Biometrika (2017), **103**, 1, *pp*. 1–13
Printed in Great Britain Advance Access publication on 31 July 2016

More Efficient Exact Group Invariance Testing: using a Representative Subgroup

BY N.W. KONING

Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.

n.w.koning@ese.eur.nl.

J. HEMERIK

Department Biometris, Wageningen University & Research, P.O. Box 16, 6700 AC Wageningen, The Netherlands.

jesse.hemerik@wur.nl.

SUMMARY

We consider testing invariance of a distribution under an algebraic group of transformations, such as permutations or sign-flips. As such groups are typically huge, tests based on the full group are often computationally infeasible. Hence, it is standard practice to use a random subset of transformations. We improve upon this by replacing the random subset with a strategically chosen, fixed subgroup of transformations. In a generalized location model, we show that the resulting tests are often consistent for lower signal-to-noise ratios. Moreover, we establish an analogy between the power improvement and switching from a t-test to a Z-test under normality. Importantly, in permutation-based multiple testing, the efficiency gain with our approach can be huge, since we attain the same power with much fewer permutations.

Some key words: Group-invariance test; permutation test; randomization test; resampling; subgroup.

1. INTRODUCTION

1·1. *Context*

Permutation tests, randomization tests and related testing procedures are ubiquitous in modern-day statistical research (Good, 2005; Onghena, 2018; Berry et al., 2014), for example in genomics (Tusher et al., 2001; Li & Tibshirani, 2013; Debeer & Strobl, 2020), neuroimaging (Eklund et al., 2016) and economics (Young, 2019). Such non-parametric and semi-parametric tests are useful, in part, because they require few assumptions on the data distribution (Anderson & Robinson, 2001; Hemerik et al., 2021). Additionally, they have seen recent popularity in the simultaneous testing of many hypotheses, as they are often able to take into account the dependence structure of the data in an exact way, leading to relatively high power (Westfall & Young, 1993; Tusher et al., 2001; Meinshausen, 2006; Pesarin & Salmaso, 2010; Meinshausen et al., 2011; Hemerik & Goeman, 2018b; Blanchard et al., 2020). For example, under strong positive dependence in the data, Bonferroni's multiple testing correction is very conservative and is greatly improved by a permutation method (Westfall & Young, 1993; Westfall & Troendle, 2008).

These non-parametric tests often rely on the assumption that under the null hypothesis, the data distribution is invariant under a set $\mathcal{G}$ of transformations that is a group, in the algebraic sense (Lehmann & Romano, 2022; Hemerik & Goeman, 2018a). That is, every element in $\mathcal{G}$ has an inverse in $\mathcal{G}$ and $\mathcal{G}$ is closed under composition. We will refer to tests based on a group invariance assumption as group invariance tests. One prominent example is permutation tests. Another is sign-flipping tests (Fisher, 1935; Efron, 1969; Bekker & Lawford, 2008; Davidson & Flachaire, 2008; Winkler et al., 2014; Andreella et al., 2023; Blain et al., 2022; Girardi et al., 2022). Sign-flipping is used, for instance, for testing in (generalized) linear models by sign-flipping residuals or score contributions (Hemerik et al., 2020, 2021; De Santis et al., 2022). Tests based on other groups of transformations, e.g. rotations are also used (Langsrud, 2005; Solari et al., 2014). The requirement of $\mathcal{G}$ to be a group is fundamental. Using a set of transformations that is not a group can yield an overly conservative or anti-conservative test (Southworth et al., 2009; Hemerik & Goeman, 2018a, 2021).

1·2. *Current practice*

For moderate or large sample sizes, the cardinality $|\mathcal{G}|$ is typically huge, so that it is often computationally infeasible to use the whole group. For example, the order of the permutation and sign-flipping groups is $n!$ and 2^n, respectively, where n is the number of observations. As a solution, it is universal practice among researchers to use a random finite subset of transformations (Eden & Yates, 1933; Dwass, 1957; Phipson & Smyth, 2010). This can be done in such a way that the test is still exact (Hemerik & Goeman, 2018a). We will henceforth refer to such tests as a Monte Carlo group invariance tests.

Using a small random subset of transformations results in power loss compared to using the full group of transformations. Moreover, it reduces replicability as the test outcome depends on the randomly sampled subset of transformations. For a Monte Carlo group invariance test to have good power and to obtain replicable results, it is therefore important that a large random subset of transformations is used. Typically, this number is several times larger than α^{-1}, where α is the nominal level of the test. For example, for a nominal level $\alpha = 0.05$, it is common to use 100-5000 random transformations. Unfortunately, using a large random subset of transformations can remain prohibitive, as tests and multiple testing methods based on permutations (or other transformations) can be highly computationally intensive (Gao et al., 2010; Kofler & Schlötterer, 2012; Hemerik et al., 2019; Vesely et al., 2023).

To reduce the number of transformations required, a few methods have been proposed in the literature. For example, Good (2005) approximates the permutation reference distribution using moment matching. Winkler et al. (2016) review and propose additional methods for obtaining high-resolution p-values based on a limited number of random permutations. However, although the resolution of permutation p-values is improved, these approaches are not exact. Moreover, the p-values will still depend on the particular random sample of permutations that has been drawn, which also reduces the replicability of the results. Further, it is not clear how to generally combine the methods in e.g. Winkler et al. (2016) with permutation-based multiple testing methods, which are often not p-value based. Thus, most of the drawbacks of the use of a limited set of random transformations have remained unresolved.

1·3. *Contribution*

In this paper, we propose an alternative approach to group invariance testing: we replace the random subset of transformations with a deterministic subgroup of transformations, i.e., a subset that is also a group. This still results in an exact test. We henceforth refer to such tests as subgroup invariance tests. The main idea of this paper is that the subgroup can be chosen in a clever way, to

obtain good power. This approach works without any additional assumptions and yields a fully deterministic test.

We illustrate this idea in a generalized location-shift model, which also contains the important two-sample comparison of means as a special case. As the group $\mathcal{G}$ we consider (subgroups of) the orthogonal group. This group contains the rotation, permutation and sign-flipping groups as subgroups, and can be conveniently represented as a set of orthonormal matrices.

In this model, we compare the power of subgroup invariance tests to the commonly used Monte Carlo group invariance tests that are based on a random subset. We prove novel consistency results that link the consistency of a subgroup invariance test to a real value $\delta_{\mathcal{S}}$ that depends on the subgroup $\mathcal{S}$ of $\mathcal{G}$. Intuitively speaking, if $\delta_{\mathcal{S}}$ is 'large' then the elements of $\mathcal{S}$ are more alike. We find that the smaller $\delta_{\mathcal{S}}$ is, the smaller the signal-to-noise ratio is that is required for the test to be consistent. We prove an analogous result about the consistency of Monte Carlo group invariance tests. There, the consistency turns out to depend on the full group $\mathcal{G}$ through $\delta_{\mathcal{G}}$. As we have $\delta_{\mathcal{S}} \leq \delta_{\mathcal{G}}$ for any subgroup $\mathcal{S}$ of $\mathcal{G}$, it follows that a lower signal-to-noise ratio is required for consistency of subgroup invariance tests than for the consistency of Monte Carlo group invariance tests.

Moreover, we provide a detailed power analysis in a normal location model, by linking the tests to Z-tests and t-tests. In particular, we consider subgroups $\mathcal{S}$ for which $\delta_{\mathcal{S}}^{\text{abs}} = 0$, a variant of $\delta_{\mathcal{S}}$ we use for the analysis of two-sided tests. We show that subgroup invariance tests based on such subgroups have the same size and power as Monte Carlo Z-tests. In addition, we show that a Monte Carlo orthogonal group invariance test has the same size and power properties as a Monte Carlo t-test (regardless of the normality assumption). Under normality, Monte Carlo Z-tests are more powerful than Monte Carlo t-tests, if the same number of draws is used. As a consequence, if the random sample and subgroup have the same number of elements, the subgroup invariance tests are more powerful.

