Parametric and nonparametric two-sample tests for feature screening in class comparison: a simulation study

Elena Landoni [(1)], *Federico Ambrogi* [(2)], *Luigi Mariani* [(1)], *Rosalba Miceli* [(1)]

(1) Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
(2) University of Milan, Milan, Italy

CORRESPONDING AUTHOR: *Elena Landoni, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy email: elena.landoni@istitutotumori.mi.it*

***DOI:* 10.2427/11808**
Accepted on May 3, 2016

ABSTRACT

Background: The identification of a location-, scale- and shape-sensitive test to detect differentially expressed features between two comparison groups represents a key point in high dimensional studies. The most commonly used tests refer to differences in location, but general distributional discrepancies might be important to reveal differential biological processes.
Methods: A simulation study was conducted to compare the performance of a set of two-sample tests, *i.e.* Student's t, Welch's t, Wilcoxon-Mann-Whitney (WMW), Podgor-Gastwirth PG2, Cucconi, Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), Anderson-Darling (AD) and Zhang tests (Z_K, Z_C and Z_A) which were investigated under different distributional patterns. We applied the same tests to a real data example.
Results: AD, CvM, Z_A and Z_C tests proved to be the most sensitive tests in mixture distribution patterns, while still maintaining a high power in normal distribution patterns. At best, the AD test showed a power loss of ~ 2% in the comparison of two normal distributions, but a gain of ~ 32% with mixture distributions with respect to the parametric tests. Accordingly, the AD test detected the greatest number of differentially expressed features in the real data application.
Conclusion: The tests for the general two-sample problem introduce a more general concept of 'differential expression', thus overcoming the limitations of the other tests restricted to specific moments of the feature distributions. In particular, the AD test should be considered as a powerful alternative to the parametric tests for feature screening in order to keep as many discriminative features as possible for the class prediction analysis.

Key words: high-dimensional data; class comparison; location-scale problem; general two-sample problem; mixtures.

INTRODUCTION

Parametric and nonparametric two-sample tests are applied to a large number of high-dimensional continuous data for explorative studies in order to detect differentially expressed (DE) features (genes, miRNAs, metabolites) between different biomedical conditions, such as diseased (cases) and healthy subjects (controls). In particular, the tests are exploited as univariable feature ranking methods in class comparison [1], as well as a preliminary step - feature screening - in class prediction [2]. Such a screening is intended to identify promising features to be possibly included in a multivariable model, as the predictor or classifier, which aims at accurately predicting the class membership of a new sample based on a combination of expression levels of the selected features.

The most commonly used tests are the t test and the nonparametric WMW test, which refer to differences in terms of location and therefore are classified as location tests. However, feature distributions in the comparison classes may differ according to other aspects such as scale or, more generally, shape. One could test for location or scale changes (location-scale problem) or look for any changes in location, scale or shape (general two-sample problem) [3]. Even small signals of general differences between the two classes could reveal discriminative features that should not be filtered out in the first phases of bioinformatics analyses, but further investigated in the following step of class prediction. Moreover, the asymptotic normality of the t-test statistic is often not fulfilled when dealing with some types of genomic data, mainly due to the small sample size [4], producing skewed, heavy-tailed or multimodal distributions of expression values.

In presence of such distributions, nonparametric alternatives to location tests, e.g. the Kolmogorov-Smirnov filter [5], could be more sensitive in feature screening, thus leading to a small number of false negative results.

In the field of high dimensional data, feature screening should not be tailored on specific distributional characteristics but rather be a flexible procedure, i.e. able to detect general differences between feature distributions under different patterns. Thus, a desirable test should prove to be robust in terms of Type I error control and powerful in a wide family of distributional patterns, even if not being the best one in every single situation.

The goal of this work was to compare via simulations different tests for class comparison of continuous data to draw suggestions for possible improvements with respect to the tests most commonly used in high-throughput data analysis. In particular, we conducted an extensive simulation study with sample sizes as small as those frequently encountered in the high dimensional data context [6] and including non normal distributions; a wide set of parametric and nonparametric tests for two class comparison was investigated according to size (i.e. type I error rate) and power, with the aim of possibly identifying a test to be used in the screening phase of high dimensional studies. Student's t and Welch's t tests [7] were used even if their assumptions are violated as they are the standard reference in practical applications. We investigated a series of nonparametric tests considering different alternatives versus the null hypothesis of equality between the Cumulative Distribution Functions (CDFs). Being aware that when the parametric tests assumptions are violated their power is deflated, our aim was to assess possible nonparametric alternatives and comparatively draw indications of possible improvements over the parametric tests. In particular, we implemented the following nonparametric tests:

- the Wilcoxon-Mann-Whitney (WMW) test [8], detecting shifts in location between the CDFs;
- two tests for the location-scale problem, i.e. the PG2 Podgor-Gastwirth (PG2) [9] and the Cucconi test [10][11]; the PG2 test has been recognised as the most powerful among the PG efficiency robust tests, while the Cucconi test represents the simplest and best performing alternative to PG2 [11];
- three chi-squared statistic-based tests, i.e. the Kolmogorov-Smirnov (KS) [12], the Cramer-von Mises (CvM) [13] and the Anderson-Darling (AD) [14] test; the KS test refers to the CDF maximum difference, the CvM test considers differences over the entire CDF range, while the AD test takes into account global CDF differences, granting more importance to the observations in the tails; the latter characteristic makes the AD test valuable when one is interested in finding also signals that are only present in a subset of patients;
- the Zhang Z_K, Z_C and Z_A tests, which are 'likelihood-ratio' based analogs of the 'traditional' KS, CvM and AD tests, respectively [3].

