Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Computing and approximating multivariate chi-square probabilities

Jens Stange, Nina Loginova, Thorsten Dickhaus

submitted: August 28, 2014

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
E-Mail: Jens.Stange@wias-berlin.de
Nina.Loginova@wias-berlin.de
Thorsten.Dickhaus@wias-berlin.de

No. 2005
Berlin 2014

2010 *Mathematics Subject Classification.* 60E05; 60E15; 62E17; 65D20.

Key words and phrases. Bonferroni inequalities, chain factorization, correlation matrix, effective number of tests, linkage disequilibrium, m-factorial matrix, product-type probability approximations, sub-Markovian, Wishart matrix.

We thank Thomas Royen for many valuable suggestions. This research was partly supported by the Deutsche Forschungsgemeinschaft via grant No. DI 1723/3-1.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

We consider computational methods for evaluating and approximating multivariate chi-square probabilities in cases where the pertaining correlation matrix or blocks thereof have a low-factorial representation. To this end, techniques from matrix factorization and probability theory are applied. We outline a variety of statistical applications of multivariate chi-square distributions and provide a system of MATLAB programs implementing the proposed algorithms. Computer simulations demonstrate the accuracy and the computational efficiency of our methods in comparison with Monte Carlo approximations, and a real data example from statistical genetics illustrates their usage in practice.

1 Introduction

Multivariate chi-square distributions play a pivotal role in many applications of modern statistics. For example, they arise as (limiting) null distributions of vectors of test statistics in the context of simultaneous inference for Gaussian variances, in multi-sample problems regarding multinomial distributions, in multiple comparisons of vectors of regression coefficients, or in simultaneous categorical data analysis; see [3] for a recent overview.

Despite this importance, computational methods for M-variate chi-square probabilities with ν degrees of freedom are up to now only available for a limited number of special cases, for instance in some cases where $M = 2$ (see, e. g., [7–10, 12, 19, 22]) or if $\nu = 1$, where such probabilities can be calculated by means of multivariate normal probabilities; see [6] for a comprehensive overview of computational methods for multivariate Student's t and normal distributions.

In this work, we present computational methods for evaluating and approximating M-variate chi-square probabilities for (in principle) arbitrary dimension M and arbitrary degrees of freedom $\nu \geq 2$, provided that the correlation structure in the k-variate marginal distributions is given (or can be approximated well) by low-rank correlation matrices, where $k \in \{2, 3, 4\}$.

The paper is structured as follows. In Section 2, we outline two approximation methods which are based on results from probability theory. Section 3 is concerned with matrix factorization techniques and resulting computational methods for multivariate chi-square probabilities. The numerical study in Section 4 assesses the accuracy and the computational efficiency of the proposed methods and their MATLAB implementations in comparison with Monte Carlo methods. A real data example from the field of statistical genetics is presented in Section 5, and we conclude with a discussion in Section 6.

2 Multivariate χ^2-distributions

2.1 Notation and preliminaries

On a suitable but not further specified probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let Gaussian vectors $\mathbf{Z}_1, \ldots,$ $\mathbf{Z}_\nu \overset{i.i.d.}{\sim} \mathcal{N}_M(0, R)$ with correlation matrix $R \in \mathbb{R}^{M\times M}$ be defined for $\nu \in \mathbb{N}$. Throughout the work and without loss of generality (w.l.o.g.) we assume that the variance of each component Z_{1j} of $\mathbf{Z}_1$ is equal to 1, where $1 \leq j \leq M$, such that R is the covariance matrix of $\mathbf{Z}_1$.

Then the random vector $\mathbf{X} = (X_1, \ldots, X_M)^\mathsf{T}$ with components

$$X_j := \sum_{i=1}^{\nu} Z_{ij}^2,\ j = 1, \ldots, M, \tag{1}$$

follows the M-variate χ^2-distribution in the sense of Definition 3.5.7 in [20] with ν degrees of freedom and pertaining correlation matrix R, denoted by $\chi^2_M(\nu, R)$. The distribution of $\mathbf{X}$ is equal to the joint distribution of the diagonal elements of an M-dimensional Wishart matrix with ν degrees of freedom and pertaining correlation matrix R, $\mathcal{W}_M(\nu, R)$ for short. The stochastic representation in (1) is convenient for Monte Carlo approximations of the distribution of $\mathbf{X}$.

We are concerned with computing and approximating the cumulative distribution function (cdf) of $\max_{1\leq j\leq M} X_j$, given by

$$F_M(x) \equiv F_M(x, \nu, R) := \mathbb{P}\left(\bigcap_{j=1}^{M}\{X_j \leq x\}\right), \tag{2}$$

or equivalently

$$\bar{F}_M(x) \equiv \bar{F}_M(x, \nu, R) := 1 - F_M(x) = \mathbb{P}\left(\bigcup_{j=1}^{M}\{X_j > x\}\right),\ x > 0. \tag{3}$$

We define for $\kappa \geq 1$ and $\delta > 0$

$$G_\kappa(x; \delta) := \exp(-\delta)\sum_{k=0}^{\infty}\frac{\delta^k}{k!}\gamma(\kappa + k; x), x \geq 0, \tag{4}$$

with the (regularized) incomplete Gamma function, given by

$$\gamma(\kappa; z) = \int_0^z \psi_\kappa(t)\, dt,\ \psi_\kappa(x) := x^{\kappa-1}\exp(-x)/\Gamma(\kappa).$$

2.2 Approximations

The exact computation of $F_M(x)$, or $\bar{F}_M(x)$, respectively, is infeasible for larger dimensions M. We may remark here that this even holds true for $\nu = 1$. For example, the `R` package `mvtnorm` which is based on [6] gives an error message whenever M exceeds 1000. Therefore, we present two basic ideas for approximating $F_M(x)$, in which only the computation of lower-dimensional marginal distributions, i. e., $F_k(x)$ for some $k < M$, is required.

