Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 4, 677–686
DOI: 10.61822/amcs-2025-0048

AN ALTERNATIVE POWERFUL APPROACH TO MULTIPLE HYPOTHESES TESTING: APPLICATION IN A CORRELATION STUDY

KATARZYNA STAPOR [a,*], GRZEGORZ KOŃCZAK [b]

[a]Department of Applied Informatics
Silesian University of Technology
Akademicka 2A, 44-100 Gliwice, Poland
e-mail: katarzyna.stapor@polsl.pl

[b]Department of Statistics, Econometrics and Mathematics
University of Economics in Katowice
1 Maja 50, 42-287 Katowice, Poland
e-mail: grzegorz.konczak@ue.katowice.pl

The multiple hypothesis testing problem occurs when a number of individual hypothesis tests are considered simultaneously. The larger the number of inferences (tests) made, the more likely erroneous inferences become. Several classical statistical techniques have been developed to address this problem. Unfortunately, very often, in studies that aim to examine the significance of multiple effects (for example, correlations), the effect of multiple testing is ignored. This leads to an inflated number of significant findings (e.g., correlation coefficients). On the other hand, in the case of a large number of tests about relatively small effects, and thus a large family of inferences, the power of individual tests decreases rapidly when classical procedures for controlling multiplicity are applied, often resulting in too few significant findings. In this paper, we propose an alternative and powerful approach to face the problem of testing multiple hypotheses and develop a new MCPerm method for testing the significance of multiple correlations. The proposed solution is definitely more effective than Holm's, which proved to be more conservative in the presence of several dependencies and tended to indicate fewer of those.

Keywords: multiple hypothesis testing, linear correlations, permutation tests.

1. Introduction

When doing research, scientists as well as practitioners commonly face the problem of multiple hypothesis testing (MHT). It arises when statistical analysis involves conducting several tests simultaneously on a single dataset, leading to multiple decisions.

In economics, biology, medicine or psychology, for example, it is frequent to search for relationships (correlations) between variables assessed by diagnostic instruments or other measurement tools. The researcher tries to detect as many correlations as possible between the studied features (variables) and formulates dozens of hypotheses in search of statistically significant relationships. The correlation coefficients are often used to measure the strength of those relationships. Significance tests for equality of means of many measured biological characteristics between the two groups of patients are a routine practice in medicine. Another interesting and very common example where we deal with multiple hypotheses is testing the statistical significance of differences in the performance of multiple classifiers or different machine learning methods in order to demonstrate the superiority of the newly introduced one (Demšar, 2006; Żelasko and Pławiak, 2021; Kowal *et al.*, 2021). The paper by Stapor *et al.* (2021) proposes a method based on a series of pairwise tests for this purpose.

However, performing multiple statistical tests is not without pitfalls. When more than one test of significance is performed based on a given sample, the probability of making at least one Type I error (i.e., obtaining significant results by chance only) rises rapidly as the number of tests increases. Assuming that k independent null hypotheses

*Corresponding author

are tested separately, each at significance level α, the probability of rejecting at least one true null hypothesis is $1-(1-\alpha)^k$. This means that, for example, testing more correlation coefficients or pairs of means causes more wrong rejections of null hypotheses (i.e., we then obtain apparently significant correlations or differences in means).

For instance, in the case of 20 independent, true null hypotheses, each tested at a significance level of $\alpha = 0.05$, the probability of rejecting at least one true null hypothesis is 0.64. Moreover, tests are rarely independent in practice, which further complicates proper control of the multiple testing effect.

Unfortunately, in studies examining the significance of multiple effects, such as correlations, the issue of multiple testing is often overlooked. Researchers should, therefore, apply the proper methodology to enable the control of multiplicity and avoid erroneously concluding the presence of significant effects (here correlations). The proper control of the multiplicity effect is difficult as well as controversial. Probably the most crucial problem is about what hypotheses should constitute a family. Some scientists, for example, are inclined to let the family be defined by all tests performed within a single experiment.

There are many error measures that can be used when testing a family focusing on either Type I or Type II errors. One of the popular ways to control the Type I error rate (in particular, when the family of hypotheses is not very large, which is the case in applications in economy or psychology) is by controlling the family-wise error rate (FWER) metric. The FWER check for a family of inferences means that the probability of rejecting at least one true null hypothesis is no greater than a predetermined value α. In cases of very large families, it is worth considering the control of the false discovery rate (FDR) measure (Benjamini and Hochberg, 1995) to control the Type I error rate. The FDR represents the expected proportion of false discoveries (incorrectly rejected null hypotheses) among all discoveries (rejected null hypotheses). There are many multiple testing procedures for controlling the FWER, FDR or other types of errors nowadays; for an overview, see, for example, the works of Dickhaus (2014) or Hochberg and Tamkane (1987). The existing methods can, for example, be divided into two classes:

- marginal,
- joint (or multivariate).

Marginal methods only model the marginal distributions of the involved test statistics explicitly and combine these test statistics or, equivalently, corresponding p-values following probabilistic calculations. Different marginal methods employ different qualitative assumptions on the dependency structure between test statistics or p-values. Joint methods consider the full joint distribution of all test statistics and rely on calculating or approximating quantiles of this joint distribution, for instance, by resampling. Based on their structure, the methods can be divided into

- single-step,
- stepwise rejective.

