Online resource 1: Supplementary material

1 Information

- Article title: Are the statistical tests the best way to deal with the biomarker selection problem?
- Journal name: Knowledge and Information Systems.
- Author names: Ari Urkullu, Aritz Pérez and Borja Calvo.
- Affiliation and e-mail address of the corresponding author:
  - Affiliation: Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU
  - E-mail address: ari.urkullu@ehu.eus

2 Introduction

In this online resource 1, we gather additional documentation which is not presented in the manuscript. Briefly, this additional documentation gathers:

- Details of the experimentation carried out.
- Descriptions of other alternative methods.

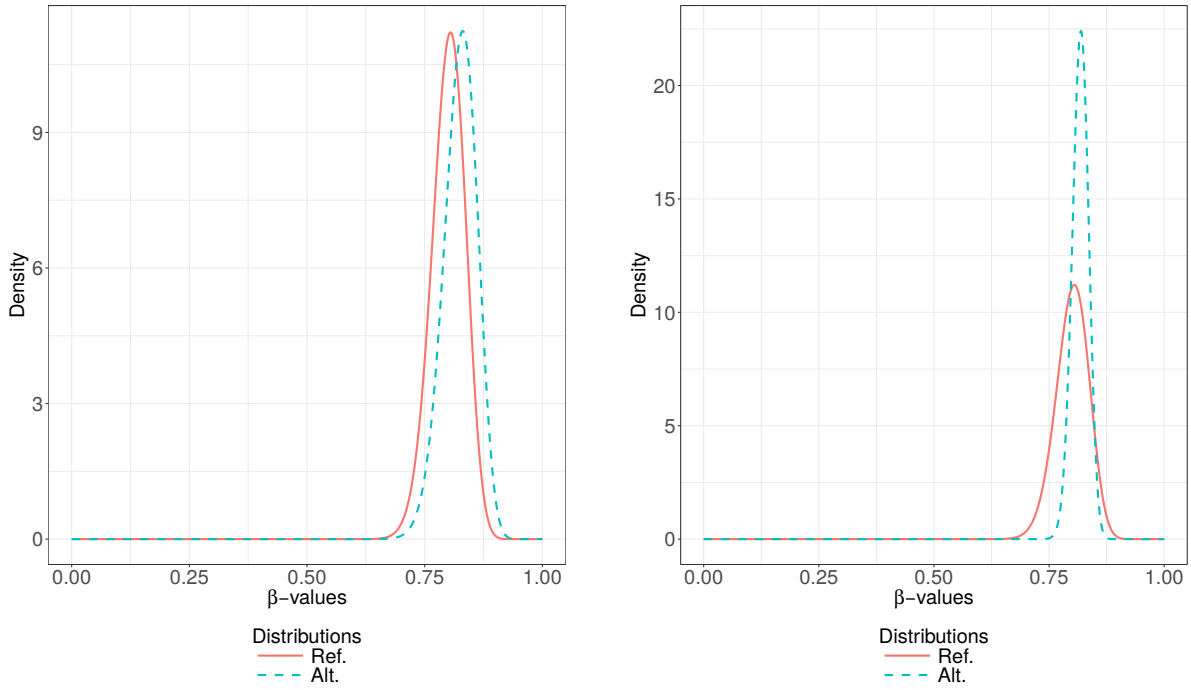
3 Details of the experimentation

In this section we explain the remaining details of the experimentation conducted. In order to do so, we have divided this section into four subsections, each one dedicated to a different stage of the experimental framework.

However, before going into the details of each stage, we first expose a common property of the first, second and third stages. That property consists of, for each configuration, the values of the parameters for the reference and alternative distributions in the first stage and the reference and fourth alternative distributions in the second and third stages are based on the experimentation conducted by Chen et al [1].

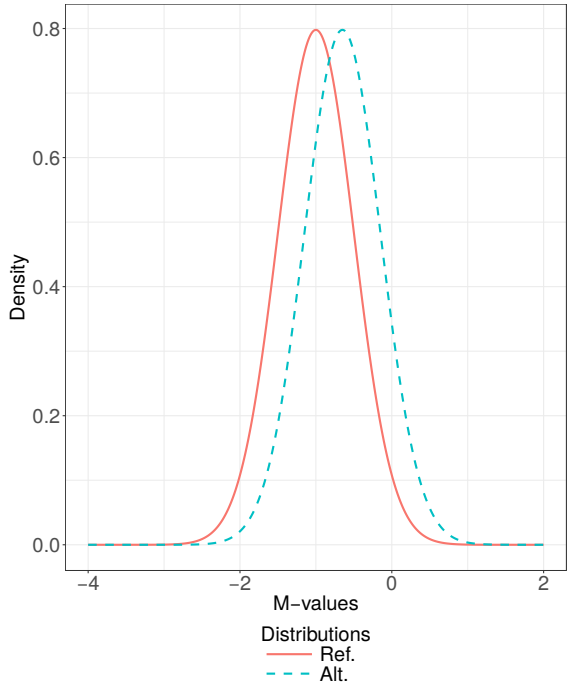
3.1 First stage

The specific values of the parameters of the reference and alternative distributions for each possible configuration depend on its associated type of distributions and its associated nature of differences. These specific values of the parameters are gathered in Table 1 and in Table 2. In addition, in Figure 1 a visual representation of all these distributions is offered.

