Exact and efficient multivariate two-sample tests through adaptive linear multi-rank statistics

Dan Daniel Erdmann-Pham

Department of Statistics, Stanford University
erdpham@stanford.edu

Abstract

So-called linear rank statistics provide a means for distribution-free (even in finite samples), yet highly flexible, two-sample testing in the setting of univariate random variables. Their flexibility derives from a choice of weights that can be adapted to any given (simple) alternative hypothesis to achieve efficiency in case of correct specification of said alternative, while their non-parametric nature guarantees well-calibrated p-values even under misspecification. By drawing connections to (generalized) maximum likelihood estimation, and expanding on recent work on ranks in multiple dimensions, we extend linear rank statistics both to multivariate random variables and composite alternatives. Doing so yields non-parametric, multivariate two-sample tests that mirror efficiency properties of likelihood ratio tests, while remaining robust against model misspecification and computationally tractable. We prove non-parametric versions of the classical Wald and score tests facilitating hypothesis testing in the asymptotic regime, and relate these generalized linear rank statistics to linear spacing statistics enabling exact p-value computations in the small to moderate sample setting. Moreover, viewing rank statistics through the lens of likelihood ratios affords applications beyond fully efficient two-sample testing, of which we demonstrate three: testing in the presence of nuisance alternatives, simultaneous detection of location and scale shifts, and K-sample testing.

Introduction

The task of deciding whether two samples $\mathcal{X}_k = \{X_1, \ldots, X_k\} \overset{iid}{\sim} F$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\} \overset{iid}{\sim} G (\mathcal{X}_k, \mathcal{Y}_n \subset \mathbb{R}^p)$ have arisen from the same ($F = G$) or distinct ($F \neq G$) distributions is known as two-sample testing, and is a staple in statistical analysis [see, e.g., LD75; Tha10; Wei60; JK12, and references therein], with applications ranging from genomics [CY15] to econometrics [GK09] to physics [AZ05] among many. When parametric assumptions on $\mathcal{X}_k$ and $\mathcal{Y}_n$ are imposed; e.g., $F, G \in \mathcal{P}$ for some sufficiently regular, finite-dimensional family of absolutely continuous distributions $\mathcal{P} = \{p_\theta : \theta \in \Theta \subset \mathbb{R}^m\}$, then a test based on the generalized likelihood ratio

$$\lambda_{\mathcal{P}}(\mathcal{X}_k, \mathcal{Y}_n) = \frac{\max_{\theta_1, \theta_2 \in \Theta} p_{\theta_1}(\mathcal{X}_k) p_{\theta_2}(\mathcal{Y}_n)}{\max_{\theta \in \Theta} p_\theta(\mathcal{X}_k) p_\theta(\mathcal{Y}_n)}$$

(1)

displays various optimality properties [Van00; Bah67; ZZM92], and is easily computed when min{$k, n$} is large as $-2\log \lambda_{\mathcal{P}}(\mathcal{X}_k, \mathcal{Y}_n) \to \chi^2_m$ (which we refer to as Wilks phenomenon, following analogous behavior
in the one-sample situation derived in [Wi38]), rendering it the typical method of choice whenever a high-confidence candidate for \( \mathcal{P} \) is available. However, when confidence in \( \mathcal{P} \) is low; that is, the user may expect \( F, G \) to be reasonably close to (in some suitable sense), yet not exactly contained within \( \mathcal{P} \), then care must be taken with tests based on \( \lambda, \mathcal{P} \): it is known that Wilks phenomenon does not persist under model misspecification, and a limiting misspecification-dependent non-central \( \chi^2 \) distribution obtains instead [Whi82]. Therefore, \( p \)-values based on \( \lambda, \mathcal{P} \) are, in general, not well calibrated in this situation. A second difficulty commonly encountered with any maximum likelihood approach lies in its computational complexity: although computing resources are becoming ever more elaborate, so do data sets and their associated models, often rendering the two optimizations involved in (1) challenging. The goal of this article is to propose an alternative to likelihood-ratio type tests that retains the favourable optimality properties of \( \lambda, \mathcal{P} \), while simultaneously being computationally feasible and robust to model misspecification.

A wealth of two-sample tests addressing different aspects of this goal have been developed: in the univariate case \( \mathcal{X}_k, \mathcal{Y}_n \subset \mathbb{R} \), statistics based on the ranks of \( \mathcal{X}_k \) and \( \mathcal{Y}_n \) in the pooled sample \( \mathcal{Z}_{k,n} = \mathcal{X}_k \cup \mathcal{Y}_n \) are distribution-free under \( H_0 \) and therefore automatically robust against misspecification of any kind. Popular examples of such tests include the Anderson-Darling and Cramér-von Mises tests, as well as the Mann-Whitney- \( U \), van-der-Waerden, Siegel-Tukey (equivalent to Bradley-Freund-Ansari), Mood, and Klotz tests [see, e.g., LD75, for overviews and detailed references]. While the former two are based on notions of distance between the empirical distribution functions \( F_k, G_n \) of \( \mathcal{X}_k \) and \( \mathcal{Y}_n \), the latter fall under the umbrella of so-called linear rank statistics, using test statistics of the form

\[
T_k^{w} = \frac{1}{N} \sum_{j=1}^{n} w \left( \frac{r_j}{N} \right) = (1 - \alpha_{k,n}) \int w \circ H_{k,n} \, dG_n,
\]

where \( H_{k,n} = \alpha_{k,n} F_k + (1 - \alpha_{k,n}) G_n \) with \( N = k + n \) and \( \alpha_{k,n} = k/N \), is the empirical distribution of \( \mathcal{Z}_{k,n} \), \( r_j \) is the rank of \( Y_j \) in \( \mathcal{Z}_{k,n} \), and \( w : [0, 1] \to \mathbb{R} \) is a weight function chosen by the user. The shape of \( w \) dictates the alternatives against which \( T_k^{w} \) is powerful, with the Mann-Whitney- \( U \) (\( w(x) = x \)) and van-der-Waerden (\( w = \Phi^{-1} \), \( \Phi \) is the Gaussian CDF) tests designed to detect differences in location (i.e., \( G(x) = F(x - \theta) \)), while Siegel-Tukey (\( w(x) = |x - 1/2| \)), Mood (\( w(x) = (x - 1/2)^2 \)), and Klotz (\( w = \Phi^{-2} \)) seek to differentiate shifts in scale \( G(x) = F(x/\theta) \). Despite their non-parametric nature, linear rank statistics often perform surprisingly well compared to their parametric counterparts: e.g., [HL56] showed that the Pitman efficiency against local contiguous alternatives [for relevant definitions see, e.g., Van00, Chapter 14] of Mann-Whitney’s \( U \) with respect to the \( t \)-test is lower bounded by \( 108/125 \approx 0.864 \) over all choices of \( F \). More strikingly yet, the same efficiency of van-der-Waerden’s test against the \( t \)-test is always at least \( 1 \), as proven in [CS58]; that is, at least asymptotically in the Pitman sense, \( T_k^{w} \) is always preferable to the \( t \)-test. Practical implementations of linear rank tests typically rely on the asymptotic normality of \( T_k^{w} \) to determine rejection thresholds, though [ETS22] has shown that exact \( p \)-values in the finite-sample setting can be obtained for so called linear rank spacings, which are closely related and asymptotically equivalent to using linear ranks [HR80].

The primary difficulty in extending results such as these to the multivariate situation \( \mathcal{X}_k, \mathcal{Y}_n \subset \mathbb{R}^p, p > 1 \) lies in the absence of a canonical ordering, leading to ambiguity of the notion of ranks in higher dimensions. Various extensions have been proposed [Bic65; Cha96; Mar99; ZS00; HP02; HP04; Ran89; PR90], though two-sample tests associated with them typically do not exhibit the same distribution-freeness that their univariate counterparts do, and can pose computational challenges. Multivariate two-sample tests that are not based on generalized ranks generally leverage either geometric graphs [FR79; Hen88; CF17; Ros05] or
embeddings into reproducing kernel Hilbert spaces [Gre+12], of which energy-distances are a special case
[BF04]. The latter usually lack distribution-freeness (even asymptotically), necessitating generic (often
conservative) tail bounds or permutation schemes, while the former are known to have 0 Pitman efficiency in
most settings [Bha19]. Recently, through the theory of optimal transport, [Hal+20] introduced a multivariate
generalization of the rank map that does inherit most desirable properties of its univariate relative, which
[DBS21] used (and expanded) effectively in the design of a multivariate two-sample test that is exactly
distribution-free and Pitman efficient against the Hotelling \( T^2 \) test [the suitable multivariate analogue of the
t-test, Hot92] in the location setting. There, the authors start from the interpretation of Mann-Whitney’s
\( U \) as a "rank version" of the \( t \)-test; that is, \( U = k^{-1} \sum_{j=1}^{k} H_{k,n}(X_j) - n^{-1} \sum_{j=1}^{n} H_{k,n}(Y_j)/n \) compared with
\( i = k^{-1} \sum_{j=1}^{k} X_j - n^{-1} \sum_{j=1}^{n} Y_j \) (in the situation where \( \text{Var} X \) and \( \text{Var} Y \) are known and identical), and proceed
to similarly define a rank version of Hotelling’s \( T^2 \). The multivariate ranks involved in this construction,
however, require computation of optimal transport maps between two samples of size \( N = k + n \), which
generically are expensive to obtain: exact solvers may take \( O(pN^2 + N^3 \log^2 N) \), while even approximate
solutions may necessitate \( O(pN^2 + N^2 \log N) \) [PC19, p. 42, 65]. With \( N \) frequently ranging into the
thousands or millions in, e.g., modern biology, such computation quickly grows intractable. Our first
contribution consists in identifying alternative transport maps that, too, allow for efficient multivariate
two-sample testing, yet can be computed in \( O(pN \log N) \). Moreover, in order to achieve efficiency, [DBS21]
introduce a multivariate analogue \( J : \mathbb{R}^p \to \mathbb{R}^{p \log N} \) of the weight function \( w \) defining \( T^w_{k,n} \), and show that it
performs well (relative to Hotelling’s \( T^2 \)) as long as a so-called effective reference distribution resulting
from the multivariate rank map and \( J \) is Gaussian (mirroring van-der-Waerden’s choice of \( w = \Phi^{-1} \)). A
number of effective reference distributions other than the Gaussian one are investigated, and shown to
exhibit substantial differences in relative efficiency compared to Hotelling’s \( T^2 \), with, e.g., some displaying
significant dependence on \( p \), the dimension of \( \mathcal{X}_k \) and \( \mathcal{Y}_n \), while others do not. In the face of such
heterogeneity, it is natural to ask: do some choices of reference distributions perform better for a given
choice of \( F \), and is there an optimal such choice? Indeed, this question extends beyond the setting of
detecting location shifts, and can be asked in the more general setting of the discussion surrounding (1).
Given that every test has meaningful power against local alternatives only in finitely many directions [see,
e.g., LRC05, Chapter 14.6], providing practitioners with a clear recipe to turn a subspace of alternatives
that they suspect may be relevant for the application at hand into a test powerful along said subspace should
be valuable. By tying the notion of effective reference distributions and weight functions to \( \lambda_{\mathcal{P}} \), we hope to
give some guidance on how such a recipe might look like.