We also study the power through simulations, where we find that the subgroup invariance tests substantially improve the Monte Carlo tests based on the same number of transformations. This is especially the case in multiple testing settings. Thus, our method is an effective way to improve power or reduce the computation time; these are two sides of the same coin.

Further, we provide theory on the existence of subgroups and how to construct them. We connect the problem of finding a subgroup for which $\delta_{\mathcal{S}}$ is small to group code problems (Slepian, 1968; Conway & Sloane, 1998), which relate to the spreading of points on a unit hypersphere. This connection exposes that demanding $\delta_{\mathcal{S}}$ (or its two-sided equivalent $\delta_{\mathcal{S}}^{\text{abs}}$) to be small, bounds the maximum order of the subgroup. For example, if $\delta_{\mathcal{S}}^{\text{abs}} = 0$ then $|\mathcal{S}| \leq n$, where n is the dimension of the data. In addition, we provide in-depth theory on the existence of subgroups of the sign-flipping group.

Our theory on using subgroups is also relevant to the framework of Ramdas et al. (2023), which generalizes an idea from Section 3.3 of Hemerik & Goeman (2018a). In Ramdas et al. (2023), the user can choose any subset of permutations or, more generally, any distribution on permutations and the method will still provide an exact test. However, the question remains how to choose a good subset of permutations, which is the topic of the current paper.

The Supplementary Material contains appendices A–E, which include an example of the two-sample comparison problem, a description of an algorithm for constructing subgroups of the sign-flipping group, as well as proofs of the results, an analysis of real fMRI data, and some additional simulation results. The algorithm described in the Supplementary Material is implemented in `https://github.com/nickwkoning/NOS`, and `https://github.com/nickwkoning/NOSdata` provides a readily available library of subgroups of the sign-flipping

groups that were obtained using the algorithm, as well as a short implementation of a subgroup based test.

2. BACKGROUND: TESTING INVARIANCE

2·1. *Invariance hypothesis*

Let $\mathbb{R}^n$ be our sample space and $\mathcal{G}$ be a compact group of $n \times n$ orthonormal matrices under matrix multiplication. For $\mathcal{G}$ to be a group, it means that it contains the product of every pair of its elements, the inverse of every element, as well as the identity matrix I. If $\mathcal{G}$ is finite, for example, then it is also compact. More generally, we can allow our sample space to be a topological space, and the group to be a compact topological group that acts continuously on the sample space (see e.g. Eaton, 1989).

We observe a realization of a random variable X in our sample space. The random variable X is said to be $\mathcal{G}$-invariant if $X \overset{d}{=} GX$, for all $G \in \mathcal{G}$. Equivalently, X is $\mathcal{G}$-invariant if $X \overset{d}{=} \overline{G}X$, where $\overline{G}$ is uniform on $\mathcal{G}$ independent of X (see Appendix D·1 for a proof). Our goal is to test whether X is $\mathcal{G}$-invariant at some level $\alpha \in (0,1)$:

$$H_0 : X \text{ is } \mathcal{G}\text{-invariant},$$
$$H_1 : X \text{ is not } \mathcal{G}\text{-invariant}.$$

Example 1. An important example is invariance under the permutation group: the collection of all permutation matrices $\mathcal{P}$. Invariance under the permutation group sometimes referred to as 'exchangeability'. As an example, if X has i.i.d. elements then X is exchangeable (though exchangeable vectors need not be i.i.d.). Another example is the collection of 'sign-flipping' matrices $\mathcal{R}$, which are diagonal matrices with diagonal elements in $\{-1, 1\}$. Invariance under $\mathcal{R}$ includes, for example, random vectors with independent elements and marginal distributions symmetric about 0. A further important example is the orthogonal group $\mathcal{H}$: the collection of all orthonormal matrices, which includes all rotations. Random vectors that are $\mathcal{H}$-invariant are also called 'spherical'.

2·2. *Group invariance tests*

In order to test invariance, it is standard practice to use a group invariance test. Let $T : \mathcal{X} \to \mathbb{R}$ act as a test statistic and $\overline{G}$ denote a random variable that is uniformly distributed on $\mathcal{G}$. Then, a level α $\mathcal{G}$-invariance test is defined as

$$\phi_\alpha^{\mathcal{G}}(X) = \mathbb{I}[\mathbb{P}_{\overline{G}}\{T(\overline{G}X) \geq T(X)\} \leq \alpha],$$

where $\mathbb{P}_{\overline{G}}[T(\overline{G}X) \geq T(X)]$ is the p-value of the test. It can be equivalently written in a form where $T(X)$ should exceed some threshold, as

$$\phi_\alpha^{\mathcal{G}}(X) = \mathbb{I}\{T(X) > q_X^\alpha(\mathcal{G})\},$$

where $q_X^\alpha(\mathcal{G})$ is the α-upper quantile of the distribution of $T(\overline{G}X)$, where $\overline{G}$ is uniform on $\mathcal{G}$. Note here that the critical value $q_X^\alpha(\mathcal{G})$ depends on the data.

Group invariance tests are popular as they yield a test with exact size control (as opposed to approximate), regardless of the chosen test statistic. This is captured by the following well-known result.

THEOREM 1. *If X is $\mathcal{G}$-invariant, then $\mathbb{E}_X \phi_\alpha^{\mathcal{G}}(X) \leq \alpha$.*

For completeness, a proof can be found in Appendix D, together with all proofs of results that are not included in the text.

2·3. *Practice*

In practice, the groups under which invariance is tested are typically huge. This often makes it infeasible to compute the p-value $\mathbb{P}_{\overline{G}}\{T(\overline{G}X) \geq T(X)\}$. Therefore, it is universal practice to use a Monte Carlo (MC) test, instead. In particular, let $\mathcal{G}_M$ be a set containing the identity I, and $M-1$ independent draws uniformly from the group $\mathcal{G}$. Moreover, let $\overline{G}_M$ be uniformly distributed on $\mathcal{G}_M$. Then, an M-draw MC $\mathcal{G}$-invariance test is defined as

$$\phi_\alpha^{\mathcal{G}_M}(X) = \mathbb{I}[\mathbb{P}_{\overline{G}_M}\{T(\overline{G}_M X) \geq T(X)\} \leq \alpha].$$

This test is still exact (Hemerik & Goeman, 2018a). However, it comes with two issues. First, if M is small, then one would expect the power to be low, as the Monte Carlo draws yield a crude approximation of the p-value based on the entire group (Dwass, 1957; Hope, 1968). Additionally, the test is random as it relies on the random draw of $\mathcal{G}_M$, leading to worsened replicability when M is small.

3. Subgroup-invariance tests

The key idea in this paper is the use of subgroup invariance tests as an alternative to Monte Carlo group invariance tests. The construction of these tests relies on Theorem 2.

Theorem 2. *If $\mathcal{S}$ is a subgroup of $\mathcal{G}$ and X is $\mathcal{G}$-invariant, then X is also $\mathcal{S}$-invariant.*

The result follows immediately from the definition of $\mathcal{G}$-invariance. If the subgroup $\mathcal{S}$ is compact, we can use it to construct a subgroup invariance test $\phi_\alpha^{\mathcal{S}}$. Such a subgroup invariance test controls size, by Corollary 1.

Corollary 1. *If X is $\mathcal{G}$-invariant, then $\mathbb{E}_X \phi_\alpha^{\mathcal{S}}(X) \leq \alpha$.*

Proof. From Theorem 2, we know X is $\mathcal{S}$-invariant. Theorem 1 then yields $\mathbb{E}_X \phi_\alpha^{\mathcal{S}}(X) \leq \alpha$.□

As any compact subgroup yields a test with exact size control, we can choose a subgroup that yields a test with desirable power properties. In particular, we will consider choosing a finite subgroup of order M, say, that yields a test with desirable power properties. This way, we aim to obtain tests that are more powerful than Monte Carlo group invariance tests.

In contrast to a Monte Carlo group invariance test, such a subgroup invariance test has the additional benefit that it is completely deterministic given the data. If desirable, one can also construct Monte Carlo tests based on the subgroup, by sampling independently from a uniform distribution on the subgroup.