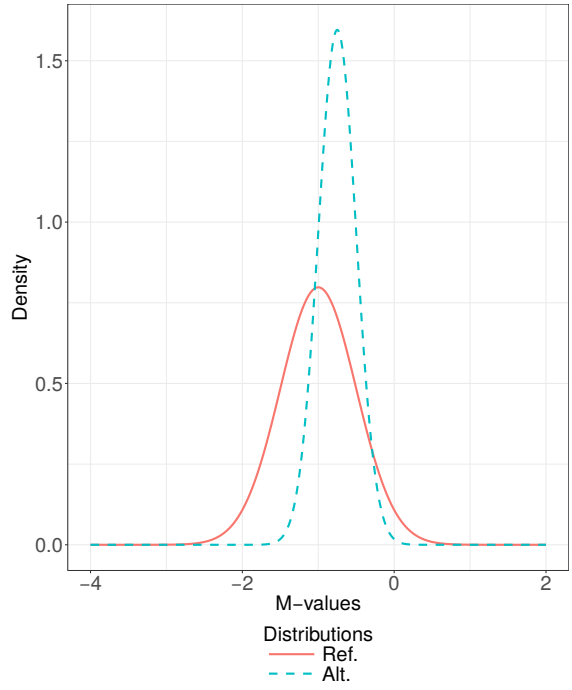


(a)

(b)



(c)



(d)

Figure 1: Probability distributions used in the first stage for the combinations 1(a) beta distributions and differences in location, 1(b) beta distributions and differences in location and spread, 1(c) normal distributions and differences in location and 1(d) normal distributions and differences in location and spread.

Table 1: Parameters of the distributions when Beta distributions are used.

Concept	Location	Location and spread
Reference distribution, α parameter	100	100
Reference distribution, β parameter	25	25
Alternative distribution, α parameter	92.99085	383.0044
Alternative distribution, β parameter	19.73293	85.32067

Table 2: Parameters of the distributions when Normal distributions are used.

Concept	Location	Location and spread
Reference distribution, μ parameter	-1	-1
Reference distribution, σ^2 parameter	0.5^2	0.5^2
Alternative distribution, μ parameter	-0.65	-0.75
Alternative distribution, σ^2 parameter	0.5^2	0.25^2

3.2 Second stage

As in the previous stage, once again, the specific values of the parameters of the reference and alternative distributions for each possible configuration depend on its associated type of distributions and its associated nature of differences. These specific values of the parameters are gathered in Table 3 and in Table 4. It is convenient to mention that in each configuration the alternative distributions 1, 2 and 3 are equally spaced between the reference distribution and the alternative distribution 4 in terms of mean and standard deviation. Finally, in Figure 2 all these distributions are represented graphically.

Table 3: Parameters of the distributions when Beta distributions are used.

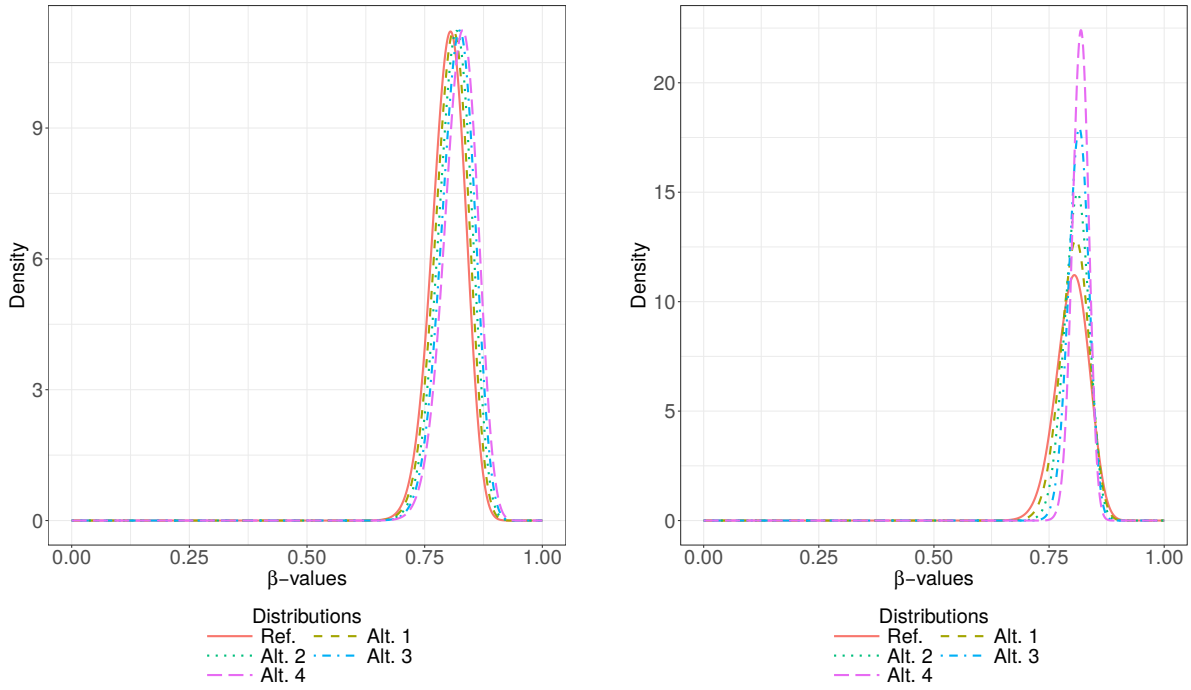
Concept	Location	Location and spread
Reference distribution, α parameter	100	100
Reference distribution, β parameter	25	25
Alternative distribution 1, α parameter	98.3792	129.3579
Alternative distribution 1, β parameter	23.64362	31.44414
Alternative distribution 2, α parameter	96.67151	174.2434
Alternative distribution 2, β parameter	22.31288	41.16213
Alternative distribution 3, α parameter	94.87577	248.105
Alternative distribution 3, β parameter	21.00894	56.93099
Alternative distribution 4, α parameter	92.99085	383.0044
Alternative distribution 4, β parameter	19.73293	85.32067

Table 4: Parameters of the distributions when Normal distributions are used.

