Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(5): 4971–4986.
DOI: 10.3934/mbe.2021253
Received: 24 February 2021
Accepted: 19 May 2021
Published: 07 June 2021

Research article

An efficient and flexible multiplicity adjustment for chi-square endpoints

Amy Wagler[1]* **and Melinda McCann**[2]

[1] Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX 79968, USA

[2] Department of Statistics, Oklahoma State University, Stillwater, OK 74701, USA

* **Correspondence:** Email: awagler2@utep.edu; Tel: +1-915-747-6847; Fax: +1-915-747-5022.

Abstract: This manuscript proposes a fast and efficient multiplicity adjustment that strictly controls the type I error for a family of high-dimensional chi-square distributed endpoints. The method is flexible and may be efficiently applied to chi-square distributed endpoints with any positive definite correlation structure. Controlling the family-wise error rate ensures that the results have a high standard of credulity due to the strict limitation of type I errors. Numerical results confirm that this procedure is effective at controlling familywise error, is far more powerful than utilizing a Bonferroni adjustment, is more computationally feasible in high-dimensional settings than existing methods, and, except for highly correlated data, performs similarly to less accessible simulation-based methods. Additionally, since this method controls the family-wise error rate, it provides protection against reproducibility issues. An application illustrates the use of the proposed multiplicity adjustment to a large scale testing example.

Keywords: multiple comparisons; simultaneous inference; Type I error control

1. Introduction

Familywise error rate (FWER) control is necessary when a set of multiple inferences are simultaneously evaluated and investigators want to obtain a small set of results that warrant further examination. When the test endpoints are correlated, it is preferable to use a procedure that takes into account the dependency structure in order to improve power. This can be done easily for multivariate normal outcomes, but an accessible and powerful method is not available for multiple chi-square endpoints even though this is a common set of test endpoints encountered in biomedical research contexts. For example, generalized linear models may be used to model mortality using a large predictor variable set. Additionally, multiple gene mutations may be simultaneously assessed for association with a particular disease outcome. Both of these common research contexts would warrant use of a multiplicity control

that can account for dependency in an efficient and effective manner. In particular, in high-dimensional data settings, use of multiple chi-square tests is quite common, but the use of, and need for, multiplicity adjustments may be unclear to many practitioners. However, either misusing or omitting multiplicity adjustments in these settings affects the reproducibility of research results, and can easily lead to spurious statistical significance declarations.

The purpose of this manuscript is to present a flexible, closed-form method for controlling the family-wise error rate for high-dimensional chi-square endpoints utilized in simultaneous testing or intervals. This is a critical setting for strict multiplicity control of type I error since false positives among large sets of associations can easily produce research results that are ultimately non-reproducible. Additionally, this procedure is adaptive to any positive definite correlation structure, thus is flexible to most any correlation structure encountered in practice. The method may be used to construct simultaneous intervals that control type I error and provides information about the practical significance of the test results. Note that the Bonferroni adjustment is currently the only easily accessible method for this general case. The proposed procedure, however, is theoretically and empirically shown to be less conservative than Bonferroni and easily attains FWER control even with complex correlation structures among the test endpoints. This proposed method may be used to complement data reduction methods to achieve reliable, reproducible and powerful research results when many chi-square test endpoints are simultaneously evaluated. The application in the conclusion of this manuscript illustrates one such use. Additionally, the proposed method is fast, simple, and completely general. Thus, it may be utilized for settings where the chi-square endpoints are correlated or uncorrelated for both high and low dimensions.

While resampling or simulation-based methods are a viable alternative for some scenarios, these methods can be difficult to implement in practice. This is particularly true for high-dimensional settings where resampling or simulation-based methods can be time consuming. The strength of resampling and simulation-based methods is that they provide a joint, rather than margin-based, solution. Two accessible options for implementing resampling-based multiplicity control in *R* [**?**] are the multtest [**?**], multcomp [**?**], and SiMaFlex [**?**] packages. Using a resampling or simulation-based approach is a gold standard, but may not be practical in every setting or circumstance.

Our goal is to provide a method that is simple to implement, even for practitioners with little statistical programming experience, but that still provides close to optimal family-wise error rate control for most settings. (We note that in some of these areas multiplicity adjustments of any kind, although warranted, are often omitted, so a method that is simple to utilize may be easier to promote.) Additionally, many data-intensive applications could require computation of a large number of critical points, justifying the need for a more efficient approach. This would be warranted if multiple independent subsets of test endpoints are selected for separate analysis. As the simulations demonstrate, the proposed critical values are often close to the resampling or simulation-based critical values when the correlations between endpoints is moderate to low. Thus, we recommend the proposed method for settings when a fast or simple method is warranted. However, resampling or simulation-based methods may be preferred for other scenarios.

In review, the purpose of this manuscript is to propose an easily implemented approach for controlling the familywise error rate for many chi-square distributed test endpoints. The remainder of the manuscript is organized as follows: in Section 2 we provide a review of probability-based multiplicity adjustments, in Section 3 we describe the proposed methodology, in Section 4 we present the

simulation results, and in Section 5 we discuss implications for research and practice.

2. Background

When conducting multiple tests of association, there are many approaches a practitioner may take. First, control of the familywise type I error rate (FWER) or the false discovery rate (FDR) are both reasonable approaches. Generally, the choice of which error rate to control depends on the nature of the study. Confirmatory studies that require overall conclusions involving multiple parameters generally utilize FWER control, while exploratory studies with many possible signals more often utilize FDR control. Control of the FDR, as first proposed by [**?**], is widely used for studies involving a large number of association tests. The FDR methodology is flexible as it is applicable to cases where the test statistics are dependent [**?**]. When there are a large number of test statistics and results are exploratory, rather than confirmatory, in nature, controlling the FDR for tests of association is very reasonable and effective. Many have worked on improving the performance of the FDR method with a focus on association tests.

