Statistics and Probability Letters 111 (2016) 32–40

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Multiplicity- and dependency-adjusted p-values for control of the family-wise error rate

Jens Stange [a], Thorsten Dickhaus [b,*], Arcadi Navarro [c,d,e], Daniel Schunk [f]

[a] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany
[b] University of Bremen, Institute for Statistics, P.O. Box 330 440, 28344 Bremen, Germany
[c] Institute of Evolutionary Biology, Universitat Pompeu Fabra, Dr. Aiguader, 88, 08003 Barcelona, Spain
[d] Institució Catalana de Recerca i Estudis Avançats (ICREA), Dr. Aiguader, 88, 08003 Barcelona, Spain
[e] Center for Genomic Regulation (CRG), Dr. Aiguader, 88, 08003 Barcelona, Spain
[f] University of Mainz, Department of Economics, 55099 Mainz, Germany

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form 4 January 2016
Accepted 4 January 2016
Available online 13 January 2016

MSC:
62J15
62P10

Keywords:
Dependency structure
Effective number of tests
Genetic epidemiology
Multiple testing
Probability approximations
Šidák correction

ABSTRACT

Under the multiple testing framework, we propose the multiplicity- and dependency-adjustment method (MADAM) which transforms test statistics into adjusted p-values for control of the family-wise error rate. For demonstration, we apply the MADAM to data from a genetic association study.

1. Introduction

Dependency plays a crucial role in virtually all modern applications of high-dimensional data analysis, at least for two reasons. On the one hand, data generated with nowadays' high-throughput measurements typically exhibit strong temporal, spatial, or spatio-temporal dependencies due to the underlying (neuro-)biological or technological mechanisms. In biology, linkage disequilibrium for alleles and co-regulation for levels of expression of genes are two prominent examples. Hence, these dependencies have to be taken into account in any realistic statistical model for such data. On the other hand, such dependencies induce an intrinsically low-dimensional structure in the sample and/or the parameter space, thus facilitating or enabling valid statistical inference even for moderate sample sizes.

Here, we focus on the multiple testing context where $M > 1$ null hypotheses $H_1, \ldots, H_M$ are to be tested simultaneously based on one and the same data sample $x \in \mathcal{X}$. We assume that the considered multiple test procedure φ relies on test

* Corresponding author.
E-mail addresses: jens.stange.85@gmail.com (J. Stange), dickhaus@uni-bremen.de (T. Dickhaus), arcadi.navarro@upf.edu (A. Navarro), daniel.schunk@uni-mainz.de (D. Schunk).

http://dx.doi.org/10.1016/j.spl.2016.01.005
0167-7152/

statistics $T_1, \ldots, T_M$ which are computed from x and compared with multiplicity-adjusted rejection thresholds. In prior work (see Dickhaus and Stange, 2013 and Stange et al., 2016) it has been demonstrated that classical multiple testing approaches for control of the family-wise error rate (FWER) like the Bonferroni or the Šidák correction can be improved if the distribution of the vector $\mathbf{T} = (T_1, \ldots, T_M)^\top$ exhibits strong dependencies.[1] The possible relaxation of the necessary correction for multiplicity was described by the concept of the "effective number of tests" of order i, $M_{\text{eff.}}^{(i)}$ for short; see also Section 4.3.3 of Dickhaus (2014). Roughly speaking, $M_{\text{eff.}}^{(i)}$ approximates the number of stochastically independent tests which lead to the same FWER as φ. Hence, $M_{\text{eff.}}^{(i)}$ equals M if all components $T_1, \ldots, T_M$ are stochastically independent, and it equals one if $T_1, \ldots, T_M$ are totally dependent in the sense that all of them essentially assess exactly the same information from the data sample x. Computing $M_{\text{eff.}}^{(i)}$ for $1 \leq i \leq M$ requires knowledge of the i-variate (marginal) distributions of $\mathbf{T}$ which are then utilized in (sum- or product-type) probability approximations of order i. Hence, $M_{\text{eff.}}^{(i)}$ is typically decreasing in i, because more and more information about the dependency structure is exploited.[2]

We may mention here that the term "effective number of tests" has already been used for a longer time and seems to have its origins in the field of genetic epidemiology (see the corresponding references in Dickhaus and Stange (2013)), but the foundations of this concept have to the best of our knowledge been made mathematically rigorous in Dickhaus and Stange (2013) for the first time. Methods for computing $M_{\text{eff.}}^{(3)}$ in the genetic epidemiology context have been provided in Stange et al. (2016) based on the theory of multivariate chi-square distributions; see also Dickhaus and Royen (2015).

Although $M_{\text{eff.}}^{(i)}$ describes the quantitative effect of the dependencies in the data x on the FWER behavior of φ in a transparent and straightforward manner, it has the undesirable property that it depends on the FWER level α. This is both counter-intuitive (the dependency structure is a feature only of the data sample x, not of the parameters of some method to analyze x) and inconvenient in practice, because iterative algorithms are required to match the probability approximation of order i and α for computing $M_{\text{eff.}}^{(i)}$. In the present work, we therefore introduce the multiplicity- and dependency-adjustment method of order i, MADAM$_i$ for short. The MADAM$_i$ transforms the vector $\mathbf{T}$ into a vector of p-values which are adjusted both for multiplicity and for ith order dependency. Hence, these p-values are typically larger than their unadjusted, marginal counterparts, but smaller than the Bonferroni- or Šidák-corrected marginal p-values. In addition, MADAM$_i$ does not require the specification of α, thus avoiding the undesirable properties of $M_{\text{eff.}}^{(i)}$. However, both methods are closely related by the fact that they exploit the same probability approximations of order i.

