arXiv: arXiv:0000.0000

Permutation p-value approximation via generalized Stolarsky invariance

Hera Y. He,

Stanford University
e-mail: hera.yu.he@gmail.com

Kinjal Basu,

LinkedIn
e-mail: kbasu@linkedin.com

Qingyuan Zhao,

University of Pennsylvania
e-mail: qyzhao@wharton.upenn.edu

Art B. Owen

Stanford University
e-mail: owen@stanford.edu

Abstract: It is common for genomic data analysis to use p-values from a large number of permutation tests. The multiplicity of tests may require very tiny p-values in order to reject any null hypotheses and the common practice of using randomly sampled permutations then becomes very expensive. We propose an inexpensive approximation to p-values for two sample linear test statistics, derived from Stolarsky's invariance principle. The method creates a geometrically derived reference set of approximate p-values for each hypothesis. The average of that set is used as a point estimate $\hat{p}$ and our generalization of the invariance principle allows us to compute the variance of the p-values in that set. We find that in cases where the point estimate is small, the variance is a modest multiple of the square of that point estimate, yielding a relative error property similar to that of saddlepoint approximations. On a Parkinson's disease data set, the new approximation is faster and more accurate than the saddlepoint approximation. We also obtain a simple probabilistic explanation of Stolarsky's invariance principle.

1. Introduction

Permutation methods are commonly used to obtain p-values in genomic applications, especially those involving gene sets. In even modestly large data sets, the exact permutation p-value becomes too expensive to compute. Then Monte Carlo sampling of random permutations becomes a standard approach. Genomic applications often test thousands of hypotheses and then multiplicity adjustment requires that some small p-values be obtained if any null hypotheses are to be

rejected. When p-values below ϵ are required to reject H_0, then Knijnenburg et al. (2009) recommend doing at least $10/\epsilon$ random permutations. As a result, even Monte Carlo sampling for permutation tests can be prohibitively expensive, and hence it pays to search for fast approximations to the permutation p-value.

In this paper we develop rapidly computable approximations to some permutation p-values. The p-values we consider are for a difference in group means. The approximations are based on ideas from spherical geometry and discrepancy, related to the Stolarsky invariance principle (Stolarsky, 1973). As described below, the resulting approximations prove to be very accurate for the tiny p-values where permutation methods are most difficult to use. The new approximations come with a numerical estimate of their own accuracy. Although they are limited to the two sample setting, that setting is very important in many applications.

We begin with some background on the genomic motivation of our work. Then we transition to spherical geometry. This section then gives an outline of the paper and a pointer to some software.

Genomic context

The specific problem that motivated us is testing for sets of genes associated with Parkinson's disease (Larson and Owen, 2015). More details about this work are given in the first author's dissertation (He, 2016). In these data sets, there are m_0 subjects without Parkinson's disease and m_1 subjects with it. One can test whether Parkinson's disease is associated with an individual gene by doing a t-test comparing gene expression levels in tissue samples from the two groups of subjects. Biological interest is often summarized more by gene sets rather than individual genes. Gene sets have two advantages: small but consistent associations of many genes with the test condition can raise power, and, the gene sets themselves often connect better to biological understanding than do individual genes.

For gene set testing, we work with a null hypothesis where a variable of interest (here a binary disease status) is statistically independent of the expression level of all the genes in the gene set. We will call the variable of interest a phenotype, though it could be any binary variable such as treatment versus control in an experiment. The most studied alternative hypotheses are those where the phenotype is associated with a location shift in some or all of the genes in a gene set. Many test statistics, some quite elaborate, have been proposed for this problem. Ackermann and Strimmer (2009) have an extensive comparison of 261 different gene set tests for this two sample setting. To cope with correlations among all of the genes within a gene set, testing is done through permutations of the phenotype values with respect to the gene expression levels. They investigated numerous ways that location shifts could manifest and judged test statistics by the resulting power of permutation tests. In practice, one does not know the precise form that the alternative will take. Fortunately, Ackermann and Strimmer (2009) identified two families of test statistics that performed well across their range of mean-shift scenarios.

To describe the best test statistics, let t_g be the ordinary t-statistic for association of gene g's expression level with a binary phenotype. It is a difference of within-group means normalized by a standard error. Now let G be a set of genes. Ackermann and Strimmer (2009) found that linear and quadratic test statistics, $L_G = \sum_{g\in G} t_g$ and $Q_G = \sum_{g\in G} t_g^2$, yielded the most powerful permutation tests, along with some simple approximations to those two statistics. These test statistics were more effective than some substantially more complicated proposals. Tian et al. (2005) use the test statistic $L_G/|G|$ where G is the cardinality of G, and the JG score of Jiang and Gentleman (2007) is similar. The statistic L_G, and approximations to it, did best when expression differences between the two conditions tended to have the same sign for each $g \in G$. When many oppositely signed treatment effects occurred, then Q_G and approximations to it did best.

In a permutation analysis, like Ackermann and Strimmer (2009) used, we consider all $N = \binom{n}{m_1}$ different ways to select a subset π containing m_1 of the $n = m_0 + m_1$ subjects. Let L_G^π be the linear test statistic recomputed as if those m_1 subjects had been the affected group. Then one-sided and two-sided permutation p-values for L_G are

$$p = \frac{1}{N}\sum_{\pi} \mathbf{1}_{L_G^\pi \geqslant L_G}, \quad \text{and} \quad p = \frac{1}{N}\sum_{\pi} \mathbf{1}_{|L_G^\pi| \geqslant |L_G|},$$

respectively. Here $\mathbf{1}_E$ takes the value 1 if the event E occurs and zero otherwise. We also use $\mathbf{1}(E)$ in places where we find it more readable than $\mathbf{1}_E$. Under the null hypothesis of independence, permutation tests derived from these p-values are exact by symmetry (Lehmann and Romano, 2005, Chapter 15.2). Note that the smallest possible value for p is $1/N$ which we call the granularity limit.

When N is too large for a permutation test to be computationally feasible, a standard practice is to estimate p via randomly sampled permutations of the treatment label, as proposed by Barnard (1963). We randomize the binary treatment label $M-1$ times, letting $\pi(\ell)$ be the affected group in randomization ℓ for $\ell = 1, \ldots, M-1$. We let $\pi(0)$ be the original allocation. Then the average

$$\hat{p} = \frac{1}{M}\sum_{\ell=0}^{M-1} \mathbf{1}_{|L_G^{\pi(\ell)}| \geqslant |L_G|}$$

is used as an estimate of p (for the two-sided linear case). In this Monte Carlo computation, the true permutation p-value p is the unknown parameter and $\hat{p}$ is the sample estimate of p. Note that $\hat{p} \geqslant 1/M$ because we have included the original allocation in the numerator. Failure to include the original allocation $\pi(0)$ can lead to $\hat{p} = 0$ which is very undesirable. Note that the Monte Carlo granularity limit $1/M$ can be much larger than the permutation limit $1/N$. When p is quite small, an enormous number M of simulations may be required to get an accurate estimate of it. For instance, in genome wide association studies (GWAS), the customary threshold for significance is $\epsilon = 5 \times 10^{-8}$, making permutation methods prohibitively expensive, or even infeasible. For a recent discussion of p-value thresholds in GWAS, see Fadista et al. (2016).

In this paper, we work with an approximation to L_G from Ackermann and Strimmer (2009). Let $X_i = 1$ if subject i is in condition 1 with $X_i = 0$ for condition 0, and let Y_{gi} be the expression level of gene g for subject i. These variables have sample averages $\bar{X}$ and $\bar{Y}_g$ respectively. Ackermann and Strimmer (2009) found that

$$\sum_{g\in G}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\bar{X}}{s_X}\frac{Y_{gi}-\bar{Y}_g}{s_g} \tag{1.1}$$

was in the same winning set of test statistics as L_G, where s_X and s_g are standard deviations of X_i and Y_{gi} respectively.

To understand why the statistic in (1.1) performs similarly to L_G, let $\hat{\rho}_g$ be the sample correlation between X_i and Y_{gi}. Then (1.1) is $\sum_{g\in G}\hat{\rho}_g$ times a constant $(n-1)/n$ that does not affect the permutation p-value. Now $t_g = \sqrt{n-2}\hat{\rho}_g/\sqrt{1-\hat{\rho}_g^2}$ and a Taylor approximation gives $t_g \doteq \sqrt{n-2}(\hat{\rho}_g + \hat{\rho}_g^3/2)$. When many small correlations $\hat{\rho}_g$ contribute to the signal, then summing $\hat{\rho}_g$ as in (1.1) gives a test statistic that is almost equivalent to summing t_g because each $\hat{\rho}_g^3$ is then very small.

Let $Y_i = Y_{Gi} \equiv \sum_{g\in G} Y_{gi}/s_g$, $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$ and let 1_n be a column vector of n ones. Then we may rewrite (1.1) as

$$\sum_{i=1}^{n}\frac{X_i-\bar{X}}{\|X-1_n\bar{X}\|}\frac{Y_i-\bar{Y}}{\|Y-1_n\bar{Y}\|} \tag{1.2}$$

multiplied by a constant that only depends on n. Equation (1.2) describes a test statistic that is a plain Euclidean inner product of two unit vectors $\boldsymbol{x}_0, \boldsymbol{y}_0 \in \mathbb{R}^n$. Here $\boldsymbol{x}_0$ has i'th component $(X_i - \bar{X})/\|X - 1_n\bar{X}\|$ and $\boldsymbol{y}_0$ is similar.

There are $N = \binom{n}{m_1}$ distinct vectors found by permuting the entries in $\boldsymbol{x}_0$. We label them $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}$ with $\boldsymbol{x}_0$ being the original one. Letting $\hat{\rho} = \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{y}_0$ we find that one and two-sided p-values for a linear statistic are

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbf{1}_{\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{y}_0\geqslant\hat{\rho}} \quad\text{and}\quad \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{1}_{|\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{y}_0|\geqslant|\hat{\rho}|} \tag{1.3}$$

repectively. We prefer inferences based on two-sided tests, but it is simpler to study one-sided tests first and then translate the results to two-sided ones.

Spherical geometry

We are now ready to make a geometric interpretation. Let $\mathbb{S}^d = \{\boldsymbol{z} \in \mathbb{R}^{d+1} \mid \boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = 1\}$ be the d-dimensional unit sphere. Our data $\boldsymbol{x}_0, \boldsymbol{y}_0$ are in a subset of $\mathbb{S}^{n-1}$ orthogonal to 1_n. That subset is isomorphic to $\mathbb{S}^{n-2}$ and so we work mostly with $d = n-2$. In our motivating setting, the vector $\boldsymbol{x}_0$ has m_1 identical positive values and m_0 identical negative values. The geometry here applies for

an arbitrary unit vector $\boldsymbol{x}_0$, but we only develop practically usable tests for binary $\boldsymbol{x}_0$.

For $\boldsymbol{y} \in \mathbb{S}^d$ and $t \in [-1, 1]$, the spherical cap of center $\boldsymbol{y}$ and height t is $C(\boldsymbol{y}; t) = \{\boldsymbol{z} \in \mathbb{S}^d \mid \langle \boldsymbol{y}, \boldsymbol{z}\rangle \geqslant t\}$, where $\langle \cdot, \cdot\rangle$ is the usual Euclidean inner product. By symmetry, $\boldsymbol{z} \in C(\boldsymbol{y}; t)$ if and only if $\boldsymbol{y} \in C(\boldsymbol{z}; t)$. The one-sided linear p-value is the fraction of $\boldsymbol{x}_k$ for $0 \leqslant k < N$ that belong to $C(\boldsymbol{y}_0; \hat{\rho})$. Viewing $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$ as approximately uniformly distributed over $\mathbb{S}^d$ by symmetry it is then natural to estimate p by

$$\hat{p}_1 = \hat{p}_1(\hat{\rho}), \quad \text{where} \quad \hat{p}_1(t) \equiv \frac{\mathbf{vol}(C(\boldsymbol{y}_0; t))}{\mathbf{vol}(\mathbb{S}^d)}. \tag{1.4}$$

We use spherical geometry to investigate the accuracy of the uniformity-based estimate $\hat{p}_1$ from (1.4) and also to motivate sharper estimates.

