South African Statistical Journal
Vol. 57, No. 1, pp. 1–12
https://doi.org/10.37920/sasj.2023.57.1.1

Some permutation symmetric multiple hypotheses testing rules under dependent setup

Anupam Kundu[1] *and Subir Kumar Bhandari*[2]

[1]Yale School of Public Health
[2]Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata

The problem of multiple hypothesis testing with correlated test statistics is a very important problem in statistical literature. Specifically, we consider the case when the joint distribution of the test statistics is a multivariate normal distribution with an unknown mean vector and compound symmetric correlation structure. Our goal is to identify nonzero entries of the mean vector. Bogdan et al. (2011) solved this problem when test statistics are independent normals along with the study of asymptotic optimality in a Bayesian decision theoretic sense. The case under dependence was left as a challenging open problem. The solution is intuitive and permutation invariant, does not assume sparsity unlike Bogdan et al. (2011) and is validated through simulation studies.

Keywords: Clustering, Multiple hypothesis testing, Permutation invariant, Subset selection.

1. Introduction

Multiple hypothesis testing has emerged as a very important topic of research in the last twenty years. The biggest impetus to such work came from the necessity to analyse and draw inference on data sets involving a large number of parameters. Such data sets occur, e.g., in the fields of biology, astronomy, and economics, just to name a few. Needless to say, the goal of simultaneous testing, or for that matter, simultaneous inference in general, is to ensure a good performance of the overall inference.

Over the years, various performance evaluation criteria have been developed to quantify the overall error in a simultaneous testing procedure. The most classical measure of this kind is the family-wise error rate (FWER). Well-known procedures that control the FWER are the Bonferroni procedure and its improvements, see for example, Holm (1979), Simes (1986), Hommel (1988), and Benjamini and Hochberg (1995). A nice historical account of the early works in this area can be found in Hochberg and Tamhane (1987). A great leap forward in the field of simultaneous inference was made through the introduction of the concept of false discovery rate (FDR) and a procedure called the Benjamini–Hochberg procedure. These appeared in the seminal work by Benjamini and Hochberg in 1995. FDR is obviously the more appropriate error to control in large-scale simultaneous testing problems compared to the FWER, since trying to control the probability of a single erroneous rejection seems

Corresponding author: Anupam Kundu (anupam.kundu@yale.edu)
MSC2020 subject classifications: 62F03, 62G10, 62H15

too stringent a requirement in such cases. See, for example, Benjamini and Liu (1999), Sarkar (2007), Storey (2002), and Storey et al. (2004) for further details. An excellent account on the literature on FDR can be found in Sarkar (2008) and Efron (2012)

The degree of "surprise" required in the observed data to declare a particular hypothesis to be false in a multiple hypothesis testing context should be more than what would be required to reject a hypothesis in an individual testing problem. The examples given above of multiple testing procedures belong to the frequentist domain. For the Bayesian, it is intuitive to reject a hypothesis when it is less likely aposteriori. The articles Scott and Berger (2006) and Scott and Berger (2010) beautifully explain this insight and explicitly demonstrate multiplicity adjustment in multiple testing through such Bayesian hierarchical modelling. See Bogdan et al. (2011) and Bogdan et al. (2008) for examples of Bayesian multiple testing rules derived as optimal rules with respect to an additive loss functions which are further discussed in Section 2. For other Bayesian decision theoretic approaches, see, e.g. Müller et al. (2004), Sun and Cai (2009) and Ghosh (2017).

The above examples of multiple hypothesis testing procedures are under the assumption of independence of the test statistics for the individual tests. However, in practice test statistics may often be dependent. It has been observed that when the procedures intended for the independent setup are applied unaltered under dependence, a lot of undesirable things can creep in and the performances of these procedures greatly suffer. See in this context, e.g., Qiu et al. (2006) and references of Cohen and Sackrowitz (2007) for further details. Although these issues have been raised, they have not been adequately resolved in the literature and the area of multiple testing under dependence is still very open to say the least.

The above works on dependence do not focus on the decision theoretic aspect of multiple testing. This aspect has been largely ignored except for some references like Sun and Cai (2009), Xie et al. (2011), Cohen and Sackrowitz (2005), Cohen and Sackrowitz (2007), and Cohen and Sackrowitz (2008). In Sun and Cai (2009) a decision theoretic study was carried out when the unknown parameters are assumed to be random in nature with a Markovian dependence structure among them. Under the dependent setup, in multiple hypothesis testing procedures there are some methods for estimating FDP, see e.g. Fan and Han (2017), Fan et al. (2017). In these methods it is assumed that the dependency comes into play in the form of some common factors. These methods perform well in the presence of the factor type dependence setup. There are some other methods by Efron (e.g. Efron, 2007; Efron, 2010), where the z-scores are transformed into count data. This translates the problem to the estimation of distribution of the correlations. But still these methods are very problem specific and mostly perform under the assumption of sparsity.