The rest of the work is structured as follows. In Section 2, the MADAM is introduced and two different variants of it are illustrated. Section 3 shows how to utilize the MADAM for evaluating genetic association studies, considering a real-data example from this field. We conclude with a discussion in Section 4. Tables displaying the numerical results for the considered real-data example are deferred to Appendix A.

2. Statistical methodology: the MADAM

2.1. Notation and preliminaries

Throughout, we assume a statistical model $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$. The null hypotheses $H_1, \ldots, H_M$ are identified with non-empty subsets of the parameter space Θ. The intersection hypothesis $H_0 = \bigcap_{j=1}^M H_j$ is called the global hypothesis. For a given $\vartheta \in \Theta$, we will denote the index set of true null hypotheses in $\mathcal{H} = \{H_1, \ldots, H_M\}$ by $I_0 \equiv I_0(\vartheta) = \{1 \leq j \leq M : \vartheta \in H_j\}$. A (non-randomized) multiple test is a measurable mapping $\varphi = (\varphi_j)_{1 \leq j \leq M} : \mathcal{X} \to \{0, 1\}^M$ the components of which have the usual interpretation of a statistical test for H_j versus K_j. The family-wise error rate of a multiple test φ is (for a given $\vartheta \in \Theta$) defined as

$$\text{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta\left(\bigcup_{j \in I_0(\vartheta)} \{\varphi_j = 1\}\right),$$

and φ is said to (strongly) control the FWER at a pre-specified level $\alpha \in (0, 1)$ if $\sup_{\vartheta \in \Theta} \text{FWER}_\vartheta(\varphi) \leq \alpha$.

Under this general framework, we make the following assumption.

Assumption 1. There exists a parameter value $\vartheta^* \in H_0$ such that

$$\forall \vartheta \in \Theta : \ \text{FWER}_\vartheta(\varphi) \leq \text{FWER}_{\vartheta^*}(\varphi). \tag{1}$$

Thus we may assume an overall null distribution $\mathbb{P} := \mathbb{P}_{\vartheta^*}$, under which all hypotheses are true, as the worst case with respect to control of the FWER.

[1] The FWER denotes the probability for at least one type I error among the M individual tests.

[2] Mathematical conditions guaranteeing that $M_{\text{eff.}}^{(i)}$ decreases with i are provided in Dickhaus and Stange (2013).

Remark 1. Assumption 1 corresponds to a "single-step" nature of the multiple test based on the MADAM. It is typically fulfilled if the null hypotheses refer to parameters of the marginal distributions of test statistics, while the dependency structure among these test statistics is a given nuisance parameter. One prototypical example is the well-known "Gaussian means problem", where the vector $\mathbf{T}$ of test statistics follows a multivariate normal distribution with unknown mean vector ϑ, but known covariance matrix Σ, and hypotheses relate to ϑ. The assumption of a known Σ is for example fulfilled in analysis of variance models with known error variance, where Σ only depends on the group-specific sample sizes; cf., e.g., Section III.A of Dickhaus and Gierl (2013). The case of an unknown Σ has been considered in Stange et al. (2015a).

If Assumption 1 is violated, one can extend the MADAM such that a closed test procedure is constructed, where a multiplicity- and dependency-adjusted p-value is computed for every partition element of the parameter space (i.e., for every intersection hypothesis). While this is conceptually straightforward, it is computationally much more demanding and computational shortcuts (leading to step-down tests) may be required for feasible implementation. We defer the reader to Romano and Wolf (2005a) and Romano and Wolf (2005b). Step-down variants can also be applied if Assumption 1 is fulfilled, in order to optimize the power of the multiple test based on the MADAM.

2.2. Multiplicity- and dependency-adjusted p-values

We restrict our attention to simultaneous test procedures (STPs) in the sense of Gabriel (1969). An STP φ is such that $\varphi_j(x) = 1 \Longleftrightarrow T_j(x) > c_\alpha$, $1 \leq j \leq M$, $x \in \mathcal{X}$, for a given real constant c_α which in general depends on the FWER level α. As in Eq. 1 of Dickhaus and Royen (2015), a valid p-value for the marginal test problem H_j versus K_j corresponding to such an STP is given by

$$p_j(x) = \mathbb{P}\left(\max_{1\leq k\leq M} T_k > t_j\right) = \mathbb{P}\left(\bigcup_{k=1}^{M}\{T_k > t_j\}\right) = 1 - \mathbb{P}\left(\bigcap_{k=1}^{M}\{T_k \leq t_j\}\right), \tag{2}$$

where $t_j = T_j(x)$ is the actually observed value of the jth test statistic for the data sample x. The p-value p_j is computed under the global hypothesis H_0. It takes the full joint distribution of $\mathbf{T}$ under H_0 into account.