There are also many other divisions using other criteria (Dickhaus, 2014).

The oldest and most famous marginal single-step procedure to control the FWER is the Bonferroni procedure, which rejects each hypothesis in a family when the corresponding p-value is not greater than the threshold equal to α/m, where m is the number of hypotheses in a family. This procedure can be used regardless of the type of relationship between test statistics. It provides FWER error control at the α level, but it is very conservative, i.e., it has low power. This conservatism depends on the correlation between the test statistics. A less conservative marginal method of FWER control is the stepwise iterative Holm universal procedure (Holm, 1979), which, after ordering p-values, rejects the i-th hypothesis in a family if the corresponding p-value is not greater than the threshold equal to $\alpha/(m-i+1)$. Among the marginal procedures, the stepwise approaches have the greatest power. Apart from Holm's old method, there are many other stepwise rejective tests of different level of sophistication that control FWER like, for example, the Hommel or Hochberg ones (see, e.g., Hochberg and Tamkane, 1987). The Hommel procedure is slightly more powerful than Holm's but is computationally quite complex. The requirements for dependencies between test statistics significantly limit the scope of applications of marginal procedures and complicate their application. Benjamini and Hochberg (1995) proposed a marginal procedure to ensure control of the FDR at the predetermined level α in the case of independent test statistics. The BH step-up procedure works as follows. We have $H_0(1), \ldots, H_0(m)$ null hypotheses tested and their corresponding p-values $p_1, \ldots, p_m$. We list them in ascending order $p_{(1)}, \ldots, p_{(m)}$. We find the largest k such that

$$p_{(k)} \leq \alpha \frac{k}{m}.$$

We then reject the null hypotheses for all $H_0(i)$ for $i = 1, \ldots, k$.

Joint multiple testing procedures take into account the joint distribution of the test statistics and are therefore less conservative than marginal approaches. Westfall and Young (1993) proposed joint multiple testing procedures using resampling. These procedures are based on the maxima or minima of test statistics. The use of resampling

allows multiple testing to be performed despite the lack of normality or the lack of knowledge of the covariance structure of the data. However, these procedures require the pivotality assumption to be met (this assumption implies that the distribution of test statistics for a group of null hypotheses is independent of the truth or falsity of other null hypotheses in the overall set). Unfortunately, in many research situations, and especially when testing the significance of correlation coefficients, this condition is not met. Another method is joint procedures based on resampling for the generation of the null distribution proposed by Dudoit and van der Laan (2008). However, the results of simulation experiments by Denkowska (2013) show that they do not always guarantee the control of the FWER for the set of inferences at a predetermined level. Moreover, the initial diagnosis also indicates some instability of the results depending on the number of samplings. It is worth mentioning the interesting approach based on combined permutation tests proposed by Pesarin and Salmaso (2010) as well as Bonnini and Cavallo (2021) in a bivariate regression problem. These authors promote the idea that the problem can be broken down into partial tests, to be than combined.

Correcting for multiple hypotheses can drastically decrease the power of an evaluation and, in the case of small effects (e.g., correlations) and insufficiently large samples, it will make it impossible to detect the existence of an effect (correlation).

In this paper, we propose an alternative and powerful approach to face the problem of testing multiple hypotheses and develop a new MCPerm (multiple correlations using permutation) method for testing the significance of multiple correlations (e.g., in a correlation matrix). The idea by Bonnini and Cavallo (2021) gave us a frame for our new method but based on only one permutation test and differently formulated set of hypotheses. By appropriately reformulating the problem, we pose it as that of testing one global hypothesis. The presented method, however, determines the significance of the system of linear correlation coefficients, i.e., the significance of the simultaneous occurrence of several correlations. The new testing procedure ensures the size of the whole test is not higher than the adopted significance level α.

The procedure was designed for research involving several or a dozen or so variables and several dozen or several hundred observations. These are fairly typical situations for statistical inference, e.g., in economic or social/psychological research. An example of application of the proposed method in an economic problem is presented further. Due to the use of permutation tests in individual steps of the proposed procedure (computational time), the possibility of applying it for thousands of variables was not considered.

For the purpose of preliminary comparison of our method with the existing ones, one of the marginal methods was chosen, namely, Holm's. We also considered candidates from joint methods, but due to the difficulties with joint approaches mentioned above (the need for pivotality assumption checking or the instability of FWER error checking), which raise additional problems, we decided not to include them in this initial comparison. Some of the marginal methods are more conservative (e.g., Bonferroni), while others are more liberal (e.g., Benjamini–Hochberg). The proposed method is based on a completely different idea, namely, the use of permutation tests to test simultaneous independence for several relationships between variables. In the simulation analyses, the aim was to compare the properties of the proposed MCPerm method with a technique representing the classical approach to multiple testing issues. Therefore, Holm's method, which is neither the most conservative nor the most liberal, was chosen for comparison.