**Univariate linear rank statistics as local likelihood ratios**

Pitman efficiencies measure relative performances of two tests against local alternatives; that is, against
alternatives \( G_n \) converging to \( F \) at rate \( N^{-1/2} \) in some suitable sense. To fix ideas, assume for now that \( \mathcal{P} \)
is univariate, twice continuously differentiable, \( 0 \in \Theta = (a,b) \), \( F = P_0 \), and \( G = P_0/\sqrt{N} \) for some \( \theta \in \mathbb{R} \).
This simple situation serves to illustrate main ideas, with extensions to more general settings and a more
rigorous treatment following in later sections. In this context, the log-likelihood of \( \mathcal{Y}_n \) under \( G \) reads

\[
\log p_{\theta/\sqrt{N}}(\mathcal{Y}_n) \approx \frac{\theta}{\sqrt{N}} z_0(\mathcal{Y}_n) - \frac{\theta^2}{2N} \partial_\theta z_0(\mathcal{Y}_n) + o_G(N^{-1}),
\]
where \( z_0 \) is the score associated with \( \mathcal{P} \) at \( \theta = 0 \), the notation \( f(\mathcal{Y}_n) \) is used to mean \( \sum_{Y \in \mathcal{Y}_n} f(Y) \) (occasionally it may represent \( \prod_{Y \in \mathcal{Y}_n} f(Y) \), which will be clear from context; e.g., when \( f \) represents a density), and the approximation is valid as long as \( N \) is sufficiently large and \( \alpha_{k,n} \to \alpha > 0 \). Heuristically,

\[
H_{k,n}(Y) \approx P_0(Y) + (1 - \alpha_{k,n}) \frac{\theta}{\sqrt{N}} \partial \theta P_0(Y),
\]

and so

\[
\theta T_{k,n}^{\log P_0^{-1}} = \frac{\theta}{\sqrt{N}} z_0 \circ P_0^{-1} \circ H_{k,n}(\mathcal{Y}_n) \approx \frac{\theta}{\sqrt{N}} z_0(\mathcal{Y}_n) + O_G(N^{-1})
\]

can be thought of as a first-order approximation to this log-likelihood, when \( w \) is chosen to be the effective score \( z_0 \circ P_0^{-1} \). For the location and shift families \( \mathcal{P}_\mu = \{ p_0(\cdot - \theta), \theta \in \mathbb{R} \} \) and \( \mathcal{P}_\sigma = \{ \theta^{-1} p_0(\cdot/\theta), \theta \in \mathbb{R}_+ \} \), the weights \( w(x) = x, w(x) = \Phi^{-1}(x) \) and \( w(x) = \Phi^{-2}(x) \) are precisely of this form corresponding to logistic and Gaussian \( p_0 \), respectively. Moreover, given that the sample’s randomness is largely captured in this first-order approximation (\( \partial \theta z_0(\mathcal{Y}_n) / N \) converges to a deterministic limit, a multiple of the Fisher information), it is not surprising that \( T_{k,n}^w \) is fully efficient in these situations (that is; has Pitman efficiency 1 compared to tests based on \( \log \lambda_{\mathcal{P}} \)). This analogy between \( T_{k,n}^w \) and the log-likelihood of \( \mathcal{P} \) naturally raises three questions:

(a) **Can this analogy be extended to the setting of multivariate \( \Theta \)?** When \( \Theta \subset \mathbb{R}^m \) is multivariate, then so is \( z_0 \), and the meaning of \( T_{k,n}^{z_0 \circ P_0^{-1}} \) becomes less clear. \( G_n \) can now approach \( P_0 \) from a variety of directions \( \theta \), for which the likelihood locally has the form

\[
\langle \theta, z_0(\mathcal{Y}_n) \rangle / \sqrt{N} - \langle \theta, \partial \theta z_0(\mathcal{Y}_n) \theta \rangle / N,
\]

and so one may expect that \( z_0 \circ P_0^{-1} \circ H_{k,n}(\mathcal{Y}_n) \) still yields a suitable non-parametric equivalent to \( \lambda_{\mathcal{P}} \) in this situation. What are the appropriate analogues of the score test, Wald test, and likelihood-ratio tests, and to which extent is a non-parametric equivalent of second-order term in (2) (which in the parametric setting is required to keep the log-likelihood bounded locally) needed?

(b) **Can this analogy be extended to the setting of multivariate \( X \) and \( Y \)?** As mentioned previously, classical notions of multivariate ranks \( R \) do not typically mirror most of the desirable properties of univariate ranks. The linear rank statistics discussed above rely on the probability integral transform, or CDF trick; that is, the uniformity of \( P_0(X) \) whenever \( X \sim P_0 \) is univariate, combined with \( P_0^{-1}(A) \sim X \) when \( A \) is uniform on \([0, 1]\). Recent work in [Hal+20] characterized one appropriate extension of the CDF trick to \( \mathbb{R}^m \) by associating the rank map \( R \) with the optimal transport map transferring \( P_0 \) to a suitable reference measure. With such transport map in hand, a natural candidate for a multivariate extension of \( T_{k,n}^w \) is \( z_0 \circ R^{-1} \circ R_{k,n}(\mathcal{Y}_n) \), where \( R_{k,n} \) presents an appropriate empirical version of \( R \). However, as mentioned previously, the cost of obtaining such \( R_{k,n} \) can be prohibitive, and so identifying alternative transport maps of similar properties (relevant to two-sample testing) yet lighter computational burden is desirable. We identify such alternative transport maps, and thus obtain a non-parametric two-sample test that is both statistically and computationally efficient.

(c) **Does the non-parametric nature of \( T_{k,n}^w \) afford any additional flexibility beyond \( \lambda_{\mathcal{P}} \)?** When \( \mathcal{P} \) is misspecified, then tests based on \( \lambda_{\mathcal{P}} \) are in general not well-calibrated [Whi82]. The null distribution
of $T^w_{k,n}$ however, solely depends on $w$ independent of the precise shape of $F (= G$ under $H_0)$, and thus possibly allows querying more complex hypotheses. Here we explore two circumstances in which this added flexibility may be of use:

1) **Testing in the presence of nuisance alternatives.** Often a practitioner might like to devise a test that has power against a parametric family $\mathcal{P}$, while simultaneously ignoring any changes along a second model $\mathcal{Q}$. That is, $\mathcal{Q}$ acts in some sense as a nuisance model, and may, e.g., represent known measurement inaccuracies. For instance, in the context of analyzing microscopic images, differences in the actual depicted samples are meaningful to detect, whereas rotations and translations resulting from sample placement are not. We will see that $T^w_{k,n}$ admits a natural choice of $w$ that provides large power against $\mathcal{P}$, while remaining insensitive to $\mathcal{Q}$.

2) **Combining multiple weights.** Characterizing individual choices of weights that perform well against particular, narrowly defined alternatives is usually feasible; e.g., Mann-Whitney’s and Klotz’ tests are two such instances in the setting of location and scale shifts. A common approach to querying composite hypotheses is to apply multiple such tests in succession, and then correct the resulting multiple testing burden by any of a number of multiple testing corrections [see, e.g., Bon36; BH95; Ben10; BC15; Nob09]. Depending on the correlation structure of the tests, such correction procedures may lack power, or control false discovery rates rather than false rejection rates. We will show that $T^w_{k,n}$ provides a convenient framework for combining multiple individual weights $w_j$ in a manner that does not necessitate multiple testing corrections, while delivering competitive power in each individual direction $w_j$.

The following chapters address each of these questions in turn, with additional sections discussing an extension to $K$-sample testing and exact finite sample results. Software implementing the most general $K$-sample test (and illustrating various of its optimality properties) can be found at [https://web.stanford.edu/~erdham/research.html](https://web.stanford.edu/~erdham/research.html).

**Extension to multivariate parameter spaces**

As indicated above, one might hope that $z_0 \circ P_0^{-1} \circ H_{k,n}$ behaves like the score function under $P_0$ and $G_n = P_\theta / \sqrt{n}$, even when $\mathcal{P}$ is multivariate. This is indeed true.

**Theorem 1.** For $k, n \to \infty, k/N \to \alpha > 0$, and $w \in \mathbb{R}^m$ continuous, square-integrable on $(0, 1)$ and $\int w = 0$, the statistic $T^w_{k,n} = w \circ H_{k,n}(\mathcal{F}_n) / \sqrt{n}$ is asymptotically normal under local contiguous alternatives with

$$
T^w_{k,n} \overset{H_0}{\to} \mathcal{N}(0, \alpha (1 - \alpha) H_{w}) \quad \text{and} \quad T^w_{k,n} \overset{H_1, \mathcal{P}}{\to} \mathcal{N}
\left(
\alpha (1 - \alpha) H_{w, z_0 \circ P_0^{-1} \circ \theta}, \alpha (1 - \alpha) H_{w}
\right),
$$

where $H_{w_1, w_2} = \int w_1 \otimes w_2, H_w = H_{w, w}$, and $H_{1, \mathcal{P}} : X \sim p_0, Y \sim p_\theta / \sqrt{n}$, given that $\mathcal{P}$ is quadratic mean differentiable, $\int \|z_0 \circ P_0^{-1}\|^2$ finite, and $H_{z_0 \circ P_0^{-1}}$ exists and is full-rank.

**Proof.** Set $\mathcal{A}_N = \{w(jN^{-1}), j = 1, \ldots, N\}$, then under $H_0$, $\{w \circ H_{k,n}(Y_j), \ldots, w \circ H_{k,n}(Y_n)\}$ can be obtained by sampling without replacement $n$ times from $\mathcal{A}_N$. Thus, the $H_0$ convergence of $T^w_{k,n}$ is a consequence of the multivariate central limit theorem for simple random sampling [see, e.g., Ros64, where the univariate
version is stated, the multivariate generalization is a direct consequence of applying the Cramér-Wold device], after verifying that
\[
\max_{a \in \mathcal{A}_N} \frac{\|H_w^{-1/2}a\|^2}{N} \to 0
\]
by the continuity and square-integrability assumptions on \(w\). The specific covariance structure is calculated to be
\[
\text{Var} T_{k,n}^w = \frac{1}{N} \sum_{j=1}^{n} \mathbb{E}w(r_j) \otimes w(r_j) + \frac{1}{N} \sum_{m \neq j} \mathbb{E}w(r_m) \otimes w(r_j) \overset{H_0}{\Rightarrow} (1 - \alpha)H_w - (1 - \alpha)^2H_w = \alpha(1 - \alpha)H_w,
\]
as desired.

The \(H_1,\mathcal{A}\) results is shown by way of LeCam’s third lemma. Writing the log-likelihood-ratio statistic
\[
\rho_n = \sum_{j=1}^{n} \log \frac{p_\theta / \sqrt{N}(Y_j)}{p_0(Y_j)}
\]
in its local expansion
\[
\rho_n = \left\langle \frac{z_0(\mathcal{A}_n)}{\sqrt{N}}, \theta \right\rangle - \frac{1}{2} \left\langle \theta, H_{\mathcal{A}_0}^{-1} \circ H_{\mathcal{A}_n}(\mathcal{A}_n) \right\rangle + o_{H_0}(1) = \left\langle \frac{z_0 \circ H_{k,n}^{-1} \circ H_{k,n}(\mathcal{A}_n)}{\sqrt{N}}, \theta \right\rangle - \frac{1}{2} \left\langle \theta, H_{\mathcal{A}_0} \circ p_0 \theta \right\rangle + o_{H_0}(1),
\]
and augmenting \(\mathcal{A}_N\) to
\[
\mathcal{A}_N^+ = \left\{ w \left( \frac{j}{N} \right) + e_{m+1} \otimes \left( z_0 \circ H_{k,n}^{-1} \left( \frac{j}{N} \right), \theta \right) \right\}, j = 1, ..., N \} = \left\{ w \left( \frac{j}{N} \right) + e_{m+1} \otimes (z_0(Z_j), \theta) \right\}
\]
yields the joint vector \((T_{k,n}^w, \rho_n)\) as a size-\(n\) sample without replacement from \(\mathcal{A}_N^+\) up to \(o_{H_0}(1)\) and a deterministic shift. Though \(\mathcal{A}_N^+\) depends on \(\mathcal{A}_N\) (and therefore is random),
\[
\mathbb{P}_{H_0} \left[ \left\{ N^{-1} \sum_{a \in \mathcal{A}_N^+} a \to 0 \right\} \cap \left\{ N^{-1} \sum_{a \in \mathcal{A}_N^+} a^2 \to \Sigma \right\} \right] = 1,
\]
for some to be determined covariance matrix \(\Sigma \in \mathbb{R}^{(m+1) \times (m+1)}\), and so the central limit theorem for simple random sampling applies to show the joint normality of \((T_{k,n}^w, \rho_n)\) under \(H_0\). By LeCam’s third lemma, the shift under \(H_{1,\mathcal{A}}\) is then given by
\[
\mathbb{E}T_{k,n}^w \rho_n = \mathbb{E} \left( \frac{1}{N} \sum_{j=1}^{n} w(r_j) \otimes z_0 \circ H_{k,n}^{-1}(r_j) + \frac{1}{N} \sum_{m \neq j} w(r_m) \otimes z_0 \circ H_{k,n}^{-1}(r_j) \right) \theta = (A_1 + A_2) \theta.
\]
By the law of large numbers for sampling without replacement [Ros64, section 5], \(A_1\) converges to \((1 - \alpha)H_w \circ z_0 \circ p_0^{-1}\), while
\[
A_2 = \frac{1 - \alpha}{n} \mathbb{E} \sum_{m \neq j} \mathbb{E} \left[ w(r_m) \otimes z_0 \circ H_{k,n}^{-1}(r_j) \mid \mathcal{A}_k^n \right] = - \frac{(n - 1)(1 - \alpha)}{N(N - 1)} \mathbb{E} \sum_{s=1}^{N} w(sN^{-1}) \otimes z_0 \circ H_{k,n}^{-1}(sN^{-1}),
\]
and
which converges to \((1 - \alpha)^2 H_{w, z_0 \circ p_0^{-1}}\). Consequently
\[
\mathbb{E} T_{k, n}^w \rho_n \xrightarrow{H_0} (1 - \alpha) H_{w, z_0 \circ p_0^{-1}} - (1 - \alpha)^2 H_{w, z_0 \circ p_0^{-1}} = \alpha (1 - \alpha) H_{w, z_0 \circ p_0^{-1}},
\]
and the theorem is proved. \(\square\)

