. Possible role of the higher dimensions (up to the forth) is demonstrated. It is stressed that the detailed topography of the distributions must be analysed to evaluate the real risk of classification in any particular case under analysis. The selected confidence level for discrimination between the batches and the corresponding d-dimensional space included characterize only the probability not accounted for in the analysis, false positive, that decreases if the coverage factor is increased. The volume of the critical region is only a very rough characteristic of the overlapping distributions. The false negative probability distribution is not only distance, but especially orientation, or coordinate, dependent. The probability of misclassification is essentially different near the critical region and at the opposite side of the distribution. Thus analysis of the distributions provides detailed information on the uncertainty of classification. More detailed description of the mathematical tools used for the analysis of distributions is included in the Appendix.

Extension of the applicability of the integral classification probabilities to data analysis could be desirable. As one of the approaches, it could be done by extension of applicability of the false positive, or α error, concept. Ellipses or higher dimension pictures of the space that includes the selected probability are in correspondence with this concept. Overlapping with other distributions indicates overestimation of the confidence level. The space being regarded ought to be restricted in a manner to exclude the regions where the false negative probability density is not negligible (e. g., three or ten times less) as compared to the lowest probability density of the null hypothesis distribution in any region being accounted for (naturally, it is lowest at the border of the selected false positive level space). Such a line or surfaces ought to exclude the regions where the false negative probability is significant. Then new integral false positive value must be found as the integral probability inside this space. It would be applicable to the parameter region found from the distribution analysis.

Thus the sum of the false negative and the false positive is only an approximate characteristic of the total uncertainty of classification. The false negative probability is extremely distribution coordinate dependent and analysis of the details of the overlapping distributions is needed to evaluate the real risk of classification of the real samples. We hope that the material concerning the multivariate normal distributions presented above and in the Appendix can be helpful for such analysis.

5. Appendix

Computer algebra system *Mathematica* [10] was used both for symbolic and numeric calculations. Besides general system kernel functionality a number of functions from *MultivariateStatistic.m* and *ANOVA.m* packages appeared to be very useful.

Mathematica's impressive symbolic definite integration capabilities were used to derive formula (9). The results were checked by numerical integration procedure for a number of selected values. Expansions of (9) for particular d and σ values, part of which are used in Table 4, were calculated with the system built-in command *FunctionExpand[ ].* The expansion results again were checked by high precision numerical integration routines.

Numerical integration of the multivariate PDFs over (generally overlapping) d-dimensional ellipsoidal regions was realized with the additional Boolean help function in the integrand. This Boole function was defined to have value 1 if the point under integration belonged to the interior of both ellipsoidal shapes and 0 otherwise. Because most of the integrands included fast falling functions, numerical integration limits had to be carefully adjusted to ensure reliable best precision results.

Full calculation details are available as *Mathematica* notebook with accompanying full measured concentration data text file, available for download from http://mokslasplius.lt/eksperimentai/files/eksperimentai/Notebooks/MultivariateQuantification.tar.gz.

References

[1] L.A. Berrueta, R.M. Alonso-Salces, and K. Heberger, Supervised pattern recognition in food analysis, J. Chromatogr. A **1158**, 196–214 (2007).

[2] M. Daszykowski and B. Walczak, Use and abuse of chemometrics in chromatography, Trends Anal. Chem. **25**, 1081–1096 (2006).

[3] A. Gustavo González, Use and misuse of supervised pattern recognition methods for interpreting compositional data, J. Chromatogr. A **1158**, 215–225 (2007).

[4] S.F. Møller, J. von Frese, and R. Bro, Robust methods for multivariate data analysis, J. Chemometrics **19**, 549–563 (2005).

[5] L. Petersen and K.H. Esbensen, Representative process sampling for reliable data analysis – a tutorial, J. Chemometrics **19**, 625–647 (2005).

[6] D. Howel, Multivariate data analysis of pollutant profiles: PCB levels across Europe, Chemosphere **67**, 1300–1307 (2007).

Table 4. Probability included in multivariate PDF. Space dimensions d and deviations σ.