The structure of the paper is as follows. In Section 2 the proposed, original solution to the problem of testing the significance of multiple correlations, the stepwise procedure MCPerm, is presented, and in Section 3 we discuss a simulation analysis of the power of the proposed procedure, comparing it with one of the classical methods, namely, Holm's marginal approach. Section 4 contains an illustrative practical example on data from the 2021 National Population and Housing Census in Poland. The paper ends with a discussion and conclusions.

2. Proposed MCPerm stepwise procedure at a fixed significance level α

Here, we propose a different approach to the solution of a multiple hypotheses testing problem. We will present it using the example of testing the significance of multiple correlations. The idea of the proposal is based on permutation tests.

Permutation tests are nonparametric statistical methods based on the Fisher–Pitman permutation model. Unlike traditional parametric tests, permutation ones do not depend on assumptions about the underlying data distribution, such as normality. In particular, permutation tests are generally asymptotically as powerful as their parametric counterparts in the conditions for the latter (Bonnini *et al.*, 2014). The core idea involves repeatedly permuting one or more variables to generate an empirical distribution of the test statistic under the null hypothesis (Toth, 2020). By comparing the observed value of the test statistic to this empirical distribution, one can estimate the probability (achieved significance level, ASL) of obtaining a value as extreme as or more extreme than the observed statistic, assuming the null hypothesis is true. This approach offers a robust and flexible alternative to classical statistical inference, particularly when sample

sizes are small, when there are non-random samples, or when data violate standard parametric assumptions (Berry *et al.*, 2018).

Let us assume we are interested in examining the significance of a set of k linear correlation coefficients $\rho_1, \ldots, \rho_k$ between the variable Y and each of the variables $X_1, X_2, \ldots, X_k$, based on a random sample taken from a population under study. This is a classical example of a multiple hypotheses testing situation, and it is necessary to use the appropriate methods mentioned in Introduction, allowing some control of an error at the assumed level of significance. The main idea of the proposed approach is to identify the maximum number, denoted by s, of statistically significant correlations among the k tested ones ($1 \leq s \leq k$). A distinctive feature of our method is that it searches for a set of simultaneously significant correlations. This formulation allows one to circumvent the classical multiple testing problem by focusing on the significance of a system of correlations. In each of the s ($s = 1, \ldots, k$) iterations (steps) of the proposed procedure described below, the following null hypothesis is tested:

$$H_0(s) : \bigwedge_{j=1}^{s} \rho_{(j)} = 0$$

against the alternative one:

$$H_1(s) : \bigwedge_{j=1}^{s} \rho_{(j)} \neq 0,$$

where $\rho_{(1)}, \rho_{(2)}, \ldots, \rho_{(k)}$ denote absolute population correlation coefficients in descending order.

The alternative hypothesis (which is not a logical negation of the null hypothesis) states the (simultaneous) significance of all s ($s \leq k$) correlation coefficients.

Let us first introduce the notation. The sample correlation coefficient or sample Pearson correlation coefficient for assessing the strength of linear dependency between the two variables (features) X_i and $Y (i = 1, 2, \ldots, k)$ is calculated according to the following formula:

$$r_i = \frac{\sum_{j=1}^{n}(x_{ij} - \overline{x}_i)(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \overline{x}_i)^2}\sqrt{\sum_{j=1}^{n}(y_j - \overline{y})^2}}.$$

The absolute values of the sample correlation coefficients are ordered non-increasingly, yielding the following sequence: $|r_{(1)}|, |r_{(2)}|, \ldots, |r_{(k)}|$, where $r_{(1)}$ is the sample linear correlation coefficient with the largest absolute value. In further deliberations, we denote the absolute values of those sample correlation coefficients after ordering as

$$R_1^0, R_2^0, \ldots, R_k^0.$$

The superscript 0 (zero) means that the coefficient is calculated from the raw input sample (the so-called raw correlation coefficient).

The proposed iterative procedure MCPerm is based on a null distribution generated by permutations (i.e., the null permutation-based Pearson correlation coefficients) as follows. The input data set (the sample) is permuted N times after all variables $X_1, X_2, \ldots, X_k$. Calculating Pearson correlation coefficients on such permuted data at the i-th permutation yields the following (ordered) absolute values of the so-called permutation-based correlation coefficients:

$$R_1^{(i)}, R_2^{(i)}, \ldots, R_k^{(i)}$$

for $i = 1, 2, \ldots, N$, where N is the number of permutations.

A key element in each permutation test is the assessment of the probability of the achieved significance level (ASL) (Efron and Tibshirani, 1993; Good, 2006):

$$ASL = P_{H_0}(T \geq T_0) \approx \frac{card\{i : T_i \geq T_0\} + 1}{N + 1},$$

where T_0 is the value of the test statistic calculated from the sampled data and T_i ($i = 1, 2, \ldots, N$) are the values of the test statistic determined for successive permutations of the data. The achieved significance level obtained via permutation estimates the probability of observing a test statistic as extreme as the one calculated from the actual data.

