Powerful rank verification for multivariate Gaussian data with any covariance structure

Anav Sood
Stanford University

March 1st, 2025

Abstract

Upon observing n-dimensional multivariate Gaussian data, when can we infer that the largest K observations came from the largest K means? When $K = 1$ and the covariance is isotropic, Gutmann and Maymin [1987] argue that this inference is justified when the two-sided difference-of-means test comparing the largest and second largest observation rejects. Leveraging tools from selective inference, we provide a generalization of their procedure that applies for both any K and any covariance structure. We show that our procedure draws the desired inference whenever the two-sided difference-of-means test comparing the pair of observations inside and outside the top K with the smallest standardized difference rejects, and sometimes even when this test fails to reject. Using this insight, we argue that our procedure renders existing simultaneous inference approaches inadmissible when $n > 2$. When the observations are independent (with possibly unequal variances) or equicorrelated, our procedure corresponds exactly to running the two-sided difference-of-means test comparing the pair of observations inside and outside the top K with the smallest standardized difference.

1 Introduction and results

1.1 Motivation

Having observed data, it is often natural to ask whether the best observation actually came from the best population. As motivation, we present two important real world variants of this problem.

Rank verification for large language models: Chatbot Arena [Chiang et al., 2024] is a platform that currently has over thirty-thousand daily users and ranks the performance of $n = 206$ large language models according to user preference data. Is the model at the top of the leaderboard actually the best model?

Rank verification in multi-arm clinical trials: In multi-arm clinical trials, each patient is randomly assigned to receive one of n different treatments (including a control). Is the treatment with the largest observed average treatment effect actually the best treatment?

Both of these scenarios are examples of a rank verification problem, and they can be formalized as follows. We observe a multivariate Gaussian vector $X \sim N(\mu, \Sigma)$ and see that $W = \text{argmax}_{1 \leq i \leq n} X_i$ is the index of the largest observation. Can we claim that $\mu_W > \max_{j \neq W} \mu_j$, i.e., that the largest observation came from the largest mean? We elaborate on how each of the above scenarios reduces to solving this problem.

Rank verification for large language models: The Chatbot Arena dataset is constructed by asking users which of two models performed better on a prompt. Treating these responses as i.i.d samples, the leaderboard fits a Bradley-Terry model [Bradley and Terry, 1952] and uses the resulting coefficients $\hat{\beta} \in \mathbb{R}^n$ to rank the models. These coefficients follow a central limit theorem when the number of samples is large, i.e., $\hat{\beta} \dot{\sim} N(\beta, \Sigma)$ for some non-diagonal covariance Σ. Letting W be the index of the model with the largest fitted coefficient, we want to know, is $\beta_W > \min_{j \neq W} \beta_j$?

Rank verification in multi-arm clinical trials: If there are enough participants in the clinical trial, then the the treatments' average observed effects $X \in \mathbb{R}^n$ obey a central limit theorem, i.e., $X\dot{\sim}N(\mu,\Sigma)$ with diagonal covariance Σ. We want to know, is $\mu_W > \max_{j\neq W}\mu_j$?

Briefly, we mention that our formalization also encompasses the problem of verifying that the machine learning model with the best performance on a challenge dataset is actually the best model. In this case, the n models' observed average performances $X \in \mathbb{R}^n$ on the dataset obey a central limit theorem (provided that the challenge dataset is moderately large), i.e., $X \sim N(\mu,\Sigma)$. Because the same dataset is used to evaluate the models, the X_i are correlated and Σ may be highly non-diagonal. Again we want to know, is $\mu_W > \max_{j\neq W}\mu_j$?.

This paper considers a generalization of our motivating problem. Defining S to be the set containing the indices of the largest $K < n$ entries of X, we aim to draw the inference $\min_{i\in S}\mu_i - \max_{j\notin S}\mu_j > \delta$ that the largest K observations came from means that are more than δ larger than the rest. We recover our motivating problem by setting $K = 1$ and $\delta = 0$.