Concept	Location	Location and spread
Reference distribution, μ parameter	-1	-1
Reference distribution, σ^2 parameter	0.5^2	0.5^2
Alternative distribution 1, μ parameter	-0.9125	-0.9375
Alternative distribution 1, σ^2 parameter	0.5^2	0.4375^2
Alternative distribution 2, μ parameter	-0.825	-0.875
Alternative distribution 2, σ^2 parameter	0.5^2	0.375^2
Alternative distribution 3, μ parameter	-0.7375	-0.8125
Alternative distribution 3, σ^2 parameter	0.5^2	0.3125^2
Alternative distribution 4, μ parameter	-0.65	-0.75
Alternative distribution 4, σ^2 parameter	0.5^2	0.25^2

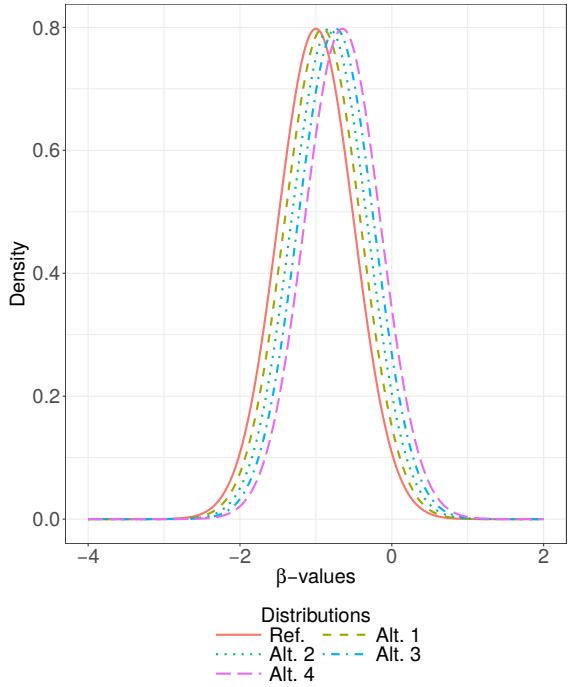
3.3 Third stage

In this stage, since only one type of distribution is used, the specific values of the parameters of the reference and alternative distributions for each possible configuration depend only on its associated nature of differences. Table 5 shows all these specific values. Besides, in each configuration, each component of the alternative distributions 1, 2 and 3 is equally spaced between the corresponding components of the reference distribution and the alternative

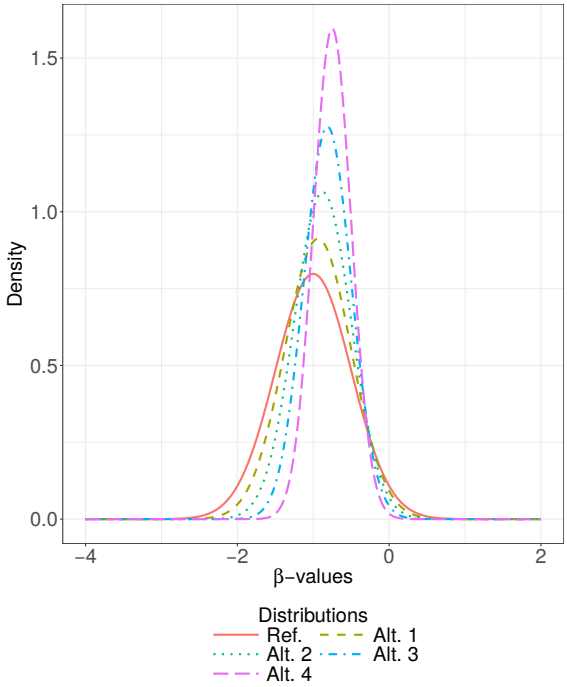


(a)

(b)



(c)



(d)

Figure 2: Probability distributions used in the second stage for the combinations 2(a) beta distributions and differences in location, 2(b) beta distributions and differences in location and spread, 2(c) normal distributions and differences in location and 2(d) normal distributions and differences in location and spread.

distribution 4 in terms of mean and standard deviation. Finally, all these distributions are displayed in Figures 3, 4 and 5.

Table 5: Parameters of the distributions.

Concept	Location	Location and spread
Reference distribution, component 1, μ parameter	-3	-3
Reference distribution, component 1, σ^2 parameter	0.5^2	0.5^2
Reference distribution, component 2, μ parameter	2	2
Reference distribution, component 2, σ^2 parameter	0.5^2	0.5^2
Alternative distribution 1, component 1, μ parameter	-2.9125	-2.9375
Alternative distribution 1, component 1, σ^2 parameter	0.5	0.4375
Alternative distribution 1, component 2, μ parameter	2.0875	2.0625
Alternative distribution 1, component 2, σ^2 parameter	0.5	0.4375
Alternative distribution 2, component 1, μ parameter	-2.825	-2.875
Alternative distribution 2, component 1, σ^2 parameter	0.5	0.375
Alternative distribution 2, component 2, μ parameter	2.175	2.125
Alternative distribution 2, component 2, σ^2 parameter	0.5	0.375
Alternative distribution 3, component 1, μ parameter	-2.7375	-2.8125
Alternative distribution 3, component 1, σ^2 parameter	0.5	0.3125
Alternative distribution 3, component 2, μ parameter	2.2625	2.1875
Alternative distribution 3, component 2, σ^2 parameter	0.5	0.3125
Alternative distribution 4, component 1, μ parameter	-2.65	-2.75
Alternative distribution 4, component 1, σ^2 parameter	0.5	0.25
Alternative distribution 4, component 2, μ parameter	2.35	2.25
Alternative distribution 4, component 2, σ^2 parameter	0.5	0.25