4. Generalized location model

4·1. *The model*

From this section onwards, we study subgroup invariance tests in a generalized location model. Suppose that X can be decomposed as

$$X = \iota\mu + \varepsilon,$$

where ι is a unit n-vector, $\mu \in \mathbb{R}$ is the parameter of interest and ε is a $\mathcal{G}$-invariant random n-vector, where $\mathcal{G} \subseteq \mathcal{H}$ is some compact group of orthonormal matrices. As ε is $\mathcal{G}$-invariant, we

have $\varepsilon \overset{d}{=} \overline{G}\varepsilon$ where $\overline{G}$ is uniform on $\mathcal{G}$, independent from ε. Hence, we can equivalently define X as

$$X = \iota\mu + \overline{G}\varepsilon.$$

We sometimes write $X_{\overline{G}}$ to emphasize the dependence of X on $\overline{G}$.

We are interested in testing $H_0 : \mu = 0$ at level α. If $\mu = 0$ then $X = \varepsilon$, so that X is $\mathcal{G}$-invariant under H_0. As a consequence, if we test for $\mathcal{G}$-invariance of X and reject, we can also reject H_0. We consider both one-sided and two-sided alternatives $H_1 : \mu \neq 0$ and $H_1^+ : \mu > 0$, for which we use the test statistics $X \mapsto |\iota' X|$ and $X \mapsto \iota' X$, respectively.

This model is more general than it may seem at first glance. In particular, choosing $\iota = n^{-1/2}(1, 1, \ldots, 1)'$ yields a 'standard' location model. But a different choice of ι can, for example, be used to model two-group comparisons of means: see Appendix A. In addition, the applicability of our results is not limited to the generalized location model. Indeed, many complex testing problems can be approximated by testing a location, such as inference in generalized linear models with nuisance covariates. There, a recent approach is to first compute residuals or score contributions and apply sign-flipping to those (De Santis et al., 2022; Hemerik et al., 2020, 2021). Such approaches are asymptotically exact. Thus, $\iota\mu + \varepsilon$ does not necessarily need to represent the 'full' data distribution. It can also represent a vector of test statistics as in Andreella et al. (2023), or a vector of residuals as in e.g. Winkler et al. (2014), or a vector of score contributions as in e.g. Hemerik et al. (2020).

4·2. *The leak*

The one-sided group invariance test compares $\iota' X$ to the α upper-quantile of the distribution of

$$\iota'\overline{G}X = \iota'\overline{G}\iota\mu + \iota'\overline{G}\varepsilon,$$

for fixed X. This distribution depends on μ through the term $\iota'\overline{G}\iota\mu$, which drops out under the null hypothesis, where $\mu = 0$. Hence, this term can be viewed as a distortion of the distribution under the null. This term will play an important role in the remainder, and we follow an earlier version of Dobriban (2022) in referring to it as a "leak" of signal into noise.

To study the impact of this leak, we consider the following quantification of its magnitude, for a given collection $\mathcal{S} \neq \{I\}$ of orthonormal matrices:

$$\delta_{\mathcal{S}} = \sup_{S \in \mathcal{S} \setminus \{I\}} \iota' S \iota,$$

which can be interpreted as a quantification of the degree to which the reference distribution is different from the reference distribution under the null. Note that we define this for an arbitrary collection of orthonormal matrices, as it will sometimes be useful to quantify the leak of a subset that is not a group.

As an example, if each $S \in \mathcal{S} \setminus \{I\}$ is a diagonal matrix with $n/2$ diagonal entries equal to 1 and $n/2$ diagonal entries equal to -1, and $\iota = n^{-1/2}(1, 1, \ldots, 1)'$, then $\delta_{\mathcal{S}} = 0$. Subgroups of this form are studied in Section 7.

4·3. *Consistency and the leak*

In this section, we describe conditions for non-asymptotic consistency of subgroup invariance tests. In particular, Theorem 3 shows that the magnitude of the leak has a negative impact on the power of the test. Specifically, conditional on $\|\varepsilon\|_2$, a test based on a subgroup $\mathcal{S}$ with a larger

leak $\delta_{\mathcal{S}}$ requires a larger signal-to-noise ratio $\mu/\|\varepsilon\|_2$ to be consistent. As a consequence, we would like to select a subgroup $\mathcal{S}$ for which $\delta_{\mathcal{S}}$ is minimized.

THEOREM 3. *Let ε be $\mathcal{G}$-invariant and let $\mathcal{S}$ be a finite subset of $\mathcal{G}$. If $\alpha \geq 1/|\mathcal{S}|$ and $\mu(1-\delta_{\mathcal{S}})^{1/2} > 2^{1/2}\|\varepsilon\|_2$, then $\mathbb{E}_{\overline{G}}\phi_\alpha^{\mathcal{S}}(X_{\overline{G}}) = 1$. If $\alpha = 1/|\mathcal{S}|$, ε is $\mathcal{H}$-invariant and $\delta_{\mathcal{S}} < 1$, then $\mu(1-\delta_{\mathcal{S}})^{1/2} \geq 2^{1/2}\|\varepsilon\|_2$ if and only if $\mathbb{E}_{\overline{H}}\phi_\alpha^{\mathcal{S}}(X_{\overline{H}}) = 1$.*

The second claim shows that the result is 'sharp', in the sense that there exists an $\alpha \geq 1/|\mathcal{S}|$ and a group $\mathcal{G}$, namely the orthogonal group $\mathcal{H}$, such that there is equivalence. In Remark 6 in Appendix D we further discuss the sharpness of the first claim. There, we also provide a 'sharper' claim that depends on $\mathcal{G}$, but which is unfortunately harder to interpret as it no longer explicitly depends on the $\delta_{\mathcal{S}}$. Furthermore, in Remark 5 in Appendix D, we also include a technical discussion of the claims of Theorem 3, which also explains why the conditioning on $\|\varepsilon\|_2$ is quite natural in this setting.

Interestingly, Theorem 3 does not require $\mathcal{S}$ to be a subgroup of $\mathcal{G}$. However, if $\mathcal{S}$ is not a subgroup of $\mathcal{G}$, we would generally not expect $\phi_\alpha^{\mathcal{S}}$ to control size.

A result about asymptotic consistency is easily obtained from Theorem 3.

COROLLARY 2. *Let $X_{\overline{G}_m}^m$, ι_m, μ_m, ε_m, $\mathcal{S}_m$, α_m, $\mathcal{G}_m$, $\overline{G}_m$ be sequences of the corresponding objects in Theorem 3 indexed by $m \in \mathbb{N}$. Suppose $\alpha_m \geq 1/|\mathcal{S}_m|$ and $\frac{\mu_m}{\|\varepsilon_m\|_2}(1-\delta_{\mathcal{S}_m})^{1/2} \to c \in (2^{1/2}, \infty]$, as $m \to \infty$. Then, $\mathbb{E}_{\overline{G}_m}\phi_{\alpha_m}^{\mathcal{S}_m}(X_{\overline{G}_m}^m) \to 1$.*

For the two-sided test, an analogous result holds. Here, we can define a two-sided version of the leak $\delta_{\mathcal{S}}^{\text{abs}} = \sup_{S\in\mathcal{S}\setminus\{I\}}|\iota' S \iota|$. Denoting the two-sided analogue of $\phi_\alpha^{\mathcal{S}}$ by $\overline{\phi}_\alpha^{\mathcal{S}}$.

THEOREM 4. *Let ε be $\mathcal{G}$-invariant and $\mathcal{S}$ be a finite subset of $\mathcal{G}$. If $\alpha \geq 1/|\mathcal{S}|$ and $|\mu|(1-\delta_{\mathcal{S}}^{abs})^{1/2} > 2^{1/2}\|\varepsilon\|_2$, then $\mathbb{E}_{\overline{G}}\overline{\phi}_\alpha^{\mathcal{S}}(X_{\overline{G}}) = 1$. If $\alpha = 1/|\mathcal{S}|$, ε is $\mathcal{H}$-invariant and $\delta_{\mathcal{S}}^{abs} < 1$, then $|\mu|(1-\delta_{\mathcal{S}}^{abs})^{1/2} \geq 2^{1/2}\|\varepsilon\|_2$ if and only if $\mathbb{E}_{\overline{H}}\overline{\phi}_\alpha^{\mathcal{S}}(X_{\overline{H}}) = 1$.*

5. GENERALIZED LOCATION MODEL: COMPARING THE POWER OF SUBGROUP AND MONTE CARLO TESTS

5·1. *Comparison based on consistency*

In this section, we study and compare the power properties of the subgroup invariance and Monte Carlo group invariance tests. We first compare the (non-asymptotic) consistency of the tests, through a consistency result for Monte Carlo group invariance tests that is analogous to Theorem 3.

