S1 – A more detailed explanation of how the selected classification methods work.

There are many various approaches to the final female/male classifier design. We applied three of them, which represent traditional and efficient learning techniques of various origins.

Linear Support Vector Machine

The first of them was the so-called Linear Support Vector Machine (SVM) [1], which is a linear classifier that is not based on any statistical theory and supposes the existence of outliers (incorrectly classified samples) during classifier learning. This approach increases the classifier quality during any cross-validation procedure. The SVM is called a linear one because the optimization task can be converted to a Linear Programming task [2].

The Linear SVM is a very simple but efficient classifier which has output response $y = \text{sign}(\sum_{k=0}^{H} w_k\, x_k), x_0 = 1$. Here, every logarithmic ratio x_k has a weight w_k which represents the polarity and magnitude of x_k for $k > 0$. The primary learning conditions are $\sum_{k=0}^{H} w_k x_{i,k} y_i^* \geq 1 - s_i, s_i \geq 0$ for $i = 1, \ldots, m$ where a positive value of s_i indicates the false classification of i-th sample. Using the parameter $C > 0$ we minimize the criterion $F = \sum_{k=1}^{H} |w_k| + C \sum_{i=1}^{m} s_i$ to obtain the optimal weights w_k including w_0. The weight w_0 is called bias and only stabilizes the classifier. The parameter C adjusts a rate between achieving a low training error (a number of outliers) and the small number of involved compounds (H). The high value of the C parameter causes the reduction of learning errors but a large number of involved compounds. On the other hand, a too low value of the C parameter causes a small number of involved compounds but, unfortunately, a high occurrence of the outliers. Finally, a compromise value of the C parameter increases the cross-validation accuracy and sensitivity; therefore, an ability to generalize the classifier to unknown data. It follows from the above that the linear SVM can reduce the initial number of the significant

compounds (their logarithmic ratios) to a new value *H* (in other words, it reduces the original *H* from the WMW test even more).

Ridge Regression with Thresholding

The second classification approach is semi-statistical. Ridge Regression (RR) [3] followed by thresholding was used to obtain another classifier with the same output response as the linear SVM but with another learning strategy driven by residues $r_i = y_i^* - \sum_{k=0}^{H} w_k x_k$. Using the parameter $\mu > 0$, which is a ratio estimate of the data noise variance and prior weight variance, we minimized a regularized sum of squares as $G = \sum_{i=1}^{m} r_i^2 + \mu \sum_{k=1}^{H} w_k^2$ to obtain the optimal weights w_k including bias w_0. Independently of the original statistical meaning of the μ parameter, this parameter is determined experimentally to obtain the maximally possible critical sensitivity of the classifier in the case of a cross-validation procedure. However, the number of involved compounds (*H*) remains unchanged.

Quadratic Discriminant Analysis with Data Whitening

The third method used is based on two pure statistical approaches. First, the Data Whitening [4], as an improvement of Principal Component Analysis (PCA) [5], is used for dimensionality reduction, data decorrelation, and standardization. The resulting classifier is based on statistical characteristics of the individual classes (females, males) and the following a comparison of their probability density functions (PDFs) as is usual in discriminant analysis.

The Quadratic Discriminant Analysis (QDA) [6] supposes a multidimensional normal distribution of descriptors (in our case compounds ratios) in every class. The learning is based on an estimation of the mean values and the corresponding covariance matrices for every class. Using the mean values and the covariance matrices it is possible to calculate adequate class densities (PDFs) and compare them for an unknown sample, which is classified according to the maximal density value. But the QDA is very sensitive to the number of involved compounds (*H*) and the class

degeneracy (hidden correlations). Therefore, we applied PCA to obtain a number D of uncorrelated components $PCA_1, PCA_2, \dots, PCA_D$ with adequate eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_D$. Then the Data Whitening provides D new descriptors $c_k = PCA_k/\lambda_k^{1/2}$ which enable more efficient QDA.

References

1. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science+Business Media: New York, 2008.
2. Sierksma, G.; Zwols, Y. *Linear and Integer Optimization: Theory and Practice, third edition*; CRC Press: Boca Raton, London, New York, 2015.
3. Golub, G.; Heath, M.; Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **1979**, *21*, 215–223, doi:10.1080/00401706.1979.10489751.
4. Kessy, A.; Lewin, A.; Strimmer, K. Optimal Whitening and Decorrelation. *Am. Stat.* **2018**, *72*, 309-314, doi:10.1080/00031305.2016.1277159.
5. Jolliffe, I.T. *Principal Component Analysis*; Springer: New York, 2002.
6. Etemad, K.; Chellappa, R. Discriminant analysis for recognition of human face images. *J. Opt. Soc. Am. A* **1997**, *14*, 1727-1733, doi:10.1007/BFb0015988.