Lemma 1

a) (Bonferroni inequalities)

Let $A_1, \ldots, A_M$ be arbitrary events. Then

$$\forall p \geq 1: \ \sum_{k=1}^{2p} (-1)^{k-1} S_k \leq \mathbb{P}\left(\bigcup_{j=1}^{M} A_j\right) \leq b_{2p-1} := \sum_{k=1}^{2p-1} (-1)^{k-1} S_k, \qquad (5)$$

where

$$S_k = \sum_{1 \leq j_1 < j_2 < \ldots < j_k \leq M} \mathbb{P}(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_k}), \ 1 \leq k \leq M, \ S_k = 0, \ k > M.$$

A bivariate variant of the aforementioned upper Bonferroni bounds is due to [21] and is given by

$$\mathbb{P}\left(\bigcup_{j=1}^{M} A_j\right) \leq b_2 := \sum_{j=1}^{M} \mathbb{P}(A_j) - \sum_{j=1}^{M-1} \mathbb{P}(A_j \cap A_{j+1}). \qquad (6)$$

For our purposes we have to consider the events $A_j = \{X_j > x\}$ so that the probability expression in (5) *and on the left-hand side of* (6) *equals $\bar{F}_M(x)$.*

b) (Product-type probability bounds)

Define the events $O_j := \{X_j \leq x\} = A_j^c$ for $1 \leq j \leq M$.

Due to chain factorization, it holds for any $1 \leq k \leq M-1$ that

$$F_M(x) = \mathbb{P}(O_1, \ldots, O_M) = \mathbb{P}(O_1, \ldots, O_k) \prod_{j=k+1}^{M} \mathbb{P}(O_j | O_{j-1}, \ldots, O_1).$$

Now assume that $\mathbf{X}$ is sub-Markovian of order $k \geq 2$ (SM_k) in the sense of Definition 2.2 in [4]. Then it holds for all $k \leq j \leq M$ that

$$\mathbb{P}(O_j | O_{j-1}, \ldots, O_1) \geq \mathbb{P}(O_j | O_{j-1}, \ldots, O_{j-k+1}) \qquad (7)$$

and, consequently,

$$F_M(x) \geq \beta_k := \mathbb{P}(O_1, \ldots, O_k) \prod_{j=k+1}^{M} \mathbb{P}(O_j | O_{j-1}, \ldots, O_{j-k+1}). \qquad (8)$$

Occasionally, we will write $b_\ell(x)$ or $\beta_k(x)$, respectively, instead of b_ℓ or β_k, respectively, to indicate the argument x at which the approximations are evaluated. Furthermore, we refer to ℓ and k, respectively, as the order of these (sum- or product-type) approximations.

Remark 1

a) We note that the complexity of computing the sums S_k in (5) is high, because $\binom{M}{k}$ k-dimensional marginal probabilities have to be evaluated. However, for some applications, for instance in multiple testing, a conservative bound on $\bar{F}_M(x)$ is required, meaning that $\bar{F}_M(x)$ is approximated from above. Such conservative bounds are provided by the right-hand side of (5). A computationally inexpensive alternative is the utilization of b_2 from (6). Under certain structural assumptions, sum-type bounds of higher order can be improved. For example, the derivations in [13, 14] are based on geometric or topological arguments.

b) In the general case the inequality relation in (7) is not fulfilled. However, β_k often yields a good approximation of $F_M(x)$ already for $k \in \{2, 3\}$, see Section 4. In the remainder, we refer to β_k as the product-type probability approximation (PTPA) of order k to $F_M(x)$.

3 Computational details

3.1 Mathematical derivations

Computation of $F_k(\cdot, \nu, R)$ is feasible if R possesses certain structural properties. In particular, low-rank factorizations of R facilitate the computation. Specifically, in [17] it is outlined how to obtain the integral representation

$$F_k(2x, 2\kappa, R) = \mathbb{E}\left[\prod_{j=1}^{k} G_\kappa\left(d_j^{-1}x; \frac{1}{2}d_j^{-1}a_j S a_j^{\mathsf{T}}\right)\right], \tag{9}$$

if R has an m-factorial representation, cf. Definition 1. The expectation in (9) refers to an $(m \times m)$-Wishart Matrix $S \sim \mathcal{W}_m(2\kappa, I_m)$.

Definition 1
A covariance matrix $R \in \mathbb{R}^{k\times k}$ has an m-factorial representation with $1 \leq m < k$, if it allows for a decomposition

$$R = D + AA^{\mathsf{T}} \tag{10}$$

with a matrix $A = (a_1, \ldots, a_k) \in \mathbb{C}^{k\times k}$ of rank m, the columns of which are real vectors $a_j \in \mathbb{R}^k$ or purely imaginary vectors $a_j \in (i\mathbb{R})^k$, and a positive definite diagonal matrix $D = diag(d_1, \ldots, d_k)$, where $d_j > 0$ for all $1 \leq j \leq k$.

Remark 2

(i) To simplify notation, we consider related decompositions $R = D + AA^{\mathsf{T}}$ with full-rank $A = (a_1, \ldots, a_m) \in \mathbb{C}^{k\times m}$.

(ii) Notice that a more general definition of m-factorial matrices is given, e. g., in Definition 1 of [17], where also negative values of the d_j are allowed. For feasible computations, however, restriction to positive values of the d_j is required.