In addition to the type of error control of interest, practitioners must consider whether the data warrants marginal or joint multiplicity control. Marginal control implies that the test statistics are assumed independent and, thus, the rejection regions are selected based only on the marginal test statistics. In contrast, joint control methods incorporate a dependency structure into the multiplicity control and the rejection region is defined based on a joint distribution with a particular dependency structure. In cases where the dependencies among the test statistics are substantial, a joint procedure is usually preferred. A related issue for consideration is also whether the multiplicity control uses a single-step or stepwise approach. A single-step approach provides a rejection region that is independent of the full set of test statistics while, in contrast, a stepwise approach takes into account the result of other test statistics in the family of inference. For stepwise approaches, testing the null hypothesis while taking into account the outcomes of related tests informs the null distribution used to define the rejection region and can lead to more powerful results.

While FDR control is advantageous in many settings, for a small set of inferences, controlling type I error provides a confirmatory set of inferences. Strict type I error control can often be achieved with reasonable power and minimal assumptions by utilizing the Bonferroni procedure. However, when the set of comparisons is large, this procedure quickly becomes too conservative, warranting more powerful bounds. When a multivariate normal distribution is an appropriate assumption, there are many methods that utilize arbitrary correlation structures to provide increased power. Some have made use of a minimal spanning tree approach ([**?**]; [**?**]) and others use tube approximation algorithms to achieve increased power ([**?**]; [**?**]; [**?**]). Additionally, many authors have considered alternatives for non-normal correlated endpoints, such as binary outcomes ([**?**]), non-normal continuous endpoints([**?**]; [**?**]), a combination of test endpoints ([**?**]; [**?**]), or generally discrete outcomes ([**?**]). When the distribution is specifically multivariate chi-square, however, the options are more limited. Resampling-based methods provide one alternative. Westfall & Tobias [**?**] and Westfall & Troendle [**?**] provide some resampling-based options. However, although these resampling-based procedures deliver strong solutions in almost any setting, they can be computationally demanding ([**?**], p.3) and computation of interval-based inferences is not always supported.

Simulation of critical values is also a reasonable approach, but implementation is often far from

direct, and can again be computationally intensive when considering large sets of inference. Stange et al. [**?**] provides another alternative that is closely related to the proposed procedure. This approach can be applied to multiple test endpoints with a multivariate chi-square distribution for situations with a low factorial correlation structure (defined in Stange et al.). This method requires a three-step process: 1) evaluation of the correlation matrix to ensure it is m-factorial, 2) numerical integration of often high-dimensional chi-square probability functions, and 3) approximation of a series to complete the multivariate probability calculations. For simple, one-factorial correlation matrices, this proposed method is efficient and highly accurate. However, the Stange method is computationally demanding and can exhibit significant approximation errors for moderately to highly complex correlation structures, is not applicable to all correlation structures, and is only available in MATLAB environments. In contrast, the proposed method may be implemented in *R* [**?**], applies to all correlation structures, is more computationally efficient than resampling or simulation-based procedures, Stange et al., or direct simulation, and is often just as accurate, although there are situations where it can behave conservatively.

A comprehensive overview and framework of approaches for error control in association testing is presented in [**?**]. While all of the above described considerations needs to be well thought out for each family of inference, we focus just on one particular type of multiplicity control. In particular, for this paper, we present a joint single-step quantile-based procedure that will control type I error and is flexible to any positive definite correlation structure derived from the test statistics.

3. Methodology

Recall that our goal is to provide a fast and simple probability point that allows evaluation of simultaneous tests or calculation of simultaneous intervals for a group of m test endpoints that follow a multivariate chi-square distribution with ν degrees of freedom and any set of off-diagonal bivariate correlations, $Corr(Y_i, Y_j) = \rho_{ij}$. A conservative solution to this probability can be achieved by utilizing the probability inequality that provides the rationale for the Hunter-Worsley method for multivariate normal or multivariate-t distributional settings. Suppose that we wish to perform m simultaneous tests based on test statistics, Y_i, i=1, . . . , m, where the joint distribution is the multivariate chi-square distribution, and where large values of Y_i, i=1, . . . , m are in favor of the alternatives. For this purpose, we denote the set of null hypotheses by $\{H_1, H_2, \ldots, H_m\}$ where each H_i is a true null hypothesis test for a test statistic Y_i using critical point c and spanning $1, \ldots, m$. Then one solution to the multiplicity problem of interest would involve finding a probability point c where for $A_i = \{|Y_i| > c\}$, $P(\cup_{i=1}^{m} A_i) \leq \alpha$ and α is the desired type I error rate. (Adjustments for other common situations are easily made.) Thus, if we find a point c where the right-hand side of Equation (**??**) is equal to α, then we have a conservative solution to the problem. This is accomplished by finding an upper (or exact) bound for $P(\cup_{i=1}^{m} A_i)$ where $A_i = \{|Y_i| > c\}$ denotes that an error has occurred with the results of the i^{th} hypothesis test or confidence interval under the null. In order to bound this probability, the following inequality is useful,

$$P(\bigcup_{i=1}^{m} A_i) \leq \sum_{i=1}^{k} P(A_i) - \sum_{(i,j)\in\tau} P(A_i \bigcap A_j). \tag{3.1}$$

Here τ is a non-cyclical tree or subgraph of G, a family of graphs formed from the set of nodes A_i with branches given by the intersections A_iA_j. While this upper bound is applicable for any tree, utilizing the

minimal spanning tree will yield the lowest upper bound with regards to the tree. Details on applying this tree to the multivariate chi-square setting are provided below.