The main contribution of this paper is to provide a powerful and computationally tractable error controlling procedure for drawing the inference $\min_{i\in S}\mu_i - \max_{j\notin S}\mu_j > \delta$. For a pre-specified level α, the probability of our procedure falsely drawing this inference will be at most α. Throughout our discussion, we will assume that the covariance Σ is known. In practice, Σ is not known but can be estimated from the data.

1.2 Method and theoretical results

In this subsection, we summarize the results of the paper. We imagine observing n-dimensional data $X \sim N(\mu,\Sigma)$ where Σ is known. Our only restriction on Σ is that we require X_i and X_j to be not perfectly correlated when $i \neq j$. Considering the null hypotheses $H^\delta_{ij} : \mu_i - \mu_j \leq \delta$, we derive an error controlling procedure for rejecting the data dependent union null $\cup_{i\in S, j\notin S} H^\delta_{ij}$, where S is the set of the largest K observations' indices. Formally, a false rejection happens when we reject $\cup_{i\in S, j\notin S} H^\delta_{ij}$ and μ_i is not more than δ larger than μ_j for all $i \in S$ and $j \notin S$. Tying back to our motivation, we can safely draw the inference $\min_{i\in S}\mu_i - \max_{j\notin S}\mu_j > \delta$ when we reject $\cup_{i\in S, j\notin S} H^\delta_{ij}$.

To simplify our exposition, we present our results in the special case that $\delta = 0$, i.e., we just consider the problem of rejecting $\cup_{i\in S, j\notin S} H^0_{ij}$ and verifying $\min_{i\in S}\mu_i > \max_{j\notin S}\mu_j$. The most general versions of our results are clearly stated in Section 2, where we provide proofs of our claims as well.

Prior to stating our simplified results, we introduce some notation. First, for a pair $i \neq j$, we define

$$D_{ij} = \frac{X_i - X_j}{v_{ij}}, \quad v^2_{ij} = \mathrm{Var}(X_i - X_j) = \Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}$$

to be the standardized difference between X_i and X_j. Considering another pair $k \neq \ell$, we use $\rho_{ij,k\ell}$ to denote the correlation between D_{ij} and $D_{k\ell}$.

Using this notation, Theorem 1 states our method. Though it may look complicated, it is easy to implement on a computer and we will soon see that its behavior is very interpretable. In our statement of the theorem, we adopt the convention that the minimum of an empty set is ∞. We provide a proof of a generalized version of Theorem 1 that applies for any $\delta \in \mathbb{R}$ in Section 2.1.

Theorem 1 (Gaussian rank verification). *If we reject $\cup_{i\in S, j\notin S} H^0_{ij}$ when*

$$\max_{i\in S, j\notin S} \frac{[1-\Phi(D_{ij})] - \left[1-\Phi\left(\min\limits_{\substack{k\in S,\ell\notin S:\\ \rho_{ij,k\ell}<0}} D_{ij} - \frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)\right]}{\left[1-\Phi\left(\max\limits_{\substack{k\in S,\,\ell\notin S:\\ \rho_{ij,k\ell}>0}} D_{ij} - \frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)\right] - \left[1-\Phi\left(\min\limits_{\substack{k\in S,\ell\notin S:\\ \rho_{ij,k\ell}<0}} D_{ij} - \frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)\right]} \leq \alpha, \tag{1}$$

then, conditional on S, the probability of making a false rejection is at most α.

Our next result, Theorem 2, helps us make sense of Theorem 1's method. We prove an analog of Theorem 2 that applies whenever $\delta \geq 0$ in Section 2.2.