Example 1

a) Every correlation matrix $R \in \mathbb{R}^{k\times k} \setminus \{I_k\}$ has a $(k-1)$-factorial representation. To see this, let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ be the diagonal matrix containing the eigenvalues of R, where w.l.o.g. $\lambda_k = \min_{1\le j\le k} \lambda_j$. Furthermore, let $\tilde{\Lambda} = \Lambda - \lambda_k I_k$. Application of the Spectral Theorem for symmetric matrices yields the decomposition

$$R = U\Lambda U^{\mathsf{T}} = U\lambda_k I_k U^{\mathsf{T}} + U\tilde{\Lambda}U^{\mathsf{T}}. \tag{11}$$

Since U is an orthogonal matrix, it is clear that a $(k-1)$-factorial representation with

$$D = \lambda_k I_k, \text{ and } A = U(\tilde{\Lambda})^{1/2}$$

is given by (11).

b) Let $\rho \in (-1, 1)$. The equi-correlation matrix $R = R(\rho)$ with entries 1 on the diagonal and ρ off the diagonal has a one-factorial representation

$$R(\rho) = \text{diag}(1-\rho, \ldots, 1-\rho) + AA^{\mathsf{T}},$$

where $A = (\sqrt{\rho}, \ldots, \sqrt{\rho})^{\mathsf{T}} \mathbb{C}^{k\times 1}$ with "rows" $a_j \equiv \sqrt{\rho} \in \mathbb{R} \cup i\mathbb{R},\ j = 1, \ldots, k.$

The following lemma provides sufficient conditions for the existence of an m-factorial representation of a given correlation matrix R.

Lemma 2

a) Let $R = (r_{ij})$ be a $(k \times k)$ correlation matrix, $k \ge 3$, with $r_{ij} \neq 0$ for all $1 \le i, j \le k$. Then R is one-factorial if and only if

$$\exists c \in \mathbb{R} \cup i\mathbb{R} \text{ with } c^2 < 1 \text{ such that } \forall 1 < j < \ell \le k : \ \frac{r_{1j} r_{1\ell}}{r_{j\ell}} = c^2. \tag{12}$$

In this case a representation is given by

$$R = D + aa^{\mathsf{T}},\ a = (a_1, \ldots, a_k), a_1 = c, a_j = \frac{r_{1j}}{c} \text{ for } j = 2, \ldots, k$$

and $D = \text{diag}(1 - a_1^2, \ldots, 1 - a_k^2)$.

b) Let $R \in \mathbb{R}^{k\times k}$ be a correlation matrix with spectral decomposition $R = U\Lambda U^{\mathsf{T}}$, where $U \in \mathbb{R}^{k\times k}$ is an orthogonal matrix and the diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k) > 0$ contains the eigenvalues of R. W.l.o.g., assume that $\lambda_1 \le \ldots \le \lambda_k$. If there exists an integer $\tilde{k} > 1$ with $\lambda_1 = \ldots = \lambda_{\tilde{k}}$, then R has an m-factorial representation with $m = k - \tilde{k}$.

Proof.

a) Let us first assume that R has a one-factorial representation. This property is equivalent to the solvability of the following system of equations.

$$\begin{array}{lllllllll} r_{12} & = & a_1a_2, & r_{13} & = & a_1a_3, & \cdots & r_{1k} & = & a_1a_k, \\ & & & r_{23} & = & a_2a_3, & \cdots & r_{2k} & = & a_2a_k, \\ & & & & & & \ddots & \vdots & & \\ & & & & & & & r_{(k-1)k} & = & a_{k-1}a_k. \end{array} \tag{13}$$

By assumption on R there exists a vector $(a_1, \ldots, a_k) \in \mathbb{C}^k$ which solves (13). One can set $c := a_1 \neq 0$, which is a real or an imaginary constant fulfilling $c^2 < 1$. Then put $a_j = \frac{r_{1j}}{c}$ for $j = 2, \ldots, k$.

In order to check the validity of this solution one has to verify that

$$\begin{array}{lllll} r_{23} & = & \frac{r_{12}r_{13}}{c^2} & \cdots & r_{2k} & = & \frac{r_{12}r_{1k}}{c^2} \\ & & & \ddots & \vdots & & \\ & & & & r_{(k-1)k} & = & \frac{r_{1(k-1)}r_{1k}}{c^2}, \end{array} \tag{14}$$

which is equivalent to the identities

$$c^2 = \frac{r_{12}r_{13}}{r_{23}} = \frac{r_{12}r_{14}}{r_{24}} = \cdots = \frac{r_{1(k-1)}r_{1k}}{r_{(k-1)k}}, \tag{15}$$

showing the assertion.

On the other hand, if the identities in (12) hold, one can construct a one-factorial representation by setting $a_1 = c$ and using (13) for defining the values of $a_2, \ldots, a_k$.

b) Setting $D := \lambda_1 I_k$, it holds

$$R = D + U\tilde{\Lambda}U^{\mathsf{T}} \tag{16}$$

with $\tilde{\Lambda} = \Lambda - D = \text{diag}(0, \ldots, 0, \lambda_{\tilde{k}+1} - \lambda_1, \ldots, \lambda_k - \lambda_1)$. It is easy to see that $A := U\tilde{\Lambda}^{1/2}$ has rank $k - \tilde{k}$, and therefore decomposition (16) is a valid $(k-\tilde{k})$-factorial representation.

□

Remark 3

a) In dimension $k = 3$ the condition in (12) is satisfied if $r_{12}r_{13}r_{23} \neq 0$ and $r_{12}r_{13}r_{23}^{-1} < 1$.

b) Notice that the eigenvalue condition in part b) in Lemma 2 will in general not be fulfilled for correlation matrices occurring in applications. However, one can use this lemma for an approximation of a general correlation matrix R by an m-factorial covariance matrix Σ. Namely, entries in the diagonal matrix of eigenvalues of R can be substituted such that the eigenvalue condition is satisfied.