Feasible numerical methods for computing p_j only exist in a limited number of special model classes and for limited ranges of the total number M of tests, except from very time-consuming Monte Carlo approximations. For example, the R-package `mvtnorm` computes multivariate t- and normal probabilities up to dimension 1000, but not for higher dimensions. Hence, we propose to approximate p_j for $1 \leq j \leq M$ conservatively by making use of probability bounds. Following Section 4.3 of Dickhaus (2014), we refer to an upper bound of the form

$$\forall c \in \mathbb{R}: \; b^{(i)}(\mathbb{P}, c) \geq \mathbb{P}\left(\bigcup_{k=1}^{M}\{T_k > c\}\right) \tag{3}$$

as a sum-type probability bound of order i (STPB$_i$), if it takes the marginal distributions of $\mathbf{T}$ up to the ith order into account. Typically, an STPB$_i$ is obtained from a (higher-order) Bonferroni inequality. Analogously, we call a lower bound of the form

$$\forall c \in \mathbb{R}: \; \beta^{(i)}(\mathbb{P}, c) \leq \mathbb{P}\left(\bigcap_{k=1}^{M}\{T_k \leq c\}\right) \tag{4}$$

taking the marginal distributions of $\mathbf{T}$ up to the ith order into account a product-type probability bound of order i (PTPB$_i$). Based on chain factorization, in Block et al. (1992) the authors considered

$$\beta^{(i)}(\mathbb{P}, c) = \mathbb{P}\left(\bigcap_{k=1}^{i}\{T_k \leq c\}\right)\prod_{k=i+1}^{M}\mathbb{P}\left(T_k \leq c \,\middle|\, \bigcap_{\ell=k-i+1}^{k-1}\{T_\ell \leq c\}\right). \tag{5}$$

It has to be mentioned that the right-hand side of (5) is not always a PTPB$_i$, because the inequality in (4) may be violated for special dependency structures in $\mathbf{T}$. However, in Stange et al. (2016) it was demonstrated that $\beta^{(i)}(\mathbb{P}, c)$ from (5) often yields accurate approximations, even for $i = 3$, and the authors termed it a product-type probability approximation of order i (PTPA$_i$).

The effective number of tests of order i mentioned in the introduction can formally be defined as the solution of

$$M^{(i)}_{\text{eff.}}\alpha_{\text{loc.}} = b^{(i)}(\mathbb{P}, c_\alpha) = \alpha \quad \text{or} \quad (1-\alpha_{\text{loc.}})^{M^{(i)}_{\text{eff.}}} = \beta^{(i)}(\mathbb{P}, c_\alpha) = 1-\alpha,$$

respectively, where $\alpha_{\text{loc.}}$ denotes the local (marginal) significance level corresponding to the multiplicity-adjusted rejection threshold c_α. The shortcomings of $M^{(i)}_{\text{eff.}}$ discussed in Section 1 lead to the following definition of multiplicity- and dependency-adjusted p-values.

Definition 1. The MADAM$_i$ transforms the values of the test statistics $T_1, \ldots, T_M$ into one of the following multiplicity- and dependency-adjusted p-values.

$$p_{\Sigma,j}^{\text{MADAM}_i}(x) = b^{(i)}(\mathbb{P}, t_j), \tag{6}$$

$$p_{\Pi,j}^{\text{MADAM}_i}(x) = 1 - \beta^{(i)}(\mathbb{P}, t_j), \tag{7}$$

for all $1 \leq j \leq M$.

The order i can be regarded as a "tuning parameter" of the MADAM which has to be chosen in a model-specific manner. Generally, one should choose i as large as possible. However, the computation of i-variate marginal distributions of $\mathbf{T}$ is required for the MADAM$_i$, which restricts the choice of i in practice.

Obviously, neither $p_{\Sigma,j}^{\text{MADAM}_i}$ nor $p_{\Pi,j}^{\text{MADAM}_i}$ depends on α. In practice, however, one can reject H_j in favor of K_j if the jth multiplicity- and dependency-adjusted p-value does not exceed α. Theorem 1 formally shows that this decision rule constitutes an FWER-controlling multiple test procedure under Assumption 1.

Theorem 1. *Let Assumption 1 be fulfilled. Denote by F_{ϑ^*} the cumulative distribution function of $\max_{1\leq k\leq M} T_k$ under ϑ^* and assume that F_{ϑ^*} is continuous. Then, the following three assertions hold true.*

(a) *The statistic $p_{\Sigma,j}^{\text{MADAM}_i}(X)$ is a valid multiplicity-adjusted p-value for FWER control, i.e., $\mathbb{P}\left(\exists 1 \leq j \leq M : p_{\Sigma,j}^{\text{MADAM}_i}(X) \leq t\right) \leq t$ for all $t \in [0, 1]$.*

(b) *The statistic $p_{\Pi,j}^{\text{MADAM}_i}(X)$ is a valid multiplicity-adjusted p-value for FWER control.*

(c) *Assume that the random vector $\mathbf{T} = (T_1, \ldots, T_M)^\top$ is sub-Markovian of order i (SM$_i$) in the sense of Definition 2.2 of Dickhaus and Stange (2013). Then, the right-hand side of (5) fulfills (4), meaning that it is a* PTPB$_i$.