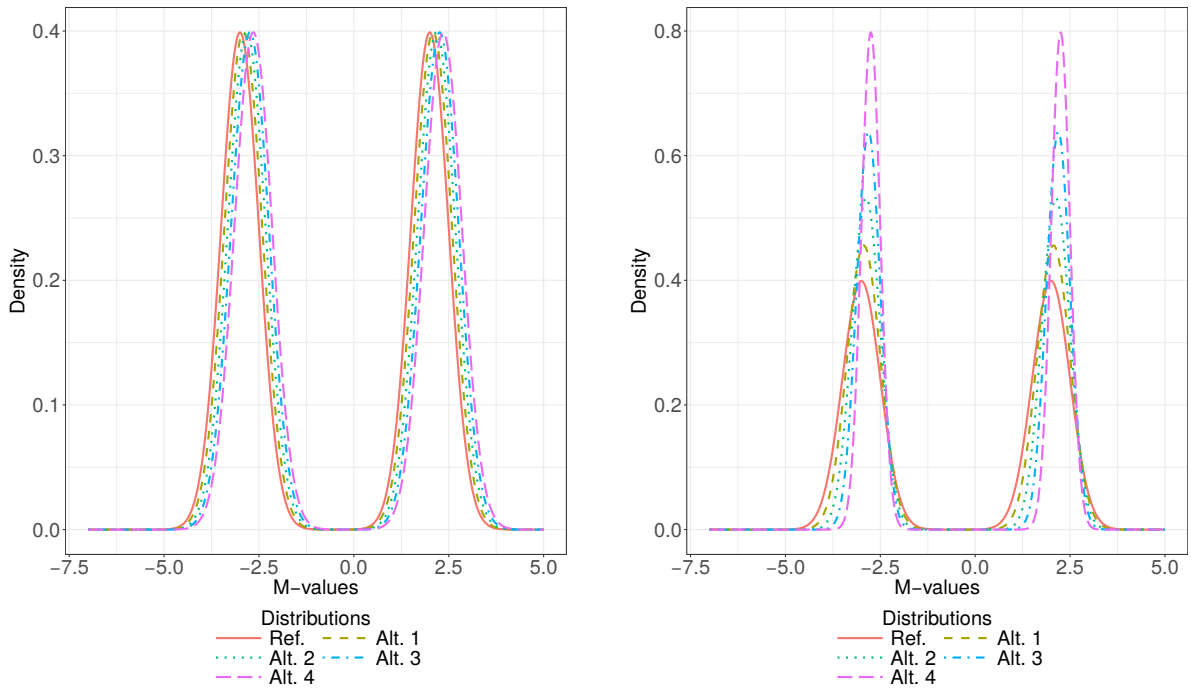
3.4 Fourth stage

We have divided this subsection into two subsections, one dedicated to the preprocessings of the ovarian cancer database and the other dedicated to the preprocessing of the nephropathy database.

3.4.1 Ovarian cancer database preprocessings

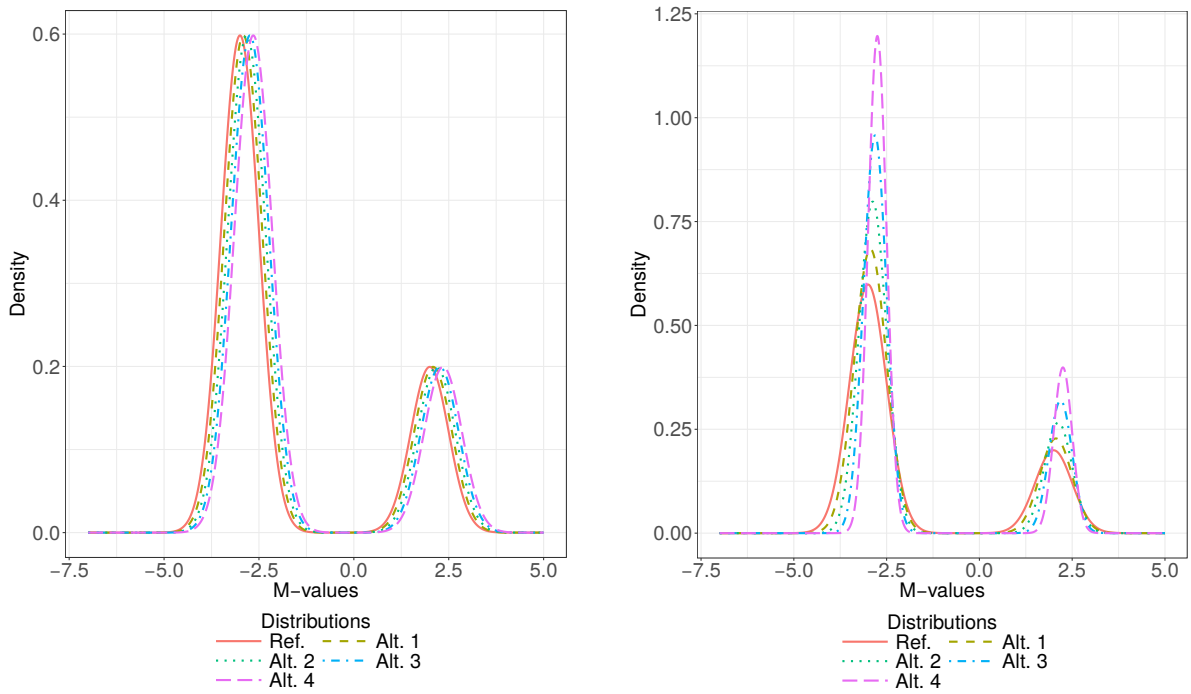
The two preprocessings done in our work, that have been applied to the ovarian cancer database, are based on what was done by Wang et al [3]. The first preprocessing which was applied, the one that does not remove every single outlier systematically, consists of applying the following steps sequentially to the matrix of β-values available in the GEO database:

1. Among all the ovarian cancer cases, only those who gave their blood at the time of their diagnosis prior to treatment have been used.
2. Samples whose bisulfite conversion efficiencies are too low (< 4000) have been removed.
3. Data from batches 10-12 have been removed.
4. In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR})$. Finally all those samples whose averages are not within that range are removed.
5. All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
6. All the CpG sites whose detection p-values are not below 0.05 in all samples are removed.
7. All the CpG sites that do not have numeric values (e.g., NA values) for at least 50 individuals per group are removed.