See the Appendix for further details on the considered tests.

As regards the simulation study, we chose to mimic the irregular pattern of the feature distributions of the two samples by sampling from mixtures of two normal distributions (NM), which should reproduce the coexisting presence of heterogeneous subpopulations underlying data, by varying the mixture parameters. We did not provide any adjustment for multiple testing. Indeed, the adjustment procedure after the tests itself is not the focus of the paper.

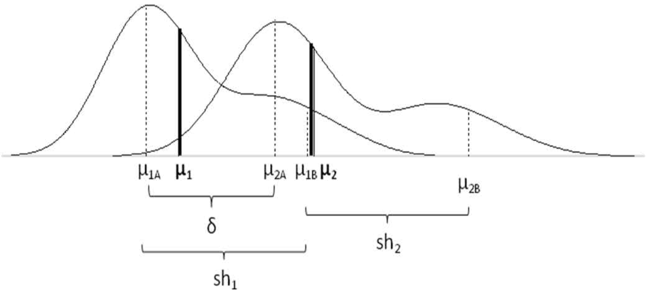
METHODS

A simulation study was conducted in order to compare the performance - in terms of size and power - of the tests described in the Appendix under different distributional patterns.

Following Burton et al. [15], data were simulated with resemblance to real continuous high dimensional data, replicating their irregular distributional patterns by sampling from mixtures of two normal distributions (NM) (Figure 1).

Let μ_{iA} and μ_{iB} be the means of the two components A and B of the mixture in the population i (i = 1,2),

FIGURE 1. Example figure of two normal mixture distributions setting adopted into the simulation.
μ_{iA} and μ_{iB} are the means of the two components A and B of the mixtures in the two samples i (i=1,2) with shifts shi = $\mu_{iB} - \mu_{iA}$ (i=1,2), $\delta = \mu_{2A} - \mu_{1A}$ is the difference between the two mixture first component means and λ_i are the two mixture weights of the A components, being their complement the mixture weights of the respective B components.



with μ_{iB} - μ_{iA}=sh_i (shift), σ^2_{iA} and σ^2_{iB} be the component variances, $\mu_i = \lambda_i\mu_{iA} + (1 - \lambda_i)\mu_{iB}$ the overall mixture mean, $\sigma_i^2= \lambda_i\ [\sigma_{iA}^2+(\ \mu_{iA}- \mu_i)] +(1-\lambda_i)[\sigma_{iB}^2+(\ \mu_{iB}- \mu_i)]$ the overall mixture variance and λi the mixture weight, which is the probability associated with the first component of the mixture.

Finally, let $\delta = \mu_{2A} - \mu_{1A}$ be the difference between the first component means in the two populations.

Three main cases were simulated: A. two normal distributions; B. one normal and one mixture distribution; C. two mixture distributions; equal shifts $sh_1 = sh_2$ were also considered. The parameters δ, sh_1, sh_2 were properly tuned over fixed ranges in order to simulate the different conditions under H_0 and H_1. We considered four mixture weights λ Є {0.80; 0.95; 0.20; 0.05} and three small sample size settings, one balanced (m=20 vs n=20) and two unbalanced (m=20 vs n=40 and m=40 vs n=20) (Table 1).

We fixed $\sigma^2_{1A} = \sigma^2_{1B} = \sigma^2_{2A} = \sigma^2_{2B} = 1$ and distinguished six different patterns, summarised in Table 2 and graphically represented in the Supplementary material (Figures S1.1-S1.4) for each combination of sample size settings and mixture weights.

We chose to consider a nominal significance level α = 0.05 and to perform B = 10000 simulations so as to obtain precise estimates derived via simulation. As an indicator of the simulation error we chose the standard error SE(p), with p indicating the nominal coverage probability [15]. Finally, we calculated the relative frequencies of H_0 rejection of the tests; under H_0 such frequencies approximate the fixed nominal significance level α, while under H_1 they correspond to the empirical power of the tests. It was possible to simulate the null hypothesis patterns only when comparing two normal distributions with equal means (pattern I) or two perfectly overlapping mixture distributions, i.e. those with equal shifts (pattern IV). The robustness of the tests under H_0 was evaluated according to the indications given by both Conover [16] and Marozzi [17]: a test is considered robust if its Maximum Estimated Significance Level (MESL), i.e. its maximum relative frequency of H_0 rejection under H_0 (Table 2 - H_0 patterns) does not exceed a given threshold, typically 2α or 1.5α to be more restrictive. As regards the nonparametric tests, exact p-values have been computed for the KS test, while for the WMW, PG2, Cucconi, CvM and AD tests we report the asymptotic p-values. For the CvM test we used the p-values tabulated by Anderson et al. [13], since it has been shown that under H_0 the two-sample statistic has the same limiting distribution as that of the one-sample statistic [18]. Moreover, using the Burr's formula, we can obtain the p-values corresponding to all the possible values of the CvM statistic and not limited to the tabulated ones; it is reported that such an approximation works to the fifth decimal place for values of the statistic between 0.42 and 2.2 and we empirically verified that, for values of the statistic greater than 0.10308, the formula approximates up to the second decimal place the p-values tabulated by Anderson et al. [13]. Moreover, values of the statistic below 0.10308 correspond to p-values higher than 0.57, which are of no interest here since they indicate