Note that application of the above bound is computationally efficient since it only requires calculation of univariate and bivariate probabilities. Moreover, the proposed algorithm is guaranteed to improve upon the Bonferroni bound in all cases except when the events are mutually exclusive, where it is equivalent to the Bonferroni bound. We note that using this procedure requires a covariance matrix with a stable estimator. In contrast to Strange et al, users will not need to evaluate whether the covariance estimate is m-factorial. Instead, for any application with a large sample size, users only need to obtain a consistent estimator for the covariance matrix. For the bivariate chi-square distribution used in the proposed method, the bivariate probability functions are increasing in ρ_{ij} for all $i, j = 1, \ldots, m$ with $i \neq j$. Thus, for these distributions, the bound given in Equation (**??**) allows for a single evaluation of the optimal τ that will maximize the second term in the left-hand side of Equation (**??**) for all possible c values. This is a significant savings in computational intensity, as evaluating the left-hand side for various values of c does not require determining a new τ. However, evaluation of the right-hand side of Equation (**??**) will require calculation of bivariate chi-square probabilities. With regards to calculation, the algorithm utilized for approximating the quantiles from the bivariate chi-square distribution will produce reliable estimates to within five decimal places until the magnitude of the pairwise correlation exceeds 0.95 [**?**]. Results for the estimation of the bivariate chi-square quantiles were evaluated prior to application of the spanning tree algorithm to the above probability distribution functions. A more detailed description of the bivariate distribution function and an algorithm to find the minimal spanning tree for this situation are provided below.

3.1. Bivariate chi-square distributions

Consider two ν-dimensional multivariate standard normal random vectors, $Z_i = (Z_{i_1}, \ldots, Z_{i_\nu})$ for $i = 1, 2$ with ν non-zero canonical correlations between Z_{1_j} and Z_{2_j}. Then $Y_i = \sum_{(j=1)}^{\nu} Z_{ij}^2, i = 1, 2$, are distributed as canonically correlated chi-square random variables with ν degrees of freedom. Krishnaiah (1980) derived the joint chi-square density function for (Y_1, Y_2). Using this density, the $(1 - \alpha)$ quantile c, i.e., $P(Y_1 \leq c, Y_2 \leq c) = 1 - \alpha$, may be calculated using numerical integration. We will utilize the cumulative distribution function given by the following, where we assume that c1=c2=c, gamma is the incomplete gamma function, $c^* = \frac{c}{(2(1-\rho_{12}^2))}$ for i=1, 2, and $Corr(Y_1, Y_2) = \rho_{12}$.

$$P(Y_1 \leq c_1, Y_2 \leq c) = (1 - \rho_{12}^2)^{(-\nu/2)} \sum_{(j=0)}^{\infty} \frac{1}{\Gamma(\nu/2 + j)\Gamma(j + 1)} \rho_{12}^{2j} \gamma(\nu/2 + i, c^*)\gamma(\nu/2 + i, c^*) \tag{3.2}$$

When the Y_i do not have the same degrees of freedom, an alternative version of the distribution function is available [**?**]. For the purposes of the proposed Hunter-Worsley approximation, the above density and algorithm for numerically approximating the quantiles were coded in *R* [**?**]. In the code, the sum in Equation (**??**) was truncated after 150 terms and the secant method was used to find the desired quantile of the bivariate distribution function. All calculations in *R* matched the quantiles given in [**?**] to within five decimal places, hence reasonable accuracy and precision was achieved.

It is important to note that, for the chi-square distribution function given in [**?**], $P(Y_1 \leq c_1, Y_2 \leq c_2)$ depends on i and j only through ρ_{ij} and is non-decreasing in $|\rho_{ij}|$. This can be proven analytically by taking the partial derivative of Equation (**??**) with respect to ρ_{ij}. Given these properties, the distribution

in Equation (**??**) may be utilized to approximate any set of intersection probabilities in the same manner utilized by both [**?**] and [**?**] for the t and normal distributions, i.e., the spanning tree need only be obtained once.

3.2. Minimal spanning tree algorithm applied to chi-square endpoints

By utilizing Equation (**??**), a conservative multiplicity-adjusted critical value appropriate in chi-square settings may be obtained. As discussed previously, the branches of τ may be solved by utilizing the maximal spanning tree algorithm of [**?**] where the edge weights are the branches ρ_{ij}. The correlation, ρ_{ij}, is the appropriate edge weight when utilizing the maximal spanning tree algorithm since the $P(Y_1 \leq c, Y_2 \leq c)$ is nondecreasing in ρ_{ij}. In particular, this set of branches will maximize the right-hand side of Equation (**??**) for any positive value c. Using the adjusted critical value c often results in considerable gains in computational efficiency, particularly since simulation or resampling-based approaches would require a non-efficient recalculation of the critical point for each set of comparisons. The algorithm for obtaining τ is as follows:

1. Choose any A_i from the set of unconnected nodes
2. Find the largest ρ_{ij} so that A_i is an unconnected node and A_j is a connected node and join these nodes
3. Repeat steps 1) and 2) until no node is unconnected and there are no cycles.

Ultimately, a root-finding algorithm (e.g., secant method) can be utilized to obtain a constant c such that the right hand side of Equation (**??**) is approximately α. For the secant method, the convergence criterion for determining the critical point c was $1x10^{-5}$. We note that the CPU time for the proposed procedure is quite low, even for large sets of comparisons. For example, running *R* 4.0.0 on a Windows-based PC for m=10,000 endpoints, the proposed procedure required 7.97 seconds of elapsed time, while the simulated critical point, using 100,000 simulated data sets to approximate the critical value, required 149.7 seconds of elapsed time. Clearly, this proposed procedure delivers considerable computational efficiency when computing even a single critical point. However, when practitioners identify multiple mutually independent sets of inference, the proposed procedure is even more helpful, due to its accessibility and greatly decreased computational cost. In addition to advantages with regards to the computational cost, the critical point obtained using this procedure will always be less than or equal to the Bonferroni adjusted critical value and is almost always nearly identical to the simulated critical point, thus providing near optimal power for the inferences.