A natural question is what would the optimal rule (Bayes Oracle) look like under an arbitrary form of dependence among test statistics and what would be its asymptotic risk properties in the asymptotic framework of Bogdan et al. (2011) and Bogdan et al. (2008). This in itself is a very challenging problem under dependence and the reason will become clear shortly. This was left as an open problem in Bogdan et al. (2011). Ours is a modest attempt to work in the direction of this challenging problem. We restrict ourselves to the setup where the test statistics jointly have a mixture multivariate normal distribution in a permutationally invariant setup. We further assume that the parameters have a joint multivariate normal distribution, given values of a vector of independent Bernoulli random variables; see Section 2. With respect to additive loss, the general form of the Bayes Oracle is very easy to derive even under dependence, but it is intractable. We tried to view

the problem in a different light. An algorithm is proposed which converges to a rule close to the ideal classifier. Since the signals and noises differ in variances, it is intuitive to propose some random threshold for separating observations with different variances. The risk from the additive loss function is tractable only through approximations. For a constant threshold C, the approximated risk becomes a continuous function of C and can be minimised mathematically.

We need an ideal classifier or 'Oracle'-type rule to provide a lower bound on the misclassification risk in the dependent setup to assess the performance of our methods, but it is hard to find due to the intractability of the naive rule; see (3). The classifier which minimises the total error, i.e. the sum of false positives and false negatives, is calculated by a grid search technique; see Section 3.4. This experiment is repeated several times and averaged to get the 'optimal' risk. Though some of our methods perform close to the ideal classifier, the ideal risk is a non-achievable lower bound even in the limit. The paper is organised as follows: In Section 2 we introduce the problem and Section 3 shows our attempts to find an easy-to-compute approximately 'optimal' rule. The methods are assessed through simulation studies in Section 4.

2. Description of the problem

Bogdan et al. (2011) deal with independent normal observations an independent normal prior. We expand the problem in the normal setup with a compound symmetric correlation structure. Suppose we have m observations $X_1, X_2, ..., X_m$ such that $\boldsymbol{X} = (X_1, X_2, ..., X_m)^\top$ and

$$\boldsymbol{X} \mid \boldsymbol{\mu} \sim N_m\left(\boldsymbol{\mu}, \sigma_\varepsilon^2 \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ represents the unknown effects and $\sigma_\varepsilon^2\boldsymbol{\Sigma}$ represents the variability of the random noise (e.g. measurement error). We assume that $\sigma_\varepsilon^2 > 0$ is known and $\boldsymbol{\Sigma}$ is a compound symmetric matrix with a known ρ. The vector of unknown effects $\boldsymbol{\mu}$ is random and its distribution is determined by the values of m unobserved independent Bernoulli(p) random variables ν_i, for some $p \in (0, 1)$. We call $H_{0i} : \nu_i = 0$ and $H_{1i} : \nu_i = 1$. We assume that, given $\boldsymbol{\nu} = \boldsymbol{\nu_0} = (\nu_{01}, \nu_{02}, \ldots \nu_{0m})^\top$, the different components of $\boldsymbol{\mu}$ are correlated (with the common correlation ρ the same as $\boldsymbol{X}$) and have the following distribution:

$$\boldsymbol{\mu}|(\boldsymbol{\nu} = \boldsymbol{\nu_0}) \sim N(0, \boldsymbol{D}_{\boldsymbol{\nu}_0}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{\nu}_0}),$$

where $\boldsymbol{D}_{\boldsymbol{\nu}_0}$ is a diagonal matrix with

$$\left(\boldsymbol{D}_{\boldsymbol{\nu}_0}\right)_{ii} = \begin{cases} \sigma_0 & \text{if } \nu_{0i} = 0, \\ \sqrt{(\sigma_0^2 + \tau^2)} & \text{if } \nu_{0i} = 1. \end{cases}$$

This implies

$$\mu_i \sim \begin{cases} N(0, \sigma_0^2) & \text{under } H_{0i}, \\ N(0, \sigma_0^2 + \tau^2) & \text{under } H_{1i}. \end{cases}$$

The interpretation is that for small $\sigma_0^2 > 0$ and substantially large $\tau^2 > 0$, the H_{0i} correspond to the insignificant signals or noises and the H_{1i} correspond to important signals. The problem of

identifying components of $\boldsymbol{\mu}$ corresponds to the signals is equivalent to test $H_{0i} : \nu_i = 0$ versus $H_{1i} : \nu_i = 1$ simultaneously for $i = 1, \ldots, m$. It is evident that when $\sigma_0^2 = 0$, H_{0i} corresponds to the point null hypothesis $\mu_i = 0$ and H_{1i} implies $P(\mu_i = 0) = 0$, which is equivalent (when $\sigma_0 = 0$) to the canonical testing problem of $\mu_i = 0$ versus $\mu_i \neq 0$.