Stolarsky's invariance principal gives a remarkable description of the accuracy of $\hat{p}_1$. Points $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1} \in \mathbb{S}^d$ have squared L_2 spherical cap discrepancy

$$L_2^2 = L_2(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1})^2 = \int_{-1}^{1} \int_{\mathbb{S}^d} |\hat{p}_1(t) - p(\boldsymbol{z}, t)|^2 \, \mathrm{d}\sigma_d(\boldsymbol{z}) \, \mathrm{d}t$$

where $p(\boldsymbol{z}, t) = (1/N) \sum_{k=0}^{N-1} \mathbf{1}_{\boldsymbol{x}_k \in C(\boldsymbol{z}, t)}$ and σ_d is the uniform (Haar) measure on $\mathbb{S}^d$. Stolarsky (1973) shows that

$$\frac{d\omega_d}{\omega_{d+1}} \times L_2^2 = \int_{\mathbb{S}^d} \int_{\mathbb{S}^d} \|\boldsymbol{x} - \boldsymbol{y}\| \, \mathrm{d}\sigma_d(\boldsymbol{x}) \, \mathrm{d}\sigma_d(\boldsymbol{y}) - \frac{1}{N^2} \sum_{k,\ell=0}^{N-1} \|\boldsymbol{x}_k - \boldsymbol{x}_\ell\| \tag{1.5}$$

where ω_d is the (surface) volume of $\mathbb{S}^d$. Equation (1.5) shows that the mean squared error of $\hat{p}_1(t)$ as an estimate of the permutation p-value $p(\boldsymbol{z}, t)$ is determined by the mean absolute Euclidean distances among the N points. In our applications, the N points will be the distinct permuted values of $\boldsymbol{x}_0$, but (1.5) holds for N arbitrary points $\boldsymbol{x}_k \in \mathbb{S}^d$.

The left side of (1.5) is, up to normalization, a mean squared discrepancy over spherical caps. The mean of $(\hat{p}_1 - p)^2$ is taken over caps of all heights from -1 to 1 corresponding to p-values of all sizes between 0 and 1. It is not then a very good accuracy measure when $\hat{p}_1(\hat{\rho})$ turns out to be very small, such as 10^{-6}. It would be more useful to get such a mean squared error taken over caps of exactly the size $\hat{p}_1(\hat{\rho})$, and no others.

Brauchart and Dick (2013) consider quasi-Monte Carlo (QMC) sampling in the sphere. They generalize Stolarsky's discrepancy formula to include a weighting function on the height t. By specializing their formula, we get an expression for the mean of $(\hat{p}_1 - p)^2$ over spherical caps of any fixed size. Discrepancy theory plays a prominent role in QMC (Niederreiter, 1992), which is about approximating an integral by a sample average. The present setting is a reversal of QMC: the discrete average p over permutations is the exact value we seek, and the integral over a continuum is the approximation $\hat{p}$. A second difference is that the QMC literature focusses on choosing N points to minimize a criterion such as (1.5), whereas here the N points are determined by the problem.

As we will show below, the estimate $\hat{p}_1$ is the average of p over all spherical caps $C(\boldsymbol{y};\hat{\rho})$ under a uniform distribution on their centers, i.e., $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$. Those caps have the same volume as $C(\boldsymbol{y}_0;\hat{\rho})$. In addition to specializing to caps $C(\boldsymbol{y};t)$ with $t = \hat{\rho}$ we make a further refinement to caps whose centers $\boldsymbol{y}$ more closely resemble $\boldsymbol{y}_0$. We find that refining to $\boldsymbol{y}$ with $\boldsymbol{y}^\mathsf{T}\boldsymbol{x}_0 = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_0$ is especially useful. Such $\boldsymbol{y}$ values have exactly the same linear test statistic that the observed data $\boldsymbol{y}_0$ had.

These restrictions on spherical caps impose the constraint $\boldsymbol{y} \in \mathbb{Y}$ for some $\mathbb{Y} \subset \mathbb{S}^d$. Then we can be sure that $p(\boldsymbol{y}_0,\hat{\rho}) \leqslant \sup_{\boldsymbol{y}\in\mathbb{Y}} p(\boldsymbol{y},\hat{\rho})$. If we could compute this supremum, then we would have a conservative permutation p-value and be sure of controlling type I errors. We are generally unable to compute this quantity but we can find both $\mathbb{E}(p(\boldsymbol{y},\hat{\rho}))$ and $\mathrm{Var}(p(\boldsymbol{y},\hat{\rho}))$ under a reference distribution $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y})$. Intuitively, taking $\mathbb{Y}$ ever closer to the ideal $\mathbb{Y} = \{\boldsymbol{y}_0\}$, should lead to a more accurate reference mean. The variance gives us a numerical measure of how accurate that reference mean is. The practical constraint on our choice of $\mathbb{Y}$ is that we must be able to compute these reference moments.

The first reference set is simply $\mathbb{Y}_1 = \mathbb{S}^d$. Our refinement of this set is $\mathbb{Y}_2 = \{\boldsymbol{y} \in \mathbb{S}^d \mid \boldsymbol{y}^\mathsf{T}\boldsymbol{x}_0 = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_0\}$. We obtain a computable expression for $\hat{p}_2 = \mathbb{E}(p(\boldsymbol{y},\hat{\rho}))$ and another one for $\mathbb{E}((\hat{p}_2 - p(\boldsymbol{y},\hat{\rho}))^2)$, both under $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_2)$, by further extending Brauchart and Dick's generalization of Stolarsky's invariance. In principle we could refine $\mathbb{Y}_2$ further in the direction of $\{\boldsymbol{y}_0\}$ by imposing additional linear constraints on $\boldsymbol{y}$. For this paper we impose just one. Our proofs and algorithms work with the more general constraint $\boldsymbol{y}^\mathsf{T}\boldsymbol{x}_c = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_c$ for any $c \in \{0, 1, \ldots, N-1\}$ that we like.

Our calculations show that $\sqrt{\mathrm{Var}(\hat{p}_2 - p(\boldsymbol{y},\hat{\rho}))}$, for $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_2)$, does not greatly exceed $\hat{p}_2$ when $\hat{p}_2$ is small and it even vanishes in the limit as $\hat{p}_2 \downarrow 1/N$. The function $p(\boldsymbol{y})$ is then nearly constant over $\boldsymbol{y} \in \mathbb{Y}_2$ in this L_2 sense, which turns out to make $\hat{p}_2$ a much better estimate of p than $\hat{p}_1$ is. We cannot rule out that the true $p(\boldsymbol{y}_0)$ could be quite different from $\hat{p}_2$ for a data generating process that differed in an adversarial way from the reference distribution, as discussed at the end of this article.

Although our results are mean square discrepancies via invariance, we can also obtain them via probabilistic arguments. As a consequence we have a probabilistic derivation of Stolarsky's formula. Bilyk et al. (2016) have independently found this connection.

Outline

The rest of the paper is organized as follows. Section 2 presents some results from spherical geometry and defines our reference distributions. In Section 3 we use Stolarsky's invariance principle as generalized by Brauchart and Dick (2013) to obtain the mean squared error between the true p-value and its continuous approximation $\hat{p}_1$, averaging over all spherical caps of volume $\hat{p}_1$. This section also has a probabilistic derivation of that mean squared error. In Section 4 we derive the refined estimate $\hat{p}_2$ and some generalizations. By construction

$\hat{p}_2 \geqslant 1/N$, respecting the true granularity of permutation testing. In Section 5 we modify the proof in Brauchart and Dick (2013), to further generalize their invariance results to include the mean squared error of $\hat{p}_2$. Section 6 extends our estimates to two-sided testing. Section 7 illustrates our p-value approximations numerically. We see that the root mean squared error in the estimate $\hat{p}_2$ is of the same order of magnitude as $\hat{p}_2$ itself. That is, $\hat{p}_2$ has a relative error property like saddlepoint estimates do. The estimate $\hat{p}_1$ is notably less accurate than $\hat{p}_2$. Section 8 makes a numerical comparison to saddlepoint methods in simulated data. The saddlepoint estimates come out more accurate than $\hat{p}_2$ but are downwardly biased in those simulated examples. Section 9 compares the accuracy of our approximations to each other and to the saddlepoint approximation for 6180 gene sets in some Parkinson's disease data sets. In the data examples, the new approximations come out closer to some gold standard estimates (based on large Monte Carlo samples) than the saddlepoint estimates do, which once again are biased low. From Table 6.3 of He (2016), the saddlepoint computations take roughly 30 times longer than $\hat{p}_2$ does. Section 10 draws some conclusions and discusses the challenges in getting a computationally feasible p-value that accounts for both sampling uncertainty of the data and the uncertainty in $\hat{p}$ as an estimate of p. At several places we refer the reader to a supplement (He et al., 2018) for additional material, including our lengthier proofs.

Software

The proposed approximations are implemented in the R package pipeGS on CRAN. Given a binary input label and a gene expression measurement matrix, it computes our p-value approximations for two sample problems. It includes the statistics, $\hat{p}_1$ and $\hat{p}_2$ mentioned above as well as a saddlepoint approximation which may be of independent interest.

2. Background and notation

Here we develop approximations to the one-sided p-value in (1.3). That is simpler than carrying the two-sided case through all of our derivations. We translate from one-sided to two-sided cases in Section 6.

The raw data contain points (X_i, Y_i) for $i = 1, \ldots, n$, where Y_i may be a composite quantity derived from all Y_{gi} for g belonging to a gene set G, such as $Y_i = Y_{Gi}$ given just before (1.2). We center and scale vectors $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$ yielding $\boldsymbol{x}_0, \boldsymbol{y}_0 \in \mathbb{S}^d$ for $d = n - 1$. Both points belong to $\{\boldsymbol{z} \in \mathbb{S}^{n-1} \mid \boldsymbol{z}^{\mathsf{T}} 1_n = 0\}$. We can use an orthogonal matrix to rotate the points of this set onto $\mathbb{S}^{n-2} \times \{0\}$. As a result, we may simply work with $\boldsymbol{x}_0, \boldsymbol{y}_0 \in \mathbb{S}^d$ where $d = n - 2$.

The sample correlation of these variables is $\hat{\rho} = \boldsymbol{x}_0^{\mathsf{T}} \boldsymbol{y}_0 = \langle \boldsymbol{x}_0, \boldsymbol{y}_0 \rangle$. We use $\langle \boldsymbol{x}_0, \boldsymbol{y}_0 \rangle$ when we find that geometrical thinking is appropriate and to conform with Brauchart and Dick (2013). We use $\boldsymbol{x}_0^{\mathsf{T}} \boldsymbol{y}_0$ to emphasize computational or algebraic connotations.

The geometry we use leads to practical algorithms when X_i takes on just two values, such as 0 and 1. When there are m_0 observations with $X_i = 0$ and m_1 with $X_i = 1$ then $\boldsymbol{x}_0$ contains m_0 components equal to $-\sqrt{m_1/(nm_0)}$ and m_1 components equal to $+\sqrt{m_0/(nm_1)}$. In our theorem statements we describe such a point $\boldsymbol{x}_0$ as a "centered and scaled binary vector".

Computational costs are often sensitive to the smaller sample size, $\underline{\mathrm{m}} \equiv \min(m_0, m_1)$. For this two-sample case, there are only $N = \binom{m_0+m_1}{m_0}$ distinct permutations of $\boldsymbol{x}_0$. We have called these $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{N-1}$ and the true p-value is $p = (1/N)\sum_{k=0}^{N-1} \mathbf{1}(\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{y}_0 \geqslant \hat{\rho}) = (1/N)\sum_{k=0}^{N-1} \mathbf{1}(\boldsymbol{x}_k \in C(\boldsymbol{y}_0; \hat{\rho}))$. In the two-sample case we can define a useful interpoint swap distance

$$r = r(\boldsymbol{x}_k, \boldsymbol{x}_\ell) = \sum_{i=1}^{n} \mathbf{1}((\boldsymbol{x}_k)_i > 0 \ \&\ (\boldsymbol{x}_\ell)_i < 0)$$

where $(\boldsymbol{x}_k)_i$ is the i'th component of $\boldsymbol{x}_k$. In the permutation taking $\boldsymbol{x}_k$ onto $\boldsymbol{x}_\ell$ there are r positive entries in $\boldsymbol{x}_k$ that have been swapped with negative ones to create $\boldsymbol{x}_\ell$. In that case we easily find that

$$u(r) \equiv \langle \boldsymbol{x}_k, \boldsymbol{x}_\ell \rangle = 1 - r\Big(\frac{1}{m_0} + \frac{1}{m_1}\Big). \tag{2.1}$$

We need some geometric properties of the unit sphere and spherical caps. The surface volume of $\mathbb{S}^d$ is $\omega_d = 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$. Recall that σ_d is the volume element in $\mathbb{S}^d$ normalized so that $\sigma_d(\mathbb{S}^d) = 1$. Henceforth 'volume' will always refer to this normalized volume. The spherical cap $C(\boldsymbol{y}; t) = \{\boldsymbol{z} \in \mathbb{S}^d \mid \boldsymbol{z}^{\mathsf{T}}\boldsymbol{y} \geqslant t\}$ has volume

$$\sigma_d(C(\boldsymbol{y}; t)) = \begin{cases} \frac{1}{2} I_{1-t^2}\left(\frac{d}{2}, \frac{1}{2}\right), & 0 \leqslant t \leqslant 1 \\ 1 - \frac{1}{2} I_{1-t^2}\left(\frac{d}{2}, \frac{1}{2}\right), & -1 \leqslant t < 0 \end{cases}$$

where $I_t(a, b)$ is the incomplete beta function

$$I_t(a, b) = \frac{1}{B(a, b)} \int_0^t x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$$

with $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$.