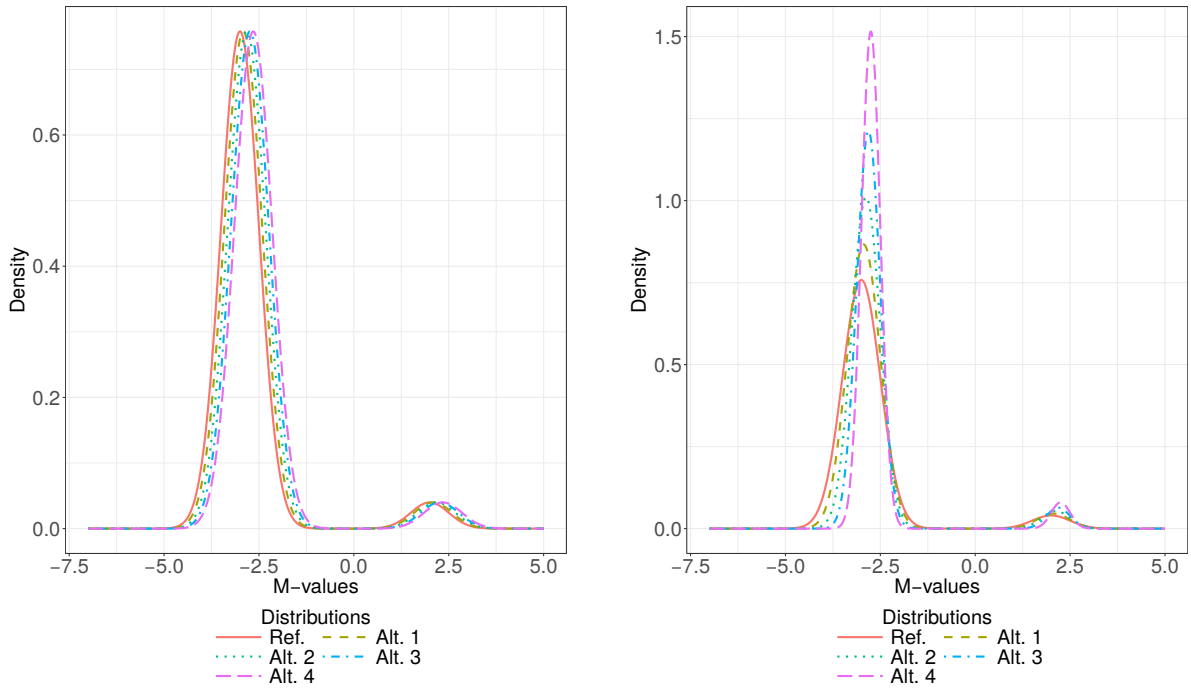
(a) (b)

Figure 3: Probability distributions used in the third stage for the combinations 3(a) weights $(50, 50)$ and differences in location and 3(b) weights $(50, 50)$ and differences in location and spread.



(a) (b)

Figure 4: Probability distributions used in the third stage for the combinations 4(a) weights $(75, 25)$ and differences in location and 4(b) weights $(75, 25)$ and differences in location and spread.



(a) (b)

Figure 5: Probability distributions used in the third stage for the combinations 5(a) weights $(95, 5)$ and differences in location and 5(b) weights $(95, 5)$ and differences in location and spread.

The other preprocessing is pretty similar to that which has already been presented. Specifically, the next steps are applied sequentially to the matrix of β-values available in the GEO database:

1. Among all the ovarian cancer cases, only those who gave their blood at the time of their diagnosis prior to treatment have been used.
2. Samples whose bisulfite conversion efficiencies are too low (< 4000) have been removed.
3. Data from batches 10-12 have been removed.
4. In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR})$. Finally all those samples whose averages are not within that range are removed.
5. For each CpG site, we have measured which values of total intensity and which β-values lie outside their corresponding ranges defined by $(Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR})$. All the corresponding β-values are erased, setting their value to NA.
6. All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
7. All the CpG sites that do not have numeric values (e.g., NA values) for at least 50 individuals per group are removed.

3.4.2 Nephropathy database preprocessing

The preprocessing done in our work to the nephropathy cancer database is based on what was done by Teschendorff et al [2]. The preprocessing we applied, consists of applying the following steps sequentially to the matrix of β-values available in the GEO database:

- Samples whose bisulfite conversion efficiencies are too low (lower values outside the range $(Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR})$) have been removed.

- In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR})$. Finally all those samples whose averages are not within that range are removed.
- All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
- All the CpG sites whose detection p-values are not below 0.05 in all samples are removed.

4 Other alternative methods

We have divided this section into two subsections, dedicating each one to a different alternative method. For a given site i, we denote the vectors of values sampled as $G_1^i = \{x_1^i, \ldots, x_M^i\}$ and as $G_2^i = \{y_1^i, \ldots, y_N^i\}$,for group 1 of M individuals and group 2 of N individuals, respectively.

4.1 Movement of distributions method

The idea behind this method is to attempt to measure somehow how much it costs to transform the estimation of the distribution of one group into the estimation of the distribution of the other group. So as to do that, we considered that at each value j of group 1 there is a portion of $1/M$ of the whole distribution density (analogously, for each value j of group 2 we considered there to be a portion of $1/N$). Namely, the method is composed of the following steps:

1. Calculate the least common multiple (lcm) between M and N.
2. Repeat each element of G_1^i lcm/M times and sort the resulting vector. Analogously, repeat each element of G_2^i lcm/N times and sort the resulting vector. We denote these two new vectors as $G_1^{'i} = \{x_1^{'i}, \ldots, x_{lcm}^{'i}\}$ and $G_2^{'i} = \{y_1^{'i}, \ldots, y_{lcm}^{'i}\}$, respectively.
3. Finally, compute the next value as the outcome of the method, $s(G_1^i)$ and $s(G_2^i)$ being the standard deviations of G_1^i and G_2^i:

$$\frac{\sum_{j=1}^{lcm} |x_j^{'i} - y_j^{'i}|}{lcm \cdot (s(G_1^i) + s(G_2^i))}$$

Since this method is sensitive to differences in both location and spread, in Figures 6 and 7 the results during the experimentation are displayed together with the results of the Tl test and the Kolmogorov-Smirnov test.

