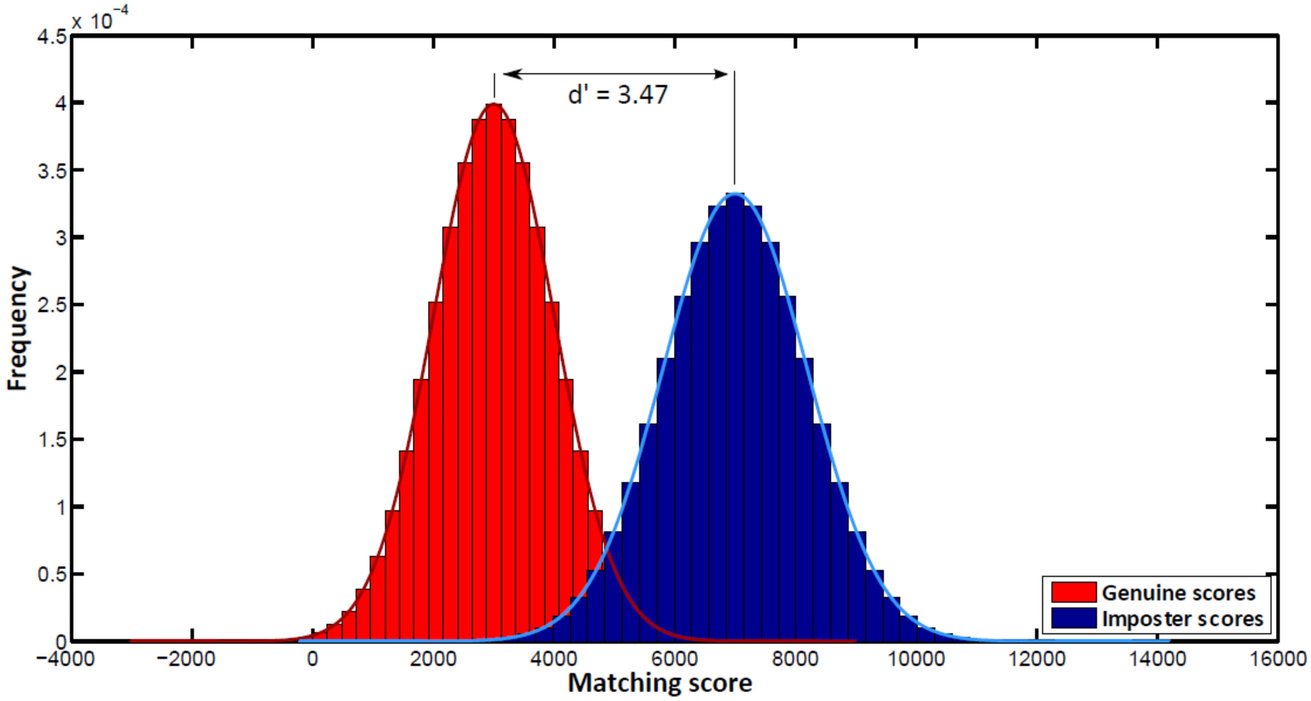
Worksheet 4 (StepScan)

Topic 1: Feature Selection – Maximum Relevance

1) The book entitled "Handbook of Biometrics", Chapter 1, [1] is the main resource on this topic.
2) Two single-valued measures, other than EER, in biometrics field are d-prime (d') and F-ratio.
   - The d-prime (d') measures the separation between the means of the genuine and imposter probability distributions in standard deviation units and is defined as,

$$d' = \frac{\sqrt{2}\,\left|\mu_{genuine} - \mu_{imposter}\right|}{\sqrt{\sigma^2_{genuine} + \sigma^2_{imposter}}}$$

where the μ's and σ's are the means and standard deviations, respectively, of the genuine and imposter distributions. A higher d-prime value indicates better performance.



   - F-ratio

$$\text{F-ratio} = \frac{\mu_{genuine} - \mu_{imposter}}{\sigma_{genuine} + \sigma_{imposter}}$$

3) Based on these measures, we can perform the simplest feature selection, i.e., filter-based univariate feature selection. This approach examines each feature individually to determine the strength of the relationship of the feature with the response variable. We can further sequentially feed the top ranked features into a classifier. All the COP and PCA features can be used as an initial feature set for feature selection.

Topic 2: Feature Selection – Minimum Redundancy

1) The paper entitled “Measures of Postural Steadiness: Differences Between Healthy Young and Elderly Adults” [2] is the main resource on this topic.
2) It is frequently observed that simply combining a “very effective” feature with another “very effective” feature often does not form a better feature set. One reason is that these two features could be highly correlated. This raises the issue of “redundancy” of feature set. The fundamental problem with redundancy is that the feature set is not a comprehensive representation of the characteristics of the target classes.
3) The most known measure for redundancy is “correlation coefficient”. As used in [2], the authors can combine several COP features into groups using the criterion: the Pearson correlation coefficients $r \geq 0.90$. We can extend from this study by including the PCA features of COP time series (as implemented in Worksheet 3) and other features you have implemented so far.
4) To illustrate their relationships, we can use a similar plot like in Fig. 6 in [3].
5) Th paper [3] also applied several simple feature ranking methods (i.e., maximum relevance) such as Chi-squared based, mutual information (MI)-based, recursive feature elimination, and L1-norm penalty-based feature selection.

   It should be noted that MI can be used not only for the maximum relevance criteria but also for the minimum redundancy criteria. MI can be used not only for discrete variables but also for continuous variables. While correlation analysis provides a quantitative means of measuring the strength of a linear relationship between two vectors of data. MI is essentially the measure of how much “knowledge” one can gain of a certain variable by knowing the value of another variable. More details about the MI for feature selection can be found in [4].

Topic 3: Feature Selection – Minimum Redundancy Maximum Relevance (mRMR)

1) The paper entitled “Minimum Redundancy Feature Selection from Microarray Gene Expression Data” [4] is the main resource on this topic.
2) Based on this paper, we can use the same concept and perform the mRMR by maximizing the d-prime (d') values and minimizing the correlations (*r*) between features.
3) There are two ways proposed in [4] to combine relevance and redundancy:
   - one is the difference criterion, i.e., max(d'- *r*); and
   - another is the quotient criterion, i.e., max(d'/ *r*).
4) Using mRMR, we can rank the COP, PCA, and other features, and then sequentially feed the top ranked features into classifiers. The cross-validation procedures and key performance metrics (from Worksheet 3) can be performed.

[1] Jain & Ross, Introduction to Biometrics, in Jain, Flynn, & Ross (Eds.). *Handbook of Biometrics* (pp. 1-22). Springer.

[2] Prieto et al., “Measures of Postural Steadiness: Differences Between Healthy Young and Elderly Adults”, IEEE Trans Biomed Eng, 43(9), pp. 956-966, Sept. 1996.

[3] Li et al., “Exploring EEG Features in Cross-Subject Emotion Recognition,” Frontiers in Neuroscience, 12(162), Mar. 2018.

[4] Ding & Peng, “Minimum Redundancy Feature Selection from Microarray Gene Expression Data,” Journal of Bioinformatics and Computational Biology, 3(2), pp. 185-205, 2005.

Checkpoint

Please find the best feature set from all the features implemented in Worksheet 1 and 3 using the mRMR feature selection method proposed in Topic 3 with their performance metrics (FAR, FRR, and accuracy).