Theorem 2 (Understanding Gaussian rank verification)**.** *Let $I \in S$ and $J \notin S$ be the indices of any pair of observations inside and outside of the top K that achieve the smallest possible standardized difference, i.e., $D_{IJ} = \min_{i \in S, j \notin S} D_{ij}$. Then the procedure from Theorem 1 is guaranteed to reject $\cup_{i \in S, j \notin S} H_{ij}^0$ whenever*

$$1 - \Phi(D_{IJ}) \leq \alpha/2. \tag{2}$$

Essentially, Theorem 2 tells us that Theorem 1's test will reject $\cup_{i \in S, j \notin S} H_{ij}^0$ whenever the two-sided difference-of-means test comparing the pair of observations inside and outside of the top K with the smallest standardized difference rejects, and possibly also in other situations as well. We mention that the time complexity of running Theorem 1's test is $O(K^2(n-K)^2)$, which can be $O(n^4)$ in the worst case (e.g. if $K = n/2$). This will still not be prohibitive for many problems, but if it is, Theorem 2 tells us that we could also safely reject $\cup_{i \in S, j \notin S} H_{ij}^0$ whenever $1 - \Phi(D_{IJ}) \leq \alpha/2$, a condition that only takes $O(K(n-K))$ time to check (which is $O(n^2)$ in the worst case). Because Theorem 1's test can reject even when this condition does not hold, however, doing so can result in a loss of power. In adversarially chosen settings, this loss of power can be very large (see Appendix A for a numerical example). We mention that, when $K = 1$, as in our original motivating problem, Theorem 1's test only takes $O(n^2)$ time to run and Theorem 2's condition only takes $O(n)$ time to check.

Theorem 2 also clarifies a sense in which Theorem 1's error control is tight. Fixing some covariance Σ, if we set $\infty = \mu_1 = \ldots \mu_{K-1} > \mu_K = \mu_{K+1} > \mu_{K+2} = \cdots = \mu_n = -\infty$, then $\cup_{i \in S, j \notin S} H_{ij}^0$ is always true and any rejection is false. In this example, it will always be the case that $I = K$, $J = K+1$, and the two-sided difference-of-means test comparing X_K to X_{K+1} will falsely reject with probability exactly α (after all, this is the setting of a vanilla two-sided test). Theorem 1's test both (1) rejects whenever this difference-of-means test does and (2) still maintains error control, so it must falsely reject with probability exactly α as well. For the generalized version of Theorem 1 that works for any $\delta \in \mathbb{R}$, the same tightness can be achieved by setting μ_K and μ_{K+1} to be exactly δ apart.

Corollary 1 tells us that, when the data is independent or equicorrelated, Theorem 1's test rejects $\cup_{i \in S, j \notin S} H_{ij}^0$ exactly when Theorem 2's condition (2) is satisfied, i.e., rejecting $\cup_{i \in S, j \notin S} H_{ij}^0$ according to Theorem 2's condition results in no power loss. Also, in Appendix B, we show for the equicorrelated case that the indices I and J from the condition (2) are always those of the K and $(K+1)$st largest observations. We provide a proof of Corollary 1 in Section 2.3. It is specific to the special case $\delta = 0$, and has no analog when $\delta \neq 0$.

Corollary 1 (Exact equivalence)**.** *If Σ is diagonal or an equicorrelation matrix (i.e., Σ_{ij} is $\rho\sigma^2$ when $i = j$ and ρ when $i \neq j$), then the procedure from Theorem 1 rejects $\cup_{i \in S, j \notin S} H_{ij}^0$ if and only if the condition (2) from Theorem 2 is satisfied.*

There are a couple other notable situations where the conclusion of Corollary 1 applies. When $K = 1$ or $K = n - 1$ and the X_i have a small amount of autocorrelation (i.e., $\Sigma_{ij} = \sigma^2 \rho^{|i-j|}$ with $|\rho| \leq 1/2$), the result of Corollary 1 still holds. It also holds when $K = 1$ and we use a multivariate Gaussian distribution to approximate the joint distribution of t multinomial trials $Y \sim \text{Multinomial}(t, \pi_1, \ldots, \pi_n)$, i.e., we define $\hat{\pi} = Y/t$ and apply our method to $\hat{\pi} \dot{\sim} N(\pi, \hat{\pi}/t - \hat{\pi}\hat{\pi}^\top/t)$. Appendix B discusses these settings in more detail.