4.2 Differences of distributions method

The essence of this method consists of computing at each of the different points (except for the first one) of $G_1^i \cup G_2^i$ taking into account the previous point, the amount of difference between the areas of the empirically estimated distributions. $G_1^i \cup G_2^i$ being $\{z_1^i, \ldots, z_K^i\}$ with $K < M + N$ and it being ordered, $z_1^i <, \ldots, < z_K^i$, the outcome of this method is summarized in the following expression:

$$\sum_{j=2}^{K} |\frac{|G_1^i < z_j^i|}{M} - \frac{|G_2^i < z_j^i|}{N}| \cdot |z_j^i - z_{j-1}^i|$$

Since this method is sensitive to differences in both location and spread, in Figures 8 and 9 the results during the experimentation are displayed together with the results of the Tl test and the Kolmogorov-Smirnov test.

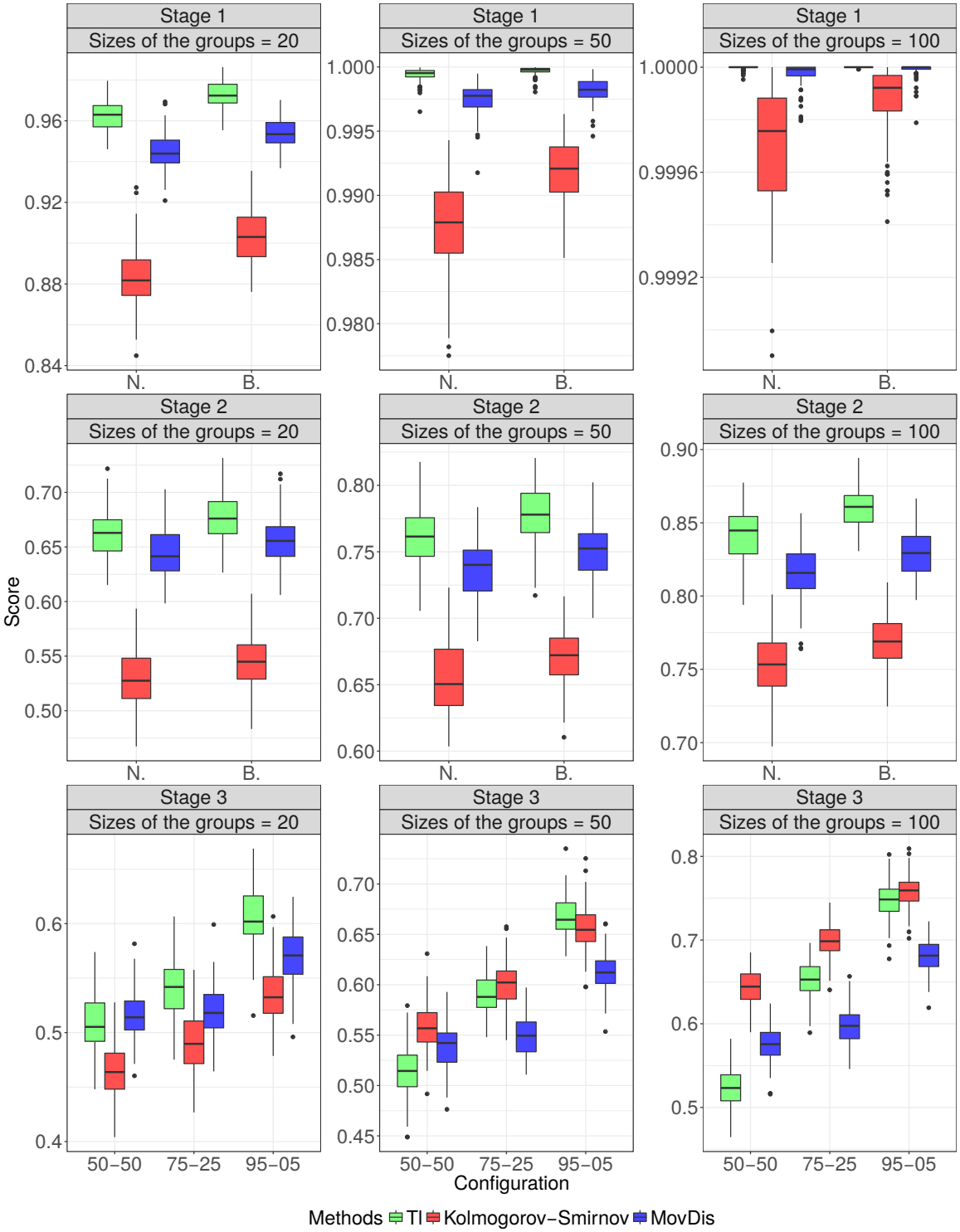
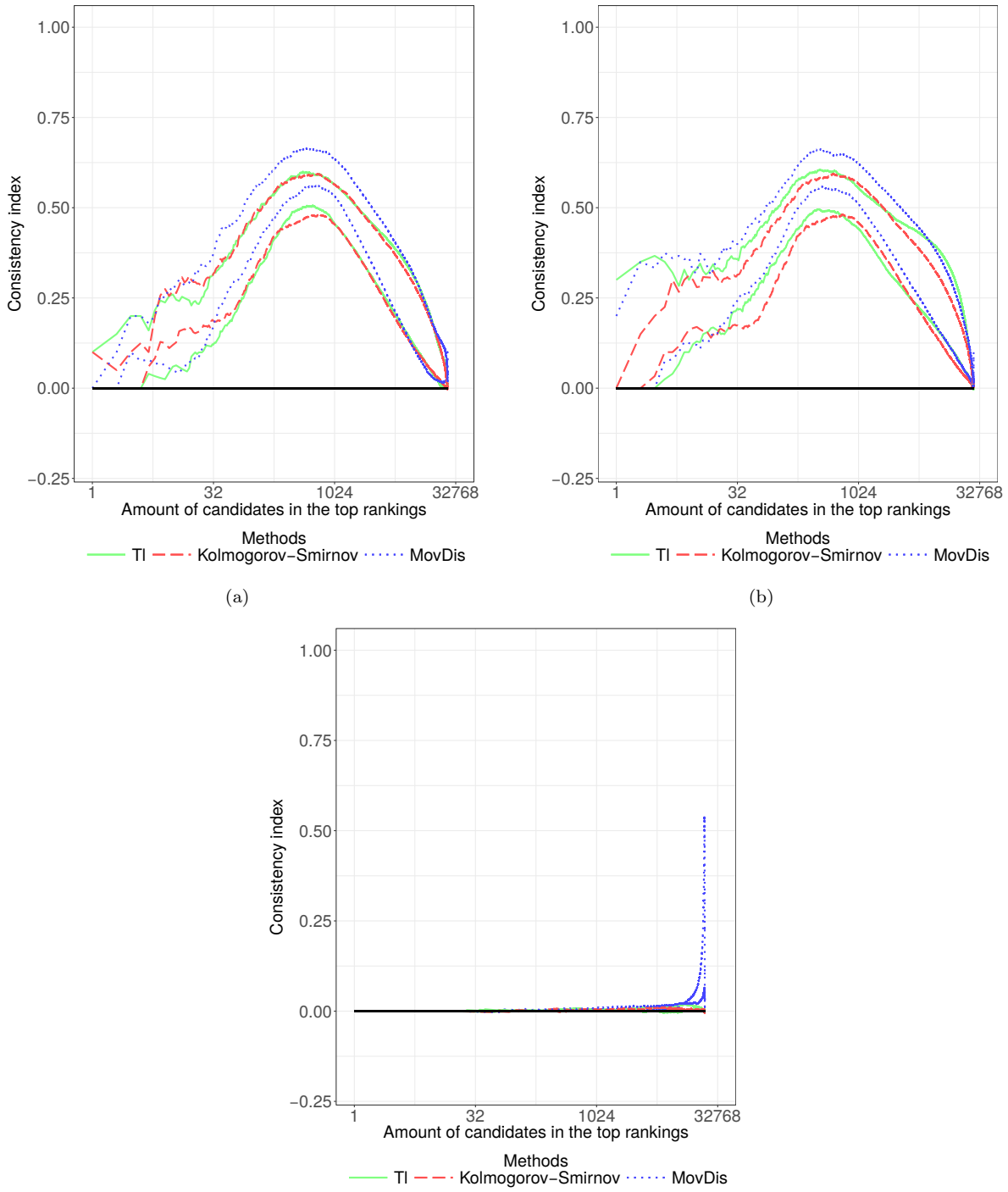


