Published in final edited form as:

Stat Probab Lett. 2009 March 1; 79(5): 664–669. doi:10.1016/j.spl.2008.10.015.

Permutation test for non-inferiority of the linear to the optimal combination of multiple tests

Hua Jin, Ph.D.¹ and Ying Lu, Ph.D.²

¹School of Mathematical Sciences, South China Normal University, Guangzhou, China

²Department of Radiology, Department of Epidemiology and Biostatistics, and UCSF Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA, U. S. A

Abstract

We proposed a permutation test for non-inferiority of the linear discriminant function to the optimal combination of multiple tests based on Mann-Whitney statistic estimate of the area under the receiver operating characteristic curve. Monte Carlo simulations showed its good performance.

Keywords

Permutation test; Non-inferiority; Receiver operating characteristic (ROC) curve; Likelihood ratio score; Mann-Whitney statistic

1. Introduction

Multiple diagnostic tests for one disease are commonly available to clinicians. Combination use of multiple tests is likely to substantially improve the statistical utility. Because sensitivity and specificity of a test depend on the threshold used to define abnormal, the receiver operating characteristic (ROC) curve is often used to assess the utility of a diagnostic test (e.g. DeLong, DeLong and Clark-Pearson, 1988). The area under the curve (AUC) is commonly used as an index of the diagnostic utility, with a larger AUC indicating a better test for discriminating between two populations.

Baker (1995) noted optimality of the likelihood ratio rule in statistical literature for combining multiple tests, and directly approximated the likelihood ratio function based on nonparametric estimation of the false positive rate and the true positive rate. McIntosh and Pepe (2002) proposed an alternative approach using standard binary regression.

From the point of theory, the likelihood ratio score leads to the uniformly most powerful test achieving the largest AUC (Neyman, and Pearson, 1933). However, it seems complex for practical use. One question is how to get good estimation of the likelihood ratio score from the observed sample when we know nothing about the distributions of the normal and abnormal populations. Furthermore, one may want to know whether a simple decision rule is non-inferior to the best. In this article, we propose a formal statistical framework to deal with this problem.

Corresponding Author: Hua Jin, Ph.D., School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China; e-mail: jinh1@163.com.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

In Section 2, we propose a permutation test for non-inferiority of a simple discriminant function to the optimal combination of multiple tests based on Mann-Whitney statistic estimate of the AUC. Section 3 conducts some simulation studies to evaluate the performance. Section 4 demonstrates an application of the proposed method to data from the Study of Osteoporotic Fractures. Discussion is presented in the last section.

2. Statistical methodology

Suppose we have K tests and/or risk factors with continuous measurements. Let D be the disease status indicator ($D=1$ for the diseased subjects and $D=0$ for the non-diseased). Denote the result of the k th test on the diseased subjects ($D=1$) by X_k , $k = 1, \dots, K$, and the vector of test results as $X = (X_1, \dots, X_K)$, and the vector of test results on the non-diseased ($D=0$) as $Y = (Y_1, \dots, Y_K)$. Furthermore, let $f_1(x)$ and $f_0(x)$ be their density functions, separately, where $x = (x_1, \dots, x_K)$. It

follows from the Neyman-Pearson lemma that the likelihood ratio $f(x) = \frac{f_1(x)}{f_0(x)}$ (or its any kind of monotonically increasing transformation) leads to the best combination of all available tests that results in a decision rule with the highest TPR for any given FPR among all possible rules based on the observed X , and thus provides the largest AUC, that is $AUC_f = P\{f(X) > f(Y)\}$. In fact, the optimum rule is $f > c(s_0)$ where $f = f(X)$ if $D=1$ and $f = f(Y)$ if $D=0$, and $c(s_0)$ satisfies $P\{f(Y) > c(s_0)\} = s_0$ so that the false positive rate is s_0 . McIntosh and Pepe (2002) called this rule the uniformly most sensitive test based on combination of multiple tests.

The likelihood ratio rule gives us the clue how to combine multiple tests to get the best decision rule in the sense of largest AUC. For example, suppose both $X = (X_1, \dots, X_K)$ and $Y = (Y_1, \dots, Y_K)$ are normally distributed multivariate random variables. If the two covariance matrices are equal, then the best rule reduces to the linear combination (Fisher, 1936). Otherwise it should be in quadratic form (Randles, Broffitt, Ramberg, et. al., 1978). However, in general case without multivariate normal assumption, we have no idea what the likelihood ratio score exactly looks like. We are interested in searching the recognizing pattern for easily clinical understanding and using.