Now that we have a good grasp of Theorem 1's test and its behavior, we can establish why it is a surprisingly powerful procedure. For our rank verification problem, the natural alternative to our selective approach is to perform simultaneous inference. The most standard example of this is Tukey's honestly significant difference (HSD) test [Tukey, 1951]. Using the random indices I and J from Equation (2), Tukey's test, once adapted to our problem, would tell us to reject $\cup_{i \in S, j \notin S} H_{ij}^0$ whenever

$$X_I \geq X_J + v_{IJ} h_{1-\alpha}, \qquad h_{1-\alpha} = \text{Quantile}\left(1 - \alpha, \max_{i \neq j} \frac{|Z_i - Z_j|}{v_{ij}}\right) \text{ with } Z = X - \mu.$$

In contrast, Theorem 1's test is guaranteed to reject $\cup_{i \in S, j \notin S} H_{ij}^0$ whenever

$$X_I \geq X_J + v_{IJ} z_{1-\alpha/2}, \qquad z_{1-\alpha} = \text{Quantile}\left(1 - \alpha, Z\right) \text{ with } Z \sim N(0,1)$$

When $n = 2$, these two approaches coincide. But as soon as $n > 2$, the quantile $h_{1-\alpha}$ becomes strictly larger than $z_{1-\alpha/2}$, and our test's rejection region becomes a strict superset of Tukey's HSD rejection region. In the case that the X_i are independent, the growth of $h_{1-\alpha}$ is at least on the order of $\sqrt{\log n}$ (see Appendix C for justification). The quantile $z_{1-\alpha/2}$ that our procedure uses, however, stays fixed. In essence, our approach avoids a multiple comparisons correction that cannot be avoided by simultaneous inference

Drawing from the prior rank verification literature [Bofinger, 1983, 1985, Hsu, 1981, 1984], there is a variant of Tukey's HSD that is more powerful for our specific rank verification problem (although, to the best of our knowledge, it is not computationally tractable when Σ is not isotropic). We provide an analogous discussion for this variant in Appendix C. The story remains is identical. When $n > 2$ our procedure avoids a multiple comparisons correction that the this more powerful simultaneous procedure cannot, and our procedure's rejection region remains a strict superset of even this more powerful simultaneous approach's rejection region.

The remainder of the article is devoted to proving more general versions of the results in this section. One of them is a generalization of Theorem 1's method that applies for any $\delta \in \mathbb{R}$, not just $\delta = 0$. We argue in Appendix D that the smallest δ for which this generalized method fails to reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ provides a $1 - \alpha$ confidence lower bound for the gap $\min_{i \in S} \mu_i - \max_{j \neq S} \mu_j$ between the smallest mean in the selected set and the largest mean in the unselected set that is valid conditional on S. In practice, this δ can be found via a binary search. Appendix D also discusses how to leverage a more general version of Theorem 2 to get a less powerful, but more computationally easier confidence lower bound for this quantity.

1.3 Related work

Gutmann and Maymin [1987] study our problem in the case that $K = 1$, $\delta = 0$, and the data $X_i \sim N(\mu_i, \sigma^2)$ are independent Gaussian samples with common variance. They show that drawing the inference $\min_{i \in S} \mu_i - \max_{j \notin S} \mu_j > \delta$ whenever the two-sided difference-of-means test comparing the largest and second largest observation rejects is an error controlling procedure. Our work provides a complete generalization of their result in the case of multivariate Gaussian data, allowing for any K, any covariance structure, and any $\delta \in \mathbb{R}$. Work prior to Gutmann and Maymin [1987] studied related rank verification problems in similar settings [Bechhofer, 1954, Bofinger, 1983, 1985, Desu, 1970, Fabian, 1962, Gupta, 1965, 1956, Hsu, 1981, 1984], but used simultaneous inference techniques and failed to avoid a multiplicity correction as Gutmann and Maymin [1987] did. Follow-up work to Gutmann and Maymin [1987] includes methods that avoid multiplicity corrections for other rank verification problems [Gutmann, 1987, Maymin and Gutmann, 1992], but in similarly restricted settings. Also, Cheng and Panchapakesan [2009] extend Gutmann and Maymin [1987]'s result to the case of independent samples from a scale family with a monotone likelihood ratio (Gutmann and Maymin [1987] themselves handle the case of independent samples from a location family with a monotone likelihood ratio).