For the evaluation of (9), it is necessary to compute integrals of the form $\mathbb{E}[g(v^{\mathsf{T}}Sv)]$ for a Wishart-distributed random matrix $S \sim \mathcal{W}_m(\nu, I_m)$, a (column) vector $v \in \mathbb{R}^m$ and some scalar function g. This amounts to an $m(m+1)/2$-dimensional integration, because integration has to be performed with respect to all non-redundant matrix entries. Proposition 1 provides explicit expressions in the cases of $m = 1$ and $m = 2$.

To this end, in addition to the notation in Section 2, we introduce two further functions. The first, namely $f_\kappa : (0, \pi) \longrightarrow [0, \infty)$, is given by

$$f_\kappa(\varphi) := \frac{\sqrt{\pi}\Gamma(\kappa - 1/2)}{\Gamma(\kappa)}(\sin^2(\varphi))^{\kappa-1} \tag{17}$$

and the second, namely $h(\cdot, \cdot, \cdot; \beta_1, \beta_2) : [0, \infty)^2 \times [0, \pi] \longrightarrow [0, \infty)$, is given by

$$h(s, t, \varphi; \beta_1, \beta_2) = \beta_1^2 s + \beta_2^2 t + 2\beta_1\beta_2 \cos(\varphi) st \tag{18}$$

for constants $\beta_1, \beta_2 \in \mathbb{C}$.

Proposition 1 (cf. [3, 15–17])
Let $R \in \mathbb{R}^{k\times k}$ be a correlation matrix.

(i) Assume that R has a one-factorial representation $R = D + aa^{\mathsf{T}}$ with a column vector $a = (a_1, \ldots, a_k)^{\mathsf{T}}$ the entries $a_j \in \mathbb{R} \cup i\mathbb{R}$ of which satisfy $a_j^2 < 1$ for $j = 1, \ldots, k$. Then the cdf of the k-variate χ^2-distribution with $\nu = 2\kappa \in \mathbb{N}$ degrees of freedom and pertaining correlation matrix R is given by

$$F_k(2x; R, 2\kappa) = \int_0^\infty \prod_{j=1}^k G_\kappa\left(\frac{x}{1-a_j^2}; \frac{a_j^2}{1-a_j^2}t\right)\psi_\kappa(t)\, dt. \tag{19}$$

(ii) Assume that R has a two-factorial representation

$$R = D + AA^{\mathsf{T}}, A \in \mathbb{C}^{p\times 2}.$$

Let $W = D^{-\frac{1}{2}}$ and $B = WA \in \mathbb{C}^{k\times 2}$, which entails $WRW = I_k + BB^{\mathsf{T}}$.

Then, with f_k from (17) *and h from* (18)*, the cdf of the $\chi_k^2(\nu, R)$-distribution with $\nu = 2\kappa \in \mathbb{N}$ degrees of freedom and pertaining correlation matrix R is given by*

$$\begin{aligned} F_k(2x; R, 2\kappa) = \int_0^\pi \int_0^\infty \int_0^\infty \prod_{j=1}^k G_\kappa(w_j^2 x; h(s, t, \varphi; b_{j1}, b_{j2}))\times \\ \psi_\kappa(s)\psi_\kappa(t) f_\kappa(\varphi)\, ds\, dt\, d\varphi. \end{aligned} \tag{20}$$

(iii) Suppose $R = (r_{ij}) \in \mathbb{R}^{4\times 4}$ is such that an index $\ell \in \{1, \ldots, 4\}$ exists for which the conditional covariance matrix

$$R_{\cdot|\ell} := R_{-\ell} - r_\ell r_\ell^{\mathsf{T}}$$

with

$$R_{-\ell} = (r_{ij})_{i,j\neq\ell} \in \mathbb{R}^{3\times 3} \text{ and } r_\ell = (r_{k\ell})_{k\neq\ell} \in \mathbb{R}^3$$

has a one-factorial representation, say

$$R_{\cdot|\ell} = W^{-2} + aa^\mathsf{T}, W = \text{diag}(w_1, w_2, w_3), a = (a_1, a_2, a_3)^\mathsf{T} \in \mathbb{C}^3.$$

Then, with $b_1 = Wa \in \mathbb{C}^3$, $b_2 = Wr_\ell \in \mathbb{R}^3$, f_k *from* (17) *and* h *from* (18)*, the cdf of the* $\chi^2_4(\nu, R)$*-distribution with* $\nu = 2\kappa \in \mathbb{N}$ *degrees of freedom and pertaining correlation matrix* R *is given by*

$$F_4(2x; R, 2\kappa) = \int_0^\pi \int_0^x \int_0^\infty \prod_{j=1}^3 G_\kappa(w_j^2 x; h(s,t,\varphi; b_{1j}, b_{2j})) \times \psi_\kappa(s)\psi_\kappa(t) f_\kappa(\varphi)\, ds\, dt\, d\varphi. \tag{21}$$

To sum up, Proposition 1 in connection with the first part of Example 1 allows for computing $F_3(x, \nu, R)$ for arbitrary $\nu \geq 2$ and R. The value $F_4(x, \nu, R)$ can be computed exactly if (i) R possesses a one-factorial representation (by means of Proposition 1.(i)), which can be checked by applying part a) of Lemma 2, or (ii) at least one three-dimensional conditional covariance matrix pertaining to R has a one-factorial representation (see Proposition 1.(iii)). If $R \in \mathbb{R}^{4\times 4}$ fulfills neither of the latter two conditions we propose two approximations to $R = U\Lambda U^\mathsf{T}$ (w.l.o.g. $\lambda_1 \geq \ldots, \geq \lambda_4$), which both rely on a substitution of the eigenvalues of R. Namely, consider

(I) $\tilde{R} = U\tilde{\Lambda}U^\mathsf{T}$, where $\tilde{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \bar{\lambda}, \bar{\lambda})$ and $\bar{\lambda} = 0.5(\lambda_3 + \lambda_4)$ is the average of the two smallest eigenvalues,

(II) $\tilde{\tilde{R}} = U\tilde{\tilde{\Lambda}}U^\mathsf{T}$, with $\tilde{\tilde{\Lambda}} = \text{diag}(\lambda_1, \lambda_2, \lambda_4, \lambda_4)$.