In this section, we propose a formal hypotheses testing framework to deal with this problem. In clinical application, we wish the discriminant function as simple as possible. On the other hand, a special form of function may be used for construction of a decision rule from medical knowledge and/or clinical experience. For example, clinicians would prefer the linear discriminant (LD) function to construct decision rules. In general, one would like to focus on a special class of functions $G = \{g : g(x; \theta) \mid \theta \in \Theta\}$, where $g(x)$ is a pre-specific known function with the parameter θ belonging to a specific parameter space Θ , from which the best one $g(x; \theta_0)$ will be selected as the discriminant function. Suppose that θ_0 is fixed (Its estimation will be discussed in subsection 2.2). We want to know whether or not $g(x; \theta_0)$ can be used in place of the likelihood ratio $f(x)$ for easily clinical application. That is to say, we want to test the following hypothesis:

$$H_0: f(x) = h(g(x; \theta_0)) \quad (1)$$

where $h(\cdot)$ is a one-dimensional, unknown monotonically increasing function. In other words, we'd like to determine if $g(x)$ is a monotonically increasing transformation of the likelihood ratio.

It seems difficult to test the composite hypothesis (1). However, if we regard f and g (where $g = g(X; \theta_0)$ if $D=1$ and $g = g(Y; \theta_0)$ if $D=0$) as two new predictors, their corresponding ROC curves are identical if and only if there exists a one-dimensional monotonically increasing

function $h(\cdot)$ such that $f(x) = h(g(x; \theta_0))$. Testing hypothesis (1) is thus equivalent to comparing ROC curves.

There are two common approaches for comparing ROC curves. One is to fit a parametric model to the data, such as the binormal model, and then test the equality of the parameters (Metz, Wang and Kronman, 1984). Another is based on nonparametric tests, such as the permutation tests (Venkatraman, 2000).

An alternative method is to test the equality of a summary measure from the ROC curves, such as the AUC. Delong, Delong, and Clarke-Pearson (1988) refined the nonparametric area test (Hanley and McNeil, 1983) with a jackknife estimate of the variance for the AUC. Noting f is uniformly better than g , we see that the two ROC curves are identical if and only if the corresponding areas under the curves (AUCs) are equal. Thus, our testing problem can be reduced to that of comparing AUCs. Specifically, we want to test the null hypothesis $H_0 : AUC_g = AUC_f$, which is equivalent to

$$H_0: \delta = AUC_f - AUC_g = 0 \quad (2)$$

where $AUC_g = P\{g(X; \theta_0) > g(Y; \theta_0)\}$ is the AUC of g .

Test statistic

In this subsection, we give a nonparametric test statistic for null hypothesis in (2) in a similar way to Delong, Delong, and Clarke-Pearson (1988). Let $X^i = (X_1^i, \dots, X_K^i)$, $i = 1, \dots, m$, be the tests results of m diseased subjects, and $Y^j = (Y_1^j, \dots, Y_K^j)$, $j = 1, \dots, n$ be the results of n non-diseased. According to Delong, Delong, and Clarke-Pearson, a good estimate of AUC_f is given by

$$\widehat{AUC}_f = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(\widehat{f}(X^i) > \widehat{f}(Y^j)) \quad (3)$$

where $I(A) = 1$ if A is true and 0 otherwise, and $\widehat{f}(x)$ is a consistent estimate of $f(x)$. Similarly,

$\widehat{AUC}_g = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(g(X^i; \widehat{\theta}_0) > g(Y^j; \widehat{\theta}_0))$ is a good estimate of AUC_g , where $\widehat{\theta}_0$ is a consistent estimate of θ_0 . Then, one natural estimate of δ is

$$\widehat{\delta} = \widehat{AUC}_f - \widehat{AUC}_g \quad (4)$$

Estimation of $f(x)$ and θ_0

Estimation of $f(x)$ can be reduced to problem of estimating multivariate density functions. A lot of literature has focused on the topic of multivariate density estimation from three major approaches: parametric, nonparametric and semi-parametric (e.g., Silverman, 1989). In this paper, we propose to use the nonparametric method because of its appealing merit that does not need specific distributional assumptions. Here we use the fixed-width kernel density estimator to estimate the likelihood ratio score. Specifically, in its most general form, the fixed-width kernel density estimator of $f_1(x)$ is