To prove their result Gutmann and Maymin [1987], condition on the index of the winning observation. This is a similar strategy to that of modern post-selection inference, a field initiated by the seminal work Lee et al. [2016]. By leveraging modern selective techniques, Hung and Fithian [2019] generalize Gutmann and Maymin [1987]'s procedure to apply for exponential families with Schur concave carrier measures (for $K = 1$ and any $\delta \in \mathbb{R}$). Schur concavity requires the carrier measure to be symmetric, so for the multivariate Gaussian case Hung and Fithian [2019] only generalize Gutmann and Maymin [1987]'s procedure from the independent to the equicorrelated setting (our Corollary 1 subsumes both cases). The main focus of Hung and Fithian [2019] is rank verification for multionmial data. If there are enough samples, then multinomial data can be approximated as correlated multivariate Gaussian data via the central limit theorem, and we show that our method behaves the same as theirs in Appendix B. Hung and Fithian [2019] also consider rank verification for the Bradley-Terry model, but their method (1) only scales to games with roughly $n = 40$ players and (2) requires each player to play the other players the same number of times (both conditions are violated in our Chatbot Arena motivating example). Work that is concurrent with and independent from ours considers multivariate Gaussian rank verification in the independent and unequal variance case [Goldwasser et al., 2025]. Indeed, once restricted to this case, Theorem 1's method matches that of Goldwasser et al. [2025]. Goldwasser et al. [2025], however, do not show that the resulting method amounts to running the two-sided difference-of-means test comparing the observation inside the top K and observation outside the top K with the smallest standardized difference (i.e., they have no analog of Theorem 2 or Corollary 1). As a consequence, they do not formally characterize the method's behavior or power.

2 Proofs

In this section, we prove more general versions of the results stated in Section 1. The more general result we are aiming to prove is stated clearly at the start of each proof. We will use

$$D_{ij}^{\delta} = \frac{(X_i - X_j) - \delta}{v_{ij}}$$

to denote the standardized distance between $X_i - X_j$ and δ. Note that $D_{ij}^0 = D_{ij}$, where D_{ij} is the standardized difference between X_i and X_j from the previous section.

2.1 Proof of Theorem 1

Considering the null hypotheses $H_{ij}^{\delta} : \mu_i - \mu_j \leq \delta$ we will show that rejecting the the data dependent union null $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ when

$$\max_{i \in S, j \notin S} \frac{\left[1 - \Phi(D_{ij}^{\delta})\right] - \left[1 - \Phi\left(\min_{\substack{k \in S,\, \ell \notin S: \\ \rho_{ij,k\ell} < 0}} D_{ij}^{\delta} - \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^{0}\right)\right]}{\left[1 - \Phi\left(\max_{\substack{k \in S,\, \ell \notin S: \\ \rho_{ij,k\ell} > 0}} D_{ij}^{\delta} - \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^{0}\right)\right] - \left[1 - \Phi\left(\min_{\substack{k \in S,\, \ell \notin S \\ \rho_{ij,k\ell} < 0}} D_{ij}^{\delta} - \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^{0}\right)\right]} \leq \alpha. \tag{3}$$

ensures that, conditional on S, the probability of a false rejection is at most α.