To compare Monte Carlo group invariance tests and subgroup invariance tests, we derive an analogue of the consistency result in Theorem 3 for Monte Carlo tests. A similar analogue can be constructed for the two-sided Theorem 4.

Theorem 5 shows that the magnitude of the leak $\delta_{\mathcal{G}}$ of the group $\mathcal{G}$ from which the Monte Carlo sample $\mathcal{G}_M$ is taken, determines whether the Monte Carlo test is consistent.

THEOREM 5. *Let ε be $\mathcal{G}$-invariant and $\|\varepsilon\|_2 > 0$. Let the set $\mathcal{G}_M$ consist of $M-1$ uniformly drawn random variables from $\mathcal{G}$ without replacement, and the identity. If $\alpha \geq 1/M$ and $\mu(1-\delta_{\mathcal{G}})^{1/2} > 2^{1/2}\|\varepsilon\|_2$, then $\mathbb{E}_{\mathcal{G}_M}\mathbb{E}_{\overline{G}}\phi_\alpha^{\mathcal{G}_M}(X_{\overline{G}}) = 1$. If $\alpha = 1/M$ and ε is $\mathcal{H}$-invariant and $\delta_{\mathcal{G}} < 1$, then $\mu(1-\delta_{\mathcal{G}})^{1/2} \geq 2^{1/2}\|\varepsilon\|_2$ if and only if $\mathbb{E}_{\mathcal{G}_M}\mathbb{E}_{\overline{H}}\phi_\alpha^{\mathcal{G}_M}(X_{\overline{H}}) = 1$.*

For any subgroup $\mathcal{S}$ of $\mathcal{G}$, we have $\delta_{\mathcal{S}} \leq \delta_{\mathcal{G}}$. Comparing Theorem 3 and 5 then tells us that a subgroup invariance test is consistent for a smaller signal-to-noise ratio than a Monte Carlo

group invariance test based on $|\mathcal{S}|$ draws from $\mathcal{G}$. A precise statement of this claim is captured in Proposition 1. As a consequence, we find that the subgroup based tests are more powerful than their Monte Carlo counterpart, as measured in terms of consistency.

PROPOSITION 1. *Suppose ε is invariant under the orthogonal group $\mathcal{H}$, $\|\varepsilon\|_2 > 0$, and $\mathcal{S}$ is a subgroup of $\mathcal{G}$ with $|\mathcal{S}| = M$. Then, $\phi^{\mathcal{S}}_{1/M}$ is consistent for a strictly smaller value of the signal-to-noise ratio $\mu/\|\varepsilon\|_2$ than $\phi^{\mathcal{G}_M}_{1/M}$ if and only if $\delta_{\mathcal{S}} < \delta_{\mathcal{G}}$. Moreover, both tests control size.*

In Appendix D·6, we include a plot of the power of a subgroup invariance test based on a subgroup $\mathcal{S}$ with $\delta_{\mathcal{S}} = 0$, and a Monte Carlo group invariance test based on M draws from $\mathcal{H}$, for $\mathcal{H}$-invariant ε.

5·2. *Power comparison under normality*

With the results discussed in the previous section, we obtain a power comparison in terms of consistency. To understand the power beyond consistency, we compare the two tests under normality. We connect Monte Carlo group invariance tests to the t-test and, under normality, subgroup invariance tests to the Z-test. This yields clean results about power of the tests, and allows for a simple comparison in the normal location model.

The following result shows that the t-test can be interpreted as a group invariance test.

THEOREM 6. *The t-test is the orthogonal group invariance test. That is, let $\widehat{\sigma} = \{X'(I - \iota\iota')X/(n-1)\}^{1/2}$, then*

$$\phi^{\mathcal{H}}_{\alpha}(X) = \mathbb{I}\{\iota' X/\widehat{\sigma} > t^{\alpha}_{n-1}\},$$

where t^{α}_{n-1} denotes the α upper-quantile of the t-distribution with $(n-1)$ degrees of freedom.

Remark 1. This result does not depend on the distribution of X: it holds conditional on X. So we do not assume X is normally distributed, unlike in Theorem 7.

We do not believe Theorem 6 is novel. However, to our surprise, we were unable to find the result in the literature or textbooks in this form, although several strongly related results exist: see Chmielewski (1981) and the final paragraph of Lehmann & Stein (1949). For example, the result does not appear in Lehmann & Romano (2022), who extensively discuss group invariance tests and their relationship to the t-test (see Chapter 15.2 and in particular Example 15.2.4). The result is straightforward to generalize to F-tests to test parameters of higher dimension.

The proof of Theorem 6 can be modified to obtain an analogous result for Monte Carlo group invariance tests.

COROLLARY 3. *The test $\phi^{\mathcal{H}_M}_{\alpha}$ has the same size and power as an M-draw Monte Carlo t_{n-1}-test.*

For subgroups with $\delta^{\text{abs}}_{\mathcal{S}} = 0$, we establish a similar connection to the Z-test.

THEOREM 7. *Let $\mathcal{S}$ be a subgroup of $\mathcal{H}$ with $\delta^{abs}_{\mathcal{S}} = 0$. Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, where $\sigma^2 > 0$. Then, an $\mathcal{S}$-invariance test has the same size and power as an $|\mathcal{S}|$-Monte Carlo Z-test.*

Intuitively, the result states that if ε happens to have distribution $\varepsilon \sim N(0, \sigma^2 I)$, $\sigma^2 > 0$, unbeknownst to the analyst, then the $\mathcal{S}$-invariance test has the same size and power as a Monte Carlo Z-test. The subgroup therefore essentially allows the analyst to sample from the unknown null distribution. For this reason, we henceforth refer to subgroups $\mathcal{S}$ with $\delta^{\text{abs}}_{\mathcal{S}} = 0$ as *oracle* subgroups. In Section 6·2 we discuss the existence and order of these subgroups.

Comparing Theorem 7 to Corollary 3, we conclude that oracle subgroup invariance tests are more powerful than MC group invariance tests in a normal location model, as captured in Corollary 4.

COROLLARY 4. *Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma^2 > 0$ and $\mathcal{S} \subset \mathcal{H}$ be a subgroup for which $\delta_{\mathcal{S}}^{abs} = 0$. Then, an $\mathcal{S}$-invariance test is more powerful than an $|\mathcal{S}|$-Monte Carlo $\mathcal{H}$-invariance test.*

6. GENERALIZED LOCATION MODEL: CHOOSING A SUBGROUP AND GROUP CODES

6·1. *Group codes*

In this section, we consider some important properties of subgroups $\mathcal{S}$ of $\mathcal{H}$ for which $\delta_{\mathcal{S}}$ is 'small'. We connect the problem of finding a subgroup for which $\delta_{\mathcal{S}}$ is small to spherical code problems that relate to packing points on a sphere.

First, notice that $-1 \leq \delta_{\mathcal{S}} \leq 1$ for every $\mathcal{S}$, so that it makes sense to focus on subgroups for which $\delta_{\mathcal{S}} < 1$. Let us write $\mathcal{S}\iota = \{S\iota \mid S \in \mathcal{S}\}$. For such subgroups, we obtain the following result.

PROPOSITION 2. *If $\delta_{\mathcal{S}} < 1$, then $\mathcal{S}$ is finite and the map $\mathcal{S} \mapsto \mathcal{S}\iota$ is bijective.*

This result allows us to store a subgroup in a matrix $\mathfrak{S}$ with columns $S\iota$, $S \in \mathcal{S}$. These columns can be interpreted as the rotations of ι by elements of $\mathcal{S}$. This matrix form of the subgroup yields the following representation of $\delta_{\mathcal{S}}$.