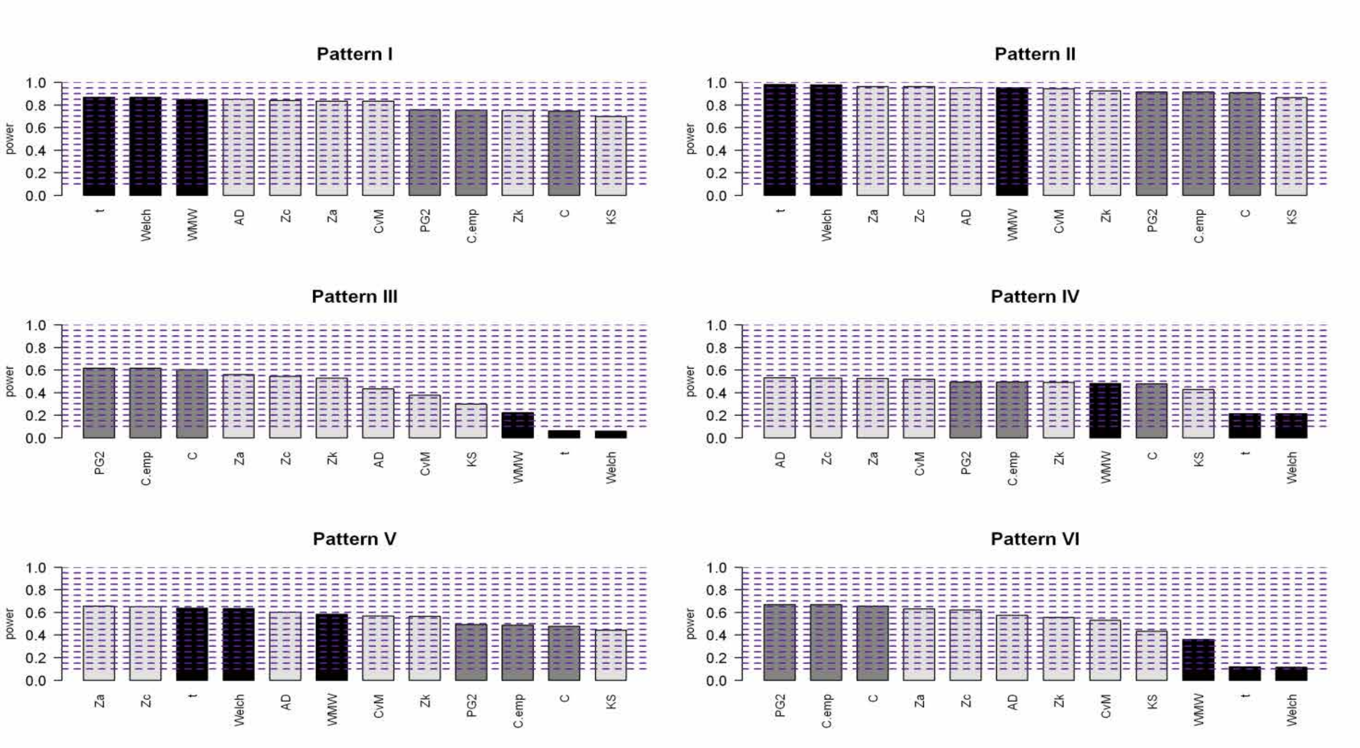
TABLE 1. Ranges of values for the parameters λ, δ, sh and (m,n) in the simulation study.

Parameter	Values		
λ	{0.8; 0.95; 0.2; 0.05}		
δ	{0; 1}		
sh	{0; 3; 6}		
(m;n)	(20; 20)	(20; 40)	(40; 20)

TABLE 2. The six patterns considered in the simulation study under null (H_0) and alternative (H_1) hypotheses: I. two normals; II. normal vs mixture (sh = 6); III. mixture (sh = 6) vs normal; IV. two mixtures with equal shifts (sh_1 = sh_2 = 6); V. two mixtures with different shifts (sh_1 = 3 < sh_2 = 6); VI. two mixtures with different shifts (sh_1 = 6 > sh_2 = 3). Abbreviations: N = Normal distribution; NM = Mixture of Normal distributions.

H_0 patterns (δ = 0)			
Pattern	**Label**	**sh_1**	**sh_2**
I	N vs N	0	0
IV	NM vs NM (sh_1 = sh_2)	3	3
IV	NM vs NM (sh_1 = sh_2)	6	6
H_1 patterns (δ = 1)			
I	N vs N	0	0
II	N vs NM	0	6
III	NM vs N	6	0
IV	NM vs NM (sh_1 = sh_2)	6	6
V	NM vs NM (sh_1 < sh_2)	3	6
VI	NM vs NM (sh_1 > sh_2)	6	3

FIGURE 2.



Barplots of power of the considered two-sample tests for the six selected scenarios with δ=1, λ=0.80 and (m,n) = (20,20): I. two normals; II. normal vs mixture (sh=6); III. mixture (sh=6) vs normal; IV. two mixtures with equal shifts (sh_1=sh_2=6); V. two mixtures with different shifts (sh_1=3 < sh_2=6); VI. two mixtures with different shifts (sh_1=6 > sh_2=3). The tests are sorted in descending order according to the power. The different colours indicate the three types of tests (location tests in black, tests for the location-scale problem in grey and tests for the general two-sample problem in light ray).