Without loss of generality, we perform our analysis conditional on the specific event $S = \{1, \ldots, K\}$ that the $X_1, \ldots, X_K$ are larger than the $X_{K+1}, \ldots X_n$. Our strategy will mimic that presented in Hung and Fithian [2019] and Sood [2024]. First, for pairs $i \leq K$ and $j > K$, we come up with a test for rejecting H_{ij}^{δ} that maintains error control conditional on $S = \{1, \ldots, K\}$. We reject $\cup_{i \leq K, j > K} H_{ij}^{\delta}$ when these tests reject for all $i \leq K$, $j > K$. Lemma 4 of Hung and Fithian [2019], which is adopted from Berger [1982], ensures that in doing so we maintain error control conditional on $S = \{1, \ldots, K\}$.

Fix some $i \leq K$ and $j > K$. To design a test for rejecting H_{ij}^{δ} that maintains error control conditional on $S = \{1, \ldots, K\}$, we will use the selective dominance machinery from Sood [2024]. Normally, to maintain marginal Type I error control, we reject the null H_{ij}^{δ} using the p-value

$$p_{ij}^{\delta} = 1 - \Phi(D_{ij}^{\delta}). \tag{4}$$

Defining

$$\epsilon_{ij,k\ell}^{\delta} = D_{k\ell}^{0} - \rho_{ij,k\ell} D_{ij}^{\delta}$$

it is straightforward to verify that the random vector ϵ_{ij}^{δ}, which consists of the entries $\epsilon_{ij,k\ell}^{\delta}$ for pairs $k \leq K$ and $\ell > K$, is independent of D_{ij}^{δ} and therefore also p_{ij}^{δ}. Since the p-value (4) corresponds to running a one-sided uniformly most powerful (UMP) test in a monotone likelihood ratio family (MLR), Example 3 in Sood [2024] tells us that it is selectively dominant (see [Sood, 2024, Definition 1]) given ϵ_{ij}^{δ}.

All that remains to do is characterize when we are selecting the p-value p_{ij}^{δ} to use for inference (i.e., when $S = \{1, \ldots, K\}$, what values is this p-value taking?). Theorem 1 from Sood [2024] then tells how to adjust the p-value to get a selective p-value. Rejecting H_{ij}^{δ} when this selective p-value is below α maintains error control conditional on $S = \{1, \ldots, K\}$.

To do so, we consider $k \leq K$ and $\ell > K$ and rewrite the event

$$\begin{aligned} X_k > X_\ell &\iff D_{k\ell}^{0} > 0 \\ &\iff \epsilon_{ij,k\ell}^{\delta} + \rho_{ij,k\ell} D_{ij}^{\delta} > 0 \end{aligned}$$

This leads to three cases:

1. If $\rho_{ij,k\ell} > 0$ then $X_k > X_\ell \iff D^\delta_{ij} > -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,k\ell}$,
2. If $\rho_{ij,k\ell} = 0$ then $X_k > X_\ell \iff \epsilon^\delta_{ij,k\ell} > 0$,
3. If $\rho_{ij,k\ell} < 0$ then $X_k > X_\ell \iff D^\delta_{ij} < -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,k\ell}$

Ultimately, we see that

$$\begin{aligned} S = \{1, \dots, K\} &\iff X_k > X_\ell \text{ for all } k \le K \text{ and } \ell > K \\ &\iff D^\delta_{ij} \in \left[\max_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} > 0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,jk}, \min_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} < 0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,jk}\right] \text{ and } \min_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} = 0}} \epsilon^\delta_{ij,k\ell} > 0. \end{aligned}$$

Essentially, the selection event $S = \{1, \dots, K\}$ corresponds to selecting p^δ_{ij} (4) to use for inference when it lives in some closed interval $[A, B]$ with bounds that are a measurable function of ϵ^δ_{ij}. In this case, Theorem 1 from Sood [2024] tells us that the selective p-value is $(p^\delta_{ij} - A)/(B - A)$. Writing everything out explicitly, the selective p-value is

$$p^\delta_{sel,ij} = \frac{\left[1 - \Phi(D^\delta_{ij})\right] - \left[1 - \Phi\left(\min\limits_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} < 0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,jk}\right)\right]}{\left[1 - \Phi\left(\max\limits_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} > 0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,jk}\right)\right] - \left[1 - \Phi\left(\min\limits_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} < 0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^\delta_{ij,jk}\right)\right]}.$$