PROPOSITION 3. $\delta_{\mathcal{S}} = \max_{i \neq j} e_i' \mathfrak{S}' \mathfrak{S} e_j$.

The matrix $\mathfrak{S}'\mathfrak{S}$ contains all the inner-products between the columns of $\mathfrak{S}$, which are all n-dimensional unit vectors and so are located on the unit hypersphere in dimension n. The value $\delta_{\mathcal{S}}$ can therefore be interpreted as the maximum inner-product between two of such points on the unit hypersphere, which has a one-to-one correspondence with the minimum angle, $\arccos(\delta_{\mathcal{S}})$, between any two points. Hence, the problem of minimizing $\delta_{\mathcal{S}}$ is equivalent to finding a group that induces points on a hypersphere that are as far away from each other as possible. Such a collection of points on a hypersphere is also known as a group code (Slepian, 1968; Conway & Sloane, 1998). The problem of finding M, say, such points on a hypersphere such that $\delta_{\mathcal{S}}$ is minimized is a group-restricted version of the so-called Tammes problem.

It is well-established that a moderately large value of M often yields a small minimum value of $\delta_{\mathcal{S}}$. For example, Sloane et al. (1996) lists a group code for $n = 16$ and $M = 256$ that is induced by a subgroup of the sign-flipping group $\mathcal{R}$, for which $\delta_{\mathcal{S}} = .25$. As a comparison, we find an average 'δ' $\approx .68$ from 10^5 random subsets of size M from $\mathcal{R}$. This suggests that a carefully chosen subgroup should typically be able to yield a much smaller leak than a Monte Carlo draw.

Remark 2. Unfortunately, we were unable to find suitable existing algorithms to construct 'good' group codes of order M, nor could we find libraries that contain them. For example, Sloane et al. (1996) lists spherical codes up to $n = 24$ for few values of M, only some of which are group codes. In addition, we may not be satisfied with any 'good' group code: the group code needs to be induced by a subgroup of the group under which invariance is assumed. Therefore, we include a simple algorithm in Appendix B for the case of the sign-flipping group.

Remark 3. Notice that $\delta_{\mathcal{S}}^{\text{abs}}$ and $\delta_{\mathcal{S}}$ are connected through the identity $\delta_{\mathcal{S}}^{\text{abs}} = \delta_{\mathcal{S} \cup (-\mathcal{S})}$, where $-\mathcal{S} = \{-S \mid S \in \mathcal{S}\}$. Hence, we can also view $\delta_{\mathcal{S}}^{\text{abs}}$ as a minumum angle between points on a hypersphere, noting that each point is accompanied by a twin on the other side of the hypersphere.

Alternatively, $\delta_{\mathcal{S}}^{\text{abs}}$ can be interpreted as a minimum angle between $|\mathcal{S}|$ lines that pass the origin and an element in $\mathcal{S}\iota$.

6·2. *Oracle subgroups*

An important special case are subgroups for which $\delta_{\mathcal{S}}^{\text{abs}} = 0$. Such subgroups are 'optimal' for two-sided testing in the sense of Theorem 4. We will refer to such subgroups as oracle subgroups, due to their power properties discussed in Section 5·2. From Propositions 2 and 3, we know that such subgroups are represented by n-row orthonormal matrices: all the off-diagonals of $\mathfrak{S}'\mathfrak{S}$ are zero. Combining this with the Gram-Schmidt Theorem proves Proposition 4.

PROPOSITION 4. *The maximum order of an oracle subgroup $\mathcal{S}$ is n.*

Furthermore, we include a result about the existence of oracle subgroups.

PROPOSITION 5. *There exists an oracle subgroup $\mathcal{S}$ of the orthogonal group $\mathcal{H}$ with respect to any unit vector ι, of any order p, with $1 \leq p \leq n$.*

Unfortunately, even for quite 'large' groups $\mathcal{G}$, there often exist values $p \leq n$ such that $\mathcal{G}$ has no oracle subgroups of order p. In addition, the existence of oracle subgroups of $\mathcal{G}$ depends intimately on the choice of ι. This will be seen in Section 7, where we characterize oracle subgroups of the sign-flipping group for $\iota = n^{-1/2}(1, 1, \ldots, 1)'$.

Remark 4. For one-sided testing, we can sometimes do slightly better. In particular, it is sometimes possible to find subgroups $\mathcal{S}$ for which $\iota' S \iota \leq 0$ for all $S \in \mathcal{S} \setminus \{S\}$ and $\iota' S \iota < 0$ for some $S \in \mathcal{S}$. Such subgroups still have a leak magnitude of $\delta_{\mathcal{S}} = 0$; see Example 2. They seem to perform slightly better for one-sided testing as seen in Section 8.

7. EXAMPLE: SIGN-FLIPPING

7·1. *Sign-flipping group*

Up to this point, we have only discussed subgroups in an abstract sense. In this section, we provide some examples by considering the sign-flipping group $\mathcal{R}$, which can be represented by all diagonal matrices with diagonal elements in $\{-1, 1\}$ under matrix multiplication. In addition, we choose $\iota = n^{-1/2}(1, 1, \ldots, 1)'$.

If ε is $\mathcal{R}$-invariant and we additionally assume that its elements are independent, then the elements of ε are marginally symmetrically distributed about the origin. The resulting test is often used in paired data, as was already proposed by Fisher (1935). There, the idea is to sign-flip differences between paired observations. Sign-flipping is also widely used in brain image analyses, see Appendix C. For additional discussions and applications, see Efron (1969); Bekker & Lawford (2008); Davidson & Flachaire (2008); Winkler et al. (2014); Andreella et al. (2023).

In order to study the leak of subgroups of $\mathcal{R}$, it is convenient to use its matrix form $\mathfrak{R} = (\iota, R_1\iota, R_2\iota, \ldots),\ R_1, R_2, \cdots \in \mathcal{R}$. The columns of $n^{1/2}\mathfrak{R}$ are the diagonals of sign-flipping matrices in $\mathcal{R}$. The same holds for a subgroup $\mathcal{S} \subset \mathcal{R}$ and its analogous matrix form $\mathfrak{S}$. In fact, the group structure is preserved by considering element-wise multiplication of the columns of $n^{1/2}\mathfrak{S}$. So, the matrix $n^{1/2}\mathfrak{S}$ fully describes the subgroup.

Taking the product between $n^{1/2}\mathfrak{S}$ and $n^{1/2}\iota$, and recalling the geometric interpretation from Section 6, the elements of the resulting vector $n\iota'\mathfrak{S}$ are inversely proportional to the cosine of the angles between ι and the columns of $\mathfrak{S}$. At the same time, notice that the distribution of the leak $n\iota'\overline{S}\iota$, where $\overline{S}$ is uniformly distributed on $\mathcal{S}$, coincides with the empirical distribution over this vector $n\iota'\mathfrak{S}$. We refer to this distribution as the 'leak distribution'. An oracle subgroup is the

special case where this leak distribution concentrates entirely on 0, except for the n^{-1} mass on n that is due to the identity element.

Example 2. If $n = 2$, then

$$n^{1/2}\mathfrak{R} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix},$$

where each column encodes an element of the group. The leak distribution of $\mathcal{R}$ is then the empirical distribution over

$$n\iota'\mathfrak{R} = \begin{bmatrix} 2 & 0 & 0 & -2 \end{bmatrix},$$

which assigns .25 mass to both 2 and -2, and .5 mass to 0. Notice that if $n = 2$, then the one-sided magnitude of the leak is $\delta_{\mathcal{R}} = 0$. However, the two-sided magnitude is $\delta^{\text{abs}}_{\mathcal{R}} = 2$, so that $\mathcal{R}$ is not an 'oracle' subgroup.

Example 3. If $n = 2$, then an example of an oracle subgroup $\mathcal{S}$ of $\mathcal{R}$ is

$$n^{1/2}\mathfrak{S} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The leak distribution is uniform on

$$n\iota'\mathfrak{S} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

Here, both $\delta_{\mathcal{S}} = 0$ and $\delta^{\text{abs}}_{\mathcal{S}} = 0$.