Theorem 1 allows for hypothesis testing as long as \(k\) and \(n\) are large enough; for small to moderate sample sizes, see the section on exact finite sample distributions. Tests based on \(T_{k, n}^w\) and \(T_{k, n}^{\mu + \sigma w}\), with \(\mu \in \mathbb{R}^m\), \(\sigma \in \mathbb{R}\) are equivalent since \((T_{k, n}^{\mu + \sigma w} - n \mu) / \sigma = T_{k, n}^w\), and so the centering \(f w = 0\) is purely to ease notation and does not impact the generality of Theorem 1. This centering will be assumed implicitly in all following results, as will be the regularity assumptions on \(\mathcal{P}\) (unless stronger conditions are needed), which guarantee local asymptotic normality of \(\rho_n\). Big-N results like these are sufficient to compare tests in the Pitman sense.

**Corollary 1.** The Pitman efficiency of \(T_{k, n}^{z_0 \circ p_0^{-1}}\) relative to the Neyman-Pearson test is \(\alpha\) for local alternatives \(p_\theta / \sqrt{n} \in \mathcal{P}\) converging to \(p_0\).

**Proof.** By local asymptotic normality, the log-likelihood ratio \(\rho_n\) converges to
\[
\rho_n \xrightarrow{H_0} \mathcal{N} \left( -\frac{\sqrt{1 - \alpha}}{2} H_{z_0 \circ p_0^{-1}} \theta, H_{z_0 \circ p_0^{-1}} \right), \quad \rho_n \xrightarrow{H_1, \theta} \mathcal{N} \left( \frac{\sqrt{1 - \alpha}}{2} H_{z_0 \circ p_0^{-1}} \theta, H_{z_0 \circ p_0^{-1}} \right).
\]
The corollary then follows from the \((\mu_1^/, \mu_2^/) (\sigma_2 / \sigma_1)\) characterization of Pitman efficiency [Corollary 15.2.1 of LRC05]. \(\square\)

The rate of \(\alpha\) is, of course, a result of not knowing \(F\) exactly. The fairer comparison is against the generalized likelihood ratio test.

**Corollary 2.** The Pitman efficiency of \(T_{k, n}^{z_0 \circ p_0^{-1}}\) against \(-2 \log \lambda_\mathcal{P}\) is \(1\) for alternatives \(p_\theta / \sqrt{n} \in \mathcal{P}\) converging to \(p_0\), assuming that the maximum-likelihood estimates involved in \(\lambda_\mathcal{P}(\mathcal{X}_k, \mathcal{Y}_n)\) are efficient in the sense of [LRC05, eq. (14.62)].

**Proof.** The regularity assumptions on \(\mathcal{P}\) guarantee that \(-2 \log \lambda_\mathcal{P}(\mathcal{X}_k, \mathcal{Y}_n)\) converges to
\[
-2 \log \lambda_\mathcal{P}(\mathcal{X}_k, \mathcal{Y}_n) \xrightarrow{H_0} \chi^2_m, \quad -2 \log \lambda_\mathcal{P}(\mathcal{X}_k, \mathcal{Y}_n) \xrightarrow{H_1, \theta} \| \mu + Z_m \|^2_2,
\]
where
\[
\mu = \left[ \Pi_{(\sqrt{\alpha}, \sqrt{1 - \alpha})} \otimes 1_m \right] \left[ \begin{pmatrix} \sqrt{\alpha} & 0 \\ 0 & \sqrt{1 - \alpha} \end{pmatrix} \otimes H_{z_0 \circ p_0^{-1}}^{1/2} \right] [e_2 \otimes \theta],
\]
with \(1_m\) the \(m \times m\) identity matrix and \(\Pi_{x^\perp}\) the projection onto \((\text{span} x)^\perp \subset \mathbb{R}^2\) for any \(x \in \mathbb{R}^2\). Consequently,
\[
\mu = \begin{pmatrix} -\sqrt{\alpha}(1 - \alpha) \\ \alpha \sqrt{1 - \alpha} \end{pmatrix} \otimes H_{z_0 \circ p_0^{-1}}^{1/2} \theta,
\]
and the result follows as
\[
\| \mu \|^2_2 = \alpha(1 - \alpha) \| H_{z_0 \circ p_0^{-1}}^{1/2} \theta \|^2_2. \quad \square
\]
Theorem 2. Set \( \hat{\theta} = \arg \max_\theta P_\theta(X_{k,n}), \hat{\omega} = z_{\hat{\theta}} \circ P_{\hat{\theta}}^{-1} \), and assume \( \hat{\theta} \to 0 \) almost surely, then

(a) \( (T_{k,n}^\hat{\omega} | \hat{\theta}) \) is exactly distribution-free under \( H_0 \).

(b) Asymptotically,

\[
T_{k,n}^\hat{\omega} \xrightarrow{H_0} \mathcal{N} \left( 0, \alpha(1 - \alpha)H_{0 \circ \theta_0} \left( 0 \right) \right), \quad T_{k,n}^\hat{\omega} \xrightarrow{H_1 \circ \theta} \mathcal{N} \left( \alpha(1 - \alpha)H_{1 \circ \theta_0} \left( \theta \right), \alpha(1 - \alpha)H_{0 \circ \theta_0} \left( \theta \right) \right),
\]

with notation as in Theorem 1.

(c) The Pitman efficiency of \( T_{k,n}^\hat{\omega} \) compared to \(-2 \log \lambda_{\not{\mathcal{P}}} \) against local alternatives \( p_{\theta^* + \theta / \sqrt{N}} \in \mathcal{P} \) is 1 for any \( \theta^* \in \Theta \).

Proof. (c) follows from (b) and the same computation as in Corollary 2, and so only (a) and (b) remain to be shown.

(a) Denoting by \( \mathbb{E}^X \) the projection onto the sigma-algebra \( \sigma(X) \) generated by \( X \), one has for any continuous, bounded \( f \) on \( \mathbb{R}^m \)

\[
\mathbb{E}_{H_0}^\hat{\theta} f(T_{k,n}^\hat{\omega}) = \mathbb{E}_{H_0}^\hat{\theta} \mathbb{E}_{H_0}^X f(T_{k,n}^\hat{\omega}) = \mathbb{E}_{H_0}^\hat{\theta} \left[ \frac{1}{(N-1)} \sum_{0 < m_1 < \ldots < m_\ell \leq N} f \left( \sum_{j=1}^n \hat{\omega}(m_j / N) \right) \right]
\]

\[
= \frac{1}{(N-1)} \sum_{0 < m_1 < \ldots < m_\ell \leq N} f \left( \sum_{j=1}^n \hat{\omega}(m_j / N) \right),
\]

since \( \hat{\theta} \) is \( \sigma(\mathcal{X}_{k,n}) \)-measurable, and \( \left\{ r_1, \ldots, r_\ell \right\} \mid \mathcal{X}_{k,n} \) is uniform over the \( n \)-subsets of \( \left\{ 1, \ldots, N \right\} \) under \( H_0 \). The right-hand side is independent of \( p_0 \) (except through \( \hat{\omega} \)), therefore proving the claim.

(b) Setting \( \mathcal{A}_N = \left\{ \hat{\omega}(jN^{-1}), j = 1, \ldots, N \right\} \) and

\[
\mathcal{A}_N^+ = \left\{ \hat{\omega} \left( \frac{j}{N} \right) + e_{m+1} \otimes \left\langle z_0 \circ H_{k,n}^{-1} \left( \frac{j}{N} \right), \theta \right\rangle, \right. \left. j = 1, \ldots, N \right\},
\]

the almost sure convergence of \( \hat{\theta} \) combined with the continuous mapping theorem imply that

\[
\mathbb{P} \left[ \mathbb{E}_{\mathcal{A}_N} X \to 0, \text{Var}_{\mathcal{A}_N} A_N \to \Sigma, \mathbb{E}_{\mathcal{A}_N^+} X \to 0, \text{Var}_{\mathcal{A}_N^+} \to \Sigma^+ \right] = 1,
\]

for some fixed choices of \( \Sigma \) and \( \Sigma^+ \), and so the same reasoning as employed in Theorem 1 carries through.

\( \square \)
Thus, similar to $T_{k,n}^\Phi$ being preferable to the $t$-test at least asymptotically in the Pitman sense, $T_{k,n}^\gamma$ recovers performance of any likelihood-ratio test in the same sense, at least for alternatives within $\mathcal{P}$, while remaining non-parametric. Moreover, unlike the generalized likelihood-ratio test, $T_{k,n}^\gamma$ behaves well for alternatives falling outside $\mathcal{P}$.

**Theorem 3.** Given a second model $\mathcal{Q}$ with $\mathcal{Q} \cap \mathcal{P}$ non-empty, local alternatives $q_0 \in \mathcal{Q} \cap \mathcal{P}$ (that is, $X \sim q_0, Y \sim q_1/\sqrt{N}$), and almost sure convergence of $\hat{\theta}$ to $\theta^* = \arg\max_\theta \mathbb{E}_{q_0} \log p_\theta(Y)$, asymptotically

$$
T_{k,n}^\gamma \xrightarrow{H_0} \mathcal{N} \left( 0, \alpha(1 - \alpha) H_{\gamma_0 \circ P_0^{-1}} \right) \quad T_{k,n}^\gamma \xrightarrow{H_1, \mathcal{Q}} \mathcal{N} \left( \alpha(1 - \alpha) K \chi, \alpha(1 - \alpha) H_{\gamma_0 \circ P_0^{-1}} \right),
$$

where $K = \int z_\theta \circ P_0^{-1} \otimes s_0(Y) \circ Q_0^{-1}$, for $s_0$ the score function associated with $\mathcal{Q}$ at $\gamma = 0$.

**Proof.** The convergence proof under $H_0$ proceeds as in part (b) of Theorem 2. The shift under $H_1, \mathcal{Q}$, is again given by

$$
\mathbb{E} T_{k,n}^\gamma \rho_n = \left( \frac{1 - \alpha}{n} \mathbb{E} \sum_{j=1}^n \hat{w}(r_j) \otimes q_0 \circ H_{k,n}^{-1}(r_j) + \frac{1 - \alpha}{n} \mathbb{E} \sum_{m \neq j} \hat{w}(r_m) \otimes s_0 \circ H_{k,n}^{-1}(r_m) \right) \gamma = (A_1 + A_2) \gamma,
$$

with the same arguments as in Theorem 1 showing

$$A_1 \xrightarrow{H_0} (1 - \alpha) H_{\gamma_0 \circ P_0^{-1}, s_0 \circ Q_0} \quad A_2 \xrightarrow{H_0} (1 - \alpha)^2 H_{\gamma_0 \circ P_0^{-1}, s_0 \circ Q_0},$$

as desired. \qed

**Remark 1.** For the location and shift models the respective effective score functions $z_\mu \circ (P^{-1} + \mu) = -p' / p$ and $z_\sigma \circ (\theta P^{-1})(x) = -xp'(x) / p(x)$ are independent of the parameters, and so $T_{k,n}^\gamma = T_{k,n}^\alpha$. In these cases, the distribution of the likelihood-ratio test under alternatives can be computed explicitly, and the efficiencies obtained from Theorem 3 reduce to the well-known efficiencies considered in, e.g., [HL56] and [Klo62].