The proposed iterative permutation-based procedure MCPerm calculates in each s-th iteration (a step) the percentage w of permutations in which at least one of the s largest permutation-based correlation coefficients assumes a value greater than the corresponding raw correlation coefficient. Given the form of hypothesis $H_0(s)$ in the proposed procedure for s linear correlation coefficients, the ASL score can be written as follows (Efron and Tibshirani, 1993):

$$ASL = P_{H_0}\left(R_1 \geq R_1^0 \vee \ldots \vee R_s \geq R_s^0\right)$$

and estimated as

$$ASL \approx \frac{card\{i : R_1^{(i)} \geq R_1^0 \vee \ldots \vee R_s^{(i)} \geq R_s^0\}}{N}.$$

The MCPerm procedure for identifying the maximum system of s simultaneously statistically significant correlations in the input k ($1 \leq s \leq k$) correlations is described by Algorithm 1.

The application of the classical significance test for the correlation coefficient requires the assumption of the normal distribution, but the adoption of such an assumption is often not unjustified in practical applications. The stepwise procedure proposed above

Algorithm 1. Pseudocode of the iterative MCPerm procedure at a fixed significance level α.

Require: $Y, X_1, \ldots, X_k$ {The data}
Ensure: : $n.sig$ {The number of significant correlations.}
1: $N = 999$ {The number of permutations.}
2: $\alpha = 0.05$ {The significance level.}
3: Calculate the k raw correlation coefficients and order their absolute values non-increasingly resulting in the following list: $R_1^0, R_2^0, \ldots, R_k^0(*)$.
4: Calculate the permutation-based correlation coefficients and order their absolute values non-increasingly resulting in the following structure: $R_1^{(i)}, R_2^{(i)}, \ldots, R_k^{(i)}(**)$, for $i = 1, 2, \ldots, N$.
5: Perform the steps of the procedure described below resulting in the number of significant correlations $n.sig$
6: $n.sig = 0$ {The number of significant correlations}
7: $w = 0$ {The percentage of permutations satisfying the key condition}
8: **for** $s \leftarrow 1$ **to** k **do**
9: $sig = 0$
10: **for** $i \leftarrow 1$ **to** N **do**
11: **if** at least one of the s largest permutation-based correlation coefficients in $(**)$ takes a value greater than or equal to the corresponding raw correlation coefficient in $(*)$ **then**
12: $sig = sig + 1$
13: **end if**
14: **end for**
15: $w = \frac{sig+1}{N+1}$
16: **if** $w < \alpha$ **then**
17: $n.sig = s$
18: **else**
19: **break**
20: **end if**
21: **end for**
22: **return** $n.sig$ {The number of significant correlations}

does not require assuming normality because it operates on the null distribution formed from permutations.

The proposed iterative permutation-based procedure MCPerm calculates in each s-th iteration (a step) the frequency w of permutations in which at least one of the s largest permutation-based correlation coefficients assumes a value greater than the corresponding raw correlation coefficient (key condition).

As the proposed iterative procedure (being a permutation test) performs testing that of only one (composite) null hypothesis based on the null distribution obtained from permutations, it ensures the size of the whole test is not higher than the adopted significance level α (Simon and Simon, 2011). Permutation tests are constructed so that their ASL values are evenly distributed on the unit interval under the null hypothesis. This means that, if the null hypothesis is true, the probability of having an ASL value less than or equal to α is just α. This ensures that the test has the right size at α, controlling the risk of an error of the first kind (Simon and Simon, 2011).

Table 1. Simulation scenarios for generating data.

Scenario	ρ_1	ρ_2	ρ_3	ρ_4
$A0$	0.00	0.00	0.00	0.00
$B2$	0.90	0.50	0.00	0.00
$C2$	0.70	0.20	0.00	0.00
$D3$	0.60	0.30	0.20	0.00
$E4$	0.40	0.30	0.20	0.10
$F4$	0.80	0.60	0.40	0.20

3. Monte Carlo study

The simulation analysis tested the significance of the relationship in a special correlation system, namely, between the explained variable Y and the four predictors X_1, X_2, X_3, X_4. The simulations included the following six variants of the multivariate normal distribution. Cases were considered where all variables were independent (variant A0), and there were two relationships (B2: two strong relationships and C2: one strong and one weak relationship). Also considered were cases where there were three relationships (D3: one strong and two weak relationships) and 4 relationships (E4: four weak relationships and F4: two strong and two weak relationships). Denoting $\rho(Y, X_i) = \rho_i$, these five random variables (Y, X_1, X_2, X_3 and X_4) were generated independently with the correlations described in Table 1. The six scenarios presented in Table 1 include the occurrence of two (B2 and C2), three (D3), or four (E4 and F4) dependencies. The analyses also included variant A0, where all variables were independent.

All simulations were carried out for sample sizes of $n = 40, 80, 120$, and 200. The number of permutations was assumed to be $N = 1000$. For each variant described above, the simulations calculated the number of significant indications (dependencies) by the proposed MCPerm method, as well as by the Holm approach (Holm, 1979) for comparison. It is noticeable that, especially for the large sample sizes, the number of significant dependencies indicated should be zero (variant A0), two (B2 and C2), three (D3) or four (E4 and F4). The detailed results from the conducted experiments are presented in Tables A1–A6 in Appendix. In all simulations, a significance level of $\alpha = 0.05$ was assumed. The graphical presentation of the obtained results is provided in the constructed modified sunflower plot (Fig. 1).