7·2. *Subgroups of the sign-flipping group*

In this section, we describe some (well-known) properties of the sign-flipping group, as well as the induced leak distributions. As $\mathcal{R}$ is isomorphic to a boolean group, its subgroups are of order 2^p, for some $p \leq n$, $p \in \mathbb{N}$, where p the called the rank of the subgroup. The subgroups of the sign-flipping group are abundant, even if n is small. The number of subgroups of rank p is equal to the pth element of the nth row of the 2-binomial coefficient triangle listed at `http://oeis.org/A022166`. The total number of subgroups is equal to the sum of the nth row of this triangle, which can be found at `http://oeis.org/A006116`. This means that if $n = 9$, say, then we have 3309747 subgroups of rank $p = 4$, and 8283458 subgroups in total.

While the number of different subgroups is large, many of them yield the same vector $n\iota'\mathfrak{S}$. This means that the leak, $\iota'\overline{S}\iota$, will have the same distribution regardless of which of these subgroups $\overline{S}$ is uniformly distributed upon. As a consequence, to find a good subgroup, we only need to search over subgroups that yield different leak distributions. The number of different leak distributions corresponding to a subgroup of rank p is equal to the pth element of the nth row of the triangle listed at `http://oeis.org/A076831`. The total number of different distributions is equal to the sum of the nth row of this triangle, which can be found at `http://oeis.org/A076766`. For example, when $n = 9$ there are 240 different leak distributions corresponding to subgroups of rank $p = 4$, and a total of 848 different leak distributions. This is a substantial reduction compared to the number of subgroups.

Example 4. For $n = 4$ and subgroups of order 4 (so $p = 2$), there exist 6 different leak distributions, which are illustrated in Figure 1. As we can see in the figure, the leak distributions are quite diverse. The fifth image corresponds to an oracle subgroup: except for the identity element, all mass is at 0 so that $\delta^{\text{abs}}_{\mathcal{S}} = 0$. The second, third and fifth image all have $\delta_{\mathcal{S}} = 0$.

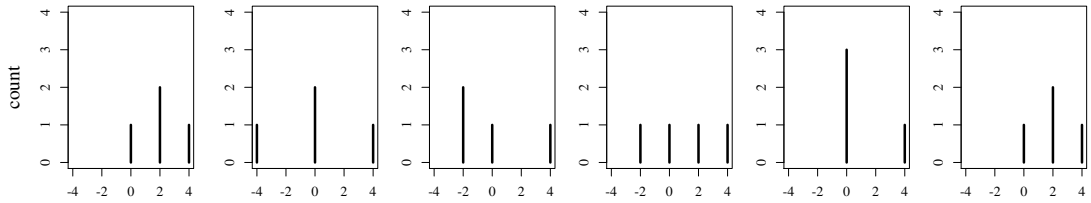

Fig. 1: Histograms of leak distributions for all subgroups of $\mathcal{R}$ of order 4 for $n = 4$. The fifth histogram corresponds to an oracle subgroup, as all its mass it at zero, except for the mass at n produced by the identity element. Notice that the leak distributions are quite diverse.

7·3. *Oracle and near-oracle subgroups*

The structure of the sign-flipping group $\mathcal{R}$ and choice $\iota = (n^{-1/2}, n^{-1/2}, \dots)$ allow us to easily characterize the order of the oracle subgroups. See the Appendix D for a constructive proof.

THEOREM 8. *Let $k \leq l$, $k \in \mathbb{N}_+$, where l is the number of 2s in the prime factorization of n. Then $\mathcal{R}$ has an oracle subgroup with respect to $\iota = (n^{-1/2}, n^{-1/2}, \dots)$ of order 2^k. Furthermore, if $\mathcal{S}$ is an oracle subgroup of $\mathcal{R}$, then it is of order 2^k for some $k \in \mathbb{N}_+$.*

Theorem 8 implies that the number of 2's in the prime factorization of n determines the maximum cardinality of its oracle subgroups. In particular if $n = 2^k$, for some $k \in \mathbb{N}$, then there exists an oracle subgroup of order n, which is maximal by Proposition 4. However, if n is an odd number, then the only oracle subgroup of $\mathcal{R}$ that exists is the trivial subgroup containing only the identity element. In Appendix B we include a simple algorithm to compute subgroups $\mathcal{S}$ of $\mathcal{R}$ for which $\delta_{\mathcal{S}}$ or $\delta^{\text{abs}}_{\mathcal{S}}$ is small. We refer to such subgroups as near-oracle subgroups.

8. SIMULATION RESULTS

8·1. *Testing a single hypothesis*

In this section, we first present simulation results for testing a single hypothesis, in order to empirically support our theoretical findings. In the subsequent section, we demonstrate that in high-dimensional multiple testing settings, our approach can lead to large power improvements.

We simulated data X using the standard normal location model

$$X = \iota\mu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0_n, I_n),$$

with $\iota = (n^{-1/2}, n^{-1/2}, \dots)'$. We tested the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu > 0$ using sign-flipping invariance, which holds for the standard normal location model.

We now describe all the tests we considered, where the abbreviation between brackets corresponds to the column names in the simulation tables. First, we include two benchmark tests to provide upper bounds on the power. The benchmark tests are: a t-test, which exploits knowledge about the orthogonal invariance of the distribution of ε (t); a Monte Carlo sign-flipping test based on 1000 draws, which serves as a proxy for the test that uses the full group (MC $\mathcal{R}$).

Next, we include tests based on $M = n$ draws: an oracle subgroup invariance test; (MC Z) an n-draw Monte Carlo Z-test (Oracle); an order n subgroup invariance test based on a subgroup with $\delta_{\mathcal{S}} = 0$ and $\iota' S \iota < 0$ for some $S \in \mathcal{S}$, as discussed in Remark 4 (Neg.); an n-draw Monte

Carlo $\mathcal{R}$-invariance test (MC $\mathcal{R}$). Furthermore, we consider tests based on $M = 2n$ draws: an order $2n$ subgroup invariance test based on a subgroup with $\delta_{\mathcal{S}} = 0$ and $\iota' S \iota < 0$ for some $S \in \mathcal{S}$ (Neg.); a $2n$-Monte Carlo $\mathcal{R}$-invariance test (MC $\mathcal{R}$). Finally, we include tests based on $M = 4n$ draws: a subgroup invariance test based on a subgroup for which $\delta_{\mathcal{S}}$ is 'small', produced by Algorithm 1 described in Appendix B (NOS); a $4n$-Monte Carlo $\mathcal{R}$-invariance test (MC $\mathcal{R}$).

We considered $n \in \{32, 64\}$. Although the methodology is not limited to a specific sample size, we chose powers of 2 to ensure the existence of oracle subgroups of $\mathcal{R}$ of order n, as stated in Theorem 8. We chose the level α such that αn is integer and $\alpha \approx .05$. This ensures all tests considered have size α (Hemerik & Goeman, 2018a). The parameter μ was chosen such that the power is sufficiently far away from both α and 1.

In tables 1 and 2, we present the power gaps compared to the 1000-draw Monte Carlo $\mathcal{R}$ test, which we treat as a proxy for the group invariance test based on the full group. The tests are grouped by the number of elements in the random subset or subgroup, both indicated by M. All values in the tables are based on 10^6 repeated simulations.

Our findings are as follows. As expected, the rejection proportion under the null hypothesis is approximately α for all tests, as all tests are exact. The benchmark t-test outperforms the other tests. This is not surprising as it makes explicit use of orthogonal invariance, which the other tests do not have access to. The 1000-Monte Carlo $\mathcal{R}$ tests are less powerful than the t-test. In line with Theorem 7, the oracle subgroup invariance tests have the same power as the Monte Carlo Z-tests, and are outperformed by the 'negative' subgroup tests, which are based on a subgroup with $\delta_{\mathcal{S}} = 0$ and $\iota' S \iota < 0$ for some $S \in \mathcal{S}$. Further, when comparing the subgroup invariance tests to the Monte Carlo group invariance tests, we found that the subgroup invariance tests have a substantially smaller power gap with the tests based on the whole group.