Both $\tilde{R}$ and $\tilde{\tilde{R}}$ possess a two-factorial representation according to part b) of Lemma 2. An approximation $F_4(x, \nu, \tilde{R})$ or $F_4(x, \nu, \tilde{\tilde{R}})$, respectively, of $F_4(x, \nu, R)$ is then computable by means of Proposition 1.(ii).

Remark 4
In principle, $F_4(x, \nu, R)$ can also be computed exactly if R itself has a two-factorial representation, see Proposition 1.(ii). However, we are not aware of any method to check the existence of a two-factorial representation (in the sense of Definition 1) of a four-variate correlation matrix. Lemma 2 in [17] offers the possibility to perform such a check, but it does not ensure positive values of the entries in the diagonal matrix D appearing in (10). Hence, an automated application of Proposition 1.(ii) to $R \in \mathbb{R}^{4\times 4}$ is only possible in special cases by our implementations.

3.2 Notes on the implementation

The computation of $F_k(\cdot, \nu, R)$ for $\nu \geq 2$ and a correlation matrix R by means of the integral representations in Proposition 1 has been implemented in `MATLAB` (version R2013b). All `MATLAB` programs are available from the first author upon request. The computations consist of three principle steps.

a) check of correlation matrices:

In order to find the suitable representation, the structure of the correlation matrix R is analyzed by means of Lemma 2. It is noted that the computation can only be performed exactly if all eigenvalues of R are greater than a certain (small) threshold. Otherwise, the variable corresponding to the smallest eigenvalue is eliminated and computation is done in dimension $k-1$.

b) numerical integration:

The one-dimensional integral displayed in (19) is approximated by the built-in `MATLAB` function `integral` on the finite interval $[0, U]$. The upper bound U is chosen automatically and only depends on the degrees of freedom ν or $\kappa = \nu/2$, respectively.

For the three-dimensional integrals in (20) or (21), respectively, integration with respect to $\varphi \in [0, \pi]$ is approximated with a composite Simpson's rule. The inner, two-dimensional integral with respect to the variables s and t is computed on a fixed grid of values $0 = \varphi_1 < \varphi_2 < \ldots < \varphi_N = \pi$ for φ with the built-in `MATLAB` function `integral2`. Integration is performed on a finite area $[0, U]^2$ or $[0, U] \times [0, x]$, respectively. Again, N and U are determined automatically.

c) approximation of the series $G_\kappa(x, \delta)$:

The recursion $T_0 = 1, T_{k+1} = T_k \delta/(k+1),\ k \geq 0$, is used to compute the partial sums

$$\exp(-\delta) \sum_{k=0}^{K} \frac{\delta^k}{k!} \gamma(\kappa + k; x) = \exp(-\delta) \sum_{k=0}^{K} T_k \gamma(\kappa + k; x),$$

where K is chosen automatically.

4 Numerical experiments

In this section we assess the numerical accuracy and the computational efficiency of the proposed implementations. We compare the analytical formulas from Section 3.1 with Monte Carlo approximations which are based on the stochastic representation (1). In this, we restrict our attention to the case of $\nu = 2$ degrees of freedom due to its relevance for the application that we are going to present in Section 5.

4.1 Validity of implemented routines for $k \in \{3, 4\}$

Here we present some results for correlation matrices in $\mathbb{R}^{3\times3}$ (Table 1) and in $\mathbb{R}^{4\times4}$ (Tables 2 and 3). As noted before, it is possible to compute $F_3(\cdot, 2, R)$ exactly up to any required numerical precision, unless the correlation matrix is close to singularity. On the contrary, it is only possible to compute $F_4(\cdot, 2, R)$ for special structures of R. Here we have chosen matrices which satisfy these structural requirements. While Table 2 compares exact computations with Monte Carlo approximations, Table 3 compares the proposed approximation methods (I) and (II)

from below Proposition 1 with the respective exact values. From our results (confirmed by further simulations not presented here) the approximation $\tilde{R}$ is recommended over the approximation $\tilde{\tilde{R}}$. Alternative approximations based on Taylor expansions can be found in Section 6.5 of [3].

x	$F_3(x,2,R)$	$\hat{F}_3(x,2,R)$
2	0.2778281	0.2779077
4	0.6646115	0.6646023
6	0.8643435	0.8643429
8	0.9478280	0.9478538
10	0.9803673	0.9803720
12	0.9926888	0.9926845
14	0.9972918	0.9972840
16	0.9989997	0.9990013
time	0.08s	554.07s

x	$F_3(x,2,R)$	$\hat{F}_3(x,2,R)$
2	0.3225955	0.3226139
4	0.6935769	0.6936257
6	0.8766919	0.8767142
8	0.9524740	0.9525019
10	0.9820150	0.9820436
12	0.9932537	0.9932654
14	0.9974814	0.9974821
16	0.9990621	0.9990631
time	8.54s	760.75s

Table 1: Numerical values of $F_3(\cdot,2,R)$. The correlation matrix R in the left table is one-factorial, hence the formula from Proposition 1.(i) is used for computation. The correlation matrix R in the right table is two-factorial, and the formula from Proposition 1.(ii) is applied. The values $\hat{F}_3(\cdot,2,R)$ correspond to Monte Carlo approximations by means of (1) with 10^8 independent repetitions. The last row indicates computing time.