For variant A0 (all variables independent), across all sample sizes, both compared methods indicated the absence of a relationship between variables with similar efficiency (see Table A1 and Fig. 1). Holm's method was more effective when there were two strong relationships (variant B2, Table A2). Still, when there was one weak relationship (variant C2, Table A3), it was the MCPerm method that more often gave a correct indication of two relationships (for all sample sizes considered). The same situation occurred for the three dependencies (variant D3, A4). In the case of the simultaneous occurrence of four dependencies (variants E4 and F4, Tables A5 and A6) for all sample sizes considered, the proposed method was definitely more effective. Holm's approach proved to be more conservative in the presence of several dependencies and tended to indicate fewer of them (see Fig. 1, rows E4 and F4).

The advantages of the proposed method are particularly evident with a larger number of dependencies present (three or four), especially for small sample sizes (Fig. 1). For sample sizes of $n = 40, 80,$ and 120 in all cases (D3, E4, and F4), MCPerm indicated the correct number of correlations significantly more often than the Holm method (Fig. 1). Only for variant D3 with a sample size of $n = 200$ was the Holm method slightly more effective (Fig. 1).

4. Correlation study on data from the 2021 National Population and Housing Census using the MCPerm method

The essence of the proposed MCPerm method is illustrated by applying it in correlation analysis on data (`https://bdl.stat.gov.pl/bdl`) from the 2021 National Population and Housing Census (NSP2021) conducted in Poland from April 1st to September 30th, 2021. The census covered various types of demographic, social, and economic information, such as age, gender, education, occupational situation, housing conditions, nationality, language spoken at home, religious beliefs, etc. It included, among other things, the number of housing units and the population in housing units by area (less than 30 m², 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99, 100–109, 110–119, and 120 m² and more), by the number of persons per room (less than 0.5, 0.5–0.99, 1–1.49, 1.50–1.99, 2.00–2.99, 3.00 and more), and the population in housing units by area. In the conducted correlation analysis, we used the data on housing units and their population on a provincial basis in the year 2021. The following six variables were included in the analyses:

- X_1: the number of people in apartments with an area of 30 or more m² per population,
- X_2: the number of dwellings with an area of less than 30 m² per population,
- X_3: the number of apartments with an area of 120 m² and more per population,
- X_4: the number of dwellings with less than 0.5 persons per room per population,
- X_5: the number of dwellings with 3.0 or more persons per room per population,
- Y: the number of total housing units per population.

The analyzed subset of all measured variables of census data consisted of 16 observations (provinces). We are interested in finding the maximum system of statistically significant (simultaneous) correlations between the variable Y and each of the variables $X_1, X_2, X_3, X_4,$ and X_5. In the presented analyses, a significance level of $\alpha = 0.05$ was adopted.

The calculated values of the sample correlation coefficients between the variable Y and each of the variables $X_1, X_2, X_3, X_4,$ and X_5 are the following: $-0.01, 0.73, -0.71, 0.40, -0.08$ (Table 2 shows all correlation coefficients from the analyzed sample). The non-decreasingly ordered sequence of absolute values of those correlation coefficients is as follows: $0.73, 0.71, 0.40, 0.08, 0.01$.

Applying Holm's method to control for multiplicity identifies two statistically significant relationships: (Y, X_2) and (Y, X_3). The permutation-based MCPerm method identifies three simultaneously statistically significant pairs: $(Y, X_2), (Y, X_3),$ and $(Y, X4)$. This indicates that the simultaneous occurrence of three correlations $0.73, -0.71,$ and 0.40 constitutes a statistically significant pattern. It can be seen that the proposed method allowed for the detection of one additional statistically significant correlation compared to the traditional Holm method.

Thus, we can conclude that Y (the number of total dwellings/population) has the strongest positive correlation with X_2 (the number of dwellings of less than 30 m²/population) with a value of 0.73. This implies that in regions with a greater number of small dwellings (less than 30 m²), the total number of dwellings per capita is higher. The variables Y and X_3 (the number of dwellings of 120 m² or more/population) are strongly negatively correlated (-0.71), suggesting that the total number of dwellings per capita is lower in provinces where a more significant number of large dwellings predominates. There is a moderate positive correlation between Y and X_4 (the number of housing units with less than 0.5 persons per room/population), with a value of 0.40, suggesting that a more significant number of spacious housing units may be associated with a higher per capita housing ratio.