$$\widehat{f}_1(x) = \frac{1}{mh_1^K} \sum_{i=1}^m W_1\left(\frac{x - X^i}{h_1}\right)$$

where the choice of kernel function W_1 and the window width h_1 determine the performance of $\widehat{f}_1$ as an estimator of f_1 . For simplicity, we take W_1 to be the Gaussian kernel

$$W_1(x) = (2\pi)^{-K/2} (\det H_1)^{-1/2} \exp(-x^T H_1^{-1} x/2) \quad (5)$$

and $H_1 = \text{cov}(X, X)$, which can be estimated by the sample covariance. This is motivated by a suggestion from Wand (1993) that one should have kernel mass with the same covariance structure as the density itself to optimally estimate a bivariate normal density. For Gaussian

kernel in (5), the optimal window width h_1^* is given by $h_1^* = \left(\frac{4}{(2K+1)m}\right)^{1/(K+4)}$. Similarly,

$f_0(x)$ can be estimated by $\widehat{f}_0(x) = \frac{1}{n(h_0^*)^K} \sum_{j=1}^n W_0\left(\frac{x - Y^j}{h_0^*}\right)$, where $h_0^* = \left(\frac{4}{(2K+1)n}\right)^{1/(K+4)}$, $W_0(x) = (2\pi)^{-K/2} (\det H_0)^{-1/2} \exp(-x^T H_0^{-1} x/2)$, and $H_0 = \text{cov}(Y, Y)$.

As for estimation of θ_0 , the parameters in g , we suggest to use logistic regression model:

$$P(D=1|x) = \frac{\exp(g(x; \theta_0))}{1 + \exp(g(x; \theta_0))} \quad (6)$$

For example, suppose that g is a linear function, i.e. $g(x) = \theta_0^T x$, we can easily get the coefficients' estimators from the linear logistic regression analysis. It is important to note that we do not need to be concerned with how well the data fits the logistic regression model (6). We just employ it to estimate θ_0 so that the resulted $g(x; \widehat{\theta}_0)$ is the best discriminant function among the class G that leads to the largest AUC. In fact, it follows from Bayes' rule and (6) that

$$\begin{aligned} g(x; \widehat{\theta}_0) &= \log f(x) + \log P(D \\ &= 1) - \log P(D=0) \end{aligned}$$

This expression shows that $g(x; \widehat{\theta}_0)$ is a monotone increasing function of the likelihood ratio $f(x)$, which should be the best discriminant function among the class G . However, the better model (6) fits, the closer the ROC curve of $g(x; \widehat{\theta}_0)$ is to $f(x)$, and the less chance to decline the null hypothesis (1).

Permutation reference distribution

In subsections 2.1 and 2.2, we have obtained a statistic $\widehat{\delta}$ in (4) to test the null hypothesis (1). It's time to determine its distribution to conduct the hypothesis test. However, neither $A\widehat{UC}_f$ nor $A\widehat{UC}_g$ is a generalized U -statistic, Delong, Delong, and Clarke-Pearson's method (1988) may not be applicable. In the following, we employ permutation method to give a reference distribution for our test statistic $\widehat{\delta}$.

We have $m + n$ observations $\{\hat{f}(X^i), i = 1, \dots, m; \hat{f}(Y^j), j = 1, \dots, n\}$ for the likelihood ratio $f(x)$.

Let $(R_{f1}, R_{f0}) = \{R_{f1}^i, i = 1, \dots, m; R_{f0}^j, j = 1, \dots, n\}$ denote the corresponding ranks. Since

ranking is a monotone increasing transformation, we have $A\widehat{UC}_f = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(R_{f1}^i > R_{f0}^j)$.

Similarly, Let $(R_{g1}, R_{g0}) = \{R_{g1}^i, i = 1, \dots, m; R_{g0}^j, j = 1, \dots, n\}$ be the corresponding ranks of the $m + n$ observations $\{g(X^i; \theta_0), i = 1, \dots, m; g(Y^j; \theta_0), j = 1, \dots, n\}$ of the supposed discriminant

function $g(x; \theta_0)$, then $A\widehat{UC}_g = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(R_{g1}^i > R_{g0}^j)$. So we can denote the test statistic as $\hat{\delta} = \hat{\delta}((R_{f1}, R_{f0}), (R_{g1}, R_{g0}))$.