x	$F_4(x,2,R)$	$\hat{F}_4(x,2,R)$
2	0.2993742	0.2993301
4	0.6731993	0.6732463
6	0.8623824	0.8623593
8	0.9446312	0.9446057
10	0.9782977	0.9782724
12	0.9916350	0.9916207
14	0.9968125	0.9968157
16	0.9987953	0.9987975
time	0.14s	539.41s

x	$F_4(x,2,R)$	$\hat{F}_4(x,2,R)$
2	0.1956750	0.1956808
4	0.5924941	0.5925014
6	0.8292380	0.8292630
8	0.9333548	0.9333743
10	0.9747319	0.9747557
12	0.9905420	0.9905466
14	0.9964819	0.9964933
16	0.9986957	0.9986995
time	6.22s	902.18s

Table 2: Numerical values of $F_4(\cdot,2,R)$. The correlation matrix R in the left table is one-factorial, hence the formula from Proposition 1.(i) is used for computation. The correlation matrix R in the right table has a pertaining one-factorial conditional covariance matrix, and the formula from Proposition 1.(iii) is applied. The values $\hat{F}_4(\cdot,2,R)$ correspond to Monte Carlo approximations by means of (1) with 10^8 independent repetitions. The last row indicates computing time.

Summarizing the results of Tables 1 and 2 we observe a clear advantage of the exact computational methods in comparison with Monte Carlo approximations. The methods from Proposition 1 are as accurate as a Monte Carlo approximation with a huge number of pseudo repetitions, but their computing time is drastically smaller. In view of applications to multiple test problems with several thousands of hypotheses as occurring frequently in modern life sciences (see our Section 5 and Part II of [1] for some examples), the proposed computational methods

x	$F_4(x,2,R)$	$F_4(x,2,\tilde{R})$	$F_4(x,2,\tilde{\tilde{R}})$
2	0.181286	0.179202	0.223520
4	0.580074	0.577999	0.637647
6	0.823400	0.821883	0.859265
8	0.931065	0.930019	0.948667
10	0.97391	0.973257	0.981698
12	0.990263	0.989887	0.993521
14	0.996390	0.996185	0.997705
16	0.998666	0.998557	0.999182

Table 3: Comparison of the approximations of R by $\tilde{R}$ or $\tilde{\tilde{R}}$, respectively. The approximation by averaging the two smallest eigenvalues yields a tighter approximation and approximates $F_4(\cdot, 2, R)$ from below. With increasing x the approximations become better. The computation of the exact probabilities is performed by means of Proposition 1.(iii).

are therefore clearly preferable. In such applications, $F_3(x)$ or $F_4(x)$ has to be evaluated for large (multiplicity-adjusted) quantiles x several thousand times.

4.2 Approximations for $M > 4$

Here we consider higher dimensionalities M and assess the numerical accuracy of the probability bounds from Lemma 1.

Table 4 corresponds to $M = 10$. We compare the PTPAs β_2 and β_3 with Bonferroni bounds of order 2 and 3, respectively. Because of the high number of terms which have to be evaluated for the computation of b_3 there is no gain in computing time in comparison with the more exact Monte Carlo approximation. However, the PTPA β_3 is very fast and yields very good approximations. The advantages of b_2 are its guaranteed conservativeness (meaning that $F_{10}(\cdot, 2, R)$ is approximated from below) and its small computing time.

In analogy to Table 4, Table 5 presents results for dimensionality $M = 20$. Due to its immense computation time we dispense with b_3 in Table 5.

5 Application to genetic association studies

Genetic association studies lead to simultaneous categorical data analysis, meaning that many (2×2) or (2×3) contingency tables have to be analyzed simultaneously, where every single contingency table summarizes study data for one position (locus) on the human genome. The aim of the statistical analysis is to carry out a test for association with a given (typically binary) phenotype at every locus under consideration. To this end often a family of χ^2 test statistics is constructed. For more details about statistical models and inferential methods we defer the reader to [2] and Section 4.1 of [4]. Here we only mention that the locus-specific χ^2 statistics

x	$\hat{F}_{10}(x,2,R)$	$1-b_3$	β_3	β_2	$1-b_2$
2	0.0934656	0	0.0609471	0.0393610	0
4	0.4509507	0	0.3929242	0.3436422	0.1080684
6	0.7368419	0.5391698	0.7023297	0.6729185	0.6287352
8	0.8862372	0.8454660	0.8718640	0.8602538	0.8530283
10	0.9534348	0.9442863	0.9482151	0.9443324	0.9432289
12	0.9815152	0.9792150	0.9797413	0.9785392	0.9783770
14	0.9927744	0.9920345	0.9922077	0.9918605	0.9918371
16	0.9972293	0.9967913	0.9970237	0.9969403	0.9969370
time	744.83s	583.73s	38.56s	0.48s	0.43s

Table 4: Comparison of sum-type approximations and product-type approximations in case of $M = 10$. The values $\hat{F}_{10}(\cdot, 2, R)$ correspond to Monte Carlo approximations by means of (1) with 10^8 independent repetitions. The last row indicates computing time.

x	$\hat{F}_{20}(x,2,R)$	β_3	β_2	$1-b_2$
2	0.0259720	0.002834	0.002494	0
4	0.2947225	0.146186	0.141662	0
6	0.6203213	0.487080	0.490405	0.334661
8	0.8239834	0.758304	0.769148	0.744267
10	0.9245446	0.899017	0.904437	0.900520
12	0.9690100	0.960167	0.962528	0.961944
14	0.9876556	0.984835	0.985681	0.985596
16	0.9952047	0.994312	0.994599	0.994587
time	1143.4s	62.71s	2.37s	2.49s

Table 5: Comparison of sum-type approximations and product-type approximations in case of $M = 20$. The values $\hat{F}_{20}(\cdot, 2, R)$ correspond to Monte Carlo approximations by means of (1) with 10^8 independent repetitions. The last row indicates computing time.

typically exhibit strong dependencies within blocks of genetic loci due to the biological mechanism of inheritance. Hence, under the global hypothesis of no genotype-phenotype associations at all, the computational methods described in Sections 2 to 4 are highly relevant for calibrating simultaneous test procedures in the sense of [5] with respect to control of the familywise error rate (FWER). If strict FWER control is targeted it is important to approximate $\bar{F}_M(x, \nu, R)$ from above. In this, M denotes the number of loci, x denotes a multiplicity-corrected χ^2 quantile, $\nu = 1$ ($\nu = 2$) in case of marginal (2×2) or (2×3) contingency tables, respectively, and the correlation matrix R encodes the dependencies among loci.