TABLE 3. Power estimates with α=0.05, δ=1, λ=0.95 and (m,n) = (20,20). a. Location tests: t = Student's t test; Welch = Welch's t test; WMW = Wilcoxon-Mann-Whitney test. b. Location-scale tests: PG2 = Podgor-Gastwirth PG2 test; C = Cucconi test (asymptotic version); C.emp = Cucconi test (empirical version). c. Tests for the general two-sample problem: KS = Kolmogorov-Smirnov test; CvM = Cramer-von Mises test; AD = Anderson-Darling test; Z_K = Zhang Z_K test; Z_C = Zhang Z_C test; Z_A = Zhang Z_A test.

Pattern	t	Welch	WMW	PG2	C	C.emp	KS	CvM	AD	Z_K	Z_C	Z_A
I	0.869	0.868	0.849	0.757	0.745	0.755	0.698	0.834	0.848	0.752	0.839	0.834
II	0.904	0.901	0.883	0.804	0.794	0.803	0.745	0.869	0.880	0.799	0.879	0.877
III	0.419	0.417	0.697	0.638	0.623	0.636	0.589	0.720	0.723	0.619	0.690	0.680
IV	0.499	0.497	0.750	0.677	0.661	0.674	0.634	0.761	0.763	0.663	0.729	0.717
V	0.729	0.726	0.770	0.682	0.666	0.680	0.636	0.768	0.772	0.669	0.745	0.737
VI	0.485	0.483	0.743	0.680	0.663	0.678	0.634	0.761	0.764	0.663	0.733	0.722

non significant features. Finally, for the three Zhang tests, as well as for the Cucconi test as an alternative to the asymptotic version, we used the Monte Carlo approach to find the corresponding approximate empirical p-values (size = 2000 simulations).

The *rnormmix* function from the *mixtools* package was used to simulate the mixtures of univariate normal distributions. Because of the computational burden of the simulation, parallel programming (using the two packages *doParallel* and *doRNG*) was implemented in order to perform simultaneous and reproducible computations.

RESULTS

The simulation study

A complete report of the simulation results is shown in the Supplementary material (Tables S2.1-S2.4). Given 10000 simulations, the chosen 5% significance level provided a SE equal to 0.2%, which is the simulation error for the H_0 patterns; as regards the power, in the worst situation when it is equal to 50%, the SE would be equal to 0.5%. Therefore, we got a precision of simulation estimates up to the third decimal. All the tests resulted robust according to both Conover and Marozzi indications, since relative frequencies of H_0 rejection under H_0 (patterns I and IV) were less than 0.1 and 0.075 respectively; the maximum frequency was 0.063 for the Z_K test, when $\lambda = 0.95$ and (m,n) = (20,20); the KS test was the only one with frequencies lower than 0.05 in most situations, reaching a minimum of 0.034, when $\lambda = 0.80$ and (m,n) = (20,20).

As regards the power, we did not find a clear winner for all the patterns; in general, within the same category (location, or location-scale, or general two-sample problem) the tests shared similar power for all the considered patterns, except for KS and Z_K which showed to be very conservative tests. Moreover, as expected, the tests for the general two-sample problem were generally more sensitive than those for the location-scale problem, especially when $\lambda = 0.95$. As an example to visualise the overall advantage brought by the general two-sample problem tests, we report the results in terms of power under the patterns I-VI, having fixed $\delta = 1$ and m = 20 vs n = 20 and with $\lambda = 0.80$ (Figure 2) and $\lambda = 0.95$ (Table 3).

With both mixture weights, the two parametric location tests (Student's t and Welch's t) headed the power ranking in case of two normal distributions (pattern I) or when the second distribution was a mixture (pattern II); in the latter case their ability lied in detecting the observations in the tail of the mixture. However, when the two ECDFs were crossing, i.e. when the two distributions overlapped at certain points, their power collapsed (see Figures S1.1-S1.4 in the Supplementary material, patterns III, IV, VI). The location tests (Student's t, Welch's t and WMW) generally showed a high power for pattern V, where the two ECDFs of the mixtures appeared mostly separated, and thus the differences between the two samples were mainly in terms of location. The nonparametric location test, i.e. the WMW, was more powerful than the parametric tests in the patterns involving two mixture distributions, except for pattern V and $\lambda = 0.80$ (Tables S2.1.6, S2.1.7). However, it did not emerge as the best alternative to the parametric tests in presence of two mixture distributions, especially when $\lambda = 0.80$, where the advantage of the location-scale and general two-sample tests was more evident.

The location-scale tests (PG2 and Cucconi tests) showed the highest power in the patterns III and VI with $\lambda = 0.80$ (Tables S2.1.6, S2.1.8), corresponding to situations of scale differences being one distribution in the middle of the other one with ECDFs overlapping for the most part. Such tests seem to be particularly able to detect the differences in the peaks of the compared distributions. In general, the PG2 test was more powerful than the Cucconi test (both asymptotic and empirical versions) and the approximate empirical version of the Cucconi test was always more powerful than the asymptotic version.

In the patterns IV and V with $\lambda = 0.80$ and the patterns III-VI with $\lambda = 0.95$ the tests for the general two-sample problem were generally the most powerful ones. The most liberal tests were the AD test, its analog Z_A test, the CvM test and its analog Z_C test; in particular, for mixtures with equal shifts (pattern IV) the AD test was the most sensitive one, together with the Z_C and CvM tests. Moreover, when the normality assumption was fulfilled, the AD, CvM, Z_A and Z_C tests had a limited power loss compared to that of the other nonparametric tests, while being very powerful in detecting any difference between the two samples in the remaining patterns. For example, the application of the AD test implied a loss in power of ~ 2% in pattern I, but a gain of ~ 32% in pattern IV with respect to the parametric tests (Table S2.1.6). It is worth to notice that, in spite of being tests for the general two-sample problem, the KS and its analog Z_K proved to be very conservative, showing low power in all the simulated patterns.