We define $P(\boldsymbol{\nu} = \boldsymbol{\nu}_0) = p_{\boldsymbol{\nu}_0} = p^{||\boldsymbol{\nu}_0||}(1-p)^{m-||\boldsymbol{\nu}_0||}$ where $||\boldsymbol{\nu_0}||$ = number of 1s in the vector $\boldsymbol{\nu_0}$. This provides the marginal distribution of $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} \mid \boldsymbol{\nu}_0 \sim N\left(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{\Sigma} + \boldsymbol{D}_{\boldsymbol{\nu}_0} \boldsymbol{\Sigma} \boldsymbol{D}_{\boldsymbol{\nu}_0}\right), \qquad \boldsymbol{X} \sim \sum_{\boldsymbol{\nu}_0} p_{\boldsymbol{\nu}_0} N(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{\Sigma} + \boldsymbol{D}_{\boldsymbol{\nu}_0} \boldsymbol{\Sigma} \boldsymbol{D}_{\boldsymbol{\nu}_0}). \tag{1}$$

The marginal distribution of $\boldsymbol{X}$ is not compound symmetric but the decision problem is still permutation invariant. Intuitively, we should reject the ith null hypothesis, i.e. $\nu_i = 1$ if $|X_i|$ is greater than some symmetric function of the observations (e.g. T_i in Section 3) including the constant function.

We should look for a procedure that identifies the signals (big signals) from the noises (insignificant signals) while reducing the expected loss. The chosen loss function is an additive one that defines the overall loss as the sum of the losses incurred in the individual testing problems. The simplest loss of this kind is the sum of the total number of type I and type II errors made by a multiple testing rule, originally proposed in Lehmann (1957a,b) and later considered by many others. See in this context Sun and Cai (2007), Bogdan et al. (2011), Bogdan et al. (2008), Datta and Ghosh (2013), and Sun and Cai (2009).

We say that a loss of $\delta_0 \in \mathbb{R}^+$ is incurred for the i-th testing problem when H_{0i} is true but it is rejected, i.e, an error of type I is made. A loss of $\delta_1 \in \mathbb{R}^+$ is said to be incurred for the i-th problem when an error of type II occurs in that problem. Here δ_0, δ_1 might depend on m. The overall loss of a multiple testing procedure is

$$L\left[\boldsymbol{\nu}(\boldsymbol{X}), \boldsymbol{\nu}\right] = \sum_{i=1}^{m} \delta_i \left(\nu_i(\boldsymbol{X}) - \nu_i\right)^2,$$

where $\boldsymbol{\nu}$ denotes the true value and $\boldsymbol{\nu}(\boldsymbol{X}) = (\nu_1(\boldsymbol{X}), \ldots, \nu_m(\boldsymbol{X}))^\top$ represents the corresponding random binary vector indicating the decisions obtained from a multiple testing procedure. More precisely, $\nu_i(\boldsymbol{X}) = 0$ if the multiple testing rule accepts H_{0i} and $\nu_i(\boldsymbol{X}) = 1$ if H_{0i} is rejected. Thus, $\delta_i = \delta_0$ when $\nu_i = 0$ but $\nu_i(\boldsymbol{X}) = 1$, whereas $\delta_i = \delta_1$ when $\nu_i = 1$ but $\nu_i(\boldsymbol{X}) = 0$. The Bayes risk is defined as $R_m = E[L(\boldsymbol{\nu}(\boldsymbol{X}), \boldsymbol{\nu})]$, where E denotes expectation with respect to the joint distribution of $(\boldsymbol{X}, \boldsymbol{\nu})$. It follows easily that

$$R_m = E^{\boldsymbol{\nu}} E[L(\boldsymbol{\nu}(\boldsymbol{X}), \boldsymbol{\nu}) | \boldsymbol{\nu}] = \sum_{i=1}^{m} \left[\delta_0 (1-p) t_{1i} + \delta_1 p t_{2i}\right], \tag{2}$$

where t_{1i} and t_{2i} denote the probabilities of type I and type II errors incurred for the i-th testing problem.