4. Simulations

In order to assess the performance of the derived critical point, the empirical family-wise error rate (eFWER) and relative efficiencies (RE) are examined for various settings. The REs, the ratio of the square of the Bonferroni and the Hunter-Worsley chi-square critical points, are reported for various correlation structures and values of m and ν. In order to facilitate comparison of a simulation-based critical value and the proposed critical point, REs are also computed to compare these values. Following the RE results, the eFWER is reported for these same settings. Both the Hunter-Worsley and Bonferroni critical points are always guaranteed to be conservative and are easily accessible options for multiplicity control in this general setting, but also require very little input from the end user. Consequently,

primary interest focuses on the REs, rather than the eFWERs of the two critical points. However, in order to compare the error rates of the proposed procedure to a simulation-based method and commonly utilized marginal multiplicity control methods (Holm, Hochberg, and Sidak), the eFWER is reported for each of these four additional procedures.

For the RE and eFWER evaluations, varied types of covariance structures are considered: an AR(1) structure, a compound symmetric structure and a block diagonal structure. These structures were considered due to their explanatory simplicity and relevance to applied settings. For example, a compound symmetric structure pertains to a case where a latent trait of a condition is determining the presence of a set of gene mutations. In this case, if any single gene mutation is activated, it increases the probability of a subset of these gene mutations as well. This may hold only for a subset of mutations, in which case, a block diagonal correlation structure is appropriate. Finally, an auto-regressive structure applies to cases where the latent variable decays over time across features. This is also a common scenario encountered in practice.

Data is generated for these specific correlation structures and, under the assumption of the null hypothesis, tested to see how many simulated data sets result in at least one type I error. We investigate sample sizes of n =100, 500, and 1000, test endpoint numbers of m =500 and 1000, for each of the three correlation structures (compound symmetric, autoregressive, and block diagonal structures). For each correlation matrix, ρ is set in the following manner: for a compound symmetric structure, $\rho_{ii} = 1$ and the off-diagonals are $\rho_{ij} = \rho$, an AR(1) structure, where the diagonal elements are $\rho_{ii} = 1$ and the off-diagonals are $\rho_{ij} = \rho^{|i-j|}$, respectively, for $i, j = 1, \ldots, m$, $i \neq j$, and a block diagonal structure where subsets of off-diagonal elements are ρ and the diagonal elements are 1. For all settings, the values ρ= 0, 0.3, 0.5, and 0.7 are considered.

The correlated multivariate chi-square random variables are generated using the following algorithm. First, using a given correlation matrix ($\boldsymbol{V}$), the eigenvalues ($\boldsymbol{l}$) and eigenvectors ($\boldsymbol{U}$) are extracted from the structure. Then using the eigenvalue diagonal matrix $\boldsymbol{L} = diag(\boldsymbol{l})$ and eigenvector matrix $\boldsymbol{U}$, $\boldsymbol{Y} = [\boldsymbol{U}'\boldsymbol{X}]'[\boldsymbol{U}'\boldsymbol{X}]$ is a multivariate set of correlated chi-square random variables provided that $\boldsymbol{X}$ is a multivariate normal set of centered random variables with correlation structure $\boldsymbol{L}$. Simulations confirm the correlation structure and mean were accurately approximated for each setting investigated to within an error of 0.001. In all cases, 10,000 replications are performed. This ensures that the eFWER is estimated accurately to within 0.001. Simulations confirm the correlation structure and mean were accurately approximated for each setting investigated to within an error of 0.001. In all cases, 1000 replications are performed. This ensures that the eFWER is estimated accurately to within 0.001. All simulations were performed in *R* [**?**].

5. Results

In all of the results presented, line plots are utilized for ease of viewing. However, these plots are not meant to imply that the statistic is linear between the respective plotting points.

5.1. Relative Efficiency Results

5.1.1. Fixed correlation structure REs

For simulating the multivariate chi-square data, the correlation structures are fixed and are also used as a basis for later random generation. Each of these correlation structures are investigated for m=500, 1000 and for ν=1. Other degrees of freedom (ν = 10 and 30) were also investigated with very similar results to when ν =1 and are omitted from presentation. Whenever REs compare the proposed point to the Bonferroni or simulated critical points, the RE is labeled *b.h* and *s.h*, respectively.

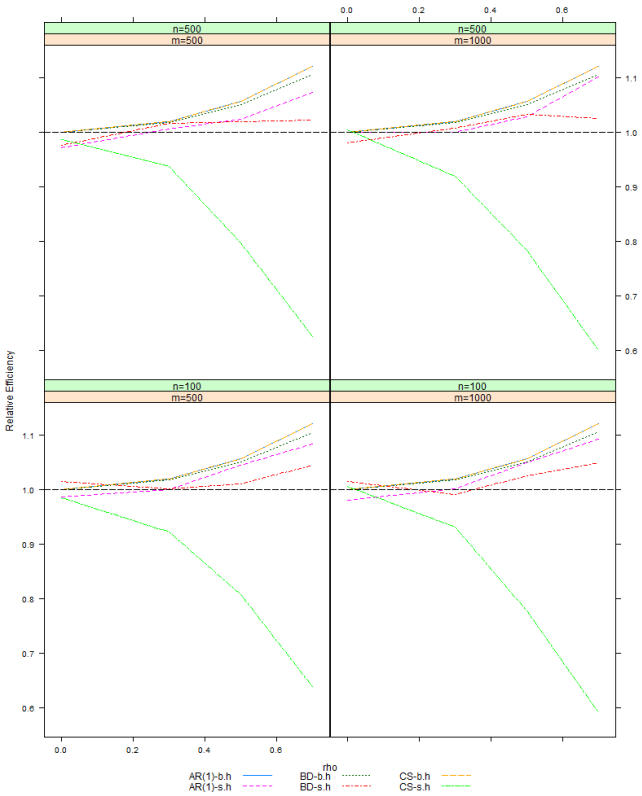

Figure 1. Mean relative efficiencies (RE) plotted against ρ with m=500 and 1000, sample sizes n =100, 500 and 1000, and line type indicating correlation structures block diagonal (BD), compound symmetric (CS), and autoregressive (AR1).