**Remark 2.** An important computational advantage of likelihood-ratio tests is their associated Wilks’ phenomenon; that is, their limiting $\chi^2_m$-distribution does not depend on any parameters of the model, and in particular does not require estimation of the Fisher information $\mathbb{E}_{p_0} z_0 \otimes z_0$. As $T_{k,n}^\gamma$ is anchored in a Taylor expansion of the log-likelihood, dispensing with the Fisher information altogether is difficult, but of course a similar limiting $\chi^2_m$ distribution can be obtained from Theorems 1 and 2. For $M_{k,n}^w = \max_\theta \langle \theta, T_{k,n}^w \rangle - \| \theta \|_H$, asymptotically

$$M_{k,n}^w \xrightarrow{H_0} \chi^2_m \quad M_{k,n}^w \xrightarrow{H_1, \mathcal{Q}} \left\| Z_m + \alpha(1 - \alpha) H_{w,\gamma_0 \circ P_0^{-1}}^{1/2} \theta \right\|_2^2,$$

where $Z_m \in \mathbb{R}^m$ is a standard normal variable. The same results hold true for $w$ replaced by $\hat{w}$.

Although the preceding analysis focused on Pitman efficiencies, and therefore local alternatives, it is useful to clarify the behavior of $T_{k,n}^\gamma$ under fixed $F, G$. As $T_{k,n}^w$ constitutes a linear approximation to the likelihood, it is not surprising that the usual non-linear KL-projection condition is replaced by a linear one. The following is essentially a restatement of Theorem 3.2 in [DBS21], and is included for completeness.
Theorem 4 ([DBS21]). For $F \neq G$ fixed, tests based on $T_{k,n}^w$ are consistent as long as $E_G w \circ (\alpha F + (1 - \alpha)G) \neq 0$. That is, under $X \sim F, Y \sim G$, $P(\text{reject } H_0) \to 1$.

Remark 3. For the Mann-Whitney test, $w(x) = x - 1/2$, and so the condition $E_G w \circ (\alpha F + (1 - \alpha)G) \neq 0$ amounts to $0 \neq E_G F - 1/2 = P(X < Y) - 1/2$, recovering a well-known result that follows from the original work in [MW47], but was pointed out explicitly by [Leh51].

Extension to multivariate sample spaces

Part of the appeal of likelihood-based methods is their straightforward extension to the multivariate setting. Once a model is specified, be it on $\mathbb{R}$ or $\mathbb{R}^p$, quantities like $\lambda, \phi$ behave similarly, at least asymptotically, with their limiting distribution solely depending on $m$ rather than $p$. Such phenomena invite similar considerations for $T_{k,n}^w$. The previous section’s analysis relied heavily on the empirical CDF $H_{k,n}$ of $\mathcal{X}_{k,n}$, whose relevant properties are most directly derived from the canonical ordering of $\mathbb{R}$. Recent work by [Hal+20] revealed that it is, in fact, not the canonical ordering that gives rise to such desired properties, but rather that they may be recovered through a certain minimization task to which the rank map constitutes a solution. This observation ties ranks to the field of optimal transport [see, e.g., PC+19], thereby enabling their generalization to the multivariate setting; see the work of [Hal+20] itself, as well as the broad exposition given in [DBS21] for how precisely it does so. The univariate behavior of $T_{k,n}^w$ readily extends to this situation.

Theorem 5. For $w$ square-integrable and continuously differentiable on $[0,1]^p$, $H_{k,n}$ the optimal transport map from $\mathcal{X}_{k,n} \subset \mathbb{R}^p$ to $\mathcal{Y}_{N} \subset [0,1]^p$, where $\mathcal{Y}_{N}$ is any set of points whose associated empirical point measure weakly to the uniform law on $[0,1]^p$ and $\max_{c \in \mathcal{Y}_{N}} \mathbb{N}^{-1} ||c||_2^2 \to 0$, define $T_{k,n}^w(\mathcal{X}_{k,n}, \mathcal{Y}_{N}) = w \circ H_{k,n}(\mathcal{Y}_{N}) / \sqrt{N}$. Then the statements of Theorems 1-4 remain true with the domain of integration defining $H = \int w \otimes w$ replaced by $[0,1]^p$, and CDFs interpreted as optimal transport maps from their respective distributions to Uniform([0,1]^p).

Proof. Statements about asymptotic distributions under $H_0$ relied solely on two properties:

(a) $\{H_{k,n}(Y_1), \ldots, H_{k,n}(Y_n)\} = \{r_1, \ldots, r_n\}$ is uniformly distributed over all $n$-subsets of $\{jN^{-1}, j = 1, \ldots, N\}$; or equivalently, it is a sample without replacement of size $n$ from $\{jN^{-1}, j = 1, \ldots, N\}$.

(b) $\{jN^{-1}, j = 1, \ldots, N\} = \{H_{k,n}(Z_1), \ldots, H_{k,n}(Z_N)\}$ converges weakly to Uniform([0,1]).

When interpreted as an optimal transport map between two point clouds in $\mathbb{R}^p$, $H_{k,n}$ still satisfies both (a) and (b) with $\{jN^{-1}, j = 1, \ldots, N\}$ and $[0,1]$ replaced by $\mathcal{Y}_{N}$ and $[0,1]^p$, respectively, and so proofs in Theorems 1-4 concerning $H_0$ may be reused essentially verbatim. To see that convergence under local contiguous alternatives remains unchanged, it suffices to observe that under $H_0$ $(T_{k,n}^w, \rho_n)$ can still be written as a simple random sample from $\mathcal{Y}_{N}$ by property (a), and $\text{Cov}(T_{k,n}^w, \rho_n)$ is identical to its univariate counterpart with $[0,1]$ replaced by $[0,1]^p$ by part (b). \hfill $\Box$

Example (Gaussian location model). For the Gaussian location model $\mathcal{P} = \{\mathcal{N}(\theta, 1_m) : \theta \in \mathbb{R}^m\}$ and $\hat{\theta} = \Phi^{-1}$ (applied component-wise), $T_{k,n}^w(\mathcal{X}_{k,n}, \mathcal{Y}_{N})$ is exactly equivalent to the rank Hotelling-$T^2$ test with Gaussian effective reference distribution, as proposed in [DBS21]. In particular, it exhibits the same favorable Pitman efficiencies compared to Hotelling’s $T^2$ against a large class of alternatives.
Theorem 5 provides theoretical insight, and is useful for sample sizes \( N \) that are small enough for optimal transport maps to be tractably computed, yet large enough for asymptotic statements to become relevant. As discussed previously, these constraints limit practical applications. The following theorem softens these constraints by accelerating computation.

**Theorem 6.** For \( w \) as in Theorem 5, \( N = a^p \) and \( \mathcal{C}_N^p = \{ j \cdot a^{-1}, j = 0, \ldots, a - 1 \}^p \), define the \( j \)-th component \( H_{k,n}^j(z) \) of \( H_{k,n}(z) \) (for \( z \) in \( \mathcal{Z}_{k,n} \)) iteratively through \( H_{k,n}^j(z) = [r_j(z)/a^{p-j}] \cdot a^{-1} \), where \( r_j(z) \) is one less than the rank of \( \langle e_j, z \rangle \) among \( \{ \langle e_j, s \rangle, s \in \mathcal{Z}_{k,n} : r_{j-1}(s) = r_{j-1}(z) \} \) and \( r_0 \equiv 0 \). Then Theorem 5 holds with this choice of \( H_{k,n} \), and the entire two-sample test can be carried out in \( O(pN \log N) \).

**Proof.** As in the proof of Theorem 5, \( H_{k,n} \) satisfies both properties (a) and (b) and thus the conclusion under \( H_0 \) follows. Similarly, the result under \( H_1, \mathcal{P} \) follows by an identical computation of \( \text{Cov}(T_{k,n}^w, \rho_n) \) and noting that \( H_{k,n} \) converges (under \( H_0 \)) to the (invertible) population map

\[
P_0 : (z_1, \ldots, z_p) \mapsto \left( \mathbb{P}_0[Z_{\ell} \leq z_\ell | Z_j = z_j, j = 1, \ldots, \ell - 1], \ell = 1, \ldots, p \right),
\]

where \( \mathbb{P}_0 \) is the measure governing \( \mathcal{Z}_{k,n} \). The computational complexity is obtained by observing that the \( j \)-th layer requires sorting \( a^{p+1-j} \) elements \( a^{j-1} \) times, resulting in \( O\left( \sum_{j=1}^p a^{j-1} \cdot a^{p+1-j} \log a^{p+1-j} \right) = O(pN \log N) \) calculations. \( \Box \)

**Remark 4.** The assumption of \( N \) being a perfect power of \( p \) is simply to streamline the proof of Theorem 6 and communicate its key ingredients more clearly. It can be relaxed to arbitrary \( N \) in a straightforward manner: instead of \( \mathcal{C}_N^p \) being a product space of ranks, it assigns \( \approx [N^{1/p}] \) along each marginal direction, balancing points as evenly as possible. \( H_{k,n} \) then performs the same ranking procedure as before. This general testing algorithm is implemented in the code provided with this paper.

**Remark 5.** When an underlying parametric family \( \mathcal{P} \) is assumed and efficiency desired, the weight computation/estimation requires inversion of the population map (3), just like efficiency in the context of Theorem 5 requires inversion of a (population) optimal transport map. While optimal transport maps are typically difficult to invert (indeed, closed-form formulae are rare even for the forward map), the explicit expression in (3) should facilitate such inversion. Moreover, even when such explicit inversion is intractable, the low computational complexity of \( H_{k,n} \) enables approximations through, e.g., simulations.

**Considerations beyond efficiency**

Although the framework presented above is fully efficient in the case of correct model specification, the more appealing feature is arguably its robustness against model misspecification. This robustness can be exploited for applications that might otherwise be challenging in a likelihood context. This section details two such instances.

**Projecting out nuisance alternatives**

The outcomes of scientific experiments are typically influenced by two sources of randomness: intrinsic fluctuations of the phenomenon of interest (e.g., particle paths behaving approximately like an ensemble of
Brownian motions) and measurement error (e.g., variations in photon densities reaching the microscope, diffraction, plate contamination, etc.); these two sources will be referred to as signal and noise. While differences in signal are usually meaningful for understanding the underlying phenomenon at hand, variation in noise is merely reflective of inconsistencies in the measurement process itself, and so is ideally discarded. That is, given a sample $X_k$ drawn from some $p = p(p_{\text{signal}}, p_{\text{noise}})$ whose details depend on signal and noise distributions $p_{\text{signal}}, p_{\text{noise}}$, respectively, it is of interest to decide whether the generating mechanism of a second sample $X_n$ differs from that of $X_k$ in the signal component $p_{\text{signal}}$, whereas changes in the noise contribution $p_{\text{noise}}$ are to be disregarded. A special case of such situation is given by models of the form $\mathcal{P} = \{p_{\theta_0, \theta_1} : \theta_j \in \Theta_j\}$ for some parameter spaces $\Theta_0, \Theta_1$, where the signal varies in $\Theta_0$ with the noise parametrized by $\Theta_1$. Writing $X_k \overset{\text{iid}}{\sim} p_{\theta_0, \theta_1}, X_n \overset{\text{iid}}{\sim} p_{\theta_0, \theta_1}$, the task then becomes to decide whether $\theta_0 = \theta_0$ against $\theta_0 = \theta_0 + a/\sqrt{N}$.