Proof. To prove part (a), let $t \in [0, 1]$ be arbitrary, but fixed. Straightforward calculation and the principle of quantile transformation yield

$$\begin{aligned}
\mathbb{P}\left(\exists 1 \leq j \leq M : p_{\Sigma,j}^{\text{MADAM}_i}(X) \leq t\right) &= \mathbb{P}\left(\exists 1 \leq j \leq M : b^{(i)}(\mathbb{P}, T_j) \leq t\right) \\
&\leq \mathbb{P}\left(\exists 1 \leq j \leq M : 1 - F_{\vartheta^*}(T_j) \leq t\right) \\
&= \mathbb{P}\left(\exists 1 \leq j \leq M : F_{\vartheta^*}(T_j) \geq 1 - t\right) \\
&= \mathbb{P}\left(F_{\vartheta^*}\left(\max_{1\leq k\leq M} T_k\right) \geq 1 - t\right) \\
&= \mathbb{P}(U \geq 1 - t) = t,
\end{aligned}$$

where U denotes a standard uniform variate. Part (b) can be proved analogously, and part (c) is an immediate consequence of the definition of SM$_i$. □

3. An application to genetic data

3.1. Testing genetic association

In genetic association studies, a (potentially very large) number M of genetic markers are simultaneously tested for associations with a given phenotype. In the case that the markers are bi-allelic, they lead to diploid genotypes with three possible realizations per genomic position (locus). Typically, single nucleotide polymorphisms (SNPs) are considered in this context. If, in addition, the phenotype is binary (e.g., a disease indicator), many (2×3) contingency tables have to be evaluated simultaneously. This is a multiple test problem. Here, for illustration, we consider chromosome-wise multiplicity, meaning that the chromosomes are treated as independent units and the methods from Section 2 are applied to each of the 22 autosomes separately (sex chromosomes require a different statistical methodology).

In the sequel, we denote by M_C, $C \in \{1, \ldots, 22\}$, the different numbers of tests (considered loci) for chromosome C. For each $1 \leq j \leq M_C$, an association test based on the contingency table data $x^{(j)}$ (see Table 1) is carried out. Notice that all quantities in Table 1 depend on the locus j, except for the row sums $n_{1\cdot}$ and $n_{2\cdot}$. This corresponds to the setup of a case-control study design; see Dickhaus and Stange (2013), Dickhaus et al. (2012), and Chapter 9 of Dickhaus (2014) for further details.

In the terminology of Section 2, we consider for all $1 \leq j \leq M_C$ the null hypothesis

$$H_j = \{\text{There is no association between the phenotype and locus } j\},$$

which can equivalently be expressed as $H_j = \{\vartheta_j = \vartheta_j^*\}$, where $\vartheta_j = \left(\pi(A_1^{(j)}A_1^{(j)}), \pi(A_1^{(j)}A_2^{(j)}), \pi(A_2^{(j)}A_2^{(j)})\right)^\top$ denotes the triple of (expected) genotype frequencies in cases at locus j and ϑ_j^* denotes the analogous triple in the entire target population. The (unique) parameter value in the global hypothesis for chromosome C is thus given by $\vartheta^* = (\vartheta_j^* : 1 \leq j \leq M_C)$.

Table 1
Genotype-phenotype counts at locus j aggregated in a (2×3)-contingency table. In the case of SNPs, the alleles $A_1^{(j)}$, $A_2^{(j)}$ are one of the nucleobases adenine (A), cytosine (C), guanine (G), or thymine (T). Cases correspond to the phenotypic value 1, while controls exhibit the phenotypic value 0.

Genotype	$A_1^{(j)}A_1^{(j)}$	$A_1^{(j)}A_2^{(j)}$	$A_2^{(j)}A_2^{(j)}$	$\sum$
Cases	$x_{10}^{(j)}$	$x_{11}^{(j)}$	$x_{12}^{(j)}$	$n_{1.}$
Controls	$x_{20}^{(j)}$	$x_{21}^{(j)}$	$x_{22}^{(j)}$	$n_{2.}$
$\sum$	$n_{.0}^{(j)}$	$n_{.1}^{(j)}$	$n_{.2}^{(j)}$	n

The null hypothesis H_j can be tested with Pearson's χ^2-test for independence (cf., e.g., Section 3.2.1 of Agresti, 2013), employing the test statistic T_j, given by

$$T_j(x) = n \sum_{r=1}^{2} \sum_{c=0}^{2} \frac{(x_{rc}^{(j)} - n_{r.} n_{.c}^{(j)}/n)^2}{n_{r.} n_{.c}^{(j)}}, \tag{8}$$

where $x = (x^{(1)}, \ldots, x^{(M_C)})^\top$ denotes the entire data sample.

If H_j is true, T_j is marginally asymptotically (with n tending to infinity) χ^2-distributed with two degrees of freedom. Notice, however, that there exist strong dependencies among the T_j, at least in blocks of markers which are in linkage disequilibrium (LD). Consequently, the vector $\mathbf{T} = (T_1, \ldots, T_{M_C})^\top$ asymptotically follows a multivariate chi-square distribution of the type considered in Section 4 of Dickhaus and Royen (2015) under H_0. Since LD can be regarded as external structural information (cf. Dickhaus et al., 2015), the multivariate methods from Section 2 are a promising approach and typically more powerful than simple Bonferroni- or Šidák-corrections.