We frequently need to use a tangent-normal decomposition (Mardia and Jupp, 2000, Chapter 9.1). The tangent-normal decomposition of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$ is

$$\boldsymbol{y} = t\boldsymbol{x} + \sqrt{1-t^2}\boldsymbol{y}^*$$

where $t = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{x} \in [-1, 1]$ and $\boldsymbol{y}^* \in \{\boldsymbol{z} \in \mathbb{S}^d \mid \boldsymbol{z}^{\mathsf{T}}\boldsymbol{x} = 0\}$ which is isomorphic to $\mathbb{S}^{d-1}$. The coordinates t and $\boldsymbol{y}^*$ are unique. We refer to $\boldsymbol{y}^*$ as the residual in this decomposition. From equation (A.1) in Brauchart and Dick (2013)

$$\mathrm{d}\sigma_d(\boldsymbol{y}) = \frac{\omega_{d-1}}{\omega_d}(1-t^2)^{d/2-1}\,\mathrm{d}t\,\mathrm{d}\sigma_{d-1}(\boldsymbol{y}^*). \tag{2.2}$$

The intersection of two spherical caps of common height t is

$$C_2(\boldsymbol{x}, \boldsymbol{y}; t) \equiv C(\boldsymbol{x}; t) \cap C(\boldsymbol{y}; t).$$

We will need the volume of this intersection. Lee and Kim (2014) give a general solution for spherical cap intersections without requiring equal heights. They enumerate 25 cases, but our case does not correspond to any single case of theirs and so we obtain the formula we need directly, below. We suspect it must be known already, but we were unable to find it in the literature.

Lemma 1. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^d$ and $-1 \leqslant t \leqslant 1$ and put $u = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}$. Let $V_2(u; t, d) = \sigma_d(C_2(\boldsymbol{x}, \boldsymbol{y}; t))$. If $u = 1$, then $V_2(u; t, d) = \sigma_d(C(\boldsymbol{x}; t))$. If $-1 < u < 1$, then*

$$V_2(u; t, d) = \frac{\omega_{d-1}}{\omega_d} \int_t^1 (1 - s^2)^{\frac{d}{2}-1} \sigma_{d-1}(C(\boldsymbol{y}^*; \rho(s)))\, \mathrm{d}s, \tag{2.3}$$

where $\rho(s) = (t - su)/\sqrt{(1 - s^2)(1 - u^2)}$. Finally, for $u = -1$,

$$V_2(u; t, d) = \begin{cases} 0, & t \geqslant 0 \\ \frac{\omega_{d-1}}{\omega_d} \int_{-|t|}^{|t|} (1 - s^2)^{\frac{d}{2}-1}\, \mathrm{d}s, & \text{else.} \end{cases} \tag{2.4}$$

Proof. See Section 11.1 of the supplement, He et al. (2018). □

When we give probabilistic arguments and interpretations we do so for a random center $\boldsymbol{y}$ of a spherical cap. That random center is taken from two reference distributions given below. Reference distribution 1 is illustrated in Figure 1. Reference distribution 2 is illustrated in Figure 2 of Section 4 where we first use it.

Reference distribution 1. *The vector $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_1)$ where $\mathbb{Y}_1 = \mathbb{S}^d$. Expectation under this distribution is denoted by $\mathbb{E}_1(\cdot)$.*

Reference distribution 2. *The vector $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_2)$ where*

$$\mathbb{Y}_2 = \mathbb{Y}_2(c) = \{\boldsymbol{z} \in \mathbb{S}^d \mid \boldsymbol{z}^{\mathsf{T}}\boldsymbol{x}_c = \tilde{\rho}\},$$

for some $c \in \{0, 1, \ldots, N-1\}$ and $\tilde{\rho} = \boldsymbol{x}_c^{\mathsf{T}}\boldsymbol{y}_0$. Then $\boldsymbol{y} = \tilde{\rho}\boldsymbol{x}_c + \sqrt{1 - \tilde{\rho}^2}\boldsymbol{y}^*$ for $\boldsymbol{y}^*$ uniformly distributed on a subset of $\mathbb{S}^d$ isomorphic to $\mathbb{S}^{d-1}$. Expectation under this distribution is denoted by $\mathbb{E}_{2,c}(\cdot)$ and $\mathbb{E}_2(\cdot)$ means $\mathbb{E}_{2,0}(\cdot)$.

We also use $\mathrm{Var}_j(\cdot)$ for a variance under reference distribution j. Reference distribution 1 holds for $\boldsymbol{y}_0$ if the Y_i are IID Gaussian random variables (with positive variance). Then $\hat{p}_1$ is the same as we would get from a t-test. Reference distribution 2 is actually a family of reference distributions indexed by c. The choice of primary interest to us has $c = 0$ corresponding to the observed treatment allocation. We write $\tilde{p}_c = \mathbb{E}_{2,c}(C(\boldsymbol{y}; \hat{\rho}))$ and $\hat{p}_2 = \tilde{p}_0$. This reference distribution represents a significant narrowing of reference distribution 1. We find good numerical performance for $\hat{p}_2$ below.

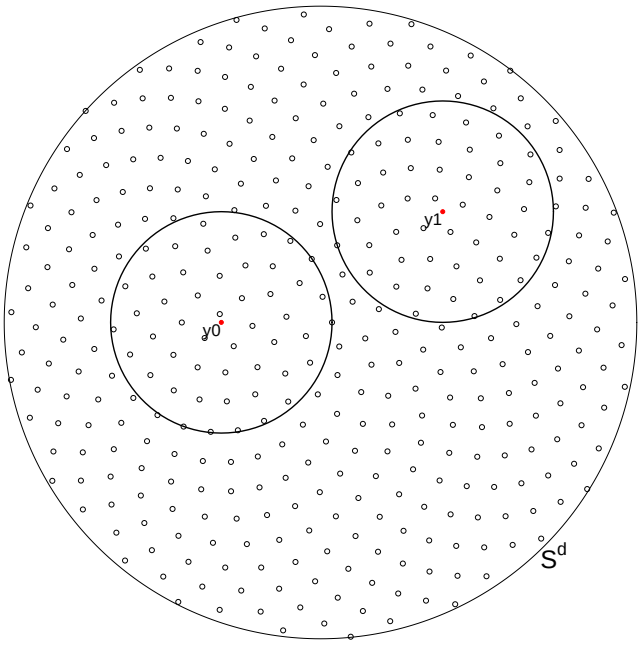


Fig 1: Reference distribution 1. Here $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$ and $\boldsymbol{y}_0$ is the observed value of $\boldsymbol{y}$. The small open circles represent permuted vectors $\boldsymbol{x}_k$. The circle around $\boldsymbol{y}_0$ goes through $\boldsymbol{x}_0$ and represents a spherical cap of height $\hat{\rho} = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_0$. A second spherical cap of equal volume is centered at $\boldsymbol{y}_1$. We study moments of the permutation p-value $p(\boldsymbol{y}, \hat{\rho})$ for random $\boldsymbol{y}$ and fixed $\hat{\rho}$.

3. Approximation via spherical cap volume

Here we study the approximate p-value $\hat{p}_1(\hat{\rho}) = \sigma_d(C(\boldsymbol{y}; \hat{\rho}))$. First we find the mean squared error of this approximation over all spherical caps of the given volume via invariance. Next we give a probabilistic interpretation which includes the conditional unbiasedness result in Proposition 2 below. Then we give two computational simplifications, first taking advantage of the permutation structure of our points, and then second for permutations of a centered and scaled binary vector.

Brauchart and Dick (2013) gave a simple proof of Stolarsky's invariance principle using reproducing kernel Hilbert spaces. They also generalized it as follows.

Theorem 1. *Let $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}$ be any points in $\mathbb{S}^d$. Let $v : [-1, 1] \to (0, \infty)$ be*

any function with an antiderivative. Then

$$\begin{aligned}&\int_{-1}^{1} v(t) \int_{\mathbb{S}^d} \left|\sigma_d(C(\boldsymbol{z};t)) - \frac{1}{N}\sum_{k=0}^{N-1} \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{x}_k)\right|^2 \mathrm{d}\sigma_d(\boldsymbol{z})\,\mathrm{d}t \\ &= \frac{1}{N^2}\sum_{k,l=0}^{N-1} K_v(\boldsymbol{x}_k,\boldsymbol{x}_\ell) - \int_{\mathbb{S}^d}\int_{\mathbb{S}^d} K_v(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\sigma_d(\boldsymbol{x})\,\mathrm{d}\sigma_d(\boldsymbol{y})\end{aligned} \tag{3.1}$$

where $K_v(\boldsymbol{x},\boldsymbol{y})$ is a reproducing kernel function defined by

$$K_v(\boldsymbol{x},\boldsymbol{y}) = \int_{-1}^{1} v(t) \int_{\mathbb{S}^d} \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{x})\mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{y})\,\mathrm{d}\sigma_d(\boldsymbol{z})\,\mathrm{d}t. \tag{3.2}$$

Proof. See Theorem 5.1 in Brauchart and Dick (2013) □

If we set $v(t) = 1$ and $K(\boldsymbol{x},\boldsymbol{y}) = 1 - C_d\|\boldsymbol{x}-\boldsymbol{y}\|$, then we recover the original Stolarsky formula (1.5). Note that the statement of Theorem 5.1 in Brauchart and Dick (2013) has a sign error in their counterpart to (3.1). The corrected statement (3.1) can be verified by comparing equations (5.3) and (5.4) of Brauchart and Dick (2013).

We would like a version of (3.1) for just one value of t such as $t = \hat{\rho} = \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{y}_0$. For $\hat{\rho} \in [-1,1)$ and $\boldsymbol{\epsilon} = (\epsilon_1,\epsilon_2) \in (0,1)^2$, let

$$v_{\boldsymbol{\epsilon}}(t) = \epsilon_2 + \frac{1}{\epsilon_1}\mathbf{1}(\hat{\rho} \leqslant t \leqslant \hat{\rho}+\epsilon_1). \tag{3.3}$$

Each $v_{\boldsymbol{\epsilon}}$ satisfies the conditions of Theorem 1 making (3.1) an identity in $\boldsymbol{\epsilon}$. We let $\epsilon_2 \to 0$ and then $\epsilon_1 \to 0$ on both sides of (3.1) for $v = v_{\boldsymbol{\epsilon}}$ yielding Theorem 2.

Theorem 2. *Let $\boldsymbol{x}_0,\boldsymbol{x}_1,\dots,\boldsymbol{x}_{N-1} \in \mathbb{S}^d$ and $t \in [-1,1]$. Then*

$$\int_{\mathbb{S}^d} |p(\boldsymbol{y},t) - \hat{p}_1(t)|^2\,\mathrm{d}\sigma_d(\boldsymbol{y}) = \frac{1}{N^2}\sum_{k=0}^{N-1}\sum_{\ell=0}^{N-1} \sigma_d(C_2(\boldsymbol{x}_k,\boldsymbol{x}_\ell;t)) - \hat{p}_1(t)^2. \tag{3.4}$$

Proof. See He et al. (2018, Section 11.2) which uses the limit argument described above. □

We now give a proposition that holds for all distributions of $\boldsymbol{y} \in \mathbb{S}^d$ including our reference distributions 1 and 2.