It may be noted that in our setup, the parameter space, the marginal distribution and conditional distribution of $(X_1, X_2, \ldots, X_m)^\top$ remain invariant with respect to permutations, which tells us to consider permutation invariant tests. This immediately implies that $t_{1i} = t_1$ and $t_{2i} = t_2$ for all i. Applying this, the risk becomes

$$R(\boldsymbol{\nu}, \boldsymbol{\nu}^*) = m\left[\delta_0 (1-p) t_1 + \delta_1 p t_2\right].$$

Our goal would be to minimise $[\delta_0(1-p)t_1 + \delta_1 p t_2]$ among permutation invariant tests to obtain a good approximate rule in this case.

It is easy to see that for this additive loss function, the optimal multiple testing rule is the one which simply applies the Bayes rule (with respect to the given δ_0, δ_1 losses) for each individual test and is given by

$$\text{Reject } H_{0i} \text{ if } \frac{f(\boldsymbol{X}|\nu_i = 1)}{f(\boldsymbol{X}|\nu_i = 0)} > \frac{(1-p)}{p}\frac{\delta_0}{\delta_1}, \text{ and accept it otherwise,} \tag{3}$$

for each $i = 1, \ldots, m$, where $f(\boldsymbol{X}|\nu_i = j)$ is the marginal density of $\boldsymbol{X}$, where $\nu_i = j$ for $j = 0, 1$. However, this rule is hard to implement because of the mathematical intractability of type I and type II errors, even asymptotically. The conditional densities of $f(\boldsymbol{X}|\nu_i = 0)$ and $f(\boldsymbol{X}|\nu_i = 1)$ are given from the mixture densities $\sum_{\boldsymbol{\nu}:\nu_i=0} f(\boldsymbol{X}|\boldsymbol{\nu})p_{\boldsymbol{\nu}}$ and $\sum_{\boldsymbol{\nu}:\nu_i=1} f(\boldsymbol{X}|\boldsymbol{\nu})p_{\boldsymbol{\nu}}$, respectively, where $f(\boldsymbol{X}|\boldsymbol{\nu})$ follows (1). The naive approach of trying to get directly the optimal rule will get us nowhere.

We observe that for both $\nu_i = 0$ and $\nu_i = 1$, the X_i come from normal distributions with zero means, but different variances, namely $\sigma_\epsilon^2 + \sigma_0^2$ and $\sigma_\epsilon^2 + \sigma_0^2 + \tau^2$. For large τ^2, the variance under $\nu_i = 1$ is much larger compared to σ_0^2 and σ_ϵ^2. Therefore an intuitive approach will be to look for coordinates of the observed data points with a large variance in order to reject the null hypothesis. The following lemma, to be proved in the appendix, would indicate that this amounts to looking for the X_i for which X_i^2 are the largest.

Lemma 1.

(a) Let $\boldsymbol{X} = (X_1, X_2, \ldots X_m)^\top$ follow a multivariate normal distribution with correlation matrix $\boldsymbol{R}$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$, respectively. Then $X_i^2 \leq_{st} X_j^2$ if and only if $\sigma_i^2 \leq \sigma_j^2$.

(b) Under the assumption of part (a), with compound symmetric correlation matrix $\boldsymbol{R}$, $X_i^2|\boldsymbol{Z} \leq_{st} X_j^2|\boldsymbol{Z}$ if and only if $\sigma_i^2 \leq \sigma_j^2$, where $\boldsymbol{Z}$ is a subset of $\{X_1, X_2, \ldots, X_m\}$ not containing X_i and X_j.

(c) Let $\boldsymbol{X} \sim N_m\left(\boldsymbol{0}, \sigma_\varepsilon^2\boldsymbol{\Sigma} + \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}\right)$, with $\boldsymbol{\Sigma}$ compound symmetric and $\boldsymbol{D} = diag\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\right)$. Then $X_i^2|\boldsymbol{Z} \leq_{st} X_j^2|\boldsymbol{Z}$ if and only if $\sigma_i^2 \leq \sigma_j^2$, where $\boldsymbol{Z}$ is a subset of $\{X_1, X_2, \ldots, X_m\}$ not containing X_i and X_j.

Note: The notation $X \leq_{st} Y$ means Y is stochastically larger than X.

The discussion above and the lemma together show that the tests corresponding to the highest K ordered statistics of the X_i^2 will be rejected. However, the question of finding an 'optimal' K remains to be answered. Put another way, this is equivalent to finding an 'optimal' threshold C (fixed or data dependent) such that we reject H_{0i} whenever $|X_i| > C$; see Section 3. The random classifier C must be a symmetric function of $X_1, X_2, \ldots, X_m$ under the permutation invariant rules.

3. Various approaches

Identification of signal from noises can be viewed as a clustering problem with two clusters. We use the risk minimisation as an optimal way to find a classifier for multiple testing. The classifier C (may be random) divides the X_i^2 into two groups which in turn rejects the corresponding tests. The methods to choose C are discussed next.