3.2. The MADAM for genetic association studies

For an approximation of $p_{\Pi,j}^{\mathrm{MADAM}_i}(x)$ from (7) for $i < M_C$, information about the i-variate (marginal) distributions of $\mathbf{T} = (T_1, \ldots, T_{M_C})^\top$ is required. Due to multivariate central limit theorems (see Section 4 in Dickhaus and Stange, 2013), it suffices to consider the correlation (i.e., LD) matrix Σ_C of the M_C markers. This LD matrix can either be obtained from publicly available databases or can be estimated from the actual study data. If the LD matrix is estimated, the (realized) family-wise error rate (FWER) of a multiple test corresponding to the MADAM is itself a random variable which is distributed around the target FWER level α, and one may construct upper confidence bounds for the FWER of the empirically calibrated multiple test. Under regularizing assumptions regarding the structure of the LD matrix, the latter case has intensively been studied in Stange et al. (2015a).

For computational convenience, we propose to replace Σ_C by one of the following schemes.

(a) Block thresholding: Submatrices of size $(b \times b)$ along the diagonal are kept, while all other entries are set to 0. This leads to the approximation

$$\tilde{\Sigma}_C = \begin{pmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & R_B \end{pmatrix}, \quad \text{where } B = M_C/b. \tag{9}$$

Since the inequality

$$\mathbb{P}\left(\bigcap_{k=1}^{M_C}\{T_k \le x\}\right) \ge \mathbb{P}\left(\bigcap_{k=1}^{b}\{T_k \le x\}\right)\mathbb{P}\left(\bigcap_{k=b+1}^{2b}\{T_k \le x\}\right)\cdots\mathbb{P}\left(\bigcap_{k=(B-1)b+1}^{Bb}\{T_k \le x\}\right) \tag{10}$$

holds true for all $x \ge 0$ due to the extended Gaussian correlation inequality proven in Royen (2014), the approximation

$$1 - \prod_{\ell=1}^{B} \mathbb{P}\left(\bigcap_{k=(\ell-1)b+1}^{\ell b}\{T_k \le t_j\}\right) \ge p_j \tag{11}$$

yields a valid p-value. The final approximation $\tilde{p}_{\Pi,j}^{\mathrm{MADAM}_i}(x)$ of $p_{\Pi,j}^{\mathrm{MADAM}_i}(x)$ is obtained by applying (5) to each of the B factors in (11), meaning that the probability approximation is applied to the sub-vector $(T_k : (\ell-1)b+1 \le k \le \ell b)$ instead of the full vector $\mathbf{T}$, for every $1 \le \ell \le B$.

(b) Neighborhood thresholding: For every marker j, only one submatrix R_j of dimension $(b \times b)$ belonging to the $b-1$ loci adjacent to j is kept, while all other correlations are set to 0. This leads to the approximated LD matrices

$$\hat{\Sigma}_{C,j} = \begin{pmatrix} I_b & 0 & \cdots & & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & \cdots & R_j & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & \cdots & & I_b \end{pmatrix}, \quad j = 1, \ldots, M_C, \tag{12}$$

where I_b denotes the identity matrix in dimension $(b \times b)$. Again this approximation induces a valid p-value, because

$$\begin{aligned} p_j &\leq 1 - \prod_{k=1}^{j-b/2} \mathbb{P}(T_k \leq t_j)\, \mathbb{P}\left(\bigcap_{k=j-b/2+1}^{j+b/2} \{T_k \leq t_j\} \right) \prod_{k=j+b/2+1}^{M_C} \mathbb{P}(T_k \leq t_j) \\ &= 1 - \mathbb{P}\left(\bigcap_{k=j-b/2+1}^{j+b/2} \{T_k \leq t_j\} \right) F_{\chi^2_2}(t_j)^{M_C - b}. \end{aligned} \tag{13}$$

The final approximation $\hat{p}^{\mathrm{MADAM}_i}_{\Pi,j}(x)$ of $p^{\mathrm{MADAM}_i}_{\Pi,j}(x)$ is obtained by applying (5) to the probability expression in (13).

Obviously, the p-value $\tilde{p}^{\mathrm{MADAM}_i}_{\Pi,j}(x)$ yields a closer approximation of p_j than $\hat{p}^{\mathrm{MADAM}_i}_{\Pi,j}(x)$, because more information is kept. On the other hand, for every j one has to apply (5) B times in order to compute $\tilde{p}^{\mathrm{MADAM}_i}_{\Pi,j}(x)$, while one single application of (5) suffices to compute $\hat{p}^{\mathrm{MADAM}_i}_{\Pi,j}(x)$.

Remark 2. (a) The MADAM_1 is equal to the Šidák-correction which is typically used in case of stochastically independent test statistics. Thus, for both approximation schemes (a) and (b) from above it holds

$$\tilde{p}^{\mathrm{MADAM}_1}_{\Pi,j}(x) = \hat{p}^{\mathrm{MADAM}_1}_{\Pi,j}(x) = p_{\check{S},j}(x) := 1 - F_{\chi^2_2}(t_j)^{M_C}.$$

(b) Application of the MADAM_i with $i > 1$ to genetic association study data requires the computation of probability bounds of order i for multivariate chi-square distributions. The computation and the computational efficiency of such bounds have been described and discussed in Stange et al. (2016), where also computer programs are available as supplementary material upon request.