	$\sigma = 1$	$\sigma = 3/2$	$\sigma = 2$
$d = 1$	$\mathrm{erf}(\frac{1}{\sqrt{2}})$	$\mathrm{erf}(\frac{3}{2\sqrt{2}})$	$\mathrm{erf}(\sqrt{2})$
$d = 2$	$1 - \frac{1}{\sqrt{e}}$	$1 - \frac{1}{e^{9/8}}$	$1 - \frac{1}{e^2}$
$d = 3$	$-\sqrt{\frac{2}{e\pi}} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{3}{e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{2\sqrt{\frac{2}{\pi}}}{e^2} + \mathrm{erf}(\sqrt{2})$
$d = 4$	$1 - \frac{3}{2\sqrt{e}}$	$1 - \frac{17}{8e^{9/8}}$	$1 - \frac{3}{e^2}$
$d = 5$	$-\frac{4}{3}\sqrt{\frac{2}{e\pi}} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{21}{4e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{14\sqrt{\frac{2}{\pi}}}{3e^2} + \mathrm{erf}(\sqrt{2})$
$d = 6$	$1 - \frac{13}{8\sqrt{e}}$	$1 - \frac{353}{128e^{9/8}}$	$1 - \frac{5}{e^2}$
$d = 7$	$-\frac{7}{5}\sqrt{\frac{2}{e\pi}} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{501}{80e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{34\sqrt{\frac{2}{\pi}}}{5e^2} + \mathrm{erf}(\sqrt{2})$
$d = 8$	$1 - \frac{79}{48\sqrt{e}}$	$1 - \frac{3067}{1024e^{9/8}}$	$1 - \frac{19}{3e^2}$
$d = 9$	$-\frac{148}{105}\sqrt{\frac{2}{e\pi}} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{14757}{2240e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{842\sqrt{\frac{2}{\pi}}}{105e^2} + \mathrm{erf}(\sqrt{2})$
$d = 10$	$1 - \frac{211}{128\sqrt{e}}$	$1 - \frac{100331}{32768e^{9/8}}$	$1 - \frac{7}{e^2}$
$d = 11$	$-\frac{1333}{945}\sqrt{\frac{2}{e\pi}} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{59757}{8960e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{1618\sqrt{\frac{2}{\pi}}}{189e^2} + \mathrm{erf}(\sqrt{2})$
$d = 12$	$1 - \frac{6331}{3840\sqrt{e}}$	$1 - \frac{4032923}{1310720e^{9/8}}$	$1 - \frac{109}{15e^2}$
$d = 13$	$-\frac{4888\sqrt{\frac{2}{e\pi}}}{3465} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{2635869}{394240e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{30346\sqrt{\frac{2}{\pi}}}{3465e^2} + \mathrm{erf}(\sqrt{2})$
$d = 14$	$1 - \frac{75973}{46080\sqrt{e}}$	$1 - \frac{64585817}{20971520e^{9/8}}$	$1 - \frac{331}{45e^2}$
$d = 15$	$-\frac{190633\sqrt{\frac{2}{e\pi}}}{135135} + \mathrm{erf}(\frac{1}{\sqrt{2}})$	$-\frac{137124237}{20500480e^{9/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{3}{2\sqrt{2}})$	$-\frac{1191686\sqrt{\frac{2}{\pi}}}{135135e^2} + \mathrm{erf}(\sqrt{2})$
$d = 16$	$1 - \frac{354541}{215040\sqrt{e}}$	$1 - \frac{3617337193}{1174405120e^{9/8}}$	$1 - \frac{155}{21e^2}$