Figure 6: Results of the Tl test, the Kolmogorov-Smirnov test and the MovDist method in the synthetic stages. The labels of the abscissa axes of the boxplots specify information about the distributions used: "N." - Normal distributions, "B." - Beta distributions, "$50-50$", "$75-25$" or "$95-05$" - mixtures of normal distributions in which the weights of the normal distributions are equal to the values specified by the corresponding label.



(a)

(b)

(c)

Figure 7: Results in the real stage, when the Tl test, the Kolmogorov-Smirnov test and the MovDist method are applied in 7(a) the ovarian cancer database (in which the first preprocessing has been executed), 7(b) the ovarian cancer database (in which the second preprocessing has been executed) and 7(c) the nephropathy database.

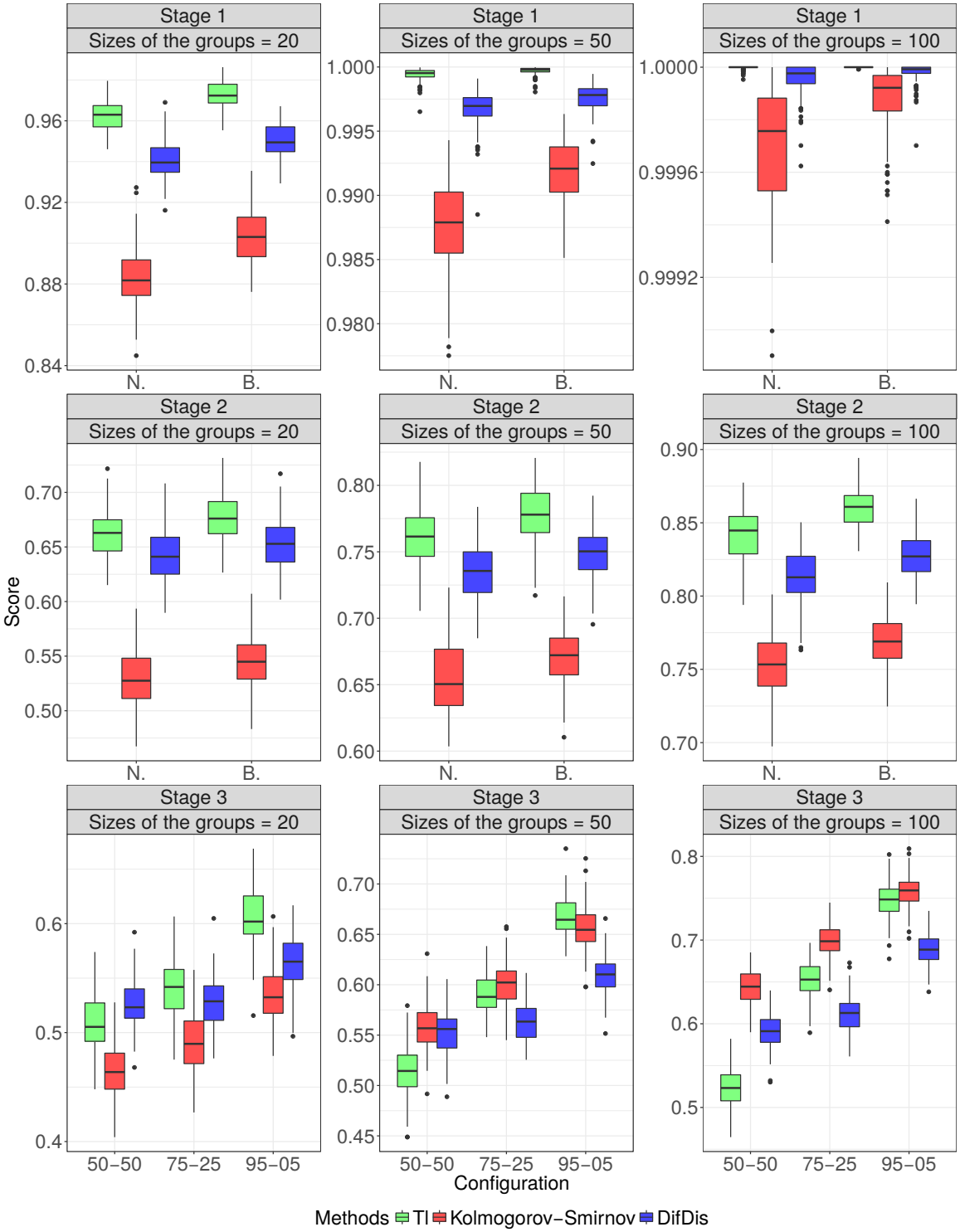
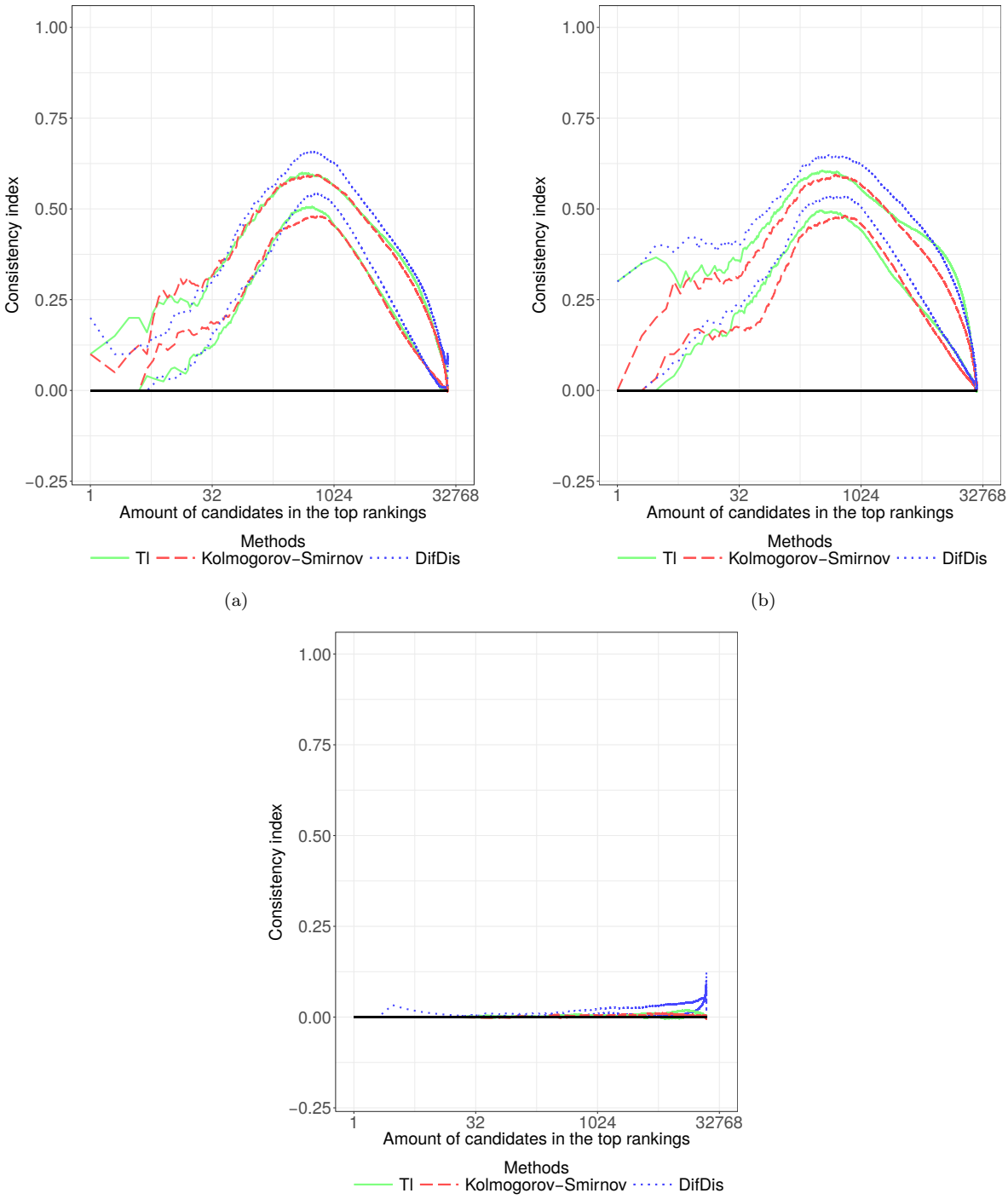


Figure 8: Results of the Tl test, the Kolmogorov-Smirnov test and the DifDist method in the synthetic stages. The labels of the abscissa axes of the boxplots specify information about the distributions used: "N." - Normal distributions, "B." - Beta distributions, "$50-50$", "$75-25$" or "$95-05$" - mixtures of normal distributions in which the weights of the normal distributions are equal to the values specified by the corresponding label.



(a)

(b)

(c)

Figure 9: Results in the real stage, when the Tl test, the Kolmogorov-Smirnov test and the DifDist method are applied in 9(a) the ovarian cancer database (in which the first preprocessing has been executed), 9(b) the ovarian cancer database (in which the second preprocessing has been executed) and 9(c) the nephropathy database.

References

[1] Chen, Yong and Ning, Yang and Hong, Chuan and Wang, Shuang: Semiparametric Tests for Identifying Differentially Methylated Loci With Case–Control Designs Using Illumina Arrays. Genetic Epidemiology **38**(1), 42–50 (2014)

[2] Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., et al.: Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome research **20**(4), 440–446 (2010)

[3] Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agustí, A., Anderson, W.H., Lomas, D.A., DeMeo, D.L.: Systemic Steroid Exposure Is Associated with Differential Methylation in Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine **186**(12), 1248–1255 (2012)