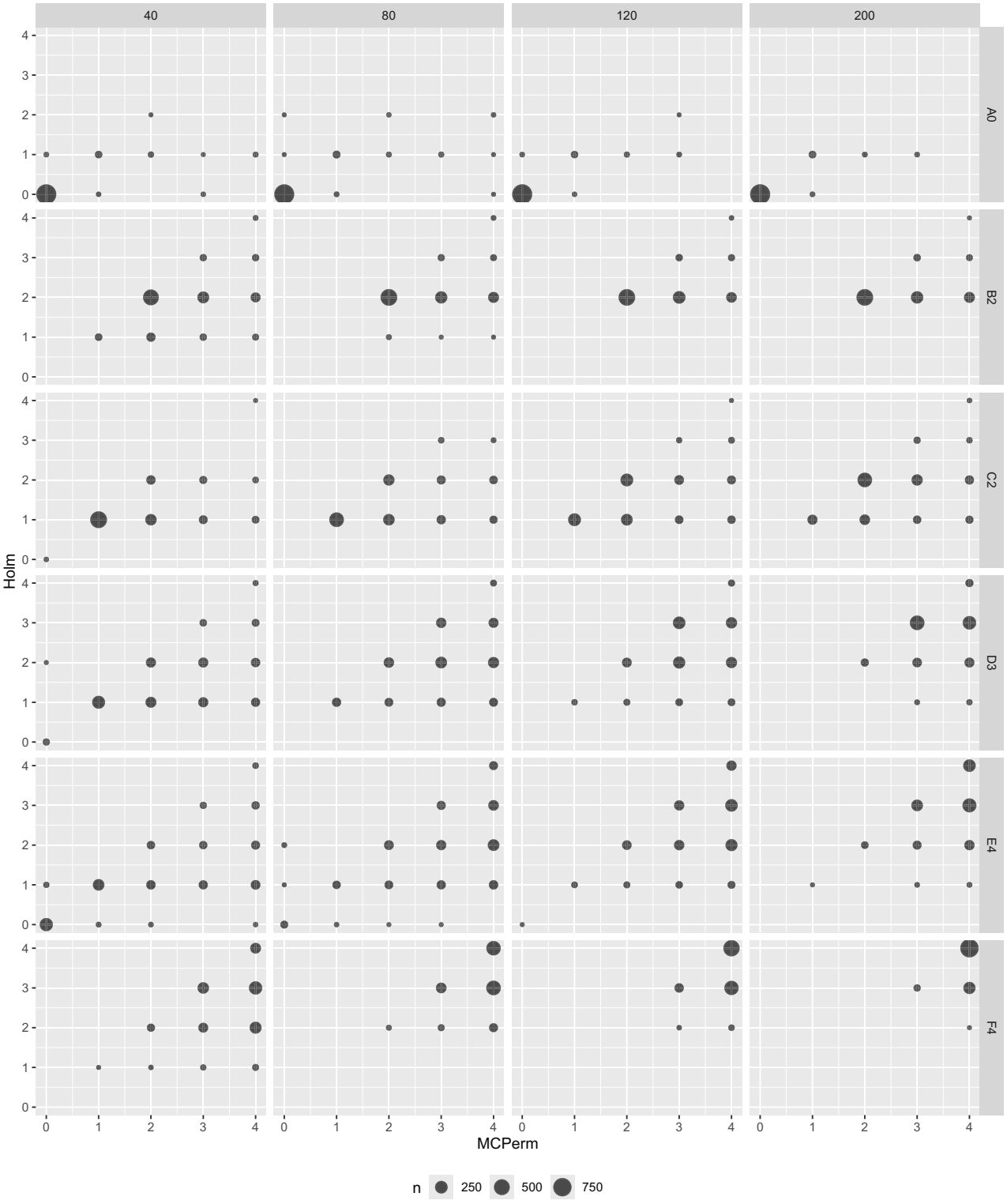

Fig. 1. Modified sunflower plot: the frequences of detecting the number of correlations for four sample sizes and six variants of variable correlations for the MCPerm and Holm methods.

Table 2. Matrix of correlation coefficients of the variables.

Variable	X_1	X_2	X_3	X_4	X_5	Y
X_1	1.00	−0.22	0.68	0.61	−0.36	−0.01
X_2	−0.22	1.00	−0.62	−0.05	0.41	0.73
X_3	0.68	−0.62	1.00	0.10	−0.13	−0.71
X_4	0.61	−0.05	0.10	1.00	−0.64	0.40
X_5	−0.36	0.41	−0.13	−0.64	1.00	−0.08
Y	−0.01	0.73	−0.71	0.40	−0.08	1.00

The detailed results for data correlations are presented in Table 2.

5. Discussion and conclusions

This paper presented a new approach to the problem of multiple hypotheses testing and application of the MCPerm method (being the implementation of this approach) to the problem of testing the significance of multiple correlations. Controlling the effect of multiple testing is undoubtedly necessary. Uncontrolled multiple testing leads to the detection of many completely random relationships. Such relationships are later often presented in scientific and popular science publications as interesting or even surprising research results. This, in turn, can arouse skepticism towards statistical methods, while the source of the misunderstanding is improperly conducted research, which does not take into account the effect of multiple testing. The purpose of the proposed MCPerm method is to effectively detect several co-occurring dependencies. Different methods will be effective in detecting strong dependencies. The proposed permutation method is expected to detect even weak dependencies effectively. As the permutation test verifying one global hypothesis, it is an alternative solution to the existing methods for testing a family of hypotheses. The presented simulation analysis for various systems of correlations and samples of different sizes shows in which situations the proposed method allows for more effective indication of the correct number of dependencies occurring simultaneously in the population. In simulation analyses, the effectiveness of the proposed method was compared with Holm's approach. In the absence of dependencies, both methods gave similar results. In the case of two strong dependencies, the Holm's method more often correctly indicates the number of existing dependencies. Still, for weak dependencies, the MCPerm approach was more effective for sample sizes of $n = 40, 60, 80$. The greatest benefits of the proposed method are observed when three or more relationships are present, regardless of their strength, especially for small or moderate sample sizes. Thus, the MCPerm approach can be recommended in such situations.