**Example** (Mixture location model). In genomics it is often of interest to detect **differential expression** of genes; that is, whether the distribution of a certain gene attribute (say, expression) varies across conditions. One approach to do so consists of collecting tissue samples across the conditions of interest, aggregating gene expression within each condition, and then comparing the resulting measurements via two-sample tests. Such procedure, however, is confounded by tissue composition; that is, distinct tissue samples may be comprised of differing proportions of cell types that constitute the tissue. Generally, genes exhibit differential expression across cell types, and consequently, most two-sample tests will detect shifts in proportions even when expression profiles themselves remain unaltered [LR14]. So-called cell-type deconvolution methods provide one attempt to overcome such confounding by inferring cell-type proportions from the data prior to any differential expression analysis [see, e.g., the discussion in Avi20]. The majority of such methods, though, rely on cell-type specific gene expression reference panels, and thereby implicitly assume the absence of any differential expression. How they interact with differential expression is generally unclear. Moreover, distributional properties of the inferred proportions are (with few exceptions [see, e.g., Erd+21]) not provided, and thus the precise manner in which proportion estimates ought to enter two-sample tests remains ambiguous.

As long as the number of cells contained in each tissue sample is sufficiently large, the central limit theorem can be invoked to describe the distribution of $g$ genes’ expression $X \in \mathbb{R}^g$ in a single tissue sample as $\mathcal{N}(M\rho, W_{(M;\Sigma M)} + 1.SP)$, where $\rho \in \Delta^{c-1}$ are the proportions of each of $c > 1$ cell types, $W_k$ is the diagonal matrix with vector $x$ on its diagonal, $\odot$ represents the entry-wise product, and $M, \Sigma \in \mathbb{R}^{g \times c}$ record the expectations and variances of every gene’s expression profile in each cell type [Erd+21]. When tissue samples share similar proportions and behave independently and identically under a given condition, one is thus interested in testing whether the signal $M$ differs across conditions, while ignoring the noise $\rho$.

Theorem 1 shows that local alternatives with associated score function $z$ induce a shift in $T_{k,n}^w$ proportional to $\mu_w(z) = H_{w:z}P^{-1}\theta = \int w(z \circ P^{-1}, \theta)$. Consequently, maximizing power along a signal $z_0$, while guaranteeing robustness against a nuisance model $z_1$ reduces to solving a linear program in $\mathcal{X} = L^2[(0,1)^p]^{m}$. 

**Theorem 7.** For $\hat{\theta} = \arg\max_{\theta} p_{\theta}(X_{k,n}),$ and 

\[
\begin{align*}
\left(\hat{w}_0 \right) &= z_0 \circ P^{-1}, \\
\left(\hat{w}_1 \right) &= z_0 \circ P^{-1},
\end{align*}
\]

12
define the effective weight \( \hat{w} = \hat{w}_0 - \hat{H}_{w_0,w_1} \hat{H}^{-1}_{w_1} \hat{w}_1 \) with \( \hat{H}_{w_0,w_1} \) and \( \hat{H}_{w_1} \) the sub-matrices of the usual estimates of the Fisher information associated with \( \mathcal{P} \) at \( \theta = 0 \). Then as long as \( \hat{\theta} \to 0 \) almost surely,

\[
T_{k,n}^{\hat{\psi}_n} \to \mathcal{N}(0, \alpha(1 - \alpha)(H_0 - H_0H_1^{-1}H_{1,0}))
\]

\[
T_{k,n}^{\hat{\psi}_n} H_{1,\mathcal{P}} \to \mathcal{N}(\alpha(1 - \alpha)(H_0 - H_0H_1^{-1}H_{1,0})(e_1 \otimes 1_m)\theta, \alpha(1 - \alpha)(H_0 - H_0H_1^{-1}H_{1,0}))
\]

where \( H_0 = H_{w_0}, H_1 = H_{w_1}, H_{j,k} = H_{w_j,w_k} \). In particular, \( \mathbb{E}_{H_{1,\mathcal{P}}} T_{k,n}^{\hat{\psi}_n} \to 0 \) whenever \( \theta = (0, \theta_1) \in \Theta_0 \times \Theta_1 \) for any \( \theta_1 \).

**Proof.** By Theorem 2,

\[
T_{k,n}^{\hat{\psi}_n} H_0 \to \mathcal{N}(0, \alpha(1 - \alpha)H_{z_0 \otimes P_0}^{-1}).
\]

Observing that \( T_{k,n}^{\hat{\psi}_n} = \hat{\Pi} T_{k,n}^{\hat{\psi}_n} \), where \( \hat{\Pi} = e_1 \otimes 1_m - e_2 \otimes \hat{H}_{0,1} \hat{H}_1^{-1} \to \Pi = e_1 \otimes 1_m - e_2 \otimes H_{0,1}H_1^{-1} \) by the continuous mapping theorem, an application of Slutsky’s theorem shows that

\[
T_{k,n}^{\hat{\psi}_n} \to \mathcal{N}(0, \alpha(1 - \alpha)\Pi H_{z_0 \otimes P_0}^{-1}\Pi^*)
\]

A quick computation verifies that \( \Pi H_{z_0 \otimes P_0}^{-1}:\Pi^* = H_0 - H_0H_1^{-1}H_{1,0} \) as desired. A similar application of Slutsky’s theorem and quick computation of \( \Pi H_{z_0 \otimes P_0}^{-1} \) establishes the convergence result under \( H_{1,\mathcal{P}} \).

Of course, the very same principle may be employed even in the absence of explicit models: For any two weights \( w_0, w_1 \), the effective weight \( w = w_0 - H_{w_0,w_1}H_{w_1}^{-1}w_1 \) will maximize power in the direction of \( w_0 \) and orthogonal to \( w_1 \).

**Example** (Mixture location model continued). To characterize when differential expression can be locally robustly distinguished from changes in proportion, it suffices by Theorem 7 to compute the weight functions \( w_{\rho} \oplus w_{\rho} = z_{\rho,M} \) corresponding to \( \rho \) and \( M \), respectively, and check whether \( w_{\rho} \) is not in the span of the components of \( w_{\rho} \). Taking \( p_{\rho,M}(x) \) to be the density of a \( \mathcal{N}(M\rho, W_{(M \oplus M)\rho + \Sigma}) \) variable, some calculations yield

\[
w_{\rho} = \sum_{\gamma \in [g]} \frac{M_{\gamma}}{\sigma(\rho, M_{\gamma}, \Sigma_{\gamma})} \Phi^{-1} \circ \Pi_{\gamma} + \frac{M^2_{\gamma} + \Sigma_{\gamma}^2 - 2 \mu(\rho, M_{\gamma})M_{\gamma}}{2\sigma^2(\rho, M_{\gamma}, \Sigma_{\gamma})} \left( \frac{\Phi^{-2} \circ \Pi_{\gamma}}{2} - 1 \right)
\]

\[
w_{\rho} = \sum_{\gamma \in [g]} e^\gamma \otimes \frac{\rho}{\sigma(\rho, M_{\gamma}, \Sigma_{\gamma})} \Phi^{-1} \circ \Pi_{\gamma} + \frac{\rho \circ M_{\gamma} - \mu(\rho, M_{\gamma})\rho}{\sigma^2(\rho, M_{\gamma}, \Sigma_{\gamma})} \left( \frac{\Phi^{-2} \circ \Pi_{\gamma}}{2} - 1 \right),
\]

where \( M_{\gamma}, \Sigma_{\gamma} \in \mathbb{R}^c \) are the expectations and variances of gene \( \gamma \) across the \( c \) cell types, \( \mu(\rho, M_{\gamma}), \sigma^2(\rho, M_{\gamma}, \Sigma_{\gamma}) \) indicate the expectation and variance of the corresponding \( \rho \)-mixture, and \( \Pi_{\gamma}: \mathbb{R}^g \to \mathbb{R} \) is the \( \gamma \)-th coordinate projection. The components of both \( w_{\rho} \) and \( w_{\rho} \) are linear combinations of the \( 2g \) linearly independent functions \( \Phi^{-1} \circ \Pi_{1}, \ldots, \Phi^{-1} \circ \Pi_{g}, \Phi^{-2} \circ \Pi_{1}, \ldots, \Phi^{-2} \circ \Pi_{g} \), and themselves (generically) linearly independent. Therefore, isolating shifts in \( M \) is possible as long as \( c < 2g \); in particular, signal and noise can be locally distinguished even when \( g + 1 < c < 2g \), in which case the deconvolution task itself is not identifiable.

**Remark 6.** Theorem 7 allows for selectively ignoring local alternatives; discarding global alternatives is, in general, more challenging. Indeed, for most models the linear closure of \( \{ z_{\theta_1, \theta_2} : \theta_2 \in \Theta_2 \} \) will have infinite rank or span all of \( \mathcal{H} \).
Qualitative and multiple testing

The early design of weight functions for linear rank statistics did not draw from the connections to likelihood functions as emphasized here, but rather attempted to qualitatively anticipate the impact of distinct shapes on the resulting test statistic. For instance, weights that are increasing and skew-symmetric around $x = 1/2$ should naturally be sensitive to shifts in location, while convex and symmetric shapes may detect perturbations in scale; [see, e.g., the discussions in Van53; AB60]. Many such tests do not naturally correspond to any single model [Moo54; ST60], yet have proven useful in various applications. The machinery developed above can typically be incorporated into this qualitative approach to two-sample testing in a straightforward manner. E.g., the same considerations around monotonicity and skew-symmetry, and convexity and evenness for detecting location and scale shifts, respectively, can be applied in order to yield corresponding two-sample tests acting on multivariate samples. Indeed, the choice of $w = J$ for some score function $J: \mathbb{R}^p \to \mathbb{R}^p$ recovers exactly the \textit{rank Hotelling} $T^2$ tests discussed in [DBS21], if $J$ is the inverse transport map from $[0, 1]^p$ to an "effective" reference distribution $\mathcal{P}$. $\mathcal{P}$ with independent components allow univariate reasoning to be transferred directly to the individual components; for example, the increasing and skew-symmetric components of distribution functions associated with laws like the uniform measure on $[0, 1]^p$ itself or $\mathcal{N}(0, 1^p)$ provide suitable generalizations of the Mann-Whitney-$U$ test and its Gaussian-score-transformed version (cf. the Example on page 10); while $w_1(x) = |x - 1/2|$, $w_2(x) = (x - 1/2)^2$ and $w_3(x) = \Phi^{-2}(x)$ (all applied component-wise) naturally extend the tests of Siegel-Tukey, Mood, and Klotz to the multivariate task of sensing scale shifts. Theorem 6 indicates that power properties of such generalized tests should behave similarly to the those of the univariate tests they generalize, at least for large classes of null distributions.

\textbf{Example} (Relative efficiencies of multivariate scale tests). Under local alternatives $p_{\theta}(x) = \theta^{-1} p_0(\theta^{-1} x)$ around $\theta = 1$ ($x \in \mathbb{R}^p$), the tests of Siegel-Tukey, Mood, and Klotz behave asymptotically as $\|Z + \sqrt{\alpha(1-\alpha)} \beta_j\|_2^2$ (cf. Remark 2), where $Z$ is a univariate standard Gaussian variable, and

\begin{align*}
\beta_1 &= -\frac{48}{p} \int_{[0,1]^p} (\|x - 1/2\|_1 - \frac{p}{4}) \langle P_0^{-1}(x), z_0 \circ P_0^{-1}(x) \rangle \\
\beta_2 &= -\frac{180}{p} \int_{[0,1]^p} (\|x - 1/2\|_2^2 - \frac{p}{12}) \langle P_0^{-1}(x), z_0 \circ P_0^{-1}(x) \rangle \\
\beta_3 &= -\frac{2}{p} \int_{[0,1]^p} (\|\Phi^{-1}(x)\|_2^2 - p) \langle P_0^{-1}(x), z_0 \circ P_0^{-1}(x) \rangle,
\end{align*}

with the usual interpretation of $P_0$ as a transport map pushing $p_0$ to Uniform$([0, 1]^p)$. From the form of the $\beta_j$, it is clear that the multivariate relative efficiencies between $w_1, w_2$ and $w_3$ reduce to their univariate analogues [reviewed, and partially worked out, in Klo62] when, e.g., $P_0$ has independent components. Moreover, if $P_0$ is elliptically symmetric (that is, $p_0(x) = Z^{-1} \exp \circ g(\|\Sigma^{-1/2} x\|_2^2)$ for some sufficiently regular function $g$, covariance matrix $\Sigma$, and normalization constant $Z$), the same conclusion holds; for in such case,

\begin{align*}
\frac{1}{2} \langle P_0^{-1}(x), z_0 \circ P_0^{-1}(x) \rangle = g'(\|\Sigma^{-1/2} P_0^{-1}(x)\|_2^2) \left\| \Sigma^{-1/2} P_0^{-1}(x) \right\|_2^2,
\end{align*}

14
and, writing \( f(x) = g'((\|\Sigma^{-1/2}P_0^{-1}(x)\|_2^2)\|\Sigma^{-1/2}P_0^{-1}(x)\|_2^2) \),

\[
\int \sum_{j=1}^{p} (c \circ \Pi_j)f = p \int (c \circ \Pi_1)f
\]

for \( j \in \{1, 2, 3\} \) and any function \( c \) due to the spherical symmetry of \( f(\text{Uniform}([0,1]^p)) \).