3.3. Data analysis

For a numerical demonstration, we consider here a study comprising genotype data of $n = 2729$ individuals. The number of markers under consideration varies between the chromosomes, ranging from $M_1 = 58{,}528$ SNPs on chromosome 1 to $M_{22} = 9563$ SNPs on chromosome 22. In this study, most markers are found on the second chromosome with $M_2 = 61{,}103$. Further, for each individual six different behavioral phenotypes were assessed in the study. The data are stored in PLINK-formatted files (see Purcell et al., 2007). Therefore, the first steps of data analysis were performed with the open-source software PLINK. For instance, with PLINK the pairwise correlations between markers were estimated. To this end, the definition of genotypic correlations as in Wellek and Ziegler (2008) or Chapter 10 in Ziegler et al. (2010) was used. Further computations were then performed with MATLAB, e.g., computation of the test statistics T_j. For the computation of the p-value approximations $\tilde{p}^{\mathrm{MADAM}_3}_{\Pi,j}(x)$ and $\hat{p}^{\mathrm{MADAM}_3}_{\Pi,j}(x)$, we employed MATLAB routines for the evaluation of two- and three-dimensional χ^2-distribution functions, which were developed in Stange et al. (2016). For the results reported in Appendix A we set the block size to $b = 100$ for scheme (a), and we used block sizes $b = 100$ and $b = 200$ in scheme (b). In general, the choice of the block size b depends on biological, technological and on computational considerations. From the biological perspective, it has long been known that high LD levels in humans rarely extend beyond 100–200 kilobases (kb). Indeed, after 100 kb most pairwise LD measures have already been reduced to $r^2 < 0.1$ (for instance, see Figure 1 in Dawson et al. (2002)). From the technological perspective, commercial arrays usually have much less than one SNP per kb, so that blocks of length $b = 100$ and $b = 200$ are likely to contain the vast majority of these dependencies. Finally, in computational terms, it would be impractical to check for full chromosomes, since the number of potential dependencies to consider would be too large.

The tables in Appendix A contain the results for three of the six phenotypes. Tables containing the results for the remaining three phenotypes are provided in Stange et al. (2016). These tables illustrate the gain in power which is possible by applying the MADAM_3, compared with a univariate Šidák-correction.

Furthermore, we also included p-values resulting from the step-down (SD) variant of the Šidák method and from the nonparametric, permutation-based "maxT" procedure proposed in Westfall and Young (1993), respectively. The columns

Table A.1
Results for the first phenotype. For the "maxT" procedure, 9999 random permutations have been employed, together with the identity permutation.

Id	C	M_C	T	p_{loc}
rs17009384	3	50 864	25.068	3.6023632e−06
rs41368544	6	46 044	24.242	5.4446072e−06
rs17076797	6	46 044	23.920	6.3963029e−06
rs2683561	10	40 184	23.906	6.4411680e−06
rs730242	16	22 704	23.082	9.7226976e−06
rs1322990	9	35 148	22.782	1.1298956e−05
rs6940980	6	46 044	22.571	1.2554344e−05
rs9320543	6	46 044	22.525	1.2844489e−05
rs4129267	1	58 528	22.282	1.4507236e−05
rs9488718	6	46 044	22.237	1.4832580e−05

Id	$p_{\check{S},j}$	$p_{\check{S}^{\text{SD}},j}$	$\hat{p}_{\Pi,j}^{\text{MADAM}_3}$ $(b = 100)$	$\hat{p}_{\Pi,j}^{\text{MADAM}_3}$ $(b = 200)$	$\tilde{p}_{\Pi,j}^{\text{MADAM}_3}$	$p_{\text{maxT},j}$
rs17009384	0.1674241	0.1674241	0.1673261	0.1672423	0.1169289	0.0892
rs41368544	0.2217381	0.2217381	0.2216367	0.2214924	0.1569391	0.1120
rs17076797	0.2551052	0.2551005	0.2549915	0.2548290	0.1816032	0.1307
rs2683561	0.2280479	0.2280479	0.2279864	0.2278656	0.1611671	0.1231
rs730242	0.1980790	0.198079	0.1978343	0.1976757	0.1431898	0.1233
rs1322990	0.3277587	0.3277587	0.3276230	0.3274678	0.2378089	0.1800
rs6940980	0.4390120	0.438998	0.4388389	0.4386177	0.3243108	0.2460
rs9320543	0.4464568	0.4464355	0.4462961	0.4460599	0.3303719	0.2510
rs4129267	0.5721941	0.5721941	0.5719520	0.5717569	0.4296207	0.3356
rs9488718	0.4948785	0.4948486	0.4947035	0.4944592	0.3704433	0.2834

corresponding to the SD variant of the Šidák method clearly reveal that changing the structure of the multiple test from single-step to step-down has a negligible effect on the adjusted p-value when compared with the effect achieved by the exploitation of (parts of) the dependency structure. The "maxT" procedure yields smaller p-values than the MADAM for phenotypes 1, 3, and 4, but larger ones for phenotypes 2, 5, and 6. However, one may be concerned about the validity of the chi-square approximation of the marginal null distributions of test statistics in the case of phenotype 6, because for this phenotype the numbers of cases (331) and controls (2398) are highly unbalanced. In such a case, the "maxT" procedure may be the better choice.