The permutation procedure is like as follows. For the diseased subjects, we first pool the observed R_{f1} and R_{g1} , and then randomly assigning half of them as the permuted data R_{f1}^* of the predictor f , while others are left as the permuted data R_{g1}^* of the predictor g . Similarly, we can obtain the permuted predictor data R_{f0}^* and R_{g0}^* for the non-diseased. Thus, such a permutation step produces an observed permuted data $(R_{f1}^*, R_{f0}^*, R_{g1}^*, R_{g0}^*)$, resulting in a permuted estimate of the parameter δ , i.e. $\hat{\delta}^* = \hat{\delta}((R_{f1}^*, R_{f0}^*), (R_{g1}^*, R_{g0}^*))$. The permutation distribution of $\hat{\delta}^*$ can be obtained either by complete implementation of all possible permutations or by just sampling a large number of permutations. Since the distributions of R_{f1} and R_{g1} , and also R_{f0} and R_{g0} , are approximately identical under the null hypothesis (2), the distributions of $\hat{\delta}$ and $\hat{\delta}^*$ are also approximately identical. Therefore, the permutation distribution of $\hat{\delta}^*$ gives us a reference distribution for $\hat{\delta}$. The p -value of the test can be estimated by the ratio of the number of times $\hat{\delta}^* \geq \hat{\delta}$ to the total number of permutations.

Under large sample sizes as required in this paper in order to get consistent estimates of density functions, our proposed permutation test should work as well as the independent sample cases because the correlations within R_{f1} and R_{g1} tend to be zero if sample sizes go to infinity.

3. Monte Carlo simulations

We conduct two Monte Carlo simulations to assess the performance of the permutation test. One is under bivariate normal distributions, typical representative of continuous densities; another is based on gamma distributions, somewhat different from normal assumptions.

For the first simulation, we suppose $X = (X_1, X_2) \sim N(\mu_{11}, \mu_{12}, \sigma_{11}, \sigma_{12}, \rho_1)$ and $Y = (Y_1, Y_2) \sim N(\mu_{01}, \mu_{02}, \sigma_{01}, \sigma_{02}, \rho_0)$, where $\mu_{11}, \mu_{12}, \sigma_{11}, \sigma_{12}$, and ρ_1 are means, standard deviations and correlation of X_1, X_2 , and $\mu_{01}, \mu_{02}, \sigma_{01}, \sigma_{02}$, and ρ_0 are means, standard deviations and correlation of Y_1, Y_2 , respectively. We consider several different cases with different parameters setting, where m and n are sample sizes for the diseased and non-diseased groups with 50 and 50 as example of small samples, 100 and 100 as moderate samples. Here we focus on testing if LD works well as the best predictor and find out when LD can be used as an alternative to the optimum (OP). That is to say, we want to test the null hypothesis $H_0: \delta = A_{OP} - A_{LD} = 0$ vs $H_1: \delta \geq \delta_1$, where A_{OP} is the largest AUC for all possible combinations of multiple predictors, A_{LD} is the corresponding AUC of LD, and δ_1 is the pre-specified threshold of non-equivalence (i.e. LD will not be regarded as good as OP if their difference in AUCs is greater than δ_1). 1,000 times is repeated to derive the permutation reference distribution of the test statistic for each experimental condition.

Table 1 shows the rejection rate of the proposed test statistic when the null hypothesis is true (i.e. $H_0: \delta = A_{OP} - A_{LD} = 0$), where the two corresponding covariance matrices are assumed equal. Each experimental condition is repeated 500 times. The type I error rate is calculated as

the percentage of a test rejected the null hypothesis in 500 repetitions. The 95% confidence interval for type I error rate under the null in 500 repetitions is (0.031, 0.069). Cases that fall outside of this interval are shown with an asterisk.

In such cases with assumption of equal covariance matrices as shown in Table 1, the best rule reduces to LD. We can see that the proposed testing algorithm works very well in all cases studied for moderate sample size $m=n=100$, where the corresponding type I error rate for 500 repetitions is just in the 95% confidence interval, while the performance for relatively small sample $m=n=50$ is also not bad.

Table 2 considers the cases where the two corresponding covariance matrices are non-equal, where the best rules are the quadratic functions, and the corresponding areas under ROC curves are 0.10 larger than those of LD. In these settings, we assess the powers of the proposed test for the null hypothesis $H_0 : \delta = A_{OP} - A_{LD} = 0$ vs $H_1 : \delta \geq 0.1$. The simulation results show that we may have enough power ($\geq 90\%$) to detect the difference between LD and the best for moderate sample size $m=n=100$.