With small weights to the first component of the mixtures ($\lambda = 0.20$ and $\lambda = 0.05$), we obtained similar results, being the AD, CvM, Z_A and Z_C the most sensitive tests in the III-VI patterns, while still maintaining a high power in the I and II patterns. Compared to $\lambda = 0.80$ and $\lambda = 0.95$, the power was higher for all the tests and, for patterns II, III and V, it often reached the 100%. Indeed, in these cases the mixture distribution density is concentrated at its second component, thus yielding well-separated distributions and easily detectable differences.

Regarding the unbalanced sample size settings (m=20 vs n=40 and m=40 vs n=20), the dominance of the general two-sample tests remained evident, except for the pattern V with m=20 vs n=40 and $\lambda = 0.80$, where the Welch's t test resulted as the most powerful test, even if the difference in power was small (~ 4%) with respect to the Z_A test (Table S2.1.7). An explanation could be the presence of a large tail of the distribution of the second sample, corresponding to an evident difference between the two means (Figure S1.1, m=20 vs n=40, pattern V).

Van't Veer data analysis

We applied the considered tests to a real dataset included in the R *Bioconductor* package *breastCancerNKI*, containing gene expression data as published in Van't Veer et al. [19] and Van de Vijver et al. [20] (24481 genes/features evaluated in 337 samples). We defined two classes according to the Estrogen Receptor (ER) status and matched 33 ER positive with 33 ER negative individuals. After a filtering at 100% level, 19264 genes remained.

We expected that the most conservative tests would detect less DE genes with respect to more liberal tests and, accordingly, the AD test identified the greatest number of DE features.

Details on the example at issue are reported in the Supplementary material, subparagraph S3.

CONCLUSION

Two-sample tests for class comparison are often used in bioinformatics and medicine for exploratory purposes, i.e. to detect DE features between different biomedical conditions.

The most commonly used are the location tests, such as the parametric t test and its variations, together with nonparametric tests such as the WMW test; however, they are able to detect only shifts in distributions and not to identify any other difference in scale or shape. These tests are often used in high dimensional studies for screening of features able to distinguish between two conditions. One drawback in applying the above tests is that the expression values may exhibit departures from normality and features could be DE in other aspects rather than location only. Indeed, they can miss features which are characterised by more general and subtle distributional discrepancies. These different signals might hide differential biological processes and have to be preserved in order to be further explored by including them in a multivariable model for class prediction.

We set a simulation study to evaluate the performance of different location tests (Student's t, Welch's t, WMW), tests for the location-scale problem (PG2 and Cucconi) and tests for the general two-sample problem (KS, CvM, AD, Z_K, Z_C, Z_A), by modelling the irregular signals by means of mixtures of two normal distributions. Although Z_C in particular was suggested as the best one among the three Z_K, Z_C and Z_A, we assessed the performance of all Zhang tests, since we wanted a complete comparison with the corresponding traditional tests (i.e. KS, CvM and AD). We did not find a clear winner for all considered distributional patterns among the tests proposed as an alternative to the most used ones. However, the simulation study and the application to Van't Veer's data showed that the tests for the general two-sample problem tend to save a greater number of DE features, with possible gain in power with respect to the location and location-scale tests. Location tests consider DE a feature with almost symmetric distributions in the two compared samples, while location-scale tests are able to detect also differences in terms of peak magnitude. The tests for the general two-sample problem go one step further introducing a more general concept of 'differential expression', thus overcoming the limitations of the above mentioned tests restricted to specific moments of the feature distributions. Specifically, the AD, CvM and their analogs Z_A and Z_C tests should be preferred since their power was very similar to that of the more efficient parametric tests when the normality assumption was fulfilled, while in all the other situations they still resulted very powerful in detecting differences between the two samples. The AD test in particular proved to be very sensitive in most of the simulated patterns; accordingly, by using the Van't Veer dataset, which represents a context similar to the simulated ones, the AD test resulted as the most 'saving-features' test to be further investigated.

In conclusion, the AD test should be considered as a powerful alternative to the parametric tests for feature screening in order to keep as many discriminative features as possible for the subsequent class prediction analysis.

APPENDIX

The investigated two-sample tests

We can classify the investigated tests into three categories:

I. location tests (Student's t, Welch's t, WMW);

II. tests for the location-scale problem (PG2, Cucconi);

III. tests for the general two-sample problem (KS, CvM, AD, Z_K, Z_C, Z_A).

All the tests were implemented using R software (http://www.r-project.org/). In the following, let x_1,... x_m and y_1,...y_n be the observations drawn from two independent random variables X and Y with continuous CDFs F and G, respectively. Also, let m + n = N.

1. Tests for the location problem

The location tests assess whether the centre of the data is the same for the two distributions.