4. Discussion

We have demonstrated how to apply sum- and product-type approximations of joint probabilities for the computation of multiplicity- and dependency-adjusted p-values for control of the FWER. As these p-values incorporate parts of the correlation structure in the data, this leads to a better exhaustion of the nominal significance level, and thus to a more powerful multiple test procedure than common generic methods, which are typically conservative (not exhausting the FWER level α). Compared to previous work on effective numbers of tests, the main advantage of the MADAM is that it can be applied without relying on a pre-specified value of α, which also facilitates the computations (no iterative algorithms are necessary). As mentioned before, one drawback of the proposed MADAM is that the choice of the order i is highly model-specific. It seems impossible to derive a generic algorithm which automatically determines the appropriate i. At least we do not see how such an automated choice of i could be implemented.

Since the methodology of effective numbers of tests has its origins in the field of genetic epidemiology and is to our knowledge mainly applied there, we illustrated the MADAM on such type of data. The p-values displayed in Tables A.1–A.3 are adjusted for chromosome-wise multiplicity and block dependency. In some genetic association studies, however, one is interested in the genome-wide association test problem. In this context, one has to deal with very large values of $M \sim 10^5$ or $M \sim 10^6$, and FWER control is considered a too conservative criterion, even if multivariate methods are applied. Instead, for problems with such massive multiplicity, control of the false discovery rate (FDR, cf. Benjamini and Hochberg, 1995) has become a standard criterion. The development of multivariate methods controlling the FDR constitutes a vivid field of modern mathematical statistics. How to apply the MADAM in the context of FDR control is an interesting and challenging direction for future research. In this, bounds or approximations for expectations of ratios of dependent random variables are needed. A hybrid two-stage approach for the analysis of whole-genome or genome-wide association studies was recommended in Dickhaus et al. (2012) (see also the references in this article for earlier developments). In the first (screening) step of such a two-stage analysis, all M markers are tested for association employing a non-stringent type I error measure like the FDR in order to identify candidate SNPs. In the second (validation) step, these candidate SNPs are then tested on an independent data sample. In this confirmatory step the FWER is the appropriate type I error measure, and the MADAM can be applied on the reduced set of candidate markers which typically has an order of magnitude of 10^3, as considered in Section 3.

Table A.2
Results for the third phenotype. For the "maxT" procedure, 9999 random permutations have been employed, together with the identity permutation.

Id	C	M_C	T	p_{loc}
rs16872525	7	39 982	25.753	2.5571790e−06
rs7628096	3	50 864	22.707	1.1728766e−05
rs13000805	2	61 103	22.659	1.2015884e−05
rs11216411	11	37 115	22.330	1.4158464e−05
rs12794686	11	37 115	22.263	1.4645703e−05
rs3757142	6	46 044	22.220	1.4964335e−05
rs3757146	6	46 044	22.220	1.4964335e−05
rs3757148	6	46 044	22.220	1.4964335e−05
rs11683516	2	61 103	22.060	1.6204048e−05
rs17731	10	40 184	21.804	1.8422075e−05

Id	$p_{\check{S},j}$	$p_{\check{S}^{\text{SD}},j}$	$\hat{p}^{\text{MADAM}_3}_{\Pi,j}$ $(b=100)$	$\hat{p}^{\text{MADAM}_3}_{\Pi,j}$ $(b=200)$	$\tilde{p}^{\text{MADAM}_3}_{\Pi,j}$	$p_{\text{maxT},j}$
rs16872525	0.0971883	0.0971883	0.0971335	0.0970476	0.0678973	0.0521
rs7628096	0.4493057	0.4493057	0.4491249	0.4489139	0.3313491	0.2648
rs13000805	0.5201161	0.5201161	0.5199587	0.5197999	0.3880328	0.3126
rs11216411	0.4087375	0.4087375	0.4084143	0.4080645	0.2957558	0.2393
rs12794686	0.4193338	0.4193253	0.4190808	0.4188688	0.3041531	0.2470
rs3757142	0.4979336	0.4979336	0.4977042	0.4974477	0.3730113	0.3008
rs3757146	0.4979336	0.4979336	0.4977117	0.4974542	0.3730113	0.3008
rs3757148	0.4979336	0.4979336	0.4977055	0.4974467	0.3730113	0.3008
rs11683516	0.6284694	0.6284633	0.6282879	0.6281430	0.4836979	0.3996
rs17731	0.5230194	0.5230194	0.5228212	0.5225375	0.3933791	0.3223

Table A.3
Results for the fourth phenotype. For the "maxT" procedure, 9999 random permutations have been employed, together with the identity permutation.

Id	C	M_C	T	p_{loc}
rs4683625	3	50 864	27.487	1.0745823e−06
rs13317804	3	50 864	26.880	1.4560467e−06
rs9831276	3	50 864	23.632	7.3854128e−06
rs4447734	3	50 864	23.438	8.1376285e−06
rs11660040	18	21 992	22.719	1.1657151e−05
rs7537401	1	58 528	22.651	1.2062033e−05
rs11071658	15	21 535	22.441	1.3395667e−05
rs7565497	2	61 103	21.256	2.4232394e−05
rs6127200	20	19 075	20.975	2.7878889e−05
rs6752766	2	61 103	20.537	3.4712161e−05