		$M = 1000$	$M = n = 32$				$M = 2n = 64$		$M = 4n = 128$	
μ	t	MC $\mathcal{R}$	Oracle	MC Z	Neg.	MC $\mathcal{R}$	Neg.	MC $\mathcal{R}$	NOS	MC $\mathcal{R}$
.0	62	63	0	0	0	0	1	1	1	0
.3	564	552	27	27	24	41	6	21	3	11
.4	767	753	30	30	27	47	7	24	3	12
.5	902	892	23	23	21	39	5	18	2	9

Table 1: Power comparison of several tests: The second and third columns report the power in thousandths for the t-test and the test based on a large number of Monte Carlo draws. The remaining columns report the power gap in thousandths, compared to the test based on a large number of Monte Carlo draws. $n = 32$. 10^6 simulations. $\alpha = 2/32 = .0625$.

8·2. *Testing multiple hypotheses*

In the previous subsection, we saw that our approach led to improvements in efficiency that were substantial but not huge. The primary reason is that, while the leak's impact on the reference distribution grows with μ, the testing problem simultaneously becomes easier as μ increases. Consequently, if μ is set sufficiently small such that the power is bounded away from 1, the impact of the leak is small in the single hypothesis testing problem. To demonstrate the practical implications of subgroup-based tests, we therefore consider a related multiple testing problem that remains difficult for much larger values of μ. We use the single-step max$T$ method by Westfall and Young, a popular permutation-based multiple testing procedure (Westfall & Young, 1993; Meinshausen et al., 2011; Goeman & Solari, 2014; Dickhaus, 2014).

		$M = 1000$	$M = n = 64$				$M = 2n = 128$		$M = 4n = 256$	
μ	t	MC $\mathcal{R}$	Oracle	MC Z	Neg.	MC $\mathcal{R}$	Neg.	MC $\mathcal{R}$	NOS	MC $\mathcal{R}$
.0	47	47	0	0	0	0	0	0	0	0
.2	470	462	15	14	14	22	3	10	1	5
.3	765	756	18	18	17	27	4	13	1	6
.4	936	931	11	11	11	18	2	8	1	3

Table 2: Power comparison of several tests: The second and third columns report the power in thousandths for the t-test and the test based on a large number of Monte Carlo draws. The remaining columns report the power gap in thousandths compared to the test based on a large number of Monte Carlo draws. 10^6 simulations. $n = 64$. $\alpha = 3/64 = .046875$.

Instead of testing one hypothesis about the location of a single n-vector X of data, we now consider testing $k \geq 1$ of such hypotheses simultaneously. In particular, we simulated an $n \times k$ matrix

$$\mathbb{X} = \iota\mu' + E,$$

where μ now represents a k-vector of means, and E has i.i.d. $\mathcal{N}(0,1)$-distributed elements. Our objective is to test the k hypotheses $H_0^j : \mu_j = 0$ against the alternatives $H_1^j : \mu_j > 0$, with $j = 1, \ldots, k$. In this multiple testing setting, the goal is to maximize the power, defined as the expected proportion of correctly rejected null hypotheses, while controlling the familywise error rate by α, defined as the probability to falsely reject at least one hypothesis.

The Max-T method for one-sided tests that we consider is defined as follows. The method rejects H_0^j in favour of H_1^j, if $\mathbb{P}_{\overline{G}}\{T(X_j) > \max_i T(\overline{G}X_i)\} \leq \alpha$, where X_j is the jth column of $\mathbb{X}$. Thus, the procedure rejects H_0^j when $T(X_j)$ exceeds a quantile of a distribution of maxima. We can analogously define Monte Carlo and subgroup tests by replacing $\overline{G}$ with a random variable that is uniformly distributed on an appropriate random subset or on a subgroup, respectively.

We took $n \in \{25, 50, 100\}$ for the number of observations and for the number of hypotheses we used 10^1, 10^2, ..., 10^5. In every setting, we chose μ such that 80% of its elements are zero, and the remaining 20% are equal to some constant $c > 0$. We manually configured c to ensure that the power is in the range 0.4-0.6. We then applied the subgroup-based and Monte Carlo-based Max-T method with $\alpha = .05$, based on a subgroup of order 256 and on 256 random draws, respectively. Moreover, we also include a comparison between subgroups of order 256 and 1000 Monte Carlo draws. The subgroups that were used are the same subgroups as we suggest for testing a single hypothesis, and can be retrieved from the R-package NOSdata `https://github.com/nickwkoning/NOSdata`.

We repeated the experiments 2000 times, and report the observed power differences in Table 3 and Table 4. In these tables, we see that the power difference grows substantially as the number of hypotheses increases. Moreover, the power difference seems to decrease as n increases. In Table 4, the Monte Carlo-based method does slightly outperform the subgroup-based method for the lowest number of hypotheses. We believe this is likely due to the fact that the number of Monte Carlo draws is much larger than the subgroup, while at the same time the number of hypotheses is too small for the benefits of the subgroup method to set in.

To investigate the performance outside of the Gaussian setting, we repeated the same experiments for a uniform distribution on $[-3^{1/2}, 3^{1/2}]$ and a scaled and symmetrized χ^2_2-distribution, constructed by multiplying a χ^2_2-distributed random variable by an independent Rademacher random variable and scaling to ensure it still has variance 1. The estimated power differences we

found were very similar to those in the Gaussian setting, and can be found in Appendix E. We also ran the same experiments as reported in Table 3 using standardized test statistics, which yielded lower power for both tests.

As already pointed out, the rather large power differences can be explained by noting the effect sizes are larger than in the previous simulations with only one hypothesis. As a consequence, the leak of signal into the reference distribution is larger. By using a suitable subgroup, we can substantially reduce the effect of this leak.

	# Hypotheses				
n	10	10^2	10^3	10^4	10^5
25	10	26	61	77	100
50	0	14	29	54	78
100	-1	7	14	23	37

Table 3: The estimated power difference in thousandths between the subgroup and Monte Carlo based Max-T method. For example, "100" indicates that the power of the former test was about 10% higher than that of the latter test. The tests were based on a subgroup and random subset of order 256 respectively. The estimates are based on 2000 repeated simulations.

	# Hypotheses				
n	10	10^2	10^3	10^4	10^5
25	-6	13	46	61	80
50	-9	6	19	41	62
100	-11	-4	5	15	28

Table 4: The estimated power difference in thousandths between the subgroup and Monte Carlo based Max-T method. The tests were based on a subgroup of order 256 and a random subset of size 1000. The estimates are based on 2000 repeated simulations.

9. Discussion

In practice, before our approach can be used, a requirement is that an appropriate subgroup is determined. Fortunately, subgroups can be pre-computed, saved and then used indefinitely. For example, our R package NOSdata (`https://github.com/nickwkoning/NOSdata`) serves as a public source for downloading a 'good' sign-flipping subgroup for $\iota = n^{-1/2}(1, 1, \ldots, 1)'$. This has many important applications, such as testing in linear models (Winkler et al., 2014; Davidson & Flachaire, 2008) and generalized linear models (Hemerik et al., 2020, 2021; De Santis et al., 2022), often combined with permutation-based multiple testing (Andreella et al., 2023; Blain et al., 2022). In addition, in Appendix A, we extend some of our results to two-sample comparisons.

One limitation is that subgroups are not available in all sizes: the subgroups of the sign-flipping group only exist in sizes that are powers of 2. In case one desires to use a specific number of sign-flips, our suggestion is to use a Monte Carlo test based on a sample of that specific size from a sufficiently large subgroup.

A relevant area for future research is the development of general algorithms to construct desired subgroups for testing invariance under other groups and for different ι. Here, it is important

to note that while it may be hard to find the subgroup that yields 'optimal' power properties, it may be quite easy to find a subgroup that yields 'good' power properties. Our consistency and simulation results suggest that such a 'good' subgroup $\mathcal{S}$, for which $\delta_{\mathcal{S}}$ is small but not minimal, is still expected to have good power properties.