Student's and Welch's t tests

Let μ_1 and μ_2 be the means and σ^2_1 and σ^2_2 be the variances of the random variables X and Y. The null and alternative hypotheses of the considered parametric tests are:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

The t test is defined as

$$t = \frac{X - Y}{s\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

and assumes that $\sigma^2_1 = \sigma^2_2$, thus estimating the pooled sample variance as

$$s^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(X_i - \underline{X})^2 + \sum_{j=1}^{n}(Y_i - \underline{Y})^2\right)$$

The Welch's t variant (Satterthwaite - Welch adjustment) assumes $\sigma^2_1 \neq \sigma^2_2$ and it is defined as

$$Welch = \frac{X - Y}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \sim t_v$$

$$v = \frac{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1}+\frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}}$$

Both the tests were performed using the t.test function included in the stats package.

Wilcoxon-Mann-Whitney (WMW) test

The WMW test considers shifts in location between the two CDFs:

$$H_0 : F(x) = G(x) \quad \forall x$$
$$H_1 : G(x) = F(x - \varDelta) \quad (\varDelta > 0 \; or \; \varDelta < 0)$$

Let R_1 and R_2 be the sums of the ranks for the observations in the two groups, $U = \min(U_1,U_2)$, where $U_1 = R_1 - [m(m+1)/2]$ and $U_2 = R_2 - [n(n+1)/2]$. The WMW test rejects the null hypothesis if there is a prevalence of high ranks (or low ranks) in one group.

For $m \geq 8$ and $n \geq 8$, a normal approximation is used to calculate the standardised WMW test statistic:

$$WMW = \frac{U - \mu_u}{\sigma_u} \qquad \mu_u = \frac{m\,n}{2} \quad \sigma_u = \sqrt{\frac{m\,n\,(N+1)}{12}}$$

The test was performed using the wilcox.test function included in the stats package.

2. Tests for the location-scale problem

The tests for the location-scale problem assess whether both the samples come from the same distribution:

$$H_0 : F(x) = G(x) \quad \forall x$$

against the location-scale alternative hypothesis:

$$H_1 : G(x) = F\left(\frac{x-\mu}{\sigma}\right)$$

with $\mu \neq 0$ or $\sigma \neq 1$.

Podgor-Gastwirth PG2 test

Let I_i, $i = 1,\dots,N$ be a group indicator so that $I_i = 1$ when the i-th element of the combined sample belongs to the first sample, $I_i = 0$ otherwise; let S_i and S^2_i be the ranks and the squared ranks of the observations in the combined sample ($i=1,\dots,N$).

The PG2 test statistic is calculated as the ordinary least squares (OLS) estimator of S_i and S^2_i on the group indicators I_i and it is distributed as a Fisher-Snedecor F with 2 and N-3 degrees of freedom:

$$PG2 = \frac{\frac{\left(b^T S^T I - \frac{m^2}{N}\right)}{2}}{\frac{(m - b^T S^T I)}{(N-3)}} \sim F_{2,N\text{-}3}$$

In the above formula, T denotes the transpose operator, b is the 3 x 1 vector of the OLS estimate of the intercept term and the regression coefficients, S the N x 3 matrix of the ranks and the squared ranks of the observations and I the N x 1 vector of the group indicators $I_1,\dots, I_N$:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$\mathrm{S} = \begin{bmatrix} 1 & \mathrm{S}_1 & \mathrm{S}_1^2 \\ 1 & \mathrm{S}_2 & \mathrm{S}_2^2 \\ . & . & . \\ . & . & . \\ 1 & \mathrm{S}_N & \mathrm{S}_N^2 \end{bmatrix}$$

$$\mathrm{I} = \begin{bmatrix} \mathrm{I}_1 \\ \mathrm{I}_2 \\ . \\ . \\ \mathrm{I}_N \end{bmatrix}$$

Large values of the statistic imply the rejection of H_0. A user-defined function was used to perform the PG2 test (see the Supplementary material, subparagraph S4).

Podgor and Gastwirth showed that asymptotically the PG2 test can be recast as a quadratic combination of the Wilcoxon rank test for location and the Mood squared rank test for scale (Mood's scores give more weight to the extreme ranks).

Cucconi test

The Cucconi test addresses the location-scale problem by using the squares of ranks, S_i, and contrary-ranks, $(N+1-S_i)$, of the observations of the sample X_i ($i=1,\dots,m$) computed in the pooled sample. The test statistic is defined as:

$$\mathrm{C} = \frac{U^2 + V^2 - 2\varrho UV}{2(1-\varrho^2)}$$

where:

$$U = \frac{6\sum_{i=1}^{m} S_i^2 - m(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)\frac{(8N+11)}{5}}}$$

$$V = \frac{6\sum_{i=1}^{m} (N+1-S_i)^2 - m(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)\frac{(8N+11)}{5}}}$$

$$\varrho = \frac{2(N^2-4)}{(2N+1)(8N+11)} - 1$$

Note that U is based on the squares of the ranks S_i, while V is based on the squares of the contrary-ranks ($N + 1 - S_i$) of the first sample.

Let U' and V' be U and V computed referring to the second sample Y_i (j=1,...,n); the aforementioned expressions of U and V become, respectively:

$$U' = \frac{6\sum_{j=1}^{n} S_j^2 - n(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)\frac{(8N+11)}{5}}}$$

$$V' = \frac{6\sum_{j=1}^{n} (N+1-S_j)^2 - n(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)\frac{(8N+11)}{5}}}$$

Since

$$\sum_{i=1}^{m} S_i^2 + \sum_{j=1}^{n} S_j^2 = \sum_{i=1}^{m} (N+1-S_i)^2 +$$

$$\sum_{j=1}^{n} (N+1-S_j)^2 = \frac{N(N+1)(2N+1)}{6}$$

then U' = -U and V' = -V. Thus, the two test statistics are equal:

$$C' = \frac{U'^2 + V'^2 - 2\varrho U'V'}{2(1-\varrho^2)} = \frac{U^2 + V^2 - 2\varrho UV}{2(1-\varrho^2)} = C$$

It makes no difference whether U and V are computed based on the data of the first or the second sample, since this choice does not modify the test statistic.