Id	$p_{\check{S},j}$	$p_{\check{S}^{\text{SD}},j}$	$\hat{p}^{\text{MADAM}_3}_{\Pi,j}$ $(b=100)$	$\hat{p}^{\text{MADAM}_3}_{\Pi,j}$ $(b=200)$	$\tilde{p}^{\text{MADAM}_3}_{\Pi,j}$	$p_{\text{maxT},j}$
rs4683625	0.0531907	0.0531907	0.0531663	0.0531312	0.0365894	0.0274
rs13317804	0.0713844	0.0713831	0.0713546	0.0713059	0.0491987	0.0371
rs9831276	0.3131594	0.3131492	0.3130476	0.3128598	0.2243635	0.1693
rs4447734	0.3389422	0.3389261	0.3388155	0.3386058	0.2440722	0.1866
rs11660040	0.2261410	0.2261410	0.2258627	0.2255321	0.1599944	0.1188
rs7537401	0.5063709	0.5063709	0.5061836	0.5059677	0.3733004	0.2720
rs11071658	0.2505964	0.2505964	0.2503414	0.2501048	0.1805887	0.1446
rs7565497	0.7725193	0.7725193	0.7723564	0.7721810	0.6269534	0.5013
rs6127200	0.4124519	0.4124519	0.4117301	0.4110675	0.3040814	0.2274
rs6752766	0.8800948	0.8800907	0.8799602	0.8798123	0.7556091	0.6316

Acknowledgments

The authors are grateful to two anonymous reviewers for their careful reading of the paper and for constructive suggestions which have lead to an improvement of the presentation. Support by the ERC "Foundations of Economic Preferences (FEP)" (grant No. ERC-2011-AdG 295642-FEP), by grant BFU2012-38236 from the Spanish Ministry of Economy and Competitiveness, and by the Deutsche Forschungsgemeinschaft via grants DI 1723/3-1 and SCHU 2828/2-1 is gratefully acknowledged.

Appendix A. Tables

In Tables A.1–A.3 the results for the most significant SNPs for the phenotypes 1, 3, and 4 are displayed. Hereby, "Id" denotes the rs-identifier of the SNP, C is the corresponding chromosome with number of SNPs equal to M_C, T refers to the value of the chi-square test statistic, and p_{loc} denotes the marginal unadjusted p-value.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spl.2016.01.005.

References

Agresti, A., 2013. Categorical Data Analysis. John Wiley & Sons.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Block, H.W., Costigan, T., Sampson, A.R., 1992. Product-type probability bounds of higher order. Probab. Engrg. Inform. Sci. 6 (3), 349–370.

Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D.R., Cardon, L.R., Dunham, I., 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature 418 (6897), 544–548.

Dickhaus, T., 2014. Simultaneous Statistical Inference with Applications in the Life Sciences. Springer-Verlag, Berlin, Heidelberg.

Dickhaus, T., Gierl, J., 2013. Simultaneous test procedures in terms of *p*-value copulae. In: Proceedings on the 2nd Annual International Conference on Computational Mathematics, Computational Geometry & Statistics, CMCGS 2013. Global Science and Technology Forum, GSTF, pp. 75–80.

Dickhaus, T., Royen, T., 2015. A survey on multivariate chi-square distributions and their applications in testing multiple hypotheses. Statistics 49 (2), 427–454.

Dickhaus, T., Stange, J., 2013. Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. Calcutta Statist. Assoc. Bull. 65 (257–260), 123–144.

Dickhaus, T., Stange, J., Demirhan, H., 2015. On an extended interpretation of linkage disequilibrium in genetic case-control association studies. Stat. Appl. Genet. Mol. Biol. 14 (5), 497–505.

Dickhaus, T., Strassburger, K., Schunk, D., Morcillo-Suarez, C., Illig, T., Navarro, A., 2012. How to analyze many contingency tables simultaneously in genetic association studies. Stat. Appl. Genet. Mol. Biol. 11 (4), Article 12.

Gabriel, K., 1969. Simultaneous test procedures—some theory of multiple comparisons. Ann. Math. Stat. 40, 224–250.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., et al., 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (3), 559–575.

Romano, J.P., Wolf, M., 2005a. Exact and approximate stepdown methods for multiple hypothesis testing. J. Amer. Statist. Assoc. 100 (469), 94–108.

Romano, J.P., Wolf, M., 2005b. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.

Royen, T., 2014. A simple proof of the Gaussian correlation conjecture extended to some multivariate gamma distributions. Far East J. Theor. Stat. 48 (2), 139–145.

Stange, J., Bodnar, T., Dickhaus, T., 2015a. Uncertainty quantification for the family-wise error rate in multivariate copula models. AStA Adv. Stat. Anal. 99 (3), 281–310.

Stange, J., Dickhaus, T., Navarro, A., Schunk, D., 2016. Supplementary material to: "Multiplicity- and dependency-adjusted *p*-values for control of the family-wise error rate". Available online at http://dx.doi.org/10.1016/j.spl.2016.01.005.

Stange, J., Loginova, N., Dickhaus, T., 2016. Computing and approximating multivariate chi-square probabilities. J. Stat. Comput. Simul. 86 (6), 1233–1247. http://dx.doi.org/10.1080/00949655.2015.1058798.

Wellek, S., Ziegler, A., 2008. A genotype-based approach to assessing the association between single nucleotide polymorphisms. Hum. Hered. 67 (2), 128–139.

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. In: Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Wiley, New York.

Ziegler, A., König, I.R., Pahlke, F., 2010. A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-Learning Platform. John Wiley & Sons.