3.1 Algorithm

Our goal is to find an algorithm that gradually converges to an 'optimal' C. Let us define three quantities first.

$$T_1 = \left(\frac{\sum_{i=1}^m |X_i|}{m}\right), \quad T_2 = \left(\frac{\sum_{i=1}^m X_i^2}{m}\right)^{\frac{1}{2}}, \quad T_4 = \left(\frac{\sum_{i=1}^m X_i^4}{m}\right)^{\frac{1}{4}}.$$

We propose an iterative algorithm for determining the classifier C, which works good with reasonable error of false positive and false negative. The iteration is as follows:

1. **Initialisation:** Start with $Z_0 = T_i$ for some $i \in \{1, 2, 4\}$. Classify the vector of coordinate-wise absolute value of X with this classifier.
2. **Loop:** The coordinates of $\boldsymbol{X}$ for which the corresponding absolute values are less than Z_0 and those which are greater than Z_0 form two groups. Call the group means A_1 and A_2, respectively, and obtain $Z_1 = (A_1 + A_2)/2$.
3. Go to Step 2 with Z_1 and obtain $Z_2, Z_3, \ldots$, respectively.
4. **Termination:** Terminate the process in the ith step if $|Z_{i+1} - Z_i| < f$, where f is a predetermined very small tolerance value.

The simulation study in Section 4 shows that the sequence $\{Z_n\}_{n\geq 0}$ converges in general and the limit of the Z_i is the 'optimal' C. Different initial statistics maintain homogeneity among the coordinates because observations only differ in variability. The following result shows the rationale behind proposing this algorithm.

Result 1. Let $w_1, w_2, \ldots, w_m$ be m positive observations. Then the within-group variance

$$V_w(C) = \sum_{w_i \leq C} (w_i - \bar{w}_1)^2 + \sum_{w_i > C} (w_i - \bar{w}_2)^2$$

is minimised for a value of C which satisfies $C = (\bar{w}_1 + \bar{w}_2)/2$, with $\bar{w}_i$ giving the mean of the ith group.

The algorithm, in each step, reduces the within-group variance and forces the limiting classifier to divide the data into two clusters, leading to less expected misclassification.

3.2 Proposed random classifier

Let us define the following general quantity

$$T_{2h} = \left(\frac{\sum_{i=1}^m X_i^{2h}}{m}\right)^{\frac{1}{2h}}.$$

Since the main interest of the observations is their measure of variability, any even-powered moment is a potential choice for the random classifier. The $2h$th root is taken to maintain homogeneity among the observations. We can find the $h = h_{opt}$ that minimises the total misclassification by brute force. We can also use any T_{2h} as our initialisation in the algorithm in Section 3.1 for various values of h.

3.3 Risk minimisation C

We will provide an almost optimal rule under additive loss within the class of permutation invariant rules. Consider a multiple testing rule that rejects H_{0i} whenever $\{|X_i| > C\}$ for $i = 1, \ldots, m$, where C is a constant. The risk is given by

$$R(C) = \delta_0 m(1-p)P\left[|Y_1| > \frac{C}{\sqrt{\sigma_\epsilon^2 + \sigma_0^2}}\right] + \delta_A mpP\left[|Y_1| < \frac{C}{\sqrt{\sigma_\epsilon^2 + \sigma_0^2 + \tau^2}}\right], \tag{4}$$

where Y_1 is standard normal random variable. This follows from the fact that the errors are identical for each i.

When the X_i and μ_i are independent for $i = 1, 2, \ldots, m$, the optimal test has a rejection region based on a threshold that is a function of the model parameters but is independent of the observations. This rule is called a Bayes Oracle. Bogdan et al. (2011, Eq. 2.4 §2) calculated the asymptotic risk of the Bayes Oracle.

Ideally we want to extend the study of optimality in the dependent setup. However, due to the intractability of the risk, our optimal test (the Bayes Oracle) is not a simple thresholding rule, unlike the independent case. The risk function for the fixed threshold test depends only on the marginal distribution of the observations. If the marginal distributions of dependent and independent cases remain identical, the risk does not change either. We can minimise the risk with respect to C to obtain a critical region for the dependent case. The implication is that the optimal fixed threshold rule for the independent case is also optimal among fixed threshold rules in the dependent case.