Proposition 1. *For a random point $\boldsymbol{y}$ from any distribution on $\mathbb{S}^d$,*

$$\mathbb{E}(p(\boldsymbol{y},t)) = \frac{1}{N}\sum_{k=0}^{N-1} \Pr(\boldsymbol{y} \in C(\boldsymbol{x}_k;t)), \quad \textit{and} \tag{3.5}$$

$$\mathbb{E}(p(\boldsymbol{y},t)^2) = \frac{1}{N^2}\sum_{k,\ell=0}^{N-1} \Pr(\boldsymbol{y} \in C_2(\boldsymbol{x}_k,\boldsymbol{x}_\ell;t)). \tag{3.6}$$

Proof. These follow directly from $p = (1/N)\sum_{k=0}^{N} \mathbf{1}(\boldsymbol{y} \in C(\boldsymbol{x}_k; t))$. □

Proposition 2. *For any $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1} \in \mathbb{S}^d$ and $t \in [-1, 1]$,*

$$\mathbb{E}_1(p(\boldsymbol{y}, t)) = \hat{p}_1(t) \quad \text{and} \quad \mathbb{E}_1(p(\boldsymbol{y}, t)^2) = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \sigma_d(C_2(\boldsymbol{x}_k, \boldsymbol{x}_\ell; t)).$$

Proof. We apply Proposition 1. Under reference distribution 1, each summand in (3.5) equals $\hat{p}_1(t)$. Similarly the k, ℓ summand in (3.6) evaluates to $\sigma_d(C_2(\boldsymbol{x}_k, \boldsymbol{x}_\ell; t))$. □

Combining Proposition 2 with Theorem 2 we find that if $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$, then $p(\boldsymbol{y}, \hat{\rho})$ is a random variable with mean $\hat{p}_1(\hat{\rho})$ and variance given by (3.4) with $t = \hat{\rho}$. The right hand side of (3.4) sums $O(N^2)$ terms. The symmetry in a permutation set allows us to use

$$\int_{\mathbb{S}^d} |p(\boldsymbol{y}, t) - \hat{p}_1(t)|^2 \, \mathrm{d}\sigma_d(\boldsymbol{y}) = \frac{1}{N} \sum_{k=0}^{N-1} \sigma_d(C_2(\boldsymbol{x}_0, \boldsymbol{x}_k; t)) - \hat{p}_1(t)^2$$

instead. This expression costs $O(N)$, the same as the full permutation analysis that we seek to avoid.

The cost becomes feasible when $\boldsymbol{x}_0$ is a centered and scaled binary vector. Then for fixed t, $\sigma_d(C_2(\boldsymbol{x}_k, \boldsymbol{x}_\ell; t))$ just depends on the swap distance r between $\boldsymbol{x}_k$ and $\boldsymbol{x}_\ell$. Let $r_{k,\ell}$ be the swap distance between $\boldsymbol{x}_k$ and $\boldsymbol{x}_\ell$ and let $N_r = \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \mathbf{1}(r_{k,\ell} = r)$ count the number of point pairs at swap distance r. Then

$$\int_{\mathbb{S}^d} |p(\boldsymbol{y}, t) - \hat{p}_1(t)|^2 \, \mathrm{d}\sigma_d(\boldsymbol{y}) = \frac{1}{N^2} \sum_{r=0}^{\mathrm{m}} N_r V_2(u(r); t, d) - \hat{p}_1(t)^2 \tag{3.7}$$

for $V_2(u(r); t, d)$ given in Lemma 1.

Theorem 3. *Let $\boldsymbol{x}_0 \in \mathbb{S}^d$ be a centered and scaled binary vector with $m_0 \geqslant 1$ negative components and $m_1 \geqslant 1$ positive components. Let $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{N-1}$ be the $N = \binom{m_0+m_1}{m_0}$ distinct permutations of $\boldsymbol{x}_0$. If $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$, then for $t \in [-1, 1]$, and with $u(r)$ defined in (2.1),*

$$\mathrm{Var}_1(p(\boldsymbol{y}, t)) = \frac{1}{N} \sum_{r=0}^{m} \binom{m_0}{r} \binom{m_1}{r} V_2(u(r); t, d) - \hat{p}_1(t)^2, \tag{3.8}$$

for $V_2(u(r); t, d)$ given in Lemma 1.

Proof. There are $\binom{m_0}{r}\binom{m_1}{r}$ permuted points $\boldsymbol{x}_k$ at swap distance r from $\boldsymbol{x}_\ell$ for each $\ell = 0, 1, \dots, N-1$. Therefore $N_r = N\binom{m_0}{r}\binom{m_1}{r}$, establishing (3.8). □

We will see in Section 7 that $\sqrt{\mathrm{Var}_1(p(\boldsymbol{y}, t))}/\mathbb{E}_1(p(\boldsymbol{y}, t))$ becomes extremely large as $t \to 1$ and $\mathbb{E}_1(p(\boldsymbol{y}, t)) \to 0$. Therefore extremely small spherical cap volumes are compatible with a wide range of permutation p-values.

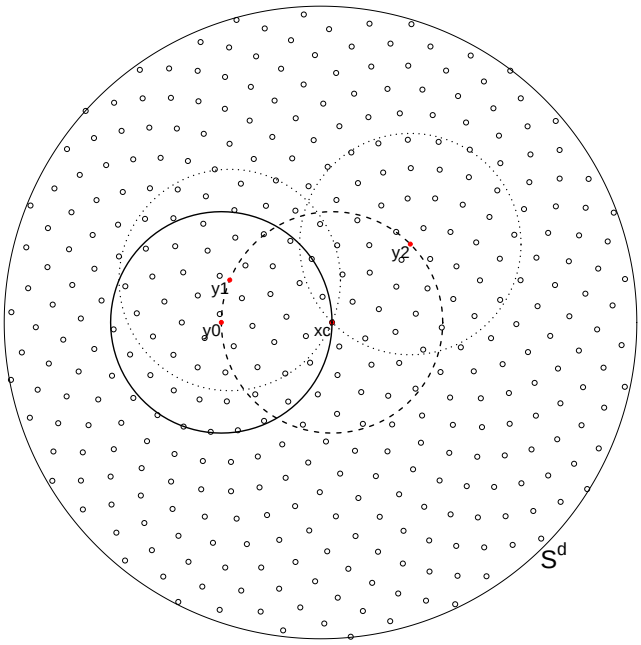


Fig 2: Reference distribution 2 with $c = 0$. The original response vector is $\boldsymbol{y}_0$ with $\boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_0 = \hat{\rho}$ and $\boldsymbol{x}_0$ marked **xc**. We consider alternative $\boldsymbol{y}$ uniformly distributed on the surface of $C(\boldsymbol{x}_0; \hat{\rho})$ (dashed circle) with examples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Around each such $\boldsymbol{y}_j$ there is a spherical cap of height $\hat{\rho}$ that just barely includes $\boldsymbol{x}_c = \boldsymbol{x}_0$. The small open circles are permutations $\boldsymbol{x}_k$ of $\boldsymbol{x}_0$. The proportion of $\boldsymbol{x}_k \in C(\boldsymbol{y}, \hat{\rho})$ is $p(\boldsymbol{y}, \hat{\rho})$. We study the mean and variance of $p(\boldsymbol{y}, \hat{\rho})$ for fixed $\hat{\rho}$ and $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_2)$ for $\mathbb{Y}_2 = \{\boldsymbol{y} \in \mathbb{S}^d \mid \boldsymbol{y}^{\mathsf{T}}\boldsymbol{x}_0 = \hat{\rho}\}$.

4. A finer approximation to the p-value

In the previous section, we studied the distribution of permutation p-values $p(\boldsymbol{y}, t)$ with spherical cap centers $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$ and heights $t = \hat{\rho}$. In this section, we use reference distribution 2 to obtain a finer approximation to $p(\boldsymbol{y}_0, \hat{\rho})$ by studying the distribution of the p-values for centers $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_2(c))$ for $c \in \{0, 1, \ldots, N-1\}$. For an index $c \in \{0, 1, \ldots, N-1\}$, conditioning as above leads to

$$\tilde{p}_c = \mathbb{E}_{2,c}(p(\boldsymbol{y}, \hat{\rho})) = \mathbb{E}_1(p(\boldsymbol{y}, \hat{\rho}) \mid \boldsymbol{y}^{\mathsf{T}}\boldsymbol{x}_c = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_c), \tag{4.1}$$

and our primary interest is in $\hat{p}_2 = \tilde{p}_0$. For an illustration of reference distribution 2 see Figure 2.

From Proposition 1, we can get our estimate $\tilde{p}_c$ and its mean squared error by finding single and double inclusion probabilities for $\boldsymbol{y}$. To compute $\tilde{p}_c$ we need

to sum N values $\Pr(\boldsymbol{y} \in C(\boldsymbol{x}_k; t) \mid \boldsymbol{y}^{\mathsf{T}}\boldsymbol{x}_c = \tilde{\rho})$ and for $\tilde{p}_c$ to be useful we must compute it in $o(N)$ time. The computations are feasible in the binary case.

Let $u_j = \boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_c$ for $j = 1, 2$, and let $u_3 = \boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{x}_2$. Let the tangent-normal decomposition of $\boldsymbol{y} \in \mathbb{Y}_2(c)$ with respect to $\boldsymbol{x}_c$ be

$$\boldsymbol{y} = \tilde{\rho}\boldsymbol{x}_c + \sqrt{1-\tilde{\rho}^2}\boldsymbol{y}^*, \quad \tilde{\rho} = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_c. \tag{4.2}$$

Then the single and double point inclusion probabilities under reference distribution 2 are

$$P_1(u_1, \tilde{\rho}, \hat{\rho}) = \int_{\mathbb{S}^{d-1}} \mathbf{1}(\langle \boldsymbol{y}, \boldsymbol{x}_1\rangle \geqslant \hat{\rho})\, \mathrm{d}\sigma_{d-1}(\boldsymbol{y}^*), \quad \text{and} \tag{4.3}$$

$$P_2(u_1, u_2, u_3, \tilde{\rho}, \hat{\rho}) = \int_{\mathbb{S}^{d-1}} \mathbf{1}(\langle \boldsymbol{y}, \boldsymbol{x}_1\rangle \geqslant \hat{\rho})\mathbf{1}(\langle \boldsymbol{y}, \boldsymbol{x}_2\rangle \geqslant \hat{\rho})\, \mathrm{d}\sigma_{d-1}(\boldsymbol{y}^*) \tag{4.4}$$

where $\hat{\rho} = \langle \boldsymbol{x}_0, \boldsymbol{y}_0\rangle$. The dependence of P_1 and P_2 on $\tilde{\rho}$ comes through $\boldsymbol{y}$ as given in equation (4.2).

Lemma 2. *Let $\boldsymbol{x}_1$ have tangent-normal decomposition $\boldsymbol{x}_1 = u_1\boldsymbol{x}_c + \sqrt{1-u_1^2}\boldsymbol{x}_1^*$ with respect to $\boldsymbol{x}_c$. Then the single point inclusion probability from (4.3) is*

$$P_1(u_1, \tilde{\rho}, \hat{\rho}) = \begin{cases} \mathbf{1}(\tilde{\rho}u_1 \geqslant \hat{\rho}), & u_1 = \pm 1 \text{ or } \tilde{\rho} = \pm 1 \\ \sigma_{d-1}(C(\boldsymbol{x}_1^*, \rho^*)), & u_1 \in (-1, 1), \tilde{\rho} \in (-1, 1) \end{cases} \tag{4.5}$$

where $\rho^* = (\hat{\rho} - \tilde{\rho}u_1)/\sqrt{(1-\tilde{\rho}^2)(1-u_1^2)}$.

Proof. Using the decomposition (4.2) of $\boldsymbol{y}$ with respect to $\boldsymbol{x}_c$,

$$\langle \boldsymbol{y}, \boldsymbol{x}_1\rangle = \begin{cases} \tilde{\rho}u_1, & u_1 = \pm 1 \text{ or } \tilde{\rho} = \pm 1 \\ \tilde{\rho}u_1 + \sqrt{1-\tilde{\rho}^2}\sqrt{1-u_1^2}\,\langle \boldsymbol{y}^*, \boldsymbol{x}_1^*\rangle, & u_1 \in (-1, 1), \tilde{\rho} \in (-1, 1) \end{cases}$$

and the result easily follows. □

Theorem 4. *Let $\boldsymbol{x}_0 \in \mathbb{S}^d$ be a centered and scaled binary vector with $m_0 \geqslant 1$ negative components and $m_1 \geqslant 1$ positive components. Let $\hat{\rho} = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_0 \in [-1, 1]$, and $\tilde{\rho} = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_c \in [-1, 1]$ for $c \in \{0, 1, \dots, N-1\}$. Then*

$$\tilde{p}_c = \mathbb{E}_{2,c}(p(\boldsymbol{y}, \hat{\rho})) = \frac{1}{N}\sum_{r=0}^{\underline{m}} \binom{m_0}{r}\binom{m_1}{r} P_1(u(r), \tilde{\rho}, \hat{\rho}) \tag{4.6}$$

where $u(r)$ is given in equation (2.1), and $P_1(u(r), \tilde{\rho}, \hat{\rho})$ is given in equation (4.5).