Large values of the statistic imply the rejection of H_0. For the asymptotic Cucconi test we used the reported critical threshold of - ln(α). User-defined functions were used to perform the asymptotic and empirical versions of the Cucconi test (see the Supplementary material, subparagraph S4).

3. Tests for the general two-sample problem

Like the WMW test and the tests for the location-scale problem, the tests for the general two-sample problem assess whether both samples come from the same distribution:

$$H_0 : F(x) = G(x) \quad \forall x$$

However, the alternative hypothesis is:

$$H_1 : F(x) \neq G(x)$$

and thus it evaluates general differences between the two CDFs. The KS test concentrates on local CDF shifts while the CvM and AD tests consider the differences all along the CDF distribution.

Kolmogorov-Smirnov (K_S) test

The KS test statistic is defined as the largest absolute value of the difference between the Empirical Cumulative Distribution Functions (ECDFs) of the two samples:

$$KS = sup\,|F_m(x) - G_n(x)|$$

For large sample sizes, i.e. m = n with n > 40 (balanced sample sizes) and m > 16 and n > 20 (unbalanced samples), the large sample approximation is used [21] and the null hypothesis is rejected at level α when:

$$KS > \frac{c(\alpha)}{\sqrt{n}}$$

for balanced sample sizes, or

$$KS > c(\alpha)\sqrt{\frac{m+n}{mn}}$$

for unbalanced sample sizes, where c(α) are tabulated values (Tables 16 and 17 [21]).

The test was performed using the ks.test function included in the stats package.

Cramer-von Mises (CvM) test

The CvM test considers the difference between the two distributions over the entire CDF range. We considered herein the L_2-norm based version of the CvM test introduced by Anderson et al. [13], which involves the quadratic distance between the two ECDFs. Let $H_N(x) = mF_m(x) + nG_n(x)/N$, being H_N the ECDF associated with the combined sample. Then the CvM test statistic is defined as:

$$CvM = \frac{mn}{N}\int_{-\infty}^{+\infty} |F_m(x) - G_n(x)|^2 dH_N(x)$$

which is equivalent to:

$$CvM = \frac{mn}{N^2}\left[\sum_{i=1}^{m}(F_m(x_i) - G_n(x_i))^2 + \sum_{j=1}^{n}(F_m(x_j) - G_n(x_j))^2\right]$$

The null hypothesis is rejected for large values of CvM; asymptotic critical values are reported by Anderson and Darling [13]. However, the distance between the two ECDFs tends to 0 when $x \to -\infty$ or $x \to +\infty$, thus the value of the CvM test statistic is rather insensitive to the differences in the distribution tails. We performed the test by implementing a user-defined function including the empirical correction formula reported by Burr [22], using the limiting distribution to approximate the exact distribution of the CvM test statistic (see the Supplementary material, subparagraph S4).

Anderson-Darling (AD) test

The AD test statistic is a modification of L_2-CvM test statistic that, in order to give more weight to the observations in the distribution tails, includes a weighting function equal to the reciprocal of the variance of the ECDF (the latter is maximal around the median and minimal in the tails):

$$A^2_{mn} = \frac{mn}{N}\int_{-\infty}^{+\infty}\frac{\left(F_m(x) - G_n(x)\right)^2}{H_N(x)\left(1 - H_N(x)\right)}dH_N(x)$$

A simplification was introduced for computational purposes:

$$A^2_{mn} = \frac{1}{mn}\sum_{i=1}^{N-1}\frac{\left(M_iN - mi\right)^2}{i\,(N - i)}$$

where M_i is defined as the number of observations in the first sample less than or equal to the i-th smallest in the pooled sample. The standardised statistic is obtained by using its exact mean (equal to 1 in case of two samples) and exact variance σ_N, which was derived by Scholz [14]:

$$AD = \frac{A^2_{mn} - 1}{\sigma_N}$$

The upper tail critical values for the aforementioned test statistic are reported by Scholz [14] and the null hypothesis is rejected for large values. The standardisation removes some of the dependence of the test on the sample size, as it was confirmed through a Monte Carlo study [14]. For not tabulated critical values, an interpolation formula may be used to obtain the percentiles of interest. The test was performed using the *adk.test* function included in the *adk package*.

Zhang tests

All the considered Zhang tests (Z_K, Z_C, Z_A) derive from two types of test statistics [23] defined as:

$$Z = \int_{-\infty}^{+\infty} Z_t\, dw(t)$$

and

$$Z_{\max} = sup\,[Z_t w(t)] \qquad t \in (-\infty, +\infty)$$

where Z_t is the likelihood ratio test statistic and w(t) is a weighting function characterising the different tests. The Zhang tests are the analogs of the traditional tests KS, CvM and AD, which are obtained using the Pearson χ^2 test statistic as Z_t.