In addition to lacking a precise generative model and therefore necessitating the design of more qualitative weights as discussed above, many applications exhibit multiple plausible directions in which a system may shift under the alternative. E.g., in the differential expression task considered in page 12’s Example, it typically is not clear a priori whether a certain gene’s abundance should be expected to vary in location across distinct conditions, or, say, spread [which has been shown to assume important roles in modulating gene networks; see, e.g., EE10, and references therein]. Testing for each such direction separately and correcting the resulting \( p \)-values for their multiple testing burden risks under-powering of the corresponding hypothesis test. As long as a suitable weight can be formulated for each query of interest, Theorem 1 allows for effective joint testing.

**Example** (Testing for location and scale shifts simultaneously). For samples \( \mathcal{X}_k, \mathcal{Y}_n \in \mathbb{R}^p \), pick \( w_\mu = \Phi^{-1}, w_\sigma = \Phi^{-2}, \) and form the combined weight \( w(x) = (w_\mu, w_\sigma) \in \mathbb{R}^{2p} \). Under \( H_0 \), the resulting \( T_{k,n}^w \) statistic is asymptotically distributed as \( \mathcal{N}(0, \alpha(1-\alpha)H) \), where \( H = p(e_1 + 2e_2) \otimes 1_p \), and so can be used to perform hypothesis testing with. Under \( p_\theta/\sqrt{\pi}(x) = W_{\sigma_N^{-1}} |p_0(W_{\sigma_N^{-1}}x - \mu_N) \), where by slight abuse of notation \( \mu_N = \mu/\sqrt{N}, \sigma_N^{-1} = (\sigma_1^{-1}, ..., \sigma_p^{-1})/\sqrt{N} \) for some \( \mu, \sigma \in \mathbb{R}^p \), and \( W_a \) is the diagonal matrix featuring \( a \) on its diagonal as before, \( \sqrt{\alpha(1-\alpha)H^{-1/2}}T_{k,n}^w \) is asymptotically distributed like

\[
\left\| Z_{2p} + \sqrt{\alpha(1-\alpha)} \left[ \int (\Phi^{-1}, \Phi^{-2}) \otimes (z_0 \circ P_0^{-1}(x), W_\chi[z_0 \circ P_0^{-1}(x)]) \right]^{1/2} (e_1 \otimes \mu + e_2 \otimes \sigma) \right\|_2. \tag{4}
\]

Assuming for simplicity that \( p_0 \) is the density of a standard normal variable in \( \mathbb{R}^p \), (4) simplifies to

\[
\left\| Z_{2p} + \sqrt{p\alpha(1-\alpha)}(e_1 \otimes \mu + \sqrt{2}e_2 \otimes \sigma) \right\|_2^2.
\]

The goal is to compare the power resulting from such expression to that of, say, a Bonferroni correction applied to \( T_{k,n}^{w_\mu} \) and \( T_{k,n}^{w_\sigma} \) separately. This is easiest computed in the large-\( p \) regime, for then

\[
\mathbb{P} \left[ \frac{p^{-1/2}}{2} \left( \left\| Z_{2p} + \sqrt{p\alpha(1-\alpha)(e_1 \otimes \mu + \sqrt{2}e_2 \otimes \sigma)} \right\|_2^2 - 2p \right) > \frac{p^{-1/2}}{2} \chi_{2p,1-\alpha}^2 - 2p \right] \\
\approx 1 - \Phi(z_{1-\alpha} - \left\| \beta \right\|_2^2 - 2\left\| \sigma \right\|_2^2), \tag{5}
\]

where \( \chi_{2p,1-\alpha}^2 \) and \( z_{1-\alpha} \) are the \((1 - \alpha)^{th}\) quantiles of a \( \chi_{2p}^2 \) and a standard Gaussian variable, respectively, and

\[
\left( \left\| \beta \right\|_2^2, \left\| \sigma \right\|_2^2 \right) = [2\alpha(1 - \alpha)]^{-1} \lim_{p \to \infty} \sqrt{p}(\left\| \mu \right\|_2^2, \left\| \sigma \right\|_2^2).
\]

15
The corresponding quantity after correcting for multiple testing of Theorem 1, and so one may without loss of generality assume that the power of tests based on $T_{k,n}^{w}$ is a natural extension of (1), and like (1) features an associated Wilks phenomenon: under the null hypothesis of $\mathcal{X}_1, \ldots, \mathcal{X}_K \overset{iid}{\sim} P_0$, $-2\log \lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K)$ behaves asymptotically like a $\chi^2_{(K-1)m}$ variable. If

$$
\lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K) = \frac{\max_{\theta_1, \ldots, \theta_K} \prod_{j=1}^{K} p_{\theta_j}(\mathcal{X}_j)}{\max_{\theta_1} \prod_{j=1}^{K} p_{\theta}(\mathcal{X}_j)},
$$

Figure 1: Comparison of simultaneously testing for location and scale shifts either jointly ($T_{k,n}^{w_{\mu},w_{\sigma}}$, cf. (5)) or by correcting for multiple testing ($\max \{z_{T_{k,n}^{w_{\mu}}}, z_{T_{k,n}^{w_{\sigma}}} \}$, cf. (6)).

Writing

$$
E_1 = \left\{ (2p)^{-1/2} \left( \left\| Z_p + \sqrt{p(1-\alpha)} \mu \right\|_2 - p \right) > (2p)^{-1/2} \left( \chi^2_{p,1-\alpha/2} - p \right) \right\}
$$

$$
E_2 = \left\{ (2p)^{-1/2} \left( \left\| Z_p + \sqrt{2p(1-\alpha)} \sigma \right\|_2 - p \right) > (2p)^{-1/2} \left( \chi^2_{p,1-\alpha/2} - p \right) \right\}
$$

The corresponding quantity after correcting for multiple testing of $T_{k,n}^{w_{\mu}}$ and $T_{k,n}^{w_{\sigma}}$ reads

$$
P[E_1 \cup E_2] \approx 1 - \Phi(z_{1-\alpha/2} - \sqrt{2}\|\mu\|_2^2) \cdot \Phi(z_{1-\alpha/2} - 2\sqrt{2}\|\sigma\|_2^2). \quad (6)
$$

For fixed choices of $\|\mu\|_2^2, \|\sigma\|_2^2$, (5) and (6) can be compared as functions of the nominal size $\alpha$. Figure 1 does so for three representative pairs $(\|\mu\|_2^2, \|\sigma\|_2^2) \in \{(0.5,0.5), (0.5,0.75), (0.75,0.5)\}$, illustrating a noticeable gain in power.

General $K$-sample testing

The power of tests based on $T_{k,n}^{w}$ is invariant under translations and scaling of $w$ (see discussion after Theorem 1), and so one may without loss of generality assume that $\sum_{j=1}^{N} w(N^{-1}j) = 0$. Under this convention, $(1 - \alpha_{k,n}) \int w \circ H_{k,n} \, dG_n = -\alpha_{k,n} \int w \circ H_{k,n} \, dF_k$, and so the roles of $\mathcal{X}_k$ and $\mathcal{Y}_n$ are entirely symmetric. That is, even though previous results were stated in terms of distributional perturbations of $\mathcal{X}_n$, they equally apply to those of $\mathcal{X}_k$. This ceases to remain true once more than two samples are considered, in which case a more general strategy is to be employed. More concretely, given $K$ samples $\mathcal{X}_1, \ldots, \mathcal{X}_K$ of sizes $n = (n_1, \ldots, n_K)$, the $K$-sample testing task consists of deciding if all samples were generated from identical distributions, and has found applications that rival those of two-sample tests in breadth [see, e.g., discussion in SS87; Tha10]. If an underlying model $\mathcal{P}$ from which the $\mathcal{X}_j$ are generated independently is known, then

$$
\lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K) = \frac{\max_{\theta_1, \ldots, \theta_K} \prod_{j=1}^{K} p_{\theta_j}(\mathcal{X}_j)}{\max_{\theta_1} \prod_{j=1}^{K} p_{\theta}(\mathcal{X}_j)},
$$

is a natural extension of (1), and like (1) features an associated Wilks phenomenon: under the null hypothesis of $\mathcal{X}_1, \ldots, \mathcal{X}_K \overset{iid}{\sim} P_0$, $-2\log \lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K)$ behaves asymptotically like a $\chi^2_{(K-1)m}$ variable. If
the null distribution is known to be uniform, then under local perturbations of \((\mathcal{X}_1, \ldots, \mathcal{X}_K)\) by \(\theta / \sqrt{N} = (\theta_1, \ldots, \theta_K) / \sqrt{N} \in \mathbb{R}^{mk}\) (with \(N = n_1 + \ldots + n_K\)) a first-order expansion of \(-2 \log \lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K)\) reads

\[
-2 \log \lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K) \approx \max_{\theta_1, \ldots, \theta_K} \frac{1}{\sqrt{N}} \sum_{k=1}^{K} (z_0(\mathcal{X}_k), \theta_k) = \max_{\theta_1, \ldots, \theta_K} \frac{1}{\sqrt{N}} \left( \sum_{k=1}^{K} e_k \otimes z_0(\mathcal{X}_k) \right) \theta,
\]
motivating the following \(K\)-sample version \(\tilde{T}_n^w\) of \(T_{k,n}^w\):

\[
\tilde{T}_n^w = \sqrt{N} \sum_{k=1}^{K} e_k \otimes \left( \alpha_k \int w \circ H_n \, dF_k \right) = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} e_k \otimes w(\mathcal{X}_k) \in \mathbb{R}^{mK},
\]

where \(H_n\) is a transport map of \(\mathcal{X}_n = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_K\) to \(\mathcal{C}_N^0\), \(\alpha_k = n_k / N\), and \(F_k\) the corresponding map for \(\mathcal{X}_k\). Given the dependence of \(\tilde{T}_n^w\)'s components \((\tilde{T}_n^w \in \ker \mathbb{1}_K \otimes 1_m = \Theta(0))\), it is more convenient to work with a transformed version \(T_n^w = \Gamma \tilde{T}_n^w \in \mathbb{R}^{(m-1)K}\), where \(\Gamma = B^* \mathbb{1}_1 / \sqrt{\alpha} \otimes 1_m : (\mathbb{R}^K \otimes \mathbb{R}^m) \to (\mathbb{R}^{K-1} \otimes \mathbb{R}^m)\), and \(B\) an orthonormal basis for \(\Theta_1 = \Theta_0^\perp\). As is the case with \(T_{k,n}^w\), the large-\(N\) behavior of \(T_n^w\) can be worked out.

**Theorem 8.** For \(\min_k n_k \to \infty, n_k / N \to \alpha_k > 0\), and the usual conditions on \(w \in \mathbb{R}^m\) and \(\mathcal{P}\), the statistic \(\sqrt{N}T_n^w\) is asymptotically normal under the null and local (contiguous) alternatives with

\[
T_n^w \overset{H_0}{\to} \mathcal{N}(0, B^* AA^* B \otimes H_w) \quad T_n^w \overset{H_1=\mathcal{P}}{\to} \mathcal{N}(B^* AA^* \sqrt{\alpha} \otimes H_w, z_0 \circ P_0^{-1} \theta, B^* AA^* B \otimes H_w),
\]

where \(A\) is an orthonormal basis for \(\ker \sqrt{\alpha} \otimes \sqrt{\alpha}\), with \(-\sqrt{\cdot}\) applied component-wise.