Proof. There are $\binom{m_0}{r}\binom{m_1}{r}$ permutations of $\boldsymbol{x}_0$ at swap distance r from $\boldsymbol{x}_c$. □

From (4.6) we see that $\tilde{p}_c$ can be computed in $O(\underline{\mathrm{m}})$ work. The mean squared error for $\tilde{p}_c$ is more complicated and more expensive. We need the double point inclusion probabilities and we need to count the number of pairs $\boldsymbol{x}_k, \boldsymbol{x}_\ell$ forming a given set of swap distances among $\boldsymbol{x}_k$, $\boldsymbol{x}_\ell$ and $\boldsymbol{x}_c$.

Lemma 3. *For $j = 1, 2$, let r_j be the swap distance of $\boldsymbol{x}_j$ from $\boldsymbol{x}_c$ and let r_3 be the swap distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Let u_1, u_2, u_3 be the corresponding inner products given by (2.1). If there are equalities among $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_c$, then the double point inclusion probability from (4.4) is*

$$P_2(u_1,u_2,u_3,\tilde\rho,\hat\rho) = \begin{cases} \mathbf{1}(\tilde\rho \geqslant \hat\rho), & \boldsymbol{x}_1 = \boldsymbol{x}_2 = \boldsymbol{x}_c \\ \mathbf{1}(\tilde\rho \geqslant \hat\rho) P_1(u_2,\tilde\rho,\hat\rho), & \boldsymbol{x}_1 = \boldsymbol{x}_c \neq \boldsymbol{x}_2 \\ \mathbf{1}(\tilde\rho \geqslant \hat\rho) P_1(u_1,\tilde\rho,\hat\rho), & \boldsymbol{x}_2 = \boldsymbol{x}_c \neq \boldsymbol{x}_1 \\ P_1(u_2,\tilde\rho,\hat\rho), & \boldsymbol{x}_1 = \boldsymbol{x}_2 \neq \boldsymbol{x}_c. \end{cases}$$

If $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_c$ are three distinct points with $\min(u_1, u_2) = -1$, then

$$P_2(u_1,u_2,u_3,\tilde\rho,\hat\rho) = \begin{cases} \mathbf{1}(-\tilde\rho \geqslant \hat\rho) P_1(u_2,\tilde\rho,\hat\rho), & u_1 = -1 \\ \mathbf{1}(-\tilde\rho \geqslant \hat\rho) P_1(u_1,\tilde\rho,\hat\rho), & u_2 = -1. \end{cases}$$

Otherwise $-1 < u_1, u_2 < 1$, and then

$$P_2(u_1,u_2,u_3,\tilde\rho,\hat\rho) = \begin{cases} \mathbf{1}(\tilde\rho u_1 \geqslant \hat\rho)\mathbf{1}(\tilde\rho u_2 \geqslant \hat\rho), & \tilde\rho = \pm 1 \\ \int_{-1}^1 \frac{\omega_{d-2}}{\omega_{d-1}}(1-t^2)^{\frac{d-1}{2}-1}\mathbf{1}(t \geqslant \rho_1)\mathbf{1}(tu_3^* \geqslant \rho_2)\,\mathrm{d}t, & \tilde\rho \neq \pm 1, u_3^* = \pm 1 \\ \int_{-1}^1 \frac{\omega_{d-2}}{\omega_{d-1}}(1-t^2)^{\frac{d-1}{2}-1}\mathbf{1}(t \geqslant \rho_1)\sigma_{d-2}\Big(C\big(\boldsymbol{x}_2^{**}, \frac{\rho_2 - tu_3^*}{\sqrt{1-t^2}\sqrt{1-u_3^{*2}}}\big)\Big)\,\mathrm{d}t, & \tilde\rho \neq \pm 1, |u_3^*| < 1 \end{cases}$$

where

$$u_3^* = \frac{u_3 - u_1u_2}{\sqrt{1-u_1^2}\sqrt{1-u_2^2}} \quad \text{and} \quad \rho_j = \frac{\hat\rho - \tilde\rho u_j}{\sqrt{1-\tilde\rho^2}\sqrt{1-u_j^2}},\ j = 1,2 \tag{4.7}$$

and $\boldsymbol{x}_2^{**}$ is the residual from the tangent-normal decomposition of $\boldsymbol{x}_2^*$ with respect to $\boldsymbol{x}_1^*$.

Proof. See He et al. (2018, Section 11.3). □

Next we consider the swap configuration among $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_c$. Let $\boldsymbol{x}_j$ be at swap distance r_j from $\boldsymbol{x}_c$, for $j = 1, 2$. We let δ_1 be the number of positive components of $\boldsymbol{x}_c$ that are negative in both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Similarly, δ_2 is the number of negative components of $\boldsymbol{x}_c$ that are positive in both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. See Figure 3. The swap distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is then $r_3 = r_1 + r_2 - \delta_1 - \delta_2$.

Let $\boldsymbol{r} = (r_1, r_2)$, $\boldsymbol{\delta} = (\delta_1, \delta_2)$ and $\underline{\mathrm{r}} = \min(r_1, r_2)$. We will study values of $r_1, r_2, r_3, \delta_1, \delta_2$ ranging over the following sets:

$$\begin{aligned} r_1, r_2 &\in R = \{1, \dots, \underline{\mathrm{m}}\} \\ \delta_1 &\in D_1(\boldsymbol{r}) = \{\max(0, r_1 + r_2 - m_0), \dots, \underline{\mathrm{r}}\} \\ \delta_2 &\in D_2(\boldsymbol{r}) = \{\max(0, r_1 + r_2 - m_1), \dots, \underline{\mathrm{r}}\}, \quad \text{and} \\ r_3 &\in R_3(\boldsymbol{r}) = \{\max(1, r_1 + r_2 - 2\underline{\mathrm{r}}), \dots, \min(r_1 + r_2, \underline{\mathrm{m}}, m_0 + m_1 - r_1 - r_2)\}. \end{aligned}$$

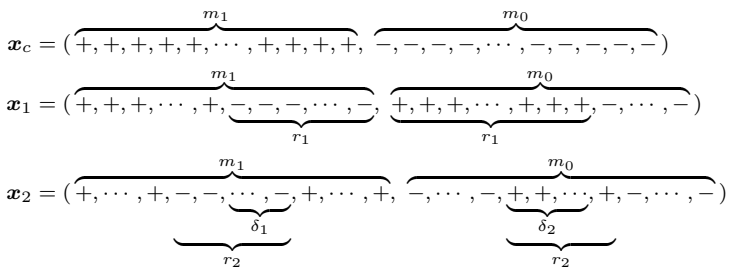


Fig 3: Illustration of r_1, r_2, δ_1 and δ_2. The points $\boldsymbol{x}_c$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ each have m_0 negative and m_1 positive components. For $j = 1, 2$ the swap distance between $\boldsymbol{x}_j$ and $\boldsymbol{x}_c$ is r_j. There are δ_1 positive components of $\boldsymbol{x}_c$ where both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are negative, and δ_2 negative components of $\boldsymbol{x}_c$ where both $\boldsymbol{x}_j$ are positive.

Whenever the lower bound for one of these sets exceeds the upper bound, we take the set to be empty, and a sum over it to be zero. Note that while $r_1 = 0$ is possible, it corresponds to $\boldsymbol{x}_1 = \boldsymbol{x}_c$ and we will handle that case specially, excluding it from R.

The number of pairs $(\boldsymbol{x}_\ell, \boldsymbol{x}_k)$ with a fixed $\boldsymbol{r}$ and $\boldsymbol{\delta}$ is

$$c(\boldsymbol{r}, \boldsymbol{\delta}) = \binom{m_0}{\delta_1}\binom{m_1}{\delta_2}\binom{m_0-\delta_1}{r_1-\delta_1}\binom{m_1-\delta_2}{r_1-\delta_2}\binom{m_0-r_1}{r_2-\delta_1}\binom{m_1-r_1}{r_2-\delta_2}. \tag{4.8}$$

Then the number of configurations given r_1, r_2 and r_3 is

$$c(r_1, r_2, r_3) = \sum_{\delta_1 \in D_1}\sum_{\delta_2 \in D_2} c(\boldsymbol{r}, \boldsymbol{\delta})\mathbf{1}(r_3 = r_1 + r_2 - \delta_1 - \delta_2). \tag{4.9}$$

The functions c are not to be confused with the index c of a permutation of $\boldsymbol{x}_0$.

We can now get an expression for the expected mean squared error under reference distribution 2 which combined with Theorem 4 for the mean provides an expression for the mean squared error of $\tilde{p}_c$.

Theorem 5. *Let $\boldsymbol{x}_0 \in \mathbb{S}^d$ be a centered and scaled binary vector with $m_0 \geqslant 1$ negative components and $m_1 \geqslant 1$ positive components. Let $\hat{\rho} = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_0 \in [-1, 1]$, and $\tilde{\rho} = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_c \in [-1, 1]$ for $c \in \{0, 1, \ldots, N-1\}$. Then*

$$\begin{aligned}\mathbb{E}_{2,c}(p(\boldsymbol{y}, \hat{\rho})^2) = \frac{1}{N^2}\Bigg[&\mathbf{1}(\tilde{\rho} \geqslant \hat{\rho}) + \sum_{r=1}^{\overline{m}}\binom{m_0}{r}\binom{m_1}{r}P_1(u(r), \tilde{\rho}, \hat{\rho}) \\ &+ 2\sum_{r=1}^{\overline{m}}\binom{m_0}{r}\binom{m_1}{r}P_2(1, u(r), u(r), \tilde{\rho}, \hat{\rho}) \\ &+ \sum_{r_1 \in R}\sum_{r_2 \in R}\sum_{r_3 \in R_3(\boldsymbol{r})} c(r_1, r_2, r_3)P_2(u_1, u_2, u_3, \tilde{\rho}, \hat{\rho})\Bigg]\end{aligned} \tag{4.10}$$

where $P_1(\cdot)$ is the single inclusion probability from Lemma 2, $P_2(\cdot)$ is the double inclusion probability from Lemma 3 and $c(r_1, r_2, r_3)$ is the configuration count in equation (4.9).

Proof. See He et al. (2018, Section 11.4). □

The function P_2 in equation (4.10) is computed by the expressions in Lemma 3. The lengthiest of these involve univariate integrals. We compute those integrals via the `integrate` function in R (R Core Team, 2015). In our experience, the cost of computing $\mathbb{E}_2(p(\boldsymbol{y}, \hat{\rho})^2)$ under reference distribution 2 is dominated by the cost of the $O(\underline{\mathrm{m}}^3)$ integrals required to get the $P_2(\cdot)$ values in (4.10). The cost also includes an $O(\underline{\mathrm{m}}^4)$ component because each $c(r_1, r_2, r_3)$ is a sum of $O(\underline{\mathrm{m}})$ terms, but it did not dominate the computation at the sample sizes we looked at (up to several hundred). See He et al. (2018, Section 13) for more details.

5. Generalized Stolarsky Invariance

Here we obtain the results for reference distribution 2 in a different way, by extending the work by Brauchart and Dick (2013). They introduced a weight on the height t of the spherical cap in the average. We now apply a weight function to the inner product $\langle \boldsymbol{z}, \boldsymbol{x}' \rangle$ between the center $\boldsymbol{z}$ of the spherical cap and a special point $\boldsymbol{x}'$.