We restrict ourselves to providing a heuristic approximation to the exact asymptotic risk. The expression of the risk of a fixed threshold in (4) can be approximated in the following way. Assuming $C/\sqrt{\sigma_\varepsilon^2 + \sigma_0^2}$ is large and $C/\sqrt{\sigma_\varepsilon^2 + \sigma_0^2 + \tau^2}$ is small, we have

$$\begin{aligned} R(C) &\approx \delta_0 m(1-p)\frac{\sqrt{2(\sigma_0^2 + \sigma_\varepsilon^2)}}{C\sqrt{\pi}} e^{-\frac{C^2}{2(\sigma_0^2+\sigma_\varepsilon^2)}} + \delta_1 mp \frac{C\sqrt{2}}{\sqrt{\pi(\sigma_0^2 + \sigma_\varepsilon^2 + \tau^2)}} \\ &= \frac{V}{C} e^{-\frac{C^2}{2(\sigma_0^2+\sigma_\varepsilon^2)}} + UC, \end{aligned} \tag{5}$$

where $V = \delta_0 m(1-p)\sqrt{\sigma_0^2 + \sigma_\varepsilon^2}\sqrt{2/\pi}$ and $U = \delta_1 mp\sqrt{2}/\sqrt{\pi(\sigma_0^2 + \sigma_\varepsilon^2 + \tau^2)}$. For the first summand, we exploit the fact that $C/\sqrt{\sigma_\varepsilon^2 + \sigma_0^2}$ is large and employ the standard approximation to normal tails using Mill's Ratio. For the second summand, we note that since $C/\sqrt{\sigma_\varepsilon^2 + \sigma_0^2 + \tau^2}$ is small and for small x, $P[|N(0,1)| < x] \approx 2x\phi(0)$. The risk function $R(C)$ in (5) is a convex function of C, since

$$R'(C) = U - Ve^{-aC^2}\left(\frac{1}{C^2} + 2a\right) = U - 2aVe^{-aC^2} - O\left(\frac{1}{e^{aC^2}C^2}\right),$$

where $a = 1/(2(\sigma_0^2 + \sigma_\varepsilon^2))$, $R'(0) > 0$ and $R'(C)$ is an increasing function of C (as $R''(C) > 0$). By

ignoring the third term in the expression of $R^{'}(C)$, the approximate 'optimal' C becomes

$$C = \sqrt{\frac{1}{a} log\left(\frac{2aV}{U}\right)} = \sqrt{2(\sigma_0^2 + \sigma_\varepsilon^2) \log\left(\frac{\delta_0(1-p)}{\delta_1 p}\sqrt{\left(1 + \frac{\tau^2}{\sigma_0^2 + \sigma_\varepsilon^2}\right)}\right)}. \tag{6}$$

3.4 Ideal classifier

It is hard to evaluate the performance of our methods without any standard oracle rule. We use a grid search to choose an ideal C and use it as a benchmark.

Let $|X|_{(j)}$ denote the jth order statistic of the absolute value of the coordinates of the $\boldsymbol{X}$ vector. Starting with $C^{(0)} = |X|_{(1)} - 1$, we classify the data, compute the error (sum of false positive and false negative) and repeat this process for each $C^{(k)} \in (|X|_{(k)}, |X|_{(k+1)})$ for $k = 0, 1, \ldots, m$. Among them, the $C^{(k^*)}$ with minimum total error is the oracle threshold or ideal C and the corresponding total error is 'optimal'. There may be multiple choices of best classifier C providing the same total error. Any of these choices is fine since we are interested in the 'optimal' total error for reference.

Remark 1. Note that, we have made classification in the ideal case with the knowledge of which observation comes from which σ_i^2. In practice, it may not be achievable as the classifier is not a function of $Y_1^2, Y_2^2, \ldots, Y_m^2$ alone. Thus the ideal case can be looked upon as some lower bound which may not be achievable even in the limit.

4. Simulation

The problem has reduced to find a suitable C for classification of the observations. We have simulated and performed the tests to validate our methods and have compared them with the independent cases. It is shown that our method with nonzero correlation coefficient is at least as good as the independent case in terms of the risk function.

4.1 Simulation setup

We have chosen $m = 200$, $\sigma_\epsilon = \sigma_0 = 1$, $\tau = 15, 90$ and $\rho = 0, 0.1, 0.5, 0.7$. We first generate the $\nu = \nu_0$ where each entry is a Bernoulli(p) variable with $p = 0.05, 0.12$. The choice of p will show that our method works for both sparse and non-sparse cases. With these parameters, we have generated the observations $\boldsymbol{X} \mid \boldsymbol{\nu}_0 = (X_1, X_2, \ldots, X_m) \mid \boldsymbol{\nu}_0$ using (1). The simulation for each set of parameter values is run 10 000 times and then the average of the sum of false positive and false negative is taken to estimate the misclassification risk.