We note that the aforementioned strong dependencies among the test statistics allow for a relaxation of the necessary adjustment for multiplicity in comparison with the case of independent test statistics. Following [11] and [4], one transparent way to express the relaxed multiplicity correction is to compute so-called effective numbers of tests based on PTPAs. Information about the so-called linkage disequilibrium (LD) matrix R is publicly available from web-based databases

for a variety of target populations. Here, we compute an effective number of tests of order 3 in the sense of Theorem 3.1 in [4] based on LD information taken from the international HapMap project (`http://hapmap.ncbi.nlm.nih.gov/`). For exemplary purposes we restrict our attention to loci on chromosome 21 in the CHD population (Chinese in Metropolitan Denver).

In total this chromosome comprises $M = 18{,}143$ loci. We divided the chromosome into $B = 363$ blocks of size 50 each (notice the last of these blocks comprises only 43 loci). We consider $\nu = 2$ degrees of freedom corresponding to (2×3) contingency tables (i. e., to diploid genotypes) and make the assumption that loci from different blocks lead to stochastically independent χ^2 test statistics. Under this assumption the effective number of tests can be calculated for every block separately, and the total effective number of tests is the sum of the block-specific ones. We may remark here that, if the assumption of independent blocks is not fulfilled, our proposed effective number of tests is conservative in the sense that it over-estimates the true value. This is due to the extension of the Gaussian correlation inequality to multivariate chi-square distributions which has recently been proved in [18]. Throughout the remainder we consider FWER control at level $\alpha = 0.05$.

Based on Theorem 3.1 in [4] the effective number of tests of order 3 in one block b is given by

$$M_{\text{eff},b}^{(3)} = \frac{\log(1-\alpha_B)}{\log(F_1(x_b,2))} = \frac{\log(1-\alpha)}{B\log(F_1(x_b,2))}, \quad 1 \leq b \leq B = 363,$$

where $\alpha_B = 1-(1-\alpha)^{1/B}$ corresponds to the assumption of independent blocks and x_b is implicitly defined by $\beta_3(x_b) = 1-\alpha_B$, where β_3 exploits the dependency structure in block b, expressed by the corresponding LD submatrix.

Figure 1 displays $M_{\text{eff},b}^{(3)}$ for $1 \leq b \leq B-1$ (we omitted the B-th block which is smaller than the others), together with their average $(B-1)^{-1}\sum_{b=1}^{B-1} M_{\text{eff},b}^{(3)}$. For the entire chromosome we obtained an effective number of tests of order 3 of $M_{\text{eff}}^{(3)} = 13{,}676.4$, meaning that effectively only approximately two third of the $M = 18{,}143$ loci contribute to the calibration with respect to the FWER level α.

Remark 5

a) At this stage of the analysis, neither the actual study data nor the precise definition of the binary phenotype to be analyzed is required. Hence, the value for the effective number of tests can be pre-computed and re-used for several association studies in the same target population.

b) Actually, the dependency structure among the χ^2 test statistics corresponding to diploid genotypes is slightly more involved than the one among the X_j from (1). Hence, $M_{\text{eff}}^{(3)}$ is only one out of several possible approximations of the true third-order effective number of tests. We defer the reader to Sections 4 and 6.4 of [3] for a detailed discussion and further approximation approaches. Anyway, all these approximations require the computation of $F_3(\cdot, 2, R)$.

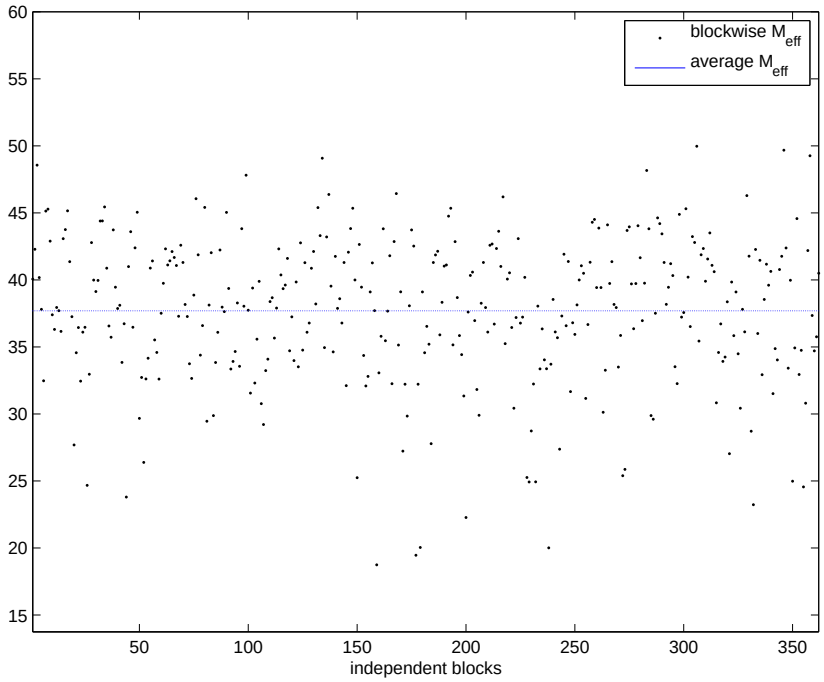

Figure 1: Block-specific effective numbers of tests for the real data example from Section 5, where the block size equals 50. The effective number of tests varies between the blocks. The stronger the dependence among the test statistics within a block the smaller is the effective number of tests. The horizontal line displays the average effective number of tests among all considered blocks.