**Proof.** To simplify notation, the proof only treats the case \(p = 1\), from which \(p > 1\) follows in a straightforward manner by reasoning as in Theorem 5. As usual, the \(H_0\) result is a consequence of the central limit theorem for sampling without replacement. More concretely, write \(T_n^w(k) = (e_k \otimes 1_m) \tilde{T}_n^w\) and set \(\mathcal{A}_N(k+1) = \{w(jN^{-1}), j = 1, \ldots, N\} \setminus \cup_{m=1}^{k} \mathcal{B}_m\), where \(\mathcal{B}_m = \{w \circ H_n(X_{m,1}), \ldots, w \circ H_n(X_{m,n_m})\}\). Then \(\tilde{T}_n^w(1)\) converges to a centered normal variable by Theorem 1, and

\[
\tilde{T}_n^w(k+1) = \sum_{j=1}^{k} \tilde{T}_n^w(j) = s
\]

is a sample without replacement of size \(n_{k+1}\) from

\[
\mathcal{A}_N(k+1) \left| N^{-1/2} \sum_{b \in \cup_{j=1}^{k} \mathcal{B}_j} b = s \right.
\]

and thus is amenable to a central limit theorem too. Therefore, \((\tilde{T}_n^w(1), \ldots, \tilde{T}_n^w(K-1))\) is jointly normal as \(N\) grows large, implying that

\[
\tilde{T}_n^w = \begin{pmatrix} \tilde{T}_n^w(1) \\ \vdots \\ \tilde{T}_n^w(K-1) \\ \tilde{T}_n^w(K) \end{pmatrix} \overset{H_0}{\to} \mathcal{N}(0, \Sigma)
\]
with \( \text{ker} \tilde{\Sigma} = \Theta_0 \). Consequently, \( T_n^w \to \mathcal{N}(0, \Sigma) \), where \( \Sigma \) is given by

\[
\text{Var} T_n^w = \mathbb{E} \sum_{k=1}^{K} \frac{1}{n_k} (B^* e_k \otimes B^* e_k) \otimes (w \circ H_{k,n}(\mathcal{X}_k) \otimes w \circ H_{k,n}(\mathcal{X}_k)) \]

\[
+ \mathbb{E} \sum_{m \neq k} \frac{1}{\sqrt{n_m n_k}} (B^* e_m \otimes B^* e_k) \otimes (w \circ H_{k,n}(\mathcal{X}_m) \otimes w \circ H_{k,n}(\mathcal{X}_k)) = A_1 + A_2.
\]

Computations similar to those used in the proof of Theorem 1 yield

\[
A_1 = \sum_{k=1}^{K} (1 - \alpha_k)(B^* e_k \otimes B^* e_k) \otimes H_w
\]

\[
A_2 = \sum_{m \neq k} \sqrt{\alpha_m \alpha_k} (B^* e_m \otimes B^* e_k) \otimes H_w,
\]

and thus

\[
\text{Var} T_n^w = B^* (1 - \sqrt{\alpha} \otimes \sqrt{\alpha}) B \otimes H_w = B^* AA^* B \otimes H_w,
\]

as advertised.

Establishing the \( H_{1,\mathcal{O}} \) convergence can be performed along similar lines as before: augment \( \mathcal{O}_N(k+1) \) to

\[
\mathcal{O}_N^+ (k+1) = \{ w(jN^{-1}) + e_{m+1} \otimes z_0 H_n^{-1}(jN^{-1}) \} \cup \bigcup_{m=1}^{k} \mathcal{O}_N^+
\]

where

\[
\mathcal{O}_N^+ = \{ w \circ H_n(X_{j,1}) + e_{m+1} \otimes z_0(X_{j,1}), \ldots, w \circ H_n(X_{j,n_j}) + e_{m+1} \otimes z_0(X_{j,n_j}) \},
\]

then the same argument showing the joint normality of \( (\tilde{T}_n^w(1), \ldots, \tilde{T}_n^w(K)) \) can be used to arrive at the joint normality of \( (\tilde{T}_n^w(1), \ldots, \tilde{T}_n^w(K), z_0(\mathcal{X}_1), \ldots, z_0(\mathcal{X}_K)) \), and therefore the joint normality of \( (T_n^w, \rho_n) \), where

\[
\rho_n = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \frac{p_{\theta_k}/\sqrt{n}}{p_{\theta_k}(X_{k,j})} = \sum_{k=1}^{K} \sqrt{\alpha_k}(z_0(\mathcal{X}_k), \theta_k) - \frac{1}{2} \left( \lambda_\alpha \otimes H_{z_0 \circ p_0^{-1}}(\theta, \theta) \right) + o_{H_0}(1)
\]

is the relevant likelihood ratio statistic. The usual appeal to LeCam’s third lemma together with the observation that

\[
\mathbb{E} T_n^w \rho_n = \sum_{k=1}^{K} \sqrt{\alpha_k} B^* e_k \otimes \mathbb{E} [w \circ H_n(\mathcal{X}_k) \otimes z_0(\mathcal{X}_k)] \theta_k + \sum_{m \neq k} \sqrt{\alpha_m \alpha_k} B^* e_k \otimes \mathbb{E} [w \circ H_n(\mathcal{X}_m) \otimes z_0(\mathcal{X}_m)] \theta_m
\]

converges to

\[
\sum_{k=1}^{K} \sqrt{\alpha_k}(1 - \alpha_k) B^* e_k \otimes H_{w,z_0 \circ p_0^{-1}}(\theta_k) - \sum_{m \neq k} \alpha_m \sqrt{\alpha_k} B^* e_k \otimes H_{w,z_0 \circ p_0^{-1}}(\theta_j)
\]

\[
= \left[ B^* (\lambda_\sqrt{\alpha} - \sqrt{\alpha} \otimes \alpha) \otimes H_{w,z_0 \circ p_0^{-1}} \right] \theta = B^* AA^* \lambda_\sqrt{\alpha} \otimes H_{w,z_0 \circ p_0^{-1}}
\]

completes the proof.

As before, the Pitman efficiency of \( T_n^w \) compared to \( \lambda_\mathcal{O}(\mathcal{X}_1, \ldots, \mathcal{X}_K) \) is favorable.
Corollary 3. The Pitman efficiency of $T_{z_0\circ P_0^{-1}}$ relative to a test based on $\lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K)$ is 1 for local alternatives $p_\theta/\sqrt{n} \in \mathcal{P}$ converging to $p_0$.

Proof. The usual regularity assumptions on $\mathcal{P}$ imply that

$$-2\log \lambda_{\mathcal{P}}(\mathcal{X}_1, \ldots, \mathcal{X}_K) \overset{H_{1,\mathcal{P}}}{\longrightarrow} \left\| Z_{(m-1)K} + \Pi \lambda_{\sqrt{\alpha}} \otimes H_{z_0\circ P_0^{-1}} \Theta_0 \lambda_{\sqrt{\alpha}} \otimes H_{z_0\circ P_0^{-1}}^{1/2} \theta \right\|_2^2,$$

(7)

where $\Pi_A : \Theta^K \rightarrow A$ for a linear subspace $A \subset \Theta^K$ is the orthogonal projection onto $A$. Here, $\Theta_0$ is the diagonal of $\Theta^K$, and so

$$\lambda_{\sqrt{\alpha}} \otimes H_{z_0\circ P_0^{-1}}^{1/2} \Theta_0 = \ker \left( I_K - \sqrt{\alpha} \otimes \sqrt{\alpha} \right) \otimes 1_m,$$

giving for the non-centrality parameter in (7) $AA^* \lambda_{\sqrt{\alpha}} \otimes H_{z_0\circ P_0^{-1}}^{1/2}$ as desired. $\square$

Likewise, results mirroring those contained in Theorems 2-4 can be obtained in this general $K$-sample setting. In particular, Corollary 3 remains true if $w = z_0\theta$ is chosen adaptively through maximum-likelihood estimation based on $\mathcal{X}_n$.

**Exact finite sample distributions**

Although typical Berry-Esseen rates of $O(N^{-1/2})$ are known for linear rank statistics under mild regularity conditions on $w$ [Fri89], modern scientific applications often must accommodate to very small sample sizes, where asymptotic results may not be of relevance yet [see, e.g., Mol+20]. Empirical simulation studies have shown that linear rank statistics continue to perform favorably in such regime [CJJ81; CGT18], and so characterizing the finite sample distribution of $T_{k,n}^w$ is therefore desirable both theoretically as well as in practice. While an explicit description appears to be currently out of reach, [ETS22] recently showed that such characterization is available for a closely related, asymptotically equivalent, statistic $S_{k,n}^w$ based on rank-spacings (i.e., $r_{j+m} - r_j$) instead of ranks, and argued that its small-sample power properties ought to be comparable to those of $T_{k,n}^w$ for most purposes. More concretely, $S_{k,n}^w$ is computed as

$$S_{k,n}^w = \frac{1}{n} \sum_{j=0}^{n-1} w \left( \frac{j}{n} \right) (r_{j+1} - r_j) = w(1) - \int H_N(x) \left( w \circ G_n(x) - w \circ (G_n - n^{-1})(x) \right) dG_n,$$

with the convention $r_0 = 0, r_{n+1} = n + 1$, and where $\mathcal{X}_k, \mathcal{Y}_n$ and $w$ are all assumed univariate [ETS22, provides a slightly different definition, though the two versions differ merely by a constant]. A first step towards adapting $S_{k,n}^w$ for our purposes is to allow $w$ be in $\mathbb{R}^m$ rather than just $\mathbb{R}$.

Theorem 9. For $\mathcal{X}_k, \mathcal{Y}_n \subset \mathbb{R}, w : [0, 1] \rightarrow \mathbb{R}^m$, the Laplace transform of $S_{k,n}^w$ is given by

$$\mathbb{E}_{H_0} e^{\langle x, S_{k,n}^w \rangle} = (k-1)(-1)^{n+1} \cdot e^{\lambda w_{\max}(s)} \sum_{j=1}^{n-1} e^t(e_j, W_s) - w_{\max}(s)) \left[ b_{k,n} \left( 1 - e^t(e_j, W_s) - w_{\max}(s) \right)^{(n+1-k)} \right] + \sum_{m=0}^{n-3} c_{k,n,m} \left( 1 - e^t(e_j, W_s) - w_{\max}(s) \right)^{n-m-2},$$

(8)
for almost every $s \in \mathbb{R}^m$, where $W \in \mathbb{R}^{n \times m}$ has rows $w(jN^{-1})$, $j = 1, \ldots, n$, $w_{\max}(s) = \|Ws\|_{\max}$, for any $r \in \mathbb{R}^n$, $a_j' = \prod_{m \neq j} (r_j - r_m)^{-1}$ and

$$b_{k,n} = \frac{(-1)^n}{n+k-1} \left( \frac{n+k-2}{n-2} \right)^{-1} c_{k,n} = \frac{(-1)^n}{k+1} \frac{k}{n+m+1},$$

as long as $\ker W = \{0\}$.

**Proof.** The proof is a direct consequence of Theorem 1 in [ETS22] after observing that $\langle s, S_{k,n}^w \rangle = S_{k,n}^{(s,w)}$. The condition on $\ker W$ guarantees distinctness of the components of $Ws$. \hfill \Box

**Remark 7.** Similar to the expression obtained in [ETS22], (8) may appear inconvenient to work with analytically, but can often be inverted numerically.

**Remark 8.** The condition on $\ker W$ can be relaxed at the expense of slightly more complicated expressions, see discussion around Theorem 1 of the univariate case.

The extension to multivariate sample spaces is more involved.