Theorem 6. *Let $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$ be arbitrary points in $\mathbb{S}^d$ and $v(\cdot)$ and $h(\cdot)$ be positive functions in $L_2([-1, 1])$. Then for any $\boldsymbol{x}' \in \mathbb{S}^d$, the following equation holds,*

$$\begin{aligned}
&\int_{-1}^{1} v(t) \int_{\mathbb{S}^d} h(\langle \boldsymbol{z}, \boldsymbol{x}' \rangle) \Big| \sigma_d(C(\boldsymbol{z}; t)) - \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{x}_k) \Big|^2 \, \mathrm{d}\sigma_d(\boldsymbol{z}) \, \mathrm{d}t \\
&= \frac{1}{N^2} \sum_{k,\ell=0}^{N-1} K_{v,h,\boldsymbol{x}'}(\boldsymbol{x}_k, \boldsymbol{x}_\ell) + \int_{\mathbb{S}^d} \int_{\mathbb{S}^d} K_{v,h,\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\sigma_d(\boldsymbol{x}) \, \mathrm{d}\sigma_d(\boldsymbol{y}) \\
&\quad - \frac{2}{N} \sum_{k=0}^{N-1} \int_{\mathbb{S}^d} K_{v,h,\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{x}_k) \, \mathrm{d}\sigma_d(\boldsymbol{x})
\end{aligned} \tag{5.1}$$

where $K_{v,h,\boldsymbol{x}'} : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}$ is a reproducing kernel defined by

$$K_{v,h,\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{y}) = \int_{-1}^{1} v(t) \int_{\mathbb{S}^d} h(\langle \boldsymbol{z}, \boldsymbol{x}' \rangle) \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{x}) \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{y}) \, \mathrm{d}\sigma_d(\boldsymbol{z}) \, \mathrm{d}t. \tag{5.2}$$

Proof. See He et al. (2018, Section 11.5). □

Remark. *This theorem holds for any $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1} \in \mathbb{S}^d$ and for any $\boldsymbol{x}' \in \mathbb{S}^d$. The result is computationally and statistically most attractive when $\boldsymbol{x}' \in \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}\}$ and those N points are permuted versions of a centered and scaled binary vector $\boldsymbol{x}_0$.*

We now show that the second moment in Theorem 5 holds as a special limiting case of Theorem 6. In addition to $v_{\boldsymbol{\epsilon}}$ from Section 3 we introduce $\boldsymbol{\eta} = (\eta_1, \eta_2) \in (0,1)^2$ and

$$h_{\boldsymbol{\eta}}(s) = \eta_2 + \frac{1}{\eta_1(\frac{\omega_{d-1}}{\omega_d}(1-s^2)^{d/2-1})}\mathbf{1}(\tilde{\rho} \leqslant s \leqslant \tilde{\rho} + \eta_1). \tag{5.3}$$

Using these results we can now establish the following theorem, which provides the second moment of $p(\boldsymbol{y}, \hat{\rho})$ under reference distribution 2.

Theorem 7. *Let $\boldsymbol{x}_0 \in \mathbb{S}^d$ be a centered and scaled binary vector with $m_0 \geqslant 1$ negative components and $m_1 \geqslant 1$ positive components. Let $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}$ be the $N = \binom{m_0+m_1}{m_0}$ distinct permutations of $\boldsymbol{x}_0$. Let $\boldsymbol{x}_c$ be one of the $\boldsymbol{x}_k$ and define $\tilde{p}_c$ by (4.1). Then*

$$\mathbb{E}_{2,c}(p(\boldsymbol{y}, \hat{\rho})^2) = \frac{1}{N^2}\sum_{k,\ell=0}^{N-1}\int_{\mathbb{S}^{d-1}} \mathbf{1}(\langle \boldsymbol{y}, \boldsymbol{x}_k\rangle \geqslant \hat{\rho})\mathbf{1}(\langle \boldsymbol{y}, \boldsymbol{x}_\ell\rangle \geqslant \hat{\rho})\,\mathrm{d}\sigma_{d-1}(\boldsymbol{y}^*)$$

where $\boldsymbol{y} = \tilde{\rho}\boldsymbol{x}_c + \sqrt{1-\tilde{\rho}^2}\boldsymbol{y}^*$ and $\mathbb{E}_{2,c}$ denotes expectation under reference distribution 2(c).

Proof. The proof uses Theorem 6 with a sequence of h defined in (5.3) and v defined in (3.3). See He et al. (2018, Section 11.6). □

This result shows that we can use the invariance principle to derive the second moment of $p(\boldsymbol{y}, \hat{\rho})$ under reference distribution 2. The mean square in Theorem 7 matches the second moment equation (3.6) in Proposition 1.

6. Two-sided p-values

In statistical applications it is more usual to report two-sided p-values. A conservative approach is to use $2\min(p, 1-p)$ where p is a one-sided p-value. A sharper choice is the two-sided p-value from (1.3) which we write here as

$$p^{(2)} = \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{1}(|\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{y}_0| \geqslant |\hat{\rho}|),$$

for $\hat{\rho} = \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{y}_0$. For this section only, we use a superscript (2) to distinguish two-sided p-values from their one-sided counterparts. We describe how to get two-sided versions $\hat{p}_1^{(2)}$ and $\hat{p}_2^{(2)}$ of our one-sided estimates as well as their respective reference variances. If $\hat{\rho} = 0$, then trivially $p^{(2)} = 1$. From here on we assume that $\hat{\rho} \neq 0$.

We begin with $\hat{p}_1$. The two-sided version of $\hat{p}_1(\hat{\rho})$ is $\hat{p}_1^{(2)} = 2\sigma_d(C(\boldsymbol{y}; |\hat{\rho}|))$. Also $\mathbb{E}_1(p^{(2)}) = \hat{p}_1^{(2)}$. We now consider the mean square for the two-sided estimate

under reference distribution 1. For $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{S}^d$ with $u = \boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{x}_2$, the two-sided double inclusion probability under reference distribution 1 is

$$\tilde{V}_2(u; t, d) = \int_{\mathbb{S}^d} \mathbf{1}(|\langle \boldsymbol{z}, \boldsymbol{x}_1\rangle| \geqslant |t|)\mathbf{1}(|\langle \boldsymbol{z}, \boldsymbol{x}_2\rangle| \geqslant |t|)\, \mathrm{d}\sigma_d(\boldsymbol{z}).$$

For $t \neq 0$, we write $\mathbf{1}(|\langle \boldsymbol{z}, \boldsymbol{x}_j\rangle| \geqslant |t|) = \mathbf{1}(\langle \boldsymbol{z}, \boldsymbol{x}_j\rangle \geqslant |t|) + \mathbf{1}(\langle \boldsymbol{z}, (-\boldsymbol{x}_j)\rangle \geqslant |t|)$ for $j = 1, 2$ and expanding the product, we get

$$\tilde{V}_2(u; t, d) = 2V_2(u; |t|, d) + 2V_2(-u; |t|, d).$$

By replacing $V_2(u; t, d)$ with $\tilde{V}_2(u; t, d)$ and $\hat{p}_1(t)$ with $2\sigma_d(C(\boldsymbol{y}; |t|))$ in equation (3.8) of Theorem 3, we get a formula for $\mathrm{Var}_1(\hat{p}_1^{(2)})$.

Next we obtain corresponding formulas under reference distribution 2. For some fixed $c \in \{0, 1, \ldots, N-1\}$, let $u_j = \boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_c$ for $j = 1, 2$, and let $u_3 = \boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{x}_2$. Let the decomposition of $\boldsymbol{y}$ with respect to $\boldsymbol{x}_c$ be $\boldsymbol{y} = \tilde{\rho}\boldsymbol{x}_c + \sqrt{1-\tilde{\rho}^2}\boldsymbol{y}^*$. The two-sided inclusion probabilities are

$$\tilde{P}_1(u_1, \tilde{\rho}, \hat{\rho}) = \int_{\mathbb{S}^{d-1}} \mathbf{1}\big(|\langle \boldsymbol{y}, \boldsymbol{x}_1\rangle| \geqslant |\hat{\rho}|\big)\, \mathrm{d}\sigma_{d-1}(\boldsymbol{y}^*), \quad \text{and,}$$
$$\tilde{P}_2(u_1, u_2, u_3, \tilde{\rho}, \hat{\rho}) = \int_{\mathbb{S}^{d-1}} \mathbf{1}\big(|\langle \boldsymbol{y}, \boldsymbol{x}_1\rangle| \geqslant |\hat{\rho}|\big)\mathbf{1}\big(|\langle \boldsymbol{y}, \boldsymbol{x}_2\rangle| \geqslant |\hat{\rho}|\big)\, \mathrm{d}\sigma_{d-1}(\boldsymbol{y}^*),$$

where $\boldsymbol{y}^*$ enters the integrands through $\boldsymbol{y}$. After writing $\mathbf{1}(|\langle \boldsymbol{y}, \boldsymbol{x}_j\rangle| \geqslant |\hat{\rho}|) = \mathbf{1}(\langle \boldsymbol{y}, \boldsymbol{x}_j\rangle \geqslant |\hat{\rho}|) + \mathbf{1}(\langle \boldsymbol{y}, -\boldsymbol{x}_j\rangle \geqslant |\hat{\rho}|)$ we find that

$$\begin{aligned}
\tilde{P}_1(u_1, \tilde{\rho}, \hat{\rho}) &= P_1(u_1, \tilde{\rho}, |\hat{\rho}|) + P_1(-u_1, \tilde{\rho}, |\hat{\rho}|), \quad \text{and} \\
\tilde{P}_2(u_1, u_2, u_3, \tilde{\rho}, \hat{\rho}) &= P_2(u_1, u_2, u_3, \tilde{\rho}, |\hat{\rho}|) + P_2(-u_1, u_2, -u_3, \tilde{\rho}, |\hat{\rho}|) \\
&\quad + P_2(u_1, -u_2, -u_3, \tilde{\rho}, |\hat{\rho}|) + P_2(-u_1, -u_2, u_3, \tilde{\rho}, |\hat{\rho}|).
\end{aligned}$$

In the expression for $\tilde{P}_2$, notice that u_3 changes to $-u_3$ if and only if exactly one of u_1, u_2 changes sign. This is because $u_3 = \boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{x}_2$. Changing $P_1(u_1, \tilde{\rho}, \hat{\rho})$ and $P_2(u_1, u_2, u_3, \tilde{\rho}, \hat{\rho})$ to $\tilde{P}_1(u_1, \tilde{\rho}, \hat{\rho})$ and $\tilde{P}_2(u_1, u_2, u_3, \tilde{\rho}, \hat{\rho})$ respectively in Theorems 4 and 5, yields expressions for $\mathbb{E}_{2,c}(\tilde{p}_c^{(2)})$ and $\mathrm{Var}_{2,c}(\tilde{p}_c^{(2)})$.

7. Numerical Results

We consider two-sided p-values in this section. The main finding is that the root mean squared error (RMSE) of $\hat{p}_2$ under reference distribution 2 is usually just a small multiple of $\hat{p}_2$ itself.

First we evaluate the accuracy of $\hat{p}_1$, the simple spherical cap volume approximate p-value. We considered $m_0 = m_1$ in a range of values from 5 to 200. The values $\hat{p}_1$ ranged from just below 1 to 2×10^{-30}. We judge the accuracy of this estimate by $\mathrm{RMSE}_1(\hat{p}_1) = (\mathbb{E}_1(\hat{p}_1(\rho) - p(\boldsymbol{y}, \rho))^2)^{1/2}$. As $\rho \to 1$, $\hat{p}_1 \to 0$ and Figure 4a shows RMSE_1 decreasing towards 0 in this limit. The RMSE also

decreases with increasing sample size, as we would expect from the central limit theorem.

As seen in Figures 4a and 4b, the RMSE is not monotone in $\hat{p}_1$. Right at $\hat{p}_1 = 1$ we know that RMSE $= 0$ and around 0.1 there is a dip. The practically interesting values of $\hat{p}_1$ are much smaller than 0.1, and the RMSE is monotone for them.

A problem with $\hat{p}_1$ is that it can approach 0 even though $p \geqslant 1/N$ must hold. The distribution 1 RMSE does not reflect this problem. By studying $\mathbb{E}_2((\hat{p}_1(\rho) - p(\boldsymbol{y}, \rho))^2)^{1/2}$, we get a different result. In Figure 4c, the RMSE of $\hat{p}_1$ under distribution 2 reaches a plateau as $\hat{p}_1$ goes to 0.