	$\sigma = 5/2$	$\sigma = 3$
$d = 1$	$\mathrm{erf}(\frac{5}{2\sqrt{2}})$	$\mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 2$	$1 - \frac{1}{e^{25/8}}$	$1 - \frac{1}{e^{9/2}}$
$d = 3$	$-\frac{5}{e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{3\sqrt{\frac{2}{\pi}}}{e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 4$	$1 - \frac{33}{8e^{25/8}}$	$1 - \frac{11}{2e^{9/2}}$
$d = 5$	$-\frac{185}{12e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{12\sqrt{\frac{2}{\pi}}}{e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 6$	$1 - \frac{1153}{128e^{25/8}}$	$1 - \frac{125}{8e^{9/2}}$
$d = 7$	$-\frac{455}{16e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{141\sqrt{\frac{2}{\pi}}}{5e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 8$	$1 - \frac{43297}{3072e^{25/8}}$	$1 - \frac{493}{16e^{9/2}}$
$d = 9$	$-\frac{53845}{1344e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{1716\sqrt{\frac{2}{\pi}}}{35e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 10$	$1 - \frac{592043}{32768e^{25/8}}$	$1 - \frac{6131}{128e^{9/2}}$
$d = 11$	$-\frac{2329045}{48384e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{489\sqrt{\frac{2}{\pi}}}{7e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 12$	$1 - \frac{16162157}{786432e^{25/8}}$	$1 - \frac{80993}{1280e^{9/2}}$
$d = 13$	$-\frac{37414535}{709632e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{33456\sqrt{\frac{2}{\pi}}}{385e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 14$	$1 - \frac{824611661}{37748736e^{25/8}}$	$1 - \frac{383021}{5120e^{9/2}}$
$d = 15$	$-\frac{552800735}{10063872e^{25/8}\sqrt{2\pi}} + \mathrm{erf}(\frac{5}{2\sqrt{2}})$	$-\frac{44907\sqrt{\frac{2}{\pi}}}{455e^{9/2}} + \mathrm{erf}(\frac{3}{\sqrt{2}})$
$d = 16$	$1 - \frac{15799652047}{704643072e^{25/8}}$	$1 - \frac{1178747}{14336e^{9/2}}$

[7] J. Neyman and E.S. Pearson, *Joint Statistical Papers* (Cambridge University Press, Cambridge, 1967) p. 99.

[8] P. Serapinas, P.R. Venskutonis, V. Aninkevičius, Ž. Ežerinskis, A. Galdikas, and V. Juzikienė, Step by step approach to multi-element data analysis in testing the provenance of wines, Food Chem. **107**, 1652–1660 (2008).

[9] M. Roesslein, M. Wolf, B. Wampfler, and W. Wegscheider, A forgotten fact about the standard deviation, Accred. Qual. Assur. **12**, 495–496 (2007).

[10] *Mathematica*, Version 5.2 (Wolfram Research, Inc., Champaign, IL, 2005).

DAUGIAMAČIŲ GAUSO SKIRSTINIŲ ANALIZĖS TAIKYMAS BANDINIŲ KLASIFIKACIJOS KOKYBEI VERTINTI

P. Serapinas, Ž. Ežerinskis, A. Acus

Vilniaus universiteto Teorinės fizikos ir astronomijos institutas, Vilnius, Lietuva

Santrauka

Duomenų klasifikacijos pasikliautinumo lygį lemia skirstinių persiklojimo laipsnis. Yra daug būdų ir patogių priemonių vienmačiams skirstiniams analizuoti, tačiau daugiamačių skirstinių analizė retai taikoma. Straipsnyje pateikiamos lentelės ir būdai daugiamačių Gauso skirstinių įskaitytajai tikimybei, kritinei sričiai, klaidingosioms teigiamosioms ir klaidingosioms neigiamosioms tikimybėms skaičiuoti. Parodoma, kad klaidingosios teigiamosios ir klaidingosios neigiamosios tikimybių suma yra tik labai apytikrė klasifikacijos pasikliautinumo charakteristika. Klaidingoji neigiamoji tikimybė yra lokalizuota skirstinyje, ir jos vaidmuo duomenų klasifikavimui iš esmės priklauso nuo to, kiek konkretūs duomenys yra toli nuo tos srities. Pateikiamas pavyzdys, kaip daugiamačių skirstinių analizė panaudojama vyno bandinių regioninei klasifikacijai pagal spektrometrinius cheminės analizės duomenis.