Figure 1 displays the RE values comparing the Hunter-Worsley chi-square critical values to the Bonferroni and simulation-based critical values. These are plotted against the ρ used to generate the correlation structure with the line type indicating the correlation structure (compound symmetric (CS), autoregressive (AR1) or block diagonal (BD)). The plots are stratified by the number of endpoints (m=500 and 1000) and sample size (n=500 and 1000). In general, as the magnitude of ρ and the number of comparisons increase, the RE for Bonferroni (*b.h*) also increases. There is also an increase in the RE values for Bonferroni as the degrees of freedom increases. The RE for Bonferroni gains ranged anywhere from 0% to 11% (when ρ = 0.7), with the largest gains in efficiency consistently occurring when ρ is greater than 0.3 in magnitude. Whenever ρ is very close to zero, the RE for Bonferroni gain is also close to 0, indicating a negligible gain in using the Hunter-Worsley as opposed to the Bonferroni

critical point. These general Bonferroni results hold for the AR(1), block diagonal (BD) and compound symmetric (CS) structures. With regard to the simulated critical point comparisons, there are two cases where the critical point using simulation has a clear advantage. When there is a block diagonal (BD) structure at lower levels of dependency ($\rho < 0.3$), there is a slight improvement with using the simulation-based critical point. Also, for the compound symmetric structure, the simulation-based critical point provides far more power than the proposed critical point. However, for the AR(1) and BD structures, the simulated and proposed critical points have similar values and, hence, the REs are close to 1.

5.1.2. Random correlation structure REs

In order to better simulate real-world covariance structures, we also use random correlation structures to see how varying levels of noise in the generation of the correlation structures affects performance. In the following, whenever correlation matrices are randomly generated, 100,000 simulated values are utilized to simulate the correlation matrix. In order to accomplish this for a specific correlation matrix (of types AR(1), BD or CS), a Wishart random variable is generated in *R* using the *rWishart* function [**?**] with either ν_W=500 or 1000 degrees of freedom for the Wishart distribution, creating high and low amounts of noise, respectively, in the correlation matrix. However, in the following, only the results using 1000 degrees of freedom are presented since the results were almost identical for both values of ν_W.

5.2. Random correlation structures

For each iteration of the simulations, a unique randomly generated correlation matrix is produced, which, correspondingly, results in a unique RE for comparing the proposed critical point to the Bonferroni and simulated critical points (labeled b.h and s.h in the plots) for that specific correlation matrix. Consequently, the RE values presented in Figures 2 and 4 are mean values. We note that the observed correlation matrices include some negative off-diagonal elements, but were positive definite and no problems arose randomly generating Wishart random variables based on these. In all the figures presented in the remaining sections, only the results for the randomly generated Wishart correlation matrices with a high level of variation are displayed, since these results are very similar to those assuming a low amount of variation. Figure 2 shows the mean RE values (b.h and s.h) for the AR(1), CS and BD structures where Wishart random variables with high noise are used to provide random correlation matrices. In general, these display a very similar pattern to the results when a fixed correlation structure is employed, as the RE increases with the number of comparisons and ρ, with the exception of the CS structure when comparing the simulation-based critical point to the proposed critical point. With the exception of this case, the RE values ranged from 0% to 12%. The results are very similar to the fixed correlation structure results and demonstrate that the method is stable even in the presence of high noise.

5.3. eFWER simulation results

In this section, the empirical family-wise error rates (eFWERs) are examined for each of the previously described data settings. The eFWER is defined to be the proportion of times that at least one of the Y_i, (i=1,...,m) exceeds the critical value. Since it is known that the proposed critical point will

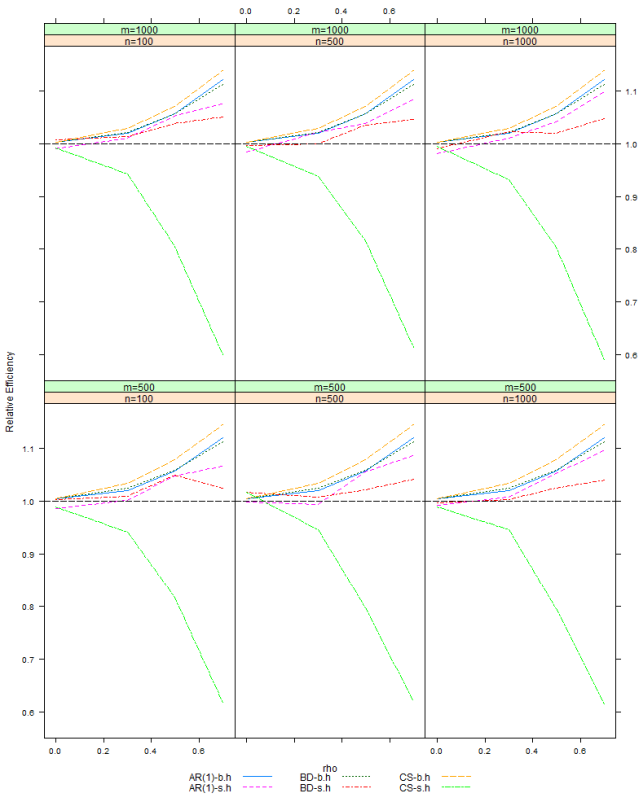

Figure 2. Mean relative efficiencies (RE) plotted against ρ with m=500 and 1000, sample sizes n =100, 500 and 1000, and line type indicating random correlation structures block diagonal (BD), compound symmetric (CS), and autoregressive (AR1).

be conservative, the eFWERs are assessed in order to gauge how conservative the proposed critical points are in varied settings. In order to evaluate the eFWER, the following procedure is utilized for generating correlated chi-square distributed endpoints. Data is generated to conform to a particular fixed correlation structure, degrees of freedom, and number of comparisons of interest. The generated vector of correlated chi-square responses was then evaluated for significance using the Hunter-Worsley chi-square critical point as well as the Bonferroni-adjusted, simulation-based, Holm, Hochberg, and Sidak based critical points.

