variables $\rho_{k_j}(Z, S_j)$. Thus, τ represents a generalization based on subsample order statistics [35] of the Wilcoxon-Mann-Whitney statistic. The statistic τ is a U-statistic and can be shown to have an asymptotically normal distribution [35]. However, for small sample sizes—and especially for unbalanced designs or unequal subset sizes k_j, in which case F_τ is skewed—inference based on the asymptotic distribution is inappropriate; the exact distribution (or a small-sample approximation thereof—see Section 3.4) is necessary. A recurrence for the exact distribution F_τ under the null hypothesis H_0 is given in [25] for $k_0 = k_1 = 1$. Classification based on the generalized Wilcoxon-Mann-Whitney statistic τ given in (4) is particularly relevant to interpoint distance-based nonparametric discriminant analysis [25]. Different choices for the parameters yield desirable power characteristics against different alternatives [35]. The issue of *adaptively* selecting the parameters r_0, r_1, k_0, k_1 will be addressed in Section 3.6.

3.2 Extension to $K > 2$ Classes

The proposed classifier has been developed for the simple two-class problem. The extension to $K > 2$ classes can be addressed in two ways. The statistic τ can be generalized to the K sample case and a recurrence for the joint distribution $F_{\tau_1,\ldots,\tau_K}$ is available [25]. Another approach is that of addressing the K class problem through consideration of a collection of two class subproblems [9], [14].

3.3 Relationship to Machine Learning

From a machine learning perspective, the classifier g based on τ is a classic example of "classification by ensemble" [11], [8], [20]. The statistic τ represents the most fundamental approach to constructing ensembles of classifiers. In a manner similar to bagging [2], subsamples S_0 and S_1 are taken (without replacement) from the training database D_n, and $I_{\{\rho_{k_1}(Z,S_1)\leq\rho_{k_0}(Z,S_0)\}}$ represents a classifier for Z based on these subsamples. Thus, all possible subsample classifiers obtained are then combined in τ via the simplest possible method for combining individual classification decisions from an ensemble of classifiers: an unweighted vote.

Observing a value of $\tau > 1/2$ means that a majority of the subclassifiers $I_{\{\rho_{k_1}(Z,S_1)\leq\rho_{k_0}(Z,S_0)\}}$ in the ensemble favors class 1. A more appropriate classification criterion is the event $\{I_{\{F_\tau(\tau(Z)>1/2)\}} = 1\}$. This event represents evidence in favor of class 1 vs. class 0 in that the vote count is *probabilistically large* (under H_0). Thus, the classifier proposed in (6), based on the unweighted ensemble vote τ, accounts for the character of the distribution F_τ. As noted above, this distribution is strongly influenced by unequal sample sizes ($n_0 \neq n_1$), unequal subset sizes ($r_0 \neq r_1$), and/or unequal rank choices ($k_0 \neq k_1$). In effect, (6) implicitly weights the ensemble votes in a probabilistically appropriate way.

The combination methodology employed here is straightforward and allows for analysis via mathematical statistics. However, more elaborate combination methods such as those presented in [20] may be beneficial in terms of classification performance. In particular, using fewer carefully selected subsets so that the ensemble does not employ so many classifiers is worthy of consideration, but makes the analysis of the statistic significantly more difficult.

3.4 Computational Considerations

For large sample sizes such as those encountered in the olfactory classification task, the calculation of the observed value of τ via (4) and of the distribution F_τ via the available recurrence, are computationally intensive exercises. For the example, results presented in Sections 4.2 and 4.3, the following estimators are used.

Let S_u be a uniform random sample of size u from the collection of subset pairs Δ. The estimator for τ is given by

$$\hat{\tau} = (1/u) \sum_{(S_0,S_1)\in S_u} I_{\{\rho_{k_1}(Z,S_1)\leq\rho_{k_0}(Z,S_0)\}}. \tag{9}$$

The estimator standard deviation $\sigma_{\hat{\tau}} \leq 1/(2\sqrt{u})$, indicating how large u must be taken (and, consequently, the required computational demand) in order to have an estimator with some prescribed accuracy. Equation (9) can be employed using either observed data or sequences generated under the null hypothesis. To obtain the quantile estimator for F_τ, we consider a collection $\{\hat{\tau}_1, \ldots, \hat{\tau}_v\}$ of such estimators taken under H_0. Then,

$$\hat{F}_\tau(t) = (1/v) \sum_{i=1}^{v} I_{\{\hat{\tau}_i\leq t\}} \tag{10}$$

with an accuracy dependent on u, v, and t.

3.5 Classifier Consistency

Our hypothesis testing approach to the two-class decision problem ([1], p. 183) allows us to address the issue of "classifier consistency" from the standpoint of consistent tests of hypotheses. Note that the null hypothesis $H_0 : F_0 = F_1$ implies $F_{\rho(z,X_i|Y_i=0)} = F_{\rho(z,X_i|Y_i=1)}$ for any fixed observation z. For simplicity, consider as alternative hypotheses *stochastic ordering*; the random variable $\rho(z, X_i|Y_i = 0)$ is defined to be stochastically smaller than $\rho(z, X_i|Y_i = 1)$, denoted as

$$\rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1),$$

if

$$F_{\rho(z,X_i|Y_i=0)}(x) \geq F_{\rho(z,X_i|Y_i=1)}(x)$$

for every x, with strict inequality for at least one x. From Fig. 3, we see that, for the left panel in which the test observation (TClf7712) is TCE-present (class 1), we have $\rho(z, X_i|Y_i = 0) >^{st} \rho(z, X_i|Y_i = 1)$. For the TCE-absent observation (Kero0203) depicted in the right panel of Fig. 3, $\rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1)$ as desired for a class 0 observation. For situations such as these, we have the following result.

Theorem. *For a fixed observation z, the classifier g given in (6) based on the statistic τ given in (4) is consistent against alternatives of stochastic ordering of the class-conditional interpoint distance distributions.*

Proof. For fixed values of r_0, r_1, k_0, k_1, as $n_0, n_1 \to \infty$ with $n_0/(n_0 + n_1) \to \zeta \in (0, 1)$, $\tau(z)$ is asymptotically normal under H_0 and $lim\ F_\tau^{-1}(1/2) = T\ a.s.$, where T is given by (8). Under $H_A : \rho(z, X_i|Y_i = 0) >^{st} \rho(z, X_i|Y_i = 1)$ (see, for example, Fig. 3, left panel) $lim\ \tau(z) > T\ a.s.$ and $lim\ I_{\{\tau(z)>F_\tau^{-1}(1/2)\}} = 1\ a.s.$, while under $H_A : \rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1)$ (see, for example, Fig. 3, right panel)