6 Discussion

The computation of multivariate χ^2 probabilities plays an important role in many fields of modern statistics, especially in multiple testing. Scientific and technological progress in the relevant application areas leads to larger and larger systems of hypotheses to be tested simultaneously, requiring fast methods to compute the necessary multiplicity-adjusted rejection thresholds for the marginal χ^2 test statistics. While one can approximate these thresholds straightforwardly by Monte Carlo methods due to the rather simple stochastic representation of a multivariate chi-square distributed random vector, our proposed numerical methods are computationally much more efficient. With them, low-dimensional probabilities can be computed with high precision in a fraction of the time required for Monte Carlo simulations. Furthermore, we have demonstrated sum-type and product-type approximations for high-dimensional χ^2 probabilities.

For practical applications, we recommend PTPAs of appropriate order. Although a rigorous mathematical analysis of their conservativity (meaning that $\bar{F}_M(x, \nu, R)$ is approximated from above) is very involved (see the corresponding references in [4]), our experience is that they often yield tight approximations, in particular for larger values of x. The latter property is important for applications in large-scale multiple testing, where small exceedance probabilities $\bar{F}_M(x, \nu, R)$, i. e., large values of x, have to be considered because of the necessary strong adjustment for multiplicity.

Future work shall aim at implementing further types of multivariate χ^2 distributions which exhibit a more complex dependency structure. For instance, the rather general type given in Definition 4.6 of [1] is relevant for certain multiple test problems with unbalanced degrees of freedom and inhomogeneous correlation matrices among the multivariate normal vectors appearing in the stochastic representation (1). Furthermore, automated matrix factorization criteria going beyond the ones presented in Lemma 2 would be helpful tools for an even more exact computation of multivariate χ^2 probabilities.

References

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

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

[3] Thorsten Dickhaus and Thomas Royen. On multivariate chi-square distributions and their applications in testing multiple hypotheses. WIAS Preprint No. 1913, Weierstrass Institute for Applied Analysis and Stochastics Berlin, 2014. Available at http://www.wias-berlin.de/preprint/1913/wias_preprints_1913.pdf.

[4] Thorsten Dickhaus and Jens Stange. Multiple Point Hypothesis Testing Problems and Effective Numbers of Tests for Control of the Family-Wise Error Rate. *Calcutta Statistical Association Bulletin*, 65(257-260):123–144, 2013.

[5] K.R. Gabriel. Simultaneous test procedures - some theory of multiple comparisons. *Ann. Math. Stat.*, 40:224–250, 1969.

[6] Alan Genz and Frank Bretz. *Computation of multivariate normal and t probabilities.* Lecture Notes in Statistics 195. Berlin: Springer, 2009.

[7] Richard F. Gunst and John T. Webster. Density functions of the bivariate chi-square distribution. *J. Stat. Comput. Simulation*, 2:275–288, 1973.

[8] Henrik Holm and Mohamed-Slim Alouini. Sum and Difference of Two Squared Correlated Nakagami Variates in Connection With the McKay Distribution. *IEEE Transactions on Communications*, 52(8):1367–1376, 2004.

[9] D.R. Jensen. An inequality for a class of bivariate chi-square distributions. *J. Am. Stat. Assoc.*, 64:333–336, 1969.

[10] D.R. Jensen. A generalization of the multivariate Rayleigh distribution. *Sankhyā, Ser. A*, 32:193–208, 1970.

[11] V. Moskvina and K. M. Schmidt. On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.*, 32(6):567–573, 2008.

[12] Saralees Nadarajah and Arjun K. Gupta. Some bivariate gamma distributions. *Appl. Math. Lett.*, 19(8):767–774, 2006.

[13] Daniel Q. Naiman and Henry P. Wynn. Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Stat.*, 20(1):43–76, 1992.

[14] D.Q. Naiman and H.P. Wynn. The algebra of Bonferroni bounds: discrete tubes and extensions. *Metrika*, 62(2-3):139–147, 2005.

[15] Thomas Royen. Expansions for the multivariate chi-square distribution. *J. Multivariate Anal.*, 38(2):213–232, 1991.

[16] Thomas Royen. On some central and non-central multivariate chi-square distributions. *Stat. Sinica*, 5(1):373–397, 1995.

[17] Thomas Royen. Integral representations and approximations for multivariate gamma distributions. *Ann. Inst. Stat. Math.*, 59(3):499–513, 2007.

[18] Thomas Royen. A simple proof of the Gaussian correlation conjecture extended to multivariate gamma distributions. Preprint arXiv:1408.1028, 2014.

[19] O. E. Smith, S. I. Adelfang, and J. D. Tubbs. A bivariate gamma probability distribution with application to gust modeling. Technical Memorandum NASA TM-82483, NASA, George C. Marshall Space Flight Center, Alabama, 1982.

[20] Neil H. Timm. *Applied multivariate analysis*. New York, NY: Springer, 2002.

[21] K.J. Worsley. An improved Bonferroni inequality and applications. *Biometrika*, 69:297–302, 1982.

[22] S. Yue, T.B.M.J. Ouarda, and B. Bobée. A review of bivariate gamma distributions for hydrological application. *Journal of Hydrology*, 246:1–18, 2001.