5.3.1. Fixed correlation structure eFWER results

For 10,000 simulated data sets, we expect a procedure with a 5% FWER to be within 0.004 of 0.05. Figure 3 presents the eFWER for the fixed AR(1), BD and CS correlation structures across various ρ and m values. The simulations confirm that the Hunter-Worsley eFWER never exceeds a margin of error of 5%. However, there are instances where the Hunter-Worsley eFWER fell below the margin of error for 5% FWER, e.g. behaved conservatively. This occurs when $\rho > 0.5$ and, as stated, the Bonferroni method always produces an even lower eFWER in each case. In general, higher correlations result in lower eFWERs. However, as confirmed by the RE results, the largest gain in using the Hunter-Worsley critical point over the Bonferroni critical point occurs at these higher correlation

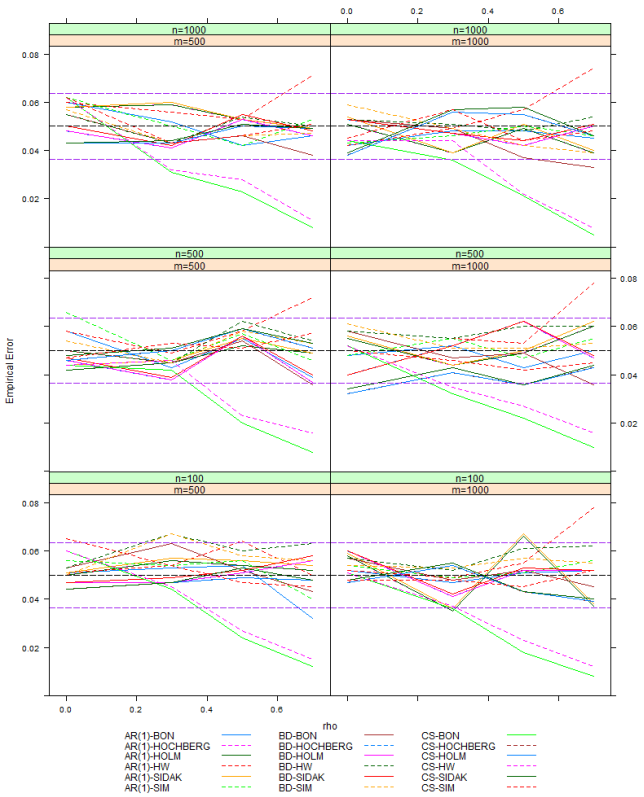

Figure 3. Empirical family-wise error rates (eFWER) plotted against ρ with m=500 and 1000, for fixed correlation matrices (BD, AR(1) and CS), line type indicating method.

settings. We observed the distribution of empirical error rates for the simulations and found that the proposed Hunter-Worsley method is on average closer to the desired error rate of 5% without becoming liberal.

At these higher correlation settings, it is also of interest to consider the use of a simulated critical point. Simulations indicate that the proposed Hunter-Worsley approximation performs similarly to the simulated critical point at low levels of correlation ($\rho < 0.5$) and for BD and AR(1) correlation structures. However, note that the marginal (Holm, Hochberg and Sidak) and simulated procedures maintain more power for CS structures at higher ($\rho > 0.5$) correlations. Simulations confirm that the proposed method is always within the error anticipated for the empirical error rates except for these cases of the CS structure (Note : We identify empirical error rates as differing from whenever they exceed the expected sampling error (2 x $\sqrt{((0.05 * 0.95)/1000)}$= 0.014). We note that only 10,000 simulated critical values were utilized for comparing the simulated critical point to the proposed critical point due to the computational burden of the evaluation. Moreover, in order to achieve similar results for the simulated critical value, we had to compute 100,000 distributions from which the maximum is retained to find the simulated critical point. This is computationally intensive and not a straightforward analysis that would be accessible to practitioners. In all cases, use of the proposed Hunter-Worsley critical point is justified since it is computationally less demanding and results in empirical error rates that are within the expected sampling error limits. Moreover, since this method is available as an *R* function, it is

accessible to end users with only an assumed overall error rate, an exact or estimated correlation or covariance matrix, and a set of planned comparisons. This analysis would fit well within existing computing structures for linear and generalized linear models in R and could easily be streamlined within a project workflow. Hence, given the evidence from these simulations that our proposed efficient critical point performs similarly to simulated values for BD and AR(1) structures, we suggest using our proposed method in place of the simulated point.

5.3.2. Random correlation structure eFWER results

When evaluating the randomly generated CS, BD and AR(1) structures, the eFWERs have a similar pattern to that observed for the fixed correlation structures. Hence, these results are not provided in a figure. However, Figure 5 displays the mean eFWERs when using the Hunter-Worsley and Bonferroni critical points for the three correlation matrices plotted against the degrees of freedom for the chi-square endpoints. The proposed critical point is always less conservative than the Bonferroni method and still controls the FWER. However, when comparing the proposed method to a simulation-based correction, the results are similar except for the higher levels of ρ with a CS dependency structure. In this case, the simulation-based or a marginal approach (Holm, Hochberg or Sidak) have more power.