**Theorem 10.** Tile the unit cube $[0,1]^n$ by a collection $\mathcal{S}$ of $[N^{1/p}]^n$ sub-cubes, each of side-length $[N^{-1/p}]$ in the natural way, and let $\pi : \{1, \ldots, |\mathcal{S}|\} \rightarrow \mathcal{S}$ be an enumeration of $\mathcal{S}$ with $\pi_{j+1}$ being adjacent to $\pi_j$ (that is, sharing a face) for all $1 \leq j < |\mathcal{S}|$. Associate with $\pi$ a non-self-intersecting curve $c : [0,1] \rightarrow [0,1]^p$ that linearly interpolates between the centers of $\pi_j$ and $c(j/|\mathcal{S}|) = \text{center}(\pi_j)$. Construct the point measure $\mathcal{C}_n^p$ as $\mathcal{C}_n^p = N^{-1} \sum_{j=1}^{N} \delta_{c(jN^{-1})}$, and optimally transport $\mathcal{X}_k$ to $\mathcal{C}_n^p$ via $H_N$ with induced transport maps $F_k, G_n$ on $\mathcal{X}_k, \mathcal{Y}_n$, define $\rho_j = \inf\{s : |c([0,s])) \cap G_n(\mathcal{Y}_n)| \geq j\}$. Then a test based on

$$S_{k,n}^{w,\text{worse}} = \frac{1}{n} \sum_{j=1}^{n} w \circ c \left( \frac{j}{n} \right) (\rho_{j+1} - \rho_j)$$

is asymptotically equivalent to one using $T_{k,n}^w$ under $H_0$, and the Laplace transform of $S_{k,n}^{w,\text{worse}}$ is given by (8) (replacing $w$ with $w \circ c$ throughout).

**Proof.** The construction essentially reduces the multivariate case to the univariate one: as $H_{k,n}(\mathcal{Y}_n)$ is uniform over $\mathcal{C}_n^p$ under $H_0$, $(\rho_1, \ldots, \rho_n)$ are uniform over subsets of size $n$ from $\{1, \ldots, N\}$. The theorem then follows from convergence results of the univariate $S_{k,n}^w$, and Theorem 9. \hfill \Box

**Remark 9.** A $\pi$ that is convenient to work with in practice can be constructed by induction: for $p = 2$, label each sub-cube in $\mathcal{S}$ by $(x_1, x_2) \in \{1, \ldots, |\mathcal{S}|^{1/2} = a_2\}^2$ and set

$$\pi_{2,j} = \left( \lceil j/a_2 \rceil, (-1)^{\lfloor j/a_2 \rfloor} \lfloor 1 + (j - 1 \mod a_2) \rfloor \right).$$

For $p > 2$, similarly label each sub-cube by a $p$-tuple $(x_1, \ldots, x_p) \in \{1, \ldots, |\mathcal{S}|^{1/p} = a_p\}^p$, and define $\pi_{p,j} = \left( \lfloor j/a_p^{-1} \rfloor, \pi_{p-1,j \mod a_p^{-1}} \right)$.

**Remark 10.** Convergence under local contiguous alternatives in the univariate setting has been worked out in [HR80]. The proofs there are established under a univariate $w$; however, it appears that this assumption can be relaxed to $n^{-1} \sum_{j=1}^{n} w(jn^{-1}) \rightarrow f w$. Therefore, as long as $c$ in Theorem 10 is chosen such that $n^{-1} \sum_{j=1}^{n} w \circ c(jn^{-1}) \rightarrow f w$ (as is, e.g., the case with the choice given in the remark above), such convergence results ought to transfer to the multivariate setting, in which case $S_{k,n}^w$ and $T_{k,n}^w$ are asymptotically equivalent even under local alternatives. Rigorously checking these conditions is left for future work.
**Conclusion**

While the local optimality of linear rank statistics for certain choices of coefficients has been known in several specific situations (e.g., location shifts), a general framework embedding such instances into maximum-likelihood-type arguments appears to not have been articulated thus far. By doing so explicitly, we hope to provide guidance to practitioners who may suspect certain generative models underlying a given application, but would like to guard against misspecification. Such situations arise frequently in modern science, where, e.g., the broad physical or biological (stochastic) mechanisms at play are understood, yet experimental measurements add less well characterized noise. With the complexity of hypotheses and experimental protocols increasing, so must the complexity of statistical models; in particular, building tests that can incorporate multivariate sample and parameter spaces is desirable. Building on recent progress tying together multivariate ranks and optimal transport, we identify alternative transport maps that extend linear rank statistics from the univariate to the fully multivariate setting while preserving all their favorable properties at improved computational cost. For problems whose complexity is well captured by the framework surrounding Pitman efficiency, the resulting adaptive linear multi-rank statistics always outperform likelihood-ratio tests, and thus should be preferred. Viewing linear rank statistics from both a non-parametric testing angle and a maximum-likelihood perspective opens up applications that might not be straightforward through either approach alone; here we showcased this flexibility on the examples of nuisance parameters, multiple testing and general $K$-sample testing. Most of the findings presented pertain to the large-sample regime, though finite-$N$ guarantees become available for the closely related linear rank-spacings statistics. A software package implementing the full $K$-sample test is made available with this paper.

**Acknowledgments**

The author warmly thanks Günther Walther for comments that inspired the section on projecting out nuisance parameters.

**References**

[AB60] Abdur Rahman Ansari and Ralph A Bradley. “Rank-sum tests for dispersions”. In: *The Annals of Mathematical Statistics* (1960), pp. 1174–1189.

[Avi+20] Francisco Avila Cobos et al. “Benchmarking of cell type deconvolution pipelines for transcriptomics data”. In: *Nature communications* 11.1 (2020), pp. 1–14.

[AZ05] B Aslan and Gunter Zech. “Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 537.3 (2005), pp. 626–636.

[Bah67] RR Bahadur. “An optimal property of the likelihood ratio statistic”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 1967, pp. 13–26.
[BC15] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: The Annals of Statistics 43.5 (2015), pp. 2055–2085.

[Ben10] Yoav Benjamini. “Simultaneous and selective inference: Current successes and future challenges”. In: Biometrical Journal 52.6 (2010), pp. 708–721.

[BF04] Ludwig Baringhaus and Carsten Franz. “On a new multivariate two-sample test”. In: Journal of multivariate analysis 88.1 (2004), pp. 190–206.

[BH95] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: Journal of the Royal statistical society: series B (Methodological) 57.1 (1995), pp. 289–300.

[Bha19] Bhaswar B Bhattacharya. “A general asymptotic framework for distribution-free graph-based two-sample tests”. In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81.3 (2019), pp. 575–602.

[Bic65] Peter J Bickel. “On Some Asymptotically Nonparametric Competitors of Hotelling’s T2 1”. In: The Annals of Mathematical Statistics (1965), pp. 160–173.

[Bon36] Carlo Bonferroni. “Teoria statistica delle classi e calcolo delle probabilita”. In: Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936), pp. 3–62.

[CF17] Hao Chen and Jerome H Friedman. “A new graph-based two-sample test for multivariate and object data”. In: Journal of the American statistical association 112.517 (2017), pp. 397–409.

[CGT18] William Jay Conover, Armando Jesús Guerrero-Serrano, and Víctor Gustavo Tercero-Gómez. “An update on a comparative study of tests for homogeneity of variance”. In: Journal of Statistical Computation and Simulation 88.8 (2018), pp. 1454–1469.

[Cha96] Probal Chaudhuri. “On a geometric notion of quantiles for multivariate data”. In: Journal of the American statistical association 91.434 (1996), pp. 862–872.

[CJJ81] William J Conover, Mark E Johnson, and Myrle M Johnson. “A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data”. In: Technometrics 23.4 (1981), pp. 351–361.

[CS58] Herman Chernoff and I Richard Savage. “Asymptotic normality and efficiency of certain nonparametric test statistics”. In: The Annals of Mathematical Statistics (1958), pp. 972–994.

[CY15] Konstantina Charmpi and Bernard Ycart. “Weighted Kolmogorov Smirnov testing: an alternative for gene set enrichment analysis”. In: Statistical Applications in Genetics and Molecular Biology 14.3 (2015), pp. 279–293.

[DBS21] Nabarun Deb, Bhaswar B Bhattacharya, and Bodhisattva Sen. “Efficiency lower bounds for distribution-free Hotelling-type two-sample tests based on optimal transport”. In: arXiv preprint arXiv:2104.01986 (2021).

[EE10] Avigdor Eldar and Michael B Elowitz. “Functional roles for noise in genetic circuits”. In: Nature 467.7312 (2010), pp. 167–173.

[Erd+21] Dan D Erdmann-Pham et al. “Likelihood-based deconvolution of bulk gene expression data using single-cell references”. In: Genome research 31.10 (2021), pp. 1794–1806.
[ETS22] Dan D Erdmann-Pham, Jonathan Terhorst, and Yun S Song. “Generalized Spacing-Statistics and a New Family of Non-Parametric Tests”. In: arXiv preprint (2022).

[FR79] Jerome H Friedman and Lawrence C Rafsky. “Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests”. In: The Annals of Statistics (1979), pp. 697–717.

[Fri89] Karl O Friedrich. “A Berry-Esseen bound for functions of independent random variables”. In: The Annals of Statistics (1989), pp. 170–183.

[GK09] Sebastian J Goerg and Johannes Kaiser. “Nonparametric testing of distributions—the Epps–Singleton two-sample test using the empirical characteristic function”. In: The Stata Journal 9.3 (2009), pp. 454–465.

[Gre+12] Arthur Gretton et al. “A kernel two-sample test”. In: The Journal of Machine Learning Research 13.1 (2012), pp. 723–773.

[Hal+20] Marc Hallin et al. “On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$: a measure transportation approach”. In: Annals of Statistics (to appear) (2020).

[Hen88] Norbert Henze. “A multivariate two-sample test based on the number of nearest neighbor type coincidences”. In: The Annals of Statistics 16.2 (1988), pp. 772–783.

[HL56] Joseph L Hodges Jr and Erich L Lehmann. “The efficiency of some nonparametric competitors of the t-test”. In: The Annals of Mathematical Statistics (1956), pp. 324–335.

[Hot92] Harold Hotelling. “The generalization of Student’s ratio”. In: Breakthroughs in statistics. Springer, 1992, pp. 54–65.

[HP02] Marc Hallin and Davy Paindaveine. “Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks”. In: Annals of Statistics (2002), pp. 1103–1133.

[HP04] Marc Hallin and Davy Paindaveine. “Rank-based optimal tests of the adequacy of an elliptic VARMA model”. In: The Annals of Statistics 32.6 (2004), pp. 2642–2678.

[HR80] Lars Holst and JS Rao. “Asymptotic theory for some families of two-sample nonparametric statistics”. In: Sankhyā: The Indian Journal of Statistics, Series A (1980), pp. 19–52.

[JK12] Jana Jurečková and Jan Kalina. “Nonparametric multivariate rank tests and their unbiasedness”. In: Bernoulli 18.1 (2012), pp. 229–251.

[Klo62] Jerome Klotz. “Nonparametric tests for scale”. In: The Annals of Mathematical Statistics 33.2 (1962), pp. 498–512.

[LD75] Erich Leo Lehmann and Howard J D’Abrera. Nonparametrics: statistical methods based on ranks. Holden-day, 1975.

[Leh51] Eric L Lehmann. “Consistency and unbiasedness of certain nonparametric tests”. In: The annals of mathematical statistics (1951), pp. 165–179.

[LR14] Robert Lowe and Vardhman K Rakyan. “Correcting for cell-type composition bias in epigenome-wide association studies”. In: Genome Medicine 6.3 (2014), pp. 1–2.

[LRC05] Erich Leo Lehmann, Joseph P Romano, and George Casella. Testing statistical hypotheses. Vol. 3. Springer, 2005.
[Wil38] Samuel S Wilks. “The large-sample distribution of the likelihood ratio for testing composite hypotheses”. In: *The annals of mathematical statistics* 9.1 (1938), pp. 60–62.

[ZS00] Yijun Zuo and Robert Serfling. “General notions of statistical depth function”. In: *Annals of statistics* (2000), pp. 461–482.

[ZZM92] Ofer Zeitouni, Jacob Ziv, and Neri Merhav. “When is the generalized likelihood ratio test optimal?” In: *IEEE Transactions on Information Theory* 38.5 (1992), pp. 1597–1602.