The application of the MCPerm method in a correlation study on real data from the 2021 National Population and Housing Census (NSP2021) in Poland showed its superiority over classical Holm's approach in detecting more statistically significant relationships. The idea behind the introduced method can be applied to develop similar procedures for simultaneous testing of multiple hypotheses (not only those concerning correlation coefficients) in medicine, biology or artificial intelligence and machine learning.

The problem of possible future adaptation of our new method to situations involving very large data sets (and/or number of variables) remains to be considered. For large datasets, the computational burden of permutation tests can be heavy, and the running time of naive implementations grows exponentially with the sample size. To manage this, several strategies can be employed. We will mention just some of them: parallel or distributed computing, Monte Carlo approximations, algorithm optimization (modern statistical software and libraries often provide optimized implementations of permutation tests), adaptive permutation procedures that can dynamically adjust the number of permutations based on the results, potentially saving computation time. feature pruning (in some cases, features that are deemed non-significant can be pruned from the analysis, reducing the computational burden).

In the future, we plan to adapt our method to problems involving large data sets. From the above mentioned possibilities, the most reasonable seems to be the use of the computational power of cost-efficient graphics processing units (GPUs), for example, based on the solutions described by Eklund *et al*. (2011) and Ekvall *et al.* (2020).

The next urgent goal is also a more comprehensive comparison of our new method with a larger number of representatives from both groups of multiple comparison methods: marginal and joint ones.

Code availability statement

The code for the MCPerm procedure and the example dataset are publicly available in our GitHub repository at `https://github.com/gkonczak/MCPerm`.

Acknowledgment

This research was supported trough statutory funds of the Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland (grant no. 02/100/BK_25/0044) and by statutory funds of the Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **57**(1): 289–300.

Berry, K.J., Johnston, J.E. and Mielke, P.W. (2018). *The Measurement of Association: A Permutation Statistical Approach*, Springer International Publishing, Cham.

Bonnini, S. and Cavallo, G. (2021). A study on the satisfaction with distance learning of university students with disabilities: Bivariate regression analysis using a multiple permutation test, *Statistica Applicata—Italian Journal of Applied Statistics* **33**(2): 143–162.

Bonnini, S., Corain, L., Marozzi, M. and Salmaso, L. (2014). *Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R*, John Wiley & Sons, Chichester.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets, *Journal of Machine Learning Research* **7**(1): 1–30.

Denkowska, S. (2013). Non-classical methods multiple testing procedures, *Statistical Review* **60**(4): 461–476 (in Polish).

Dickhaus, T. (2014). *Simultaneous Statistical Inference with Applications in the Life Sciences*, Springer, Berlin/Heidelberg.

Dudoit, S. and van der Laan, M.J. (2008). *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics, Springer, New York.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton.

Eklund, A., Andersson, M. and Knutsson, H. (2011). Fast random permutation tests enable objective evaluation of methods for single-subject FMRI analysis, *International Journal of Biomedical Imaging* **2011**: 1–15, Article ID: 627947, DOI: 10.1155/2011/627947.

Ekvall, M., Höhle, M. and Käll, L. (2020). Parallelized calculation of permutation tests, *Bioinformatics* **36**(22–23): 5392–5397.

Good, P.I. (2006). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer Science+Business Media, DOI: 10.1007/s00184-006-0088-1.

Hochberg, Y. and Tamhane, A.C. (1987). *Simultaneous Statistical Inference*, John Wiley and Sons, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**(2): 65–70.

Kowal, M., Skobel, M., Gramacki, A. and Korbicz, J. (2021). Breast cancer nuclei segmentation and classification based on a deep learning approach, *International Journal of Applied Mathematics and Computer Science* **31**(1): 135–153, DOI: 10.34768/amcs-2021-0007.

Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*, Wiley, Chichester.

Simon, R. and Simon, N.R. (2011). Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization, *Statistics & Probability Letters* **81**(7): 767–772.

Stapor, K., Ksieniewicz, P., Garcia, S. and Wozniak, M. (2021). How to design the fair experimental classifier evaluation, *Applied Soft Computing* **104**: 1–12

Toth, D. (2020). A permutation test on complex sample data, *Journal of Survey Statistics and Methodology* **8**(4): 772–791.

van Ginkel, J.R. (2019). Significance tests and estimates for R2 for multiple regression in multiply imputed datasets: A cautionary note on earlier findings, and alternative solutions, *Multivariate Behavioral Research* **54**(4): 514–529.

Westfall, P.H. and Young, S.S. (1993). *Resampling Based Multiple Testing*, Wiley, New York.

Zar, J.H. (2010). *Biostatistical Analysis*, Prentice-Hall/Pearson, Upper Saddle River.