- Zhang Z_K test

Z_K is the analog of the KS test and it is obtained from Z_{max} with w(t) = 1. The computational formula for the Z_K test statistic is:

$$Z_k = \max_{1\le k\le N}\left[m\left(F_m \ln\frac{F_m}{H_N} + (1 - F_m)\ln\frac{1 - F_m}{1 - H_N}\right)\right.$$
$$\left. + n\left(G_n \ln\frac{G_n}{H_N} + (1 - G_n)\ln\frac{1 - G_n}{1 - H_N}\right)\right]$$

where

$$F_m = F_m(x_{(k)});\quad G_n = G_n(x_{(k)});\quad H_N = H_N(x_{(k)})$$

and H_N denotes the ECDF of the pooled sample (k=1,…,N). Large values of the statistic guide to the rejection of H_0.

- Zhang Z_c test

Z_C is the analog of the CvM test and it is obtained from Z with dw(t) defined as:

$$F(t)^{-1}[1 - F(t)]^{-1}dF(t)$$

where F(t) is the common underlying distribution under H_0. Let R_1 denote the rank in the pooled sample of the i-th ordered statistic $X_{(i)}$ in the first sample X_i (i=1,…,m) and R_2 denote the rank in the pooled sample of the j-th ordered statistic $Y_{(j)}$ in the second sample Y_j (j=1,…,n). The computational formula for the Z_C test statistic is:

$$Z_c = \frac{1}{N}\left[\sum_{i=1}^{m} \ln\left(\frac{m}{i-0.5}-1\right)\ln\left(\frac{N}{R_1-0.5}-1\right) + \sum_{j=1}^{n} \ln\left(\frac{n}{j-0.5}-1\right)\ln\left(\frac{N}{R_2-0.5}-1\right)\right]$$

Small values of the statistic guide to the rejection of H_0.

- Zhang Z_A test

Z_A is the analog of the AD test and it is obtained from Z with dw(t) defined as:

$$F_m(t)^{-1}[1 - F_m(t)]^{-1}dF_m(t)$$

The computational formula for the Z_A test statistic is:

$$Z_A = -\sum_{k=1}^{N}\left[m\frac{F_m \ln F_m + \left(1 - F_m\right)\ln\left(1 - F_m\right)}{(k-0.5)(N-k+0.5)} + n\frac{G_n \ln G_n + \left(1 - G_n\right)\ln\left(1 - G_n\right)}{(k-0.5)(N-k+0.5)}\right]$$

Small values of the statistic guide to the rejection of H_0. To perform the Zhang tests we converted the S-PLUS codes reported by Zhang [3] to R functions (see the Supplementary material, subparagraph S4).

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

This work was supported by a grant from Associazione Italiana Ricerca sul Cancro (AIRC) (Project 'EDERA - Tumor microenvironment-related changes as new tools for early detection and assessment of high-risk diseases', 5X1000, No.12162).

AUTHORS' CONTRIBUTIONS

EL planned, carried out the analysis of data and wrote the manuscript. RM planned the analysis and wrote the manuscript. FA planned the analysis and LM revised the manuscript. All authors have read and approved the final manuscript.

REFERENCES

1. Troyanskaya O, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002;18:1454-61.
2. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507-17.
3. Zhang J. Powerful two-sample tests based on the likelihood ratio. Technometrics 2006;48:95-103.
4. Clarke R, Ressom HW, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008;8(1):37-49.
5. Mai Q. The kolmogorov filter for variable screening in high-dimensional binary classification. Biometrika 2013;100:229-34.
6. Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics 2010;11:447.
7. Pagano M, Gauvreau K. Principles of Biostatistics. Duxbury/Thomson Learning, Pacific Grove, Calif, 2000.
8. Mann H, Whitney D. On a test whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18:50-60.
9. Podgor M, Gastwirth J. On non-parametric and generalised tests for the two-sample problem with location and scale change alternatives. Stat Med 1994;13:747-58.
10. Cucconi O. Un nuovo test non parametrico per il confronto tra due gruppi campionari (A new nonparametric test for the comparison between two sample groups). Giornale degli Economisti 1968;27:225-48.
11. Marozzi M. Some notes on the location-scale Cucconi test. J Nonparametric Stat 2009;21:629-47.
12. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione (On the empirical determination of a distribution law). Giornale dell'Istituto Italiano degli Attuari 1933;83-91.
13. Anderson TW, Darling DA. Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. Ann Math Stat 1952;23:193-212.
14. Scholz F, Stephens M. K-sample Anderson-Darling tests. J Am

Statistic Assoc 1987;82:918-24.

15. Burton A, Altman D, Royston P, Holder R. The design of simulation studies in medical statistics. Stat Med 2006;25:4279-92.
16. Conover W, Johnson ME, Johnson MM. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 1981;23:351-61.
17. Marozzi M. Levene type tests for the ratio of two scales. J Statist Comput Simulation 2011;81:815-26.
18. Rosenblatt M. Limiting theorems associated with variants of the von Mises statistic. Ann Math Stat 1952;23:1006-16.
19. Van't Veer L, Dai H, van de Vijver M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.
20. Van de Vijver M, He YD, van 't Veer LJ, et al. A gene expression signature as a predictor of survival in breast cancer. NEJM 2002;347:1999-2009.
21. Conover W. The design of simulation studies in medical statistics. John Wiley and Sons, New York, 1971.
22. Burr E. Distribution of the two-sample Cramer-von Mises criterion for small equal samples. Ann Math Stat 1963;34:1-374.
23. Zhang J. Powerful goodeness-of-fit based on likelihood ratio. JR Stat Soc Ser B Stat Methodol 2002;64:281-94.

*