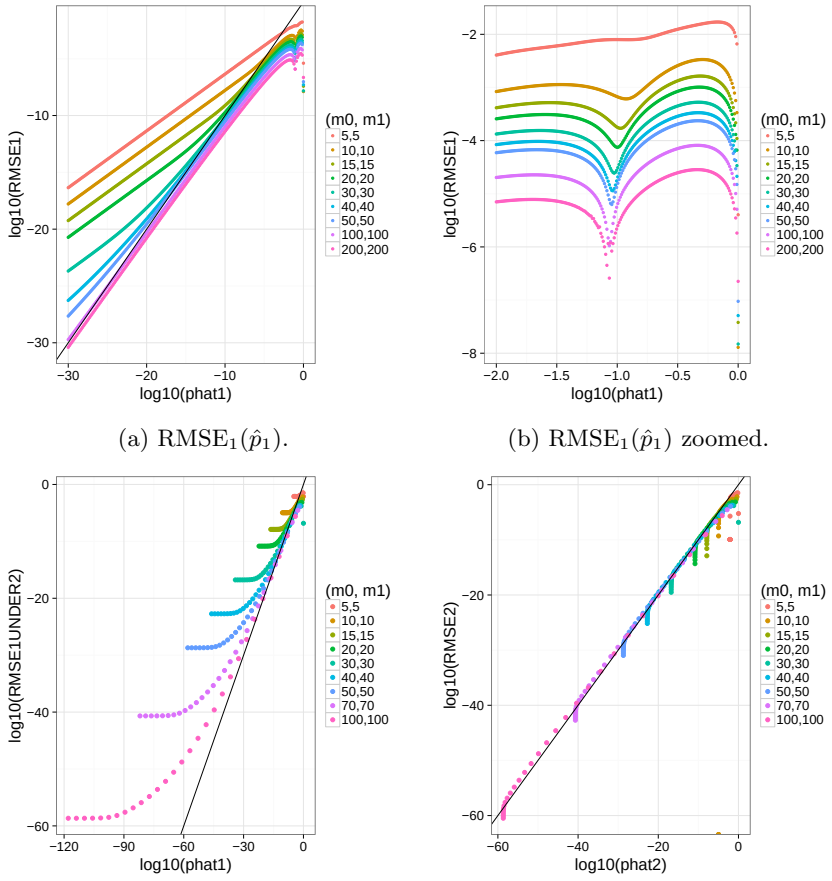
The estimator $\hat{p}_2 = \tilde{p}_0$ performs better than $\hat{p}_1$ because it makes more use of the data, and it is never below $1/N$. As seen in Figure 4d, the RMSE of $\hat{p}_2$ very closely matches $\hat{p}_2$ itself as $\hat{p}_2$ decreases to zero. That is, the relative error $|\hat{p}_2 - p|/\hat{p}_2$ is well behaved for small p-values. In rare event estimation, that property is known as strong efficiency (Blanchet and Glynn, 2008) and can be very hard to achieve. Here as $\hat{p}_2$ decreases to the granularity limit $1/N$, its RMSE actually decreases to 0. Eventually the distance from $\boldsymbol{y}_0$ to $\boldsymbol{x}_0$ is below the minimum interpoint distance among the $\boldsymbol{x}_k$ and then, for a one-sided test, $\hat{p}_2 = p = 1/N$. The estimators $\hat{p}_1$ and $\hat{p}_2$, do not differ much for larger p-values as seen in Figure 5a. But in the limit as $\rho \to 1$ we see that $\hat{p}_1 \to 0$, while $\hat{p}_2$ approaches the granularity limit $1/N$ instead.

Figure 5b compares the RMSE of the two estimators under distribution 2. As expected, $\hat{p}_2$ is more accurate. It also shows that the biggest differences occur only when $\hat{p}_1$ goes below $1/N$.

To examine the behavior of $\hat{p}_2$ more closely, we plot its coefficient of variation in Figure 6. We see that the relative uncertainty in $\hat{p}_2$ is not extremely large. Even when the estimated p-values are as small as 10^{-30} the coefficient of variation is below 5.

Our derivations for $\hat{p}_2$ extend to $\tilde{p}_c$ for any $c \in \{0, 1, \dots, N-1\}$. The estimator $\hat{p}_2 = \tilde{p}_0$ never goes below $1/N$ because $\boldsymbol{x}_0 \in C(\boldsymbol{y}_0, \hat{\rho})$. The same will hold for any c with $\langle \boldsymbol{x}_c, \boldsymbol{y}_0 \rangle \geqslant \langle \boldsymbol{x}_0, \boldsymbol{y}_0 \rangle$. Among these, we took a particular interest in $c = c_* \equiv \arg\max_k \langle \boldsymbol{x}_k, \boldsymbol{y}_0 \rangle$, the closest of all permutations of $\boldsymbol{x}_0$ to $\boldsymbol{y}_0$. This choice also minimizes $\sigma_d(\mathbb{Y}_2(c))$ and serves to illustrate the generality of our theory. For the two-sided case $c_* = \arg\max_k |\langle \boldsymbol{x}_k, \boldsymbol{y}_0 \rangle|$. We defined $\hat{p}_3 = \tilde{p}_{c_*} = \mathbb{E}(p(\boldsymbol{y}, \hat{\rho}))$ for $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y}_2(c_*))$. Figure 2.7 in He (2016) compares $\hat{p}_3$ to $\hat{p}_2$ for the same simulated cases reported here. There $\hat{p}_3$ tends to be larger (more conservative) than $\hat{p}_2$, though it does sometimes come out smaller. Figure 2.8 of He (2016) compares the RMSE of $\hat{p}_3$ to $\hat{p}_2$. The upward bias of $\hat{p}_3$ gave it a much larger RMSE.

Our simulations here all have $m_0 = m_1$. Section 12 of He et al. (2018) shows some simulations with $m_0 \neq m_1$. The results are quite similar to the ones in this section. The Parkinson's data in Section 8 have $m_0 \neq m_1$.



(a) $\mathrm{RMSE}_1(\hat{p}_1)$. (b) $\mathrm{RMSE}_1(\hat{p}_1)$ zoomed.

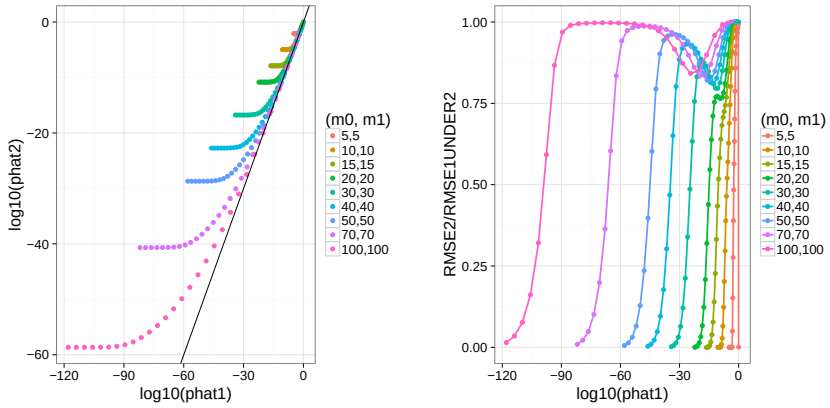
(c) $\mathrm{RMSE}_2(\hat{p}_1)$. (d) $\mathrm{RMSE}_2(\hat{p}_2)$.

Fig 4: RMSEs for $\hat{p}_1$ and $\hat{p}_2$ under reference distributions 1 and 2. The x-axis shows the estimate $\hat{p}_1$ or $\hat{p}_2$ as ρ varies from 1 to 0. Here $m_0 = m_1$.

8. Comparison to saddlepoint approximation

The small relative error property of $\hat{p}_2$ is similar to the relative error property in saddlepoint approximations, and so we compare our methods to saddlepoints approximations. Reid (1988) surveys saddlepoint approximations and Robinson (1982) develops them for permutation tests of linear statistics. When the true p-value is p, the saddlepoint approximation $\hat{p}_s$ satisfies $\hat{p}_s = p(1 + O(1/n))$. Because we do not know the implied constant in $O(1/n)$ or the n at which it takes effect, the saddlepoint approximation does not provide a computable upper bound for the true permutation p-value p.

Figure 7 compares our estimates to each other and to the saddlepoint ap-



(a) (b)

Fig 5: Comparison of $\hat{p}_1$ and $\hat{p}_2$. Panel (a) plots $\log_{10}(\hat{p}_2)$ against $\log_{10}(\hat{p}_1)$ for varying ρ with a 45 degree reference line. Panel (b) plots $\mathrm{RMSE}_2(\hat{p}_2)/\mathrm{RMSE}_2(\hat{p}_1)$ versus $\log_{10}(\hat{p}_1)$ for varying ρ.

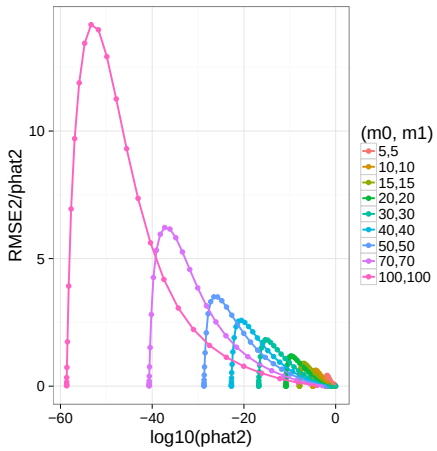


Fig 6: Coefficient of variation $\mathrm{RMSE}_2(\hat{p}_2)/\hat{p}_2$ versus $\log_{10}(\hat{p}_2)$ for varying ρ.

proximation, equation (1) from Robinson (1982). It includes the estimator $\hat{p}_3$ mentioned at the end of Section 7. The simulated data have the Exp(1) distribution under the control condition and are $2 + \mathrm{Exp}(1)$ under the affected condition. The sample sizes were $m_0 = m_1 = 10$ making it feasible to compute the exact permutation p-value. We ran 500 independent simulations, comparing two-sided p-values. Chapter 2 of He (2016) considers simulations from some additional distributions. Those had $t_{(5)}$, $\mathcal{N}(0,1)$ and $\mathbf{U}(0,1)$ for the control data

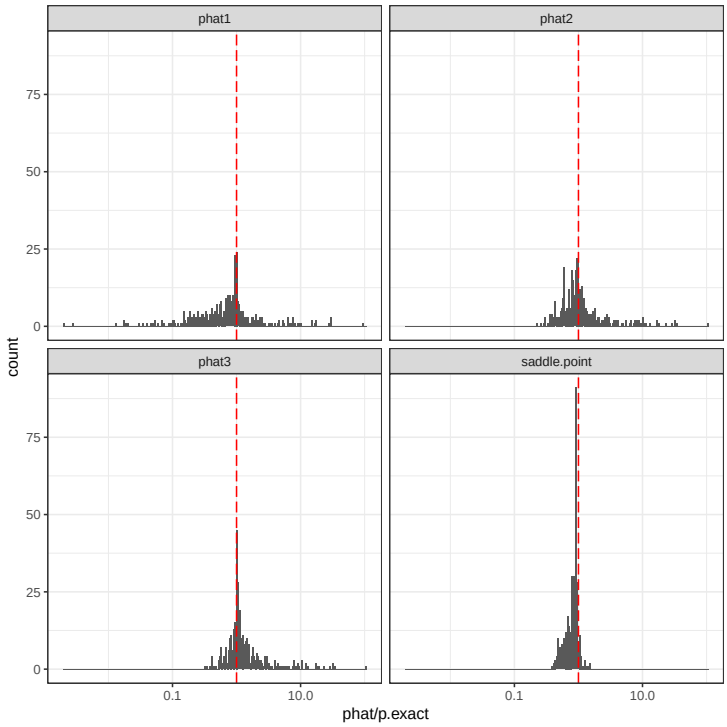


Fig 7: Simulation results $\hat{p}/p$ as described in the text, for $Y_{0,i} \overset{\text{iid}}{\sim} \text{Exp}(1)$, and $Y_{1,i} \overset{\text{iid}}{\sim} \text{Exp}(1) + 2$.

while the affected condition data are shifted versions of these distributions.

In these simulations, the naive spherical cap estimator $\hat{p}_1$, with no good relative error properties, is consistently least accurate and is often much smaller than the true p. The saddlepoint estimate is very accurate but tends to come out slightly smaller than the true p. The estimators $\hat{p}_2$ and $\hat{p}_3$ are less likely to be below p than the saddlepoint estimate, and by construction, they are never below the granularity limit. Qualitatively similar results happened for the other distributions. See He (2016, Chapter 2). The accuracy of all of these p-value estimates tends to be better for lighter tailed Y_i.

We can also construct Z scores, $Z_2 = (p - \hat{p}_2)/\text{RMSE}_2$ and a similar Z_3. If these take large values, then it means that $\hat{p}$ is too small and, moreover, that our computed RMSE does not diagnose it. The largest Z scores we observed are in Table 1. The largest Z values arose for exponential data with $p \doteq 0.89$ and $\hat{p}_2 \doteq 0.78 \doteq \hat{p}_3$. Such large p-values are not very important and so maximal Z scores are also shown among estimated p-values below 0.1.

The Z values are not very extreme. This suggests that it might be feasible to get a conservative p-value estimate by adding some multiple of RMSE_2 to $\hat{p}_2$.