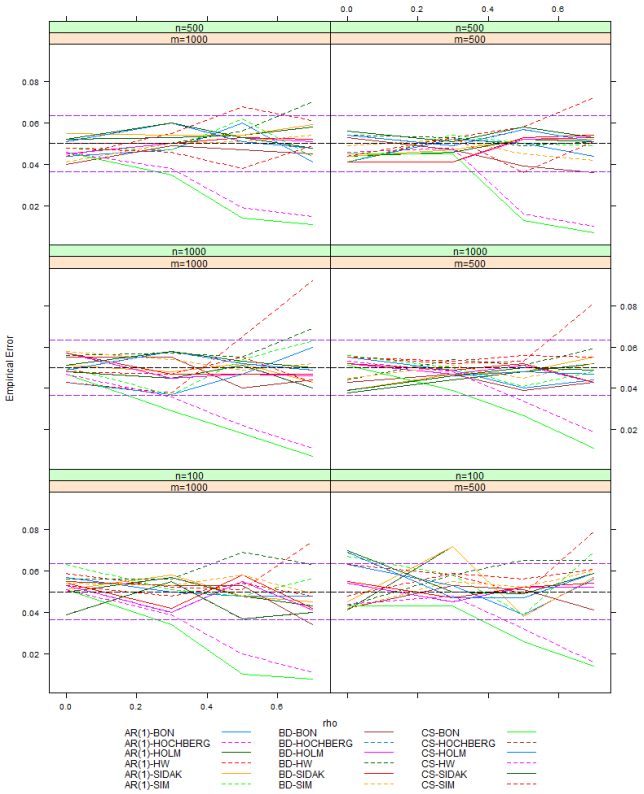

Figure 4. Mean empirical family-wise error rates (eFWER) plotted against degrees of freedom with randomly generated correlation matrices (CS, BD and AR(1)) and line type indicating method.

6. Application

This application uses data from a genome-wide association study with the objective of identifying exonic variants that play a role in a specific form of leukemia. The mode of analysis involves Fisher exact tests on 1061 variants with 30 total subjects, 20 of which served as controls and 10 with leukemia. In order to apply an appropriate multiplicity control method, a salient dependence structure must be well-defined. With this aim in mind, we consider using two possible dependency structures in order to apply the proposed multiplicity correction: Case 1) the standard covariance matrix estimated using a subset of the full data that focuses on those occurring on a selected list of cancer-related gene mutations, and Case 2) a dimension-reduced version derived from a hierarchical agglomerative clustering algorithm. We demonstrate both approaches in order to show how the proposed procedure could be combined with a dimension-reduction algorithm to further increase the power of the multiplicity correction. First, the covariance matrix for the 93 kinome variants is estimated. Second, the gene mutations were grouped using a hierarchical agglomerative algorithm as described in [**?**]. This algorithm identified 8 clusters using the linkage disequilibrium structure of the variants that are related to this form of leukemia. Following this clustering of the variants, we performed multiple testing on the aggregated predictors to test associations. Then, the covariance matrix was estimated based on this dimension-reduced set and utilized with the proposed algorithm. When using Case 1 variants, the adjusted critical value was c =11.21 and, when using the reduced set of 8 gene variants clusters from Case 2, the critical value reduced to c =6.51. Using the Case 1 critical point, 13 of the 92 variants were significant and using the Case 2 critical point, there are 13 significant variants. Note that both of these critical points are more powerful alternatives than a standard Bonferroni adjusted critical point which, in this case, would be 11.98. The computational time for obtaining the adjusted critical values was less than 3 seconds and clearly would not provide a burden to the researcher, even if there were multiple subsets of variants to assess independently.

We note that use of this proposed methodology extends well beyond the presented genomics application. There are multiple contexts that warrant use of a multiplicity adjustment for model or non-model based chi-square test endpoints being simultaneously evaluated. Just of a few of these include model re-specification tests for structural equation modeling, general model comparison tests, or any study involving categorical variables that might be jointly assessed in a log linear association model. These analytic approaches can occur across a range of disciplinary contexts and might involve smaller or large scale inference sets. The appeal of this method is that it can be used in most any context and with little computational load.

7. Conclusion

This manuscript presents a multiplicity correction suitable for many modes of analysis that involves chi-square distributed endpoints. This method is more efficient than the Bonferroni critical point, often is as effective as a simulated critical point, and is easy to employ. However, for a compound symmetric structure with higher correlation, a different marginal method or simulation-based method would be a better choice. In the above section, this proposed multiplicity correction is applied to a setting where multiple association tests could be utilized. However, researchers could alternatively present score tests resulting from an estimated generalized linear model in this setting, such as Rao's

score test ([**?**]; [**?**]), Pearson's test ([**?**]; [**?**]). The proposed method is applicable to either mode of analysis. Clearly this multiplicity correction is suitable in a wide variety of settings involving chi-square distributed endpoints, including multiple association tests, model fit tests, or score-based tests in generalized linear model (GLM) settings. The multiplicity adjustment may also be utilized to obtain both multiplicity corrections for test results or to construct simultaneous confidence intervals.

With regards to the proposed critical point itself, it is known that existing methods are either quite conservative (a.k.a. Bonferroni), difficult to employ, or computationally intensive (a.k.a., simulation or resampling-based methods). The proposed methodology is easy for practitioners to use, more powerful than the Bonferroni adjustment, and is also easily accessible via an *R* function, available upon request of the first author. The simulations performed indicate that there are many settings where the proposed Hunter-Worsley chi-square critical point is far more efficient than the Bonferroni adjusted critical point. It also performs very similarly to the simulated critical point for many settings. First, whenever there is a significant amount of correlation between the endpoints (e.g., whenever the ρ exceeds 0.3), the Hunter-Worsley adjusted critical point is more efficient than the Bonferroni critical point, regardless of correlation structure, number of comparisons, or degrees of freedom. Also, even when there is no correlation between the endpoints, the proposed critical point effectively controls the FWER. In contrast, a compound symmetric correlation structure with ρ less than 0.7 or an AR(1) correlation structure with ρ less than 0.9 results in negligible empirical error rate differences when comparing the simulated versus Hunter-Worsley approaches. Moreover, since the Hunter-Worsley critical point is faster and easier to implement, we suggest utilizing it in these settings. It is worth noting that for non-statisticians, the simulation-based critical points are more difficult to employ because they require more background knowledge and considerable computing power, particularly in high-dimensional settings. In contrast, the Hunter-Worsley procedure is available to end-users via *R* code and only requires an exact or estimated correlation matrix for the data, significance level, and degrees of freedom. Consequently, the proposed critical point is readily accessible, and may be applied to any general setting where a multiplicity adjustment is warranted and chi-square distributed endpoints are utilized. The simulations demonstrate that the Hunter-Worsley adjusted critical point is at least as effective as, and often considerably more effective than, other easily accessible multiplicity correction methods except for highly dependent compound symmetric data where this procedure lacks power. Table 1 provides a summary of recommendations for practitioners.