Żelasko, D. and Pławiak P. (2021). Ensemble learning techniques for transmission quality classification in a Pay&Require multi-layer network, *International Journal of Applied Mathematics and Computer Science* **31**(1): 135–153, DOI: 10.34768/amcs-2021-0010.

Katarzyna Stapor is a full professor at the Department of Automatic Control, Electronics and Computer Science in the Silesian University of Technology in Gliwice, Poland. Her research interests are statistical pattern recognition, multivariate statistical analysis, protein bioinformatics, and computational neuroscience based on EEG analysis in particular. Professor Stapor has published more than 120 research papers in international/national journals and conferences, including one monograph on pattern recognition and the coursebook for a lecture on statistical methods. She is an active member of several professional scientific bodies.

Grzegorz Kończak holds an MSc in mathematics from the University of Silesia. He also holds PhD and DSc (habilitation) in economy from the Faculty of Management of the University of Economics. In 2018 he received the full professorial title. He is currently employed at the Department of Statistics, Econometrics and Mathematics of the University of Economics in Katowice. His areas of expertise include data analysis, statistical inference in economic research, Monte Carlo study, permutation tests and data visualization. He is the author or a co-author of more than 100 scientific publications, 13 books and dozens of unpublished works. He has participated in more than 80 scientific conferences and seminars at home and abroad.

Appendix
Results of a Monte Carlo study

Table A1. Estimated probabilities of indicating statistically significant correlations (A0).

Sample size n	Number of significant correlations: MCPerm method Holm method				
	0	1	2	3	4
40	0.951	0.031	0.010	0.003	0.005
	0.952	0.047	0.001	0.000	0.000
80	0.942	0.041	0.007	0.006	0.004
	0.945	0.050	0.005	0.000	0.000
120	0.959	0.029	0.007	0.005	0.000
	0.958	0.041	0.001	0.000	0.000
200	0.958	0.035	0.004	0.003	0.000
	0.961	0.039	0.000	0.000	0.000

Table A2. Estimated probabilities of indicating statistically significant correlations (B2).

Sample size n	Number of significant correlations: MCPerm method Holm method				
	0	1	2	3	4
40	0.000	0.030	0.572	0.246	0.152
	0.000	0.151	0.798	0.047	0.004
80	0.000	0.000	0.584	0.245	0.171
	0.000	0.007	0.951	0.039	0.003
120	0.000	0.000	0.572	0.272	0.156
	0.000	0.000	0.953	0.045	0.002
200	0.000	0.000	0.574	0.257	0.169
	0.000	0.000	0.954	0.045	0.001

Table A3. Estimated probabilities of indicating statistically significant correlations (C2).

Sample size n	Number of significant correlations: MCPerm method Holm method				
	0	1	2	3	4
40	0.002	0.588	0.257	0.104	0.049
	0.002	0.861	0.136	0.000	0.001
80	0.000	0.395	0.343	0.160	0.102
	0.000	0.698	0.287	0.015	0.000
120	0.000	0.261	0.447	0.160	0.132
	0.000	0.559	0.417	0.023	0.001
200	0.000	0.111	0.520	0.238	0.131
	0.000	0.336	0.630	0.032	0.002

Table A4. Estimated probabilities of indicating statistically significant correlations (D3).

Sample size n	The number of significant correlations: MCPerm method Holm method				
	0	1	2	3	4
40	0.023	0.269	0.261	0.247	0.200
	0.022	0.605	0.304	0.063	0.006
80	0.000	0.077	0.189	0.384	0.350
	0.000	0.278	0.476	0.230	0.016
120	0.000	0.010	0.117	0.491	0.382
	0.000	0.094	0.477	0.409	0.020
200	0.000	0.000	0.045	0.488	0.467
	0.000	0.012	0.242	0.703	0.043

Table A5. Estimated probabilities of indicating statistically significant correlations (E4).

Sample size n	Number of significant correlations: MCPerm method Holm method				
	0	1	2	3	4
40	0.296	0.188	0.139	0.148	0.229
	0.297	0.449	0.171	0.071	0.012
80	0.044	0.053	0.160	0.267	0.476
	0.043	0.274	0.409	0.208	0.066
120	0.001	0.014	0.109	0.268	0.608
	0.001	0.099	0.425	0.358	0.117
200	0.000	0.001	0.031	0.258	0.710
	0.000	0.008	0.220	0.528	0.244

Table A6. Estimated probabilities of indicating statistically significant correlations (F4).

Sample size n	Number of significant correlations: MCPerm method Holm method				
	0	1	2	3	4
40	0.000	0.001	0.046	0.291	0.662
	0.000	0.029	0.356	0.472	0.143
80	0.000	0.000	0.005	0.155	0.840
	0.000	0.000	0.094	0.528	0.378
120	0.000	0.000	0.000	0.086	0.914
	0.000	0.000	0.011	0.448	0.539
200	0.000	0.000	0.000	0.024	0.976
	0.000	0.000	0.001	0.244	0.755

Received: 21 February 2025
Revised: 14 June 2025
Re-revised: 3 August 2025
Accepted: 7 August 2025