An application to testing exchangeability of binary sequences is discussed in the master's thesis by Clemens (2021), which is based on an early version of this paper. In particular, they test exchangeability against an alternative that generates streaks of zeros and ones. They construct subgroups geared towards 'breaking' such streaks, and find some subgroups that can indeed outperform Monte Carlo tests in a simulation study. This illustrates that our approach can be useful in others settings not considered in the present paper.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online consists of five parts: A – E. In Part A, we show how the two-sample comparison problem fits into the generalized location model we discuss in the paper. Moreover, we show how oracle subgroups for this problem can be recovered from oracle subgroups we derived for the sign-flipping setting. In Part B, we discuss an algorithm to find near-oracle subgroups of the sign-flipping group. In Part C, we discuss an application to fMRI data. Part D contains all proofs and some technical remarks that were omitted from the main text. Finally, Part E contains simulation results for our multiple testing experiments based on non-Gaussian distributions.

REFERENCES

ANDERSON, M. J. & ROBINSON, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics* **43**, 75–88.

ANDREELLA, A., HEMERIK, J., FINOS, L., WEEDA, W. & GOEMAN, J. (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine* **42**, 2311–2340.

BEKKER, P. A. & LAWFORD, S. (2008). Symmetry-based inference in an instrumental variable setting. *Journal of Econometrics* **142**, 28–49.

BERRY, K. J., JOHNSTON, J. E. & MIELKE JR, P. W. (2014). A chronicle of permutation statistical methods. *Cham: Springer* .

BLAIN, A., THIRION, B. & NEUVIAL, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage* **260**, 119492.

BLANCHARD, G., NEUVIAL, P. & ROQUAIN, E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics* **48**, 1281–1303.

CHMIELEWSKI, M. (1981). Elliptically symmetric distributions: A review and bibliography. *International Statistical Review/Revue Internationale de Statistique* , 67–74.

CLEMENS, J. (2021). *Enhancing the power of permutation tests for positive serial dependence in binary data by using streak-breaking subgroups*. Master's thesis, Erasmus University Rotterdam.

CONWAY, J. H. & SLOANE, N. J. A. (1998). *Sphere Packings, Lattices and Groups (Third Edition)*. Springer-Verlag, New York.

DARMOIS, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique* , 2–8.

DAVIDSON, R. & FLACHAIRE, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics* **146**, 162–169.

DE SANTIS, R., GOEMAN, J. J., HEMERIK, J. & FINOS, L. (2022). Inference in generalized linear models with robustness to misspecified variances.

DEBEER, D. & STROBL, C. (2020). Conditional permutation importance revisited. *BMC Bioinformatics* **21**, 1–30.

DICKHAUS, T. (2014). *Simultaneous Statistical Inference: with Applications in the Life Sciences*. Springer Science & Business Media.

DOBRIBAN, E. (2022). Consistency of invariance-based randomization tests. *The Annals of Statistics* **50**, 2443 – 2466.

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics* **28**, 181–187.

EATON, M. L. (1989). Group invariance applications in statistics. In *Regional Conference Series in Probability and Statistics*. JSTOR.

EDEN, T. & YATES, F. (1933). On the validity of fisher's z test when applied to an actual example of non-normal data.(with five text-figures.). *The Journal of Agricultural Science* **23**, 6–17.

EFRON, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64**, 1278–1302.

EKLUND, A., NICHOLS, T. E. & KNUTSSON, H. (2016). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences* **113**, 7900–7905.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507–521.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

GAO, X., BECKER, L. C., BECKER, D. M., STARMER, J. D. & PROVINCE, M. A. (2010). Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **34**, 100–105.

GIRARDI, P., VESELY, A., LAKENS, D., ALTOÈ, G., PASTORE, M., CALCAGNÌ, A. & FINOS, L. (2022). Post-selection inference in multiverse analysis (pima): an inferential framework based on the sign flipping score test. *arXiv preprint arXiv:2210.02794* .

GOEMAN, J. J. & SOLARI, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine* **33**, 1946–1978.

GOOD, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses (3rd ed.)*. Springer-Verlag, New York.

HEMERIK, J. & GOEMAN, J. J. (2018a). Exact testing with random permutations. *TEST* **27**, 811–825.

HEMERIK, J. & GOEMAN, J. J. (2018b). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 137–155.

HEMERIK, J. & GOEMAN, J. J. (2021). Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review* **89**, 367–381.

HEMERIK, J., GOEMAN, J. J. & FINOS, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 841–864.

HEMERIK, J., SOLARI, A. & GOEMAN, J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106**, 635–649.

HEMERIK, J., THORESEN, M. & FINOS, L. (2021). Permutation testing in high-dimensional linear models: an empirical investigation. *Journal of Statistical Computation and Simulation* **91**, 897–914.

HOPE, A. C. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society: Series B (Methodological)* **30**, 582–598.

HOTELLING, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)* **15**, 193–232.

KOFLER, R. & SCHLÖTTERER, C. (2012). Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* **28**, 2084–2085.

LANGSRUD, Ø. (2005). Rotation tests. *Statistics and Computing* **15**, 53–60.

LEHMANN, E. & ROMANO, J. P. (2022). *Testing statistical hypotheses*. Springer.

LEHMANN, E. L. & STEIN, C. (1949). On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics* **20**, 28–45.

LI, J. & TIBSHIRANI, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Statistical Methods in Medical Research* **22**, 519–536.

MEINSHAUSEN, N. (2006). False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics* **33**, 227–237.

MEINSHAUSEN, N., MAATHUIS, M. H., BÜHLMANN, P. et al. (2011). Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics* **39**, 3369–3391.

ONGHENA, P. (2018). Randomization tests or permutation tests? A historical and terminological clarification. *Randomization, Masking, and Allocation Concealment* , 209–227.

PESARIN, F. & SALMASO, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. John Wiley & Sons.

PHIPSON, B. & SMYTH, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* **9**, 39.

RAMDAS, A., BARBER, R. F., CANDÈS, E. J. & TIBSHIRANI, R. J. (2023). Permutation tests using arbitrary permutation distributions. *Sankhya A* , 1–22.

SKITOVITCH, V. P. (1953). On a property of the normal distribution. *Doklady Akad. Nauk SSSR (N.S)* **89**, 217–219.

SLEPIAN, D. (1968). Group codes for the gaussian channel. *Bell System Technical Journal* **47**, 575–602.

SLOANE, N. J. A., HARDIN, R., SMITH, W. et al. (1996). Tables of spherical codes. `http://neilsloane.com/packings/`. Accessed: 2021-11-19.

SMEETS, P. A., KROESE, F. M., EVERS, C. & DE RIDDER, D. T. (2013). Allured or alarmed: counteractive control responses to food temptations in the brain. *Behavioural Brain Research* **248**, 41–45.

SOLARI, A., FINOS, L. & GOEMAN, J. J. (2014). Rotation-based multiple testing in the multivariate linear model. *Biometrics* **70**, 954–961.

SOUTHWORTH, L. K., KIM, S. K. & OWEN, A. B. (2009). Properties of balanced permutations. *Journal of Computational Biology* **16**, 625–638.

TUSHER, V. G., TIBSHIRANI, R. & CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116–5121.

VESELY, A., FINOS, L. & GOEMAN, J. J. (2023). Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology* .

WESTFALL, P. H. & TROENDLE, J. F. (2008). Multiple testing with minimal assumptions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50**, 745–755.

WESTFALL, P. H. & YOUNG, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, vol. 279. John Wiley & Sons.

WINKLER, A. M., RIDGWAY, G. R., DOUAUD, G., NICHOLS, T. E. & SMITH, S. M. (2016). Faster permutation inference in brain imaging. *NeuroImage* **141**, 502–516.

WINKLER, A. M., RIDGWAY, G. R., WEBSTER, M. A., SMITH, S. M. & NICHOLS, T. E. (2014). Permutation inference for the general linear model. *Neuroimage* **92**, 381–397.

YOUNG, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics* **134**, 557–598.