The proposed procedure is efficient and controls the type I error rate with more power than a Bonferroni correction. The method is particularly well-suited for, and flexible to, many dependence structures of the test statistics. Additionally the type I error control of the method provides increased confidence

Table 1. Practical Recommendations for FWER Control.

	Low level of dependence	High level of dependence
$m < 10$	Marginal or joint methods: Bonferroni or similar	Joint methods: Resampling/Simulation or proposed
$10 \leq m \leq 100$	Joint methods: Resampling or proposed	Joint methods: Resampling/Simulation or proposed
$m > 100$	Joint or partially-joint methods: Proposed	Joint methods:

that research results will be reproducible.

Acknowledgment

This work was supported by Grant 5U54MD007592 from the National Institutes on Minority Health and Health Disparities (NIMHD), a component of the National Institutes of Health (NIH).

Conflict of interest

The authors have no conflict of interest.

References

1. R Core Team, R: A language and environment for statistical computing, *R Found Stat. Comput.*, Vienna, Austria, 2019. `http://www.R-project.org/`.
2. K. S. Pollard, S. Dudoit, M. J. van der Laan, Multiple testing procedures: R multtest package and applications to Genomics, Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, 2005.
3. T. Hothorn, F. Bretz, P. Westfall, Simultaneous inference in general parametric models, *Biomet. J.*, **50** (2008), 346–363.
4. C. C. Bartenschlager, J. O. Brunner, A new user specific multiple testing method for business applications: The SiMaFlex procedure, *J. Stat. Plan Infer.*, 2021, Online.
5. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. Royal Stat. Soc. B.*, **57** (1995), 289–300.
6. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *Ann. Stat.*, **29** (2001), 1165–1188.
7. D. Hunter, An upper bound for the probability of a union, *J. Appl. Probab.*, **13** (1976), 597–603.
8. M. McCann, D. Edwards, A path length inequality for the multivariate-t distribution, with applications to multiple comparisons, *J. Am. Stat. Assoc.*, **91** (1996), 211–216.
9. D. Q. Naiman, Simultaneous confidence bounds in multiple regression using predictor variable constraints, *J. Am. Stat. Assoc.*, **82** (1987), 214–219.
10. J. Sun, C. R. Loader, Simultaneous confidence bands for linear regression and smoothing, *Ann. Stat.*, (1994), 1328–1345.
11. K. J. Worsley, An improved Bonferroni inequality and applications, *Biometrika*, **69** (1982), 297–302.
12. M. Heo, A. C. Leon, Comparison of statistical methods for analysis of clustered binary observations, *Stat. Med.*, **24** (2005), 911–923.
13. R. B. Arani, J. J. Chen, A power study of a sequential method of p-value adjustment for correlated continuous endpoints, *J. Biopharm. Stat.*, **8** (1998), 585–598.
14. S. James, The approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials, *Stat. Med.*, **10** (1991), 1123–1135.

15. S. J. Pocock, N. L. Geller, A. A. Tsiatis, The analysis of multiple endpoints in clinical trials, *Biometrics*, (1987), 487–498.
16. P. C. O'Brien, Procedures for comparing samples with multiple endpoints, *Biometrics*, (1984), 1079–1087.
17. R. E. Tarone, A modified Bonferroni method for discrete data, *Biometrics*, (1990), 515–522.
18. P. H. Westfall, R. D. Tobias, Multiple testing of general contrasts, *J. Am. Stat. Assoc.*, **92** (2007), 299–306.
19. P. H. Westfall, J. F. Troendle, Multiple testing with minimal assumptions, *Biomet. J.*, **50** (2008), 745–755.
20. W. Pan, Asymptotic tests of association with multiple SNPs in linkage disequilibrium, *Genet. Epidemiol.*, **33** (2009), 497–507.
21. J. Stange, N. Loginova, T. Dickhaus, Computing and approximating multivariate chi-square probabilities, *J. Stat. Comput. Sim.*, **86** (2016), 1233–1247.
22. S. Dudoit, M. J. van der Laan, Multiple tests of association with biological annotation metadata, in *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics, Springer, (2008), 413–476.
23. K. Wright, W. J. Kennedy, Self-validated Computations for the Probabilities of the Central Bivariate Chi-square Distribution and a Bivariate F Distribution, *J. Stat. Comput. Sim.*, **72** (2002), 63–75.
24. P. R. Krishnaiah (Ed.), *Handbook of statistics*, Motilal Banarsidass Publisher, (1980).
25. J. B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, *Proc. Am. Math. Soc.*, **7** (1956), 48–50.
26. A. K. Bera, Y. Bilias, Rao's score, Neyman's C() and Silvey's LM tests: An essay on historical developments and some new results, *J. Stat. Plan Infer.*, **97** (2001), 9–44.
27. F. Guinot, M. Szafranski, C. Ambroise, F. Samson, Learning the optimal scale for GWAS through hierarchical SNP aggregation, *BMC Bioinform.*, **19** (2018), 1–14.
28. G. Lovison, On Rao score and Pearson χ^2 statistics in generalized linear models, *Stat. Papers*, **46** (2005), 555–574.
29. D. Pregibon, Score tests in GLIM with applications, In GLIM82: Proceedings of the International Conference on Generalized Linear Models, R Gilchrist (ed.), *Lec. Notes Stat.*, 14, Springer, New York, (1982), 87–97.
30. G. K. Smyth, Pearson's goodness of fit statistic as a score test statistic, Science and Statistics: A Festschrift for Terry Speed, D. R. Goldstein (ed.), *IMS Lec Notes.*, 40, Institute of Mathematical Statistics, Beachwood, Ohio, (2003), 115–126. `http://www.statsci.org/smyth/pubs/goodness.pdf`.

AIMS Press