Table 1 and 2 also present the mean, bias and root of mean square error (RMSE) for 500 estimates of AUC of the optimal combination. The results show that the global bandwidth kernel estimators for the joint density functions of multiple predictors may lead to good estimations of the related AUCs.

Our second simulation is based on gamma distributions. We simulate Y_1 and Y_2 from independent standard normal distribution, and X_1 and X_2 from a mixture of two bivariate gamma distributions that we generate as follows. First, we generate a random variable ξ distributed with 0-1 distribution with $P(\xi = 0) = 0.6$. If $\xi = 0$, we generate two independent $X_1 \sim \text{Gamma}(1,5)$ and $X_2 \sim \text{Gamma}(5,5)$. Otherwise, we generate $X_1 \sim \text{Gamma}(1,1)$ and $X_2 \sim \text{Gamma}(2,2)$ independently.

In this case, LD cannot be the optimum. The true values of the AUC for OP and LD are respectively 0.84 and 0.74. We want to assess the power of the proposed algorithm for the null hypothesis $H_0 : \delta = A_{OP} - A_{LD} = 0$ vs $H_1 : \delta \geq 0.1$. The nonparametric estimator of AUC for OP is 0.849 (with SD=0.0454) for small sample size $m=n=50$, and 0.839 (SD=0.0315) for moderate sample size $m=n=100$. The power to detect difference between OP and LD is 0.674 and 0.882 for small and moderate sample sizes respectively. We can see that we have enough power ($> 88\%$) to know the AUC difference of 0.10 even with not large sample size $m=n=100$.

In summary, our test performs well under studied cases even with small sample sizes $m=n=50$.

4. An Example

In this section, we apply our method to the data from the Study of Osteoporotic Fractures (SOF) (Cummings, 1998). From the study, 43 variables, including patient demographic data, bone mineral density (BMD), have been identified as significant predictors of hip fracture risk, but only a few of them are necessary to identify subjects with elevated fracture risk. Standard approaches such as logistic regression and classification tree model show that age, the femoral neck BMD and loss of height are among the most important predictors. Our simulations only focus on combinations of two variables. Here we want to show that LD may be equivalent to OP when we just use these three specified markers to predict hip fracture.

The likelihood ratio score leads to the optimum discriminant function with the largest AUC 0.80. However, its explicit expression is unknown and so inconvenient for clinical application. Fortunately, the best linear combination from the logistic regression model happens to have the estimated AUC 0.80.

Among the included 7,112 subjects, there are only 229 women with hip fracture in 5 years. To show the robustness of our test, we construct a relatively small dataset that consists of the 229 fractured women and another 229 randomly sampled women without fracture. Based on the new data, we test if the AUC of the best linear function is equal to that of the optimal combination and calculate the corresponding p -value. We repeat this procedure for 100 times and the (empirical) 95% confidence interval of the p -value is (0.35,0.95), which suggests that LD be a good alternative to OP.

5. Discussion

It's important to use all the good diagnostic predictors simultaneously to establish a new predictor with better statistical utility. Clinicians usually search among linear functions and pay little attention to the null hypothesis if LD is non-inferior to OP. Our permutation test may answer this question. If we do not reject the null hypothesis, LD can be a reasonable alternative to OP in clinical use. The simple function g is not limited to linear or polynomial functions. Any function that can be interpreted and has a closed explicit form can be tested for non-inferiority to OP. If no simple meaningful function is found non-inferior to OP, it's better to use the likelihood ratio score for higher statistical efficiency.

To apply our test, it is critical to get good estimation of the multivariate densities. We employ the global bandwidth kernel estimator, which leads to satisfactory results presented in both simulation studies and real application to the SOF data. One difficulty is the increase of dimensionalities. Higher dimensionality will yield higher inflation of the type I error. For very high dimensions, one may need to do the factor analysis first and then model the non-parametric densities on the major factors. Other more efficient estimates of the multivariate densities need further research.

Our simulation studies suggest that small sample such as 100 may produce stable results. However, the required sample size will increase with the number of markers to be combined. As for 3-5 markers, our test needs more observations (300-500) to work well. These coincide with Gürlér and Prewitt's finding (2000), where both m and n selected to be 300 are considered as moderate sample sizes to get good estimators of density functions.