We calculate the discrepancies to compare the performance of our methods using the following formula:

$$\textbf{Discrepancy in percentage} = 100 \times \frac{E_K - E_{K_0}}{E_{K_0}}$$

where E_K is the error (false positive, false negative or total) in the corresponding choice of C and E_{K_0} is the error in the ideal choice of C. The idea is similar to PRIAL in Ledoit and Wolf (2004).

Table 1. The discrepancy percentage for two cases: 1) for C in (6) and 2) for C computed using the algorithm with starting point T_1.

Parameters			$\mathbf{C_{Det}}$			$\mathbf{T_1^{Algo}}$		
p	ρ	τ^2	*F.Pos*	*F.Neg*	*Total*	*F.Pos*	*F.Neg*	*Total*
0.05	0.0	15	300.00	4.46	18.30	23 227.27	−54.02	1035.74
		90	112.50	14.52	20.45	3737.50	25.00	250.00
	0.1	15	104.76	7.89	13.22	10 047.62	−48.25	535.81
		90	−10.53	19.92	17.78	−63.16	61.75	52.96
	0.5	15	180.00	20.32	29.81	16 393.33	−50.60	880.75
		90	112.50	36.43	41.18	2462.50	46.51	189.71
	0.8	15	1000.00	51.14	90.32	83 225.00	−57.95	3523.66
		90	800.00	81.63	109.80	51 400.00	0.00	2015.69
0.12	0.0	15	73.13	3.99	9.82	1304.48	−25.31	86.90
		90	27.78	12.01	13.37	−97.22	74.93	60.14
	0.1	15	190.48	0.93	14.66	3323.81	−36.43	206.90
		90	22.22	13.35	13.88	−97.22	75.39	60.53
	0.5	15	164.44	14.88	26.17	2568.89	−27.77	168.29
		90	283.33	25.52	39.51	191.67	64.06	70.73
	0.8	15	111.11	52.96	57.94	1233.33	−9.37	92.28
		90	164.71	75.20	80.99	−52.94	88.62	79.47

4.2 Discussion

Our procedure is at least as good as in the case of the independent setup in terms of risk function irrespective of the method, see Table 1. The performance decreases as the data become more correlated. For large values of the ratio τ/σ_0 the classification is better and the expected number of both false positives and false negatives decrease.

We have chosen $\delta_0 = \delta_A = 1$ for computing C from (6). We can see from Table 1 that our method works for both sparse ($p = 0.05$) and non-sparse ($p = 0.12$) cases. The C from (6) performs better than the algorithm in general. Discrepancy in estimating expected false negative values is less than false positives, but expected value of false negatives is almost uniformly larger than its counterpart.

5. Conclusion

We have proposed methods to do multiple hypothesis testing for a specific dependent setup. Our methods are simple and computationally fast. The classifier C_{Det} performs best in both sparse and non-sparse scenarios and comes very close to the 'oracle' in terms of misclassification risk.

Acknowledgement. We are grateful to Professor Malay Ghosh for his valuable suggestions and comments.

Appendix A

1. Proof of (2):

$$\begin{aligned}
R(\boldsymbol{\nu}, \boldsymbol{\nu}^*) &= EE[L(\boldsymbol{\nu}, \boldsymbol{\nu}^*)|\boldsymbol{\nu} = \boldsymbol{\nu_0}] \\
&= \sum_{\nu_0} p_{\nu_0} E\left[\sum_{i=1}^{m} \delta_0 1_{[(\nu_{0i} - \nu_i^*)=1]} + \sum_{i=1}^{m} \delta_A 1_{[(\nu_{0i} - \nu_i^*)=-1]}\right] \\
&= \sum_{i=1}^{m} E\left[\sum_{\boldsymbol{\nu_0}} p_{\nu_0} \left(\delta_0 1_{[(\nu_{0i} - \nu_i^*)=1} + \delta_A 1_{[(\nu_{0i} - \nu_i^*)=-1]}\right)\right] \\
&= \sum_{i=1}^{m} \left[\sum_{\boldsymbol{\nu_0}} p_{\nu_0} \left(\delta_0 E(1_{(\nu_{0i}=1)}|\nu_i^* = 0)P(\nu_i^* = 0) + \delta_A E(1_{(\nu_{0i}=0)}|\nu_i^* = 1)P(\nu_i^* = 1)\right)\right] \\
&= \sum_{i=1}^{m} \delta_0(1-p)t_{1i} + \delta_A p t_{2i}.
\end{aligned}$$

2. Proof of Lemma 1:

   (a) First we consider marginals of X_i^2 and X_j^2 with their marginal variance σ_i^2 and σ_j^2 as their variances, respectively. Then
   $$X_i^2 \sim \sigma_i^2 V,$$
   where V is chi square with one degree of freedom. From this it easily follows that $X_i^2 \leq_{st} X_j^2$ if and only if $\sigma_i^2 \leq \sigma_j^2$.