Dist'n $Y_{0,i}$	$\max Z_2$	$\max\{Z_2 \mid \hat{p}_2 < 0.1\}$	$\max Z_3$	$\max\{Z_3 \mid \hat{p}_3 < 0.1\}$
$t_{(5)}$	26.7	1.91	31.5	3.87
Exp(1)	7.55	7.55	7.76	7.76
$\mathbf{U}(0,1)$	3.49	2.45	5.87	2.61
$\mathcal{N}(0,1)$	3.07	2.78	3.07	2.78

TABLE 1
Maximal Z scores observed for $\hat{p}_2$ and $\hat{p}_3$ in 500 independent replications.

First author	m_1	m_0	$N = \binom{m_1+m_0}{m_1}$
Zhang	11	18	3.5×10^7
Moran	29	14	7.9×10^{10}
Scherzer	50	22	1.8×10^{18}

TABLE 2
Sample sizes for three microarray studies.

Further work would be required to identify an appropriate multiple.

9. Data comparisons

Three data sets on Parkinson's disease were used by Larson and Owen (2015) and investigated in Chapter 6 of He (2016). They come from Scherzer et al. (2007), Moran et al. (2006) and Zhang et al. (2005). Table 2 shows their sample sizes. Section 14 of He et al. (2018) describes how to obtain this data.

For this comparison, there were 6180 gene sets from v5.1 of mSigDB's gene set collections. Curated gene sets and Gene Ontology gene sets were used. The gene sets ranged in size from 5 to 2131 genes with an average size of 93.08 genes. Slightly different versions of the gene sets were used in Larson and Owen (2015).

Estimates of two-sided p-values for linear test statistics were obtained using 10^6 random permutations, so $M = 10^6 + 1$. When the estimate was below 10^{-4}, then we increased M to $10^7 + 1$. We will use the resulting Monte Carlo estimates as the gold standard to evaluate less expensive approximate p-values. The Zhang data set had the smallest sample size and had no gene sets with gold standard p-value below 0.01 and so we do not compare estimated p-value estimates for this data set. Table 3 shows some timing data.

Table 4 gives the correlation over gene sets between $\log_{10}(\hat{p})$ and $\log_{10}(p_g)$ for each of our estimates $\hat{p}$, where p_g is our gold standard value. From Table 4, we see that $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ have nearly the same correlations with the gold stan-

Data Set	Saddle	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
Zhang	0.0631	0.0024	0.0031	0.0032
Moran	0.0894	0.0029	0.0037	0.0038
Scherzer	0.1394	0.0034	0.0045	0.0047

TABLE 3
Average, over 6180 gene sets of the running time in seconds for saddlepoint p-values and $\hat{p}_j$, $j = 1, 2, 3$. Total time ranges from under 1/4 minute to just over 14 minutes.

Data source	Corr.	# sets	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_{\text{saddle}}$
Moran small	Pearson	3594	0.9997	0.9997	0.9997	0.9934
Moran small	Kendall	3594	0.9857	0.9857	0.9866	0.9397
Moran tiny	Pearson	253	0.9684	0.9688	0.9787	0.7930
Moran tiny	Kendall	253	0.8820	0.8820	0.9033	0.6863
Scherzer small	Pearson	504	0.9997	0.9997	0.9997	0.9836
Scherzer small	Kendall	504	0.9871	0.9871	0.9871	0.8965
Scherzer tiny	Pearson	16	0.9950	0.9950	0.9956	0.8794
Scherzer tiny	Kendall	16	0.9500	0.9500	0.9500	0.7833

TABLE 4

Pearson and Kendall correlations over gene sets, between four approximations $\log_{10}(\hat{p})$ *and* $\log_{10}(p_g)$*, where* p_g *is an expensive gold standard estimate. Here 'small' refers to only the* 3594 *gene sets with* $p_g < 0.05$ *for Moran (*504 *for Scherzer), while 'tiny' refers to* 253 *gene sets with* $p_g < 10^{-4}$ *(Moran) or* 16 *gene sets with* $p_g < 10^{-3}$ *(Scherzer).*

dard; indeed they correlate highly with each other. They correlate with the gold standard estimate much more closely than the saddlepoint estimator does. The estimate $\hat{p}_3$ is frequently most correlated with the gold standard. These correlations capture the ability of a method to correctly rank the gene sets by significance. Figures in Chapter 6 of He (2016) give scatterplots that show the accuracy of each $\hat{p}$ for p_g. These show the saddlepoint estimator is biased slightly low and $\hat{p}_3$ is biased slightly high.

10. Discussion

We have constructed approximations to the permutation p-value using probability and spherical geometry. Many other approximation methods have been proposed for permutation tests. For instance, Zhou et al. (2009) fit approximations by moments in the Pearson family. Larson and Owen (2015) fit Gaussian and beta approximations to linear statistics and gamma approximations to quadratic statistics for gene set testing problems. Knijnenburg et al. (2009) fit generalized extreme value distributions to the tails of sampled permutation values.

Our proposed estimator $\hat{p}_2$ has a small relative error, as measured by RMSE in the limit as $\hat{\rho} \to 1$. The well-known saddlepoint estimator also has a small relative error of $O(1/n)$ extending to very small p-values. We found it performed well on our simulated data but not as well on the Parkinson's disease gene sets.

None of these approximations come with an all inclusive p-value that accounts for both numerical uncertainty of the estimation and sampling uncertainty behind the original data. Monte Carlo sampling of permutations has such a p-value, but it is computationally infeasible to attain very small p-values that way, and so a gap remains.

We have employed reference distributions in an effort to address this gap. We select a set $\mathbb{Y}$ containing $\boldsymbol{y}_0$ and find the first two moments of $p(\boldsymbol{y}, \hat{\rho})$ for $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y})$. If the data $\boldsymbol{y}_0$ were actually sampled from our reference distribution, then we could get an all inclusive conservative p-value via the Chebychev inequality.

To illustrate the Chebychev inequality, let $\mu = \mathbb{E}(p(\boldsymbol{y}, \hat{\rho}))$ and $\sigma^2 = \mathrm{Var}(p(\boldsymbol{y}, \hat{\rho}))$ for the observed value $\hat{\rho} = \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{y}_0$ and for random $\boldsymbol{y} \sim \mathbf{U}(\mathbb{Y})$ for some reference set $\mathbb{Y}$. Then $\Pr(p \geqslant \mu + \lambda\sigma) \leqslant 1/(1+\lambda^2)$ for any $\lambda > 0$. Under this model, $p^* = \mu + \lambda\sigma + 1/(1+\lambda^2)$ is a conservative p-value. Minimizing p^* over λ reduces to solving $2\lambda = \sigma(1+\lambda^2)^2$. For small p we anticipate $\lambda \gg 1$ and hence $\lambda' = (2/\sigma)^{1/3}$ will be almost as good as the optimal λ we could find numerically. That choice leads to $p^* \leqslant \mu + (2^{1/3} + 2^{-2/3})\sigma^{2/3}$.

For a numerical illustration, consider $\mu = 10^{-30}$ and $\sigma = 3 \times 10^{-30}$, roughly describing the small p-value estimates from the case $m_0 = m_1 = 70$. Then $p^* \leqslant 4 \times 10^{-20}$ is much larger than μ and yet still very small, likely small enough to be significant after multiplicity adjustments. This numerical illustration uses a Chebychev inequality at $\lambda' \doteq 8.7 \times 10^9$ standard deviations. We suspect that this is conservative but do not have rigorous information to support that suspicion.

A rigorous upper bound for p could be attained using L_∞ spherical cap discrepancies instead of the L_2 version, but computing such discrepancies is a major challenge. Narcowich et al. (2010) give upper bounds for the L_∞ spherical cap discrepancy, in terms of averages of a great many harmonic functions at the points $\boldsymbol{x}_i$. For our application we need bounds for spherical caps of a fixed volume (under distribution 1) and of fixed volume and constrained location (under distribution 2) and those go beyond what is in Narcowich et al. (2010).

Our permutation points fall into a lattice subset of $\mathbb{R}^d$ intersected with the unit sphere $\mathbb{S}^d$. Our problem of counting the number of such points in a subset is one that is addressed under the term 'Geometry of numbers'. According to a personal communication from Neil Sloane, the standard approach to such problems is via the volume ratio, which in our setting is $\hat{p}_1$. The new estimator $\hat{p}_2$ does much better than $\hat{p}_1$ on our simulated data and slightly better on the real data.

Of the methods we investigated, the saddlepoint approximation did best on the simulated data, while the geometric methods $\hat{p}_j$ for $j = 1, 2, 3$ were more accurate than the saddlepoint method on the Parkinson's data, especially so on the smallest p-values. The best relative error $|\hat{p} - p|/p$ for tiny p is attained by saddlepoints (from asymptotic theory) and $\hat{p}_2$ from the computations illustrated in Figure 4d. Figure 2.8 of He (2016) shows that $\hat{p}_3$ is slightly worse than $\hat{p}_2$. Both are much better than $\hat{p}_1$. Based on both accuracy and the existence of a numerical accuracy estimate, we prefer $\hat{p}_2$ for fast approximations to L_G. He (2016) includes some approximations for the quadratic statistic Q_G.

Acknowledgements

This work was supported by the US National Science Foundation under grants DMS-1407397 and DMS-1521145. We thank John Robinson and Neil Sloane for helpful comments. We also thank two anonymous referees and an AE for comments that helped us improve our presentation.

References

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:1–20.

Barnard, G. A. (1963). Discussion of the spectral analysis of point processes (by M. S. Bartlett). *Journal of the Royal Statistical Society, Series B*, 25:294.

Bilyk, D., Dai, F., and Matzke, R. (2016). Stolarsky principle and energy optimization on the sphere. Technical report, arXiv:1611.04420.

Blanchet, J. and Glynn, P. (2008). Efficient rare-event simulation for the maximum of heavy-tailed random walks. *The Annals of Applied Probability*, pages 1351–1378.

Brauchart, J. and Dick, J. (2013). A simple proof of Stolarsky's invariance principle. *Proceedings of the American Mathematical Society*, 141(6):2085–2096.

Fadista, J., Manning, A. K., Florez, J. C., and Groop, L. (2016). The (in) famous GWAS p-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8):1202–1205.

He, H. Y. (2016). *Efficient Permutation-Based P-Value Estimation for Gene Set Tests*. PhD thesis, Stanford University.

He, H. Y., Basu, K., Zhao, Q., and Owen, A. B. (2018). Supplement to: Permutation p-value approximation via generalized Stolarsky invariance. Technical report, Stanford University.

Jiang, Z. and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313.

Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., and Shmulevich, I. (2009). Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168.

Larson, J. L. and Owen, A. B. (2015). Moment based gene set tests. *BMC Bioinformatics*, 16(1):132.

Lee, Y. and Kim, W. C. (2014). Concise formulas for the surface area of the intersection of two hyperspherical caps. Technical report, Korea Advanced Institute of Science and Technology.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, 3rd edition.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. John Wiley & Sons Ltd., Chichester, UK.

Moran, L. B., Duke, D. C., Deprez, M., Dexter, D. T., Pearce, R. K. B., and Graeber, M. B. (2006). Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. *Neurogenetics*, 7(1):1–11.

Narcowich, F. J., Sun, X., Ward, J. D., and Wu, Z. (2010). Leveque type inequalities and discrepancy estimates for minimal energy configurations on spheres. *Journal of Approximation Theory*, 162(6):1256–1278.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, PA.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, pages 213–227.

Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *Journal of the Royal Statistical Society, Series B*, pages 91–101.

Scherzer, C. R., Eklund, A. C., Morse, L. J., Liao, Z., Locascio, J. J., Fefer, D., Schwarzschild, M. A., Schlossmacher, M. G., Hauser, M. A., Vance, J. M., Sudarsky, L. R., Standaert, D. G., Growdon, J. H., Jensen, R. V., and Gullans, S. R. (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc Natl Acad Sci*, 104(3):955–60.

Stolarsky, K. B. (1973). Sums of distances between points on a sphere. II. *Proceedings of the American Mathematical Society*, 41(2):575–582.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102(38):13544–13549.

Zhang, Y., James, M., Middleton, F. A., and Davis, R. L. (2005). Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *American J Med Genet B Neuropsychiatry Genet*, 137B(1):5–16.

Zhou, C., Wang, H. J., and Wang, Y. M. (2009). Efficient moments-based permutation tests. In *Advances in Neural Information Processing Systems*, pages 2277–2285.