Acknowledgments

The study was supported by grants from the National Institutes of Health R03 AR47104, R01EB004079, and a grant from National Bureau of Statistics of China 2006B45. The authors would also like to thank the editor and two reviewers for their constructive comments which much improved the paper.

References

- Baker SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics* 1995;51:330–337. [PubMed: 7539300]
- Cummings S, Browner W, Bauer D, et al. Endogenous hormones and the risk of hip and vertebral fractures among older women. Study of Osteoporotic Fractures Research Group. *New England Journal of Medicine* 1998;339:767–768. [PubMed: 9731095]
- DeLong ER, DeLong DM, Clark-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics* 1988;44:837–845. [PubMed: 3203132]
- Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936;7:179–88.
- Gürlér Ü, Prewitt K. Bivariate density estimation with randomly truncated data. *Journal of Multivariate Analysis* 2000;74:88–115.
- Hanley J, McNeil B. A method of comparing the area under two ROC curves derived from the same cases. *Radiology* 1983;148:839–843. [PubMed: 6878708]

- McIntosh M, Pepe M. Combining several screening tests: optimality of the risk score. *Biometrics* 2002;58:657–664. [PubMed: 12230001]
- Metz, DE.; Wang, PL.; Kronman, HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconick, F., editor. *Information Processing in Medical Imaging*. Nijhoff; The Hague: 1984. p. 432-445.
- Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transaction of the Royal Society of London, Series A* 1933;231:289–337.
- Randles RH, Broffitt JD, Ramberg JS, et al. Generalized linear and quadratic discriminant functions using robust estimates. *J Amer Statist Assoc* 1978;73:564–568.
- Silverman, BW. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall; London: 1989.
- Venkatraman ES. A permutation test to compare receiver operating characteristic curves. *Biometrics* 2000;56:1134–1138. [PubMed: 11129471]
- Wand MP, Jones MC. Comparison of Smoothing Parameteizations in Bivariate Kernel Density Estimation. *J Amer Statist Assoc* 1993;88:520–528.

Table 1

Assessing the type I error of the proposed algorithm for the null hypothesis $H_0 : \delta = A_{OP} - A_{LD} = 0$

$(\mu_{11}, \mu_{12}, \sigma_{11}, \sigma_{12}, \rho_1)$	$(\mu_{01}, \mu_{02}, \sigma_{01}, \sigma_{02}, \rho_0)$	A_{OP}	(m, n)	mean	Bias	RMSE	type I error
(0.34, 0.34, 1, 1, 0.8)	(0.0, 1, 1, 0.8)	0.60	(50, 50)	0.648	0.048	0.046	0.078*
			(100, 100)	0.624	0.024	0.037	0.060
(0.64, 0.64, 1, 1, 0.5)	(0.0, 1, 1, 0.5)	0.70	(50, 50)	0.725	0.025	0.047	0.066
			(100, 100)	0.709	0.009	0.034	0.044
(0.92, 0.92, 1, 1, 0.2)	(0.0, 1, 1, 0.2)	0.80	(50, 50)	0.808	0.008	0.042	0.078*
			(100, 100)	0.805	0.005	0.029	0.056
(1.28, 1.28, 1, 1, 0)	(0.0, 1, 1, 0)	0.90	(50, 50)	0.904	0.004	0.030	0.066
			(100, 100)	0.902	0.002	0.022	0.056

Table 2

Assessing the powers of the proposed algorithm for the hypothesis $H_0 : \delta = A_{OP} - A_{LD} = 0$ v.s. $H_1 : \delta \geq 0.1$

$(\mu_{11}, \mu_{12}, \sigma_{11}, \sigma_{12}, \rho_1)$	$(\mu_{01}, \mu_{02}, \sigma_{01}, \sigma_{02}, \rho_0)$	A_{OP}	(m, n)	mean	Bias	RMSE	power
(0.3, 0.3, 1, 1, 0.7)	(0, 0, 1, 1, 0.2)	0.70	(50, 50)	0.721	0.021	0.049	0.66
			(100, 100)	0.706	0.006	0.034	0.92
(0.63, 0.63, 1, 1, 0.8)	(0, 0, 1, 1, 0.05)	0.80	(50, 50)	0.816	0.016	0.041	0.90
			(100, 100)	0.813	0.013	0.030	0.998
(1.01, 1.01, 1, 1, 0.9)	(0, 0, 1, 1, 0)	0.90	(50, 50)	0.907	0.007	0.029	0.988
			(100, 100)	0.904	0.004	0.020	1