   (b) Here we shall prove $X_i^2|Z \leq_{st} X_j^2|Z$ when the inequality in variance holds as stated above. Here Z is a subset of $\{X_1, X_2, \ldots, X_m\}$ deleted by X_i and X_j, respectively. Now in the equi-correlated setup, without loss of generality, instead of $i \neq j$ we may simply work with elements 1 and 2. Define
   $$U = \left[\left(\frac{X_1}{\sigma_1}, \frac{X_2}{\sigma_2}\right) \middle| \left(\frac{X_3}{\sigma_3}, \frac{X_4}{\sigma_4}, \ldots, \frac{X_m}{\sigma_m}\right)\right].$$
   This quantity is free of σ_i^2, and $\left(\frac{X_1}{\sigma_1}, \frac{X_2}{\sigma_2}, \frac{X_3}{\sigma_3}, \frac{X_4}{\sigma_4}, \ldots, \frac{X_m}{\sigma_m}\right)$ has exchangeable distributions. Now $U = (U_1, U_2)$ which are exchangeable. $(X_1^2, X_2^2) = (\sigma_1^2 U_1^2, \sigma_2^2 U_2^2)$ which has equi-correlated matrix $\mathcal{R}^*$. Hence by part (a) the result follows.

   (c) This part follows from part (a) and part (b).

3. Proof of Result 1:
   Let us assume a continuous p.d.f. of w and call it f_w. Then the within-group variance $V_W(C)$ of the two groups obtained from w, using C, is a continuous function of C. We consider
   $$\frac{\partial V_W(C)}{\partial C} = 0$$
   and obtain the result. Now as the result holds for a continuous p.d.f., it is easy to see that it holds for the discrete case also.

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**, 289–300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, **82**, 163–170.

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Annals of Statistics*, **39**, 1551–1579.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. *In Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*. Institute of Mathematical Statistics, 211–230.

Cohen, A. and Sackrowitz, H. B. (2005). Characterization of Bayes procedures for multiple endpoint problems and inadmissibility of the step-up procedure. *Annals of Statistics*, **33**, 145–158.

Cohen, A. and Sackrowitz, H. B. (2007). More on the inadmissibility of step-up. *Journal of Multivariate Analysis*, **98**, 481–492.

Cohen, A. and Sackrowitz, H. B. (2008). Multiple testing of two-sided alternatives with dependent data. *Statistica Sinica*, **18**, 1593–1602.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, **8**, 111–132.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, **102**, 93–103.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association*, **105**, 1042–1055.

Efron, B. (2012). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge University Press, Cambridge.

Fan, J. and Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society. Series B*, **79**, 1143–1164.

Fan, J., Ke, Y., Sun, Q., and Zhou, W.-X. (2017). FARM-Test: Factor-adjusted robust multiple testing with false discovery control. *arXiv preprint arXiv:1711.05386*.

Ghosh, P. (2017). Some theoretical and methodological aspects of simultaneous inference with special emphasis on high dimensional problems under sparsity. *Ph.D. Thesis, Indian Statistical Institute, Kolkata.*

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Methods*. John Wiley, New York, NY.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383–386.

LEDOIT, O. AND WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

LEHMANN, E. L. (1957a). A theory of some multiple decision problems, I. *Annals of Mathematical Statistics*, **28**, 1–25.

LEHMANN, E. L. (1957b). A theory of some multiple decision problems, II. *Annals of Mathematical Statistics*, **28**, 547–572.

MÜLLER, P., PARMIGIANI, G., ROBERT, C., AND ROUSSEAU, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, **99**, 990–1001.

QIU, X., XIAO, Y., GORDON, A., AND YAKOVLEV, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, **7**, 50.

SARKAR, S. K. (2007). Stepup procedures controlling generalized fwer and generalized fdr. *Annals of Statistics*, **35**, 2405–2420.

SARKAR, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, **70**, 135–168.

SCOTT, J. G. AND BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, **136**, 2144 – 2162. In Memory of Dr. Shanti Swarup Gupta.

SCOTT, J. G. AND BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, **38**, 2587–2619.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, **64**, 479–498.

STOREY, J. D., TAYLOR, J. E., AND SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B*, **66**, 187–205.

SUN, W. AND CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, **102**, 901–912.

SUN, W. AND CAI, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society. Series B*, **71**, 393–424.

XIE, J., CAI, T. T., MARIS, J., AND LI, H. (2011). Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, **4**, 417.

Manuscript received 2021-11-05, revised 2022-10-05, accepted 2023-01-22.