Recalling the definition of $\epsilon_{ij,k\ell}$, we can rewrite this selective p-value as

$$p^\delta_{sel,ij} = \frac{\left[1 - \Phi(D^\delta_{ij})\right] - \left[1 - \Phi\left(\min\limits_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} < 0}} D^\delta_{ij} - \frac{1}{\rho_{ij,k\ell}}D^0_{k\ell}\right)\right]}{\left[1 - \Phi\left(\max\limits_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} > 0}} D^\delta_{ij} - \frac{1}{\rho_{ij,k\ell}}D^0_{k\ell}\right)\right] - \left[1 - \Phi\left(\min\limits_{\substack{k\le K,\, \ell > K: \\ \rho_{ij,k\ell} < 0}} D^\delta_{ij} - \frac{1}{\rho_{ij,k\ell}}D^0_{k\ell}\right)\right]}. \tag{5}$$

Our proposed procedure of rejecting $\cup^n_{i \le K, j > K} H^\delta_{ij}$ whenever (3) holds corresponds exactly to rejecting when all these selective p-values are at most α, establishing the validity of the procedure.

2.2 Proof of Theorem 2

Let $I^\delta \in S$, $J^\delta \in S$ be indices that minimize D^δ_{ij} over $i \in S$, $j \notin S$, i.e., $D^\delta_{I^\delta J^\delta} = \min_{i \in S, j \notin S} D^\delta_{ij}$. We will show that, defining $\delta^+ = \max(\delta, 0)$, the condition (3) is satisfied whenever

$$1 - \Phi(D^{\delta^+}_{I^{\delta^+} J^{\delta^+}}) \le \alpha/2. \tag{6}$$

We again without loss of generality perform our analysis conditional on the specific event $S = \{1, \dots, K\}$. Again fix $i \le K$ and $j > K$. Recall that if $a > b > c$, then $(b - c)/(a - c) \le b/a$. Using this fact, we can bound every selective p-value (5):

$$\begin{aligned}
&\frac{\left[1-\Phi(D^{\delta}_{ij})\right]-\left[1-\Phi\left(\min\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)\right]}{\left[1-\Phi\left(\max\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}} D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)\right]-\left[1-\Phi\left(\min\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)\right]}. \\
&\le \frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi\left(\max\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}} D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)} \\
&= \max_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}} \frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi\left(D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)}.
\end{aligned}$$

We restrict our attention to pairs $k \le K$ and $\ell > K$ such that $\rho_{ij,k\ell} > 0$ and further bound the term inside the maximium in two separate cases.

Case one: $\boldsymbol{D^{\delta}_{ij} \le D^{0}_{k\ell}}$. Since $\frac{1}{\rho_{ij,k\ell}} \ge 1$, in this case we have $D^{\delta}_{ij} - \frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell} \le 0$, so

$$\frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi(D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell})} \le \frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi(0)} \le 2(1-\Phi(D^{\delta}_{I^{\delta}J^{\delta}}))$$

Case two: $\boldsymbol{D^{\delta}_{ij} > D^{0}_{k\ell}}$. To handle this case we first show that

$$f(x) = \frac{1-\Phi(x)}{1-\Phi(x-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell})}$$

is a non-increasing function of x. The derivative of the function is

$$f'(x) = \frac{\phi(x)\phi\left(x-\frac{1}{\rho_{ij,j\ell}}D^{0}_{k\ell}\right)}{\left(1-\Phi\left(x-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)^{2}\right)}\left(\frac{1-\Phi(x)}{\phi(x)}-\frac{1-\Phi\left(x-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)}{\phi\left(x-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)}\right) \le 0$$

where the inequality follows from the fact that the Mills ratio $(1-\Phi(x))/\phi(x)$ is strictly decreasing [Baricz, 2008, Mills, 1926], and we always have

$$x > x - \frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}$$

because $\rho_{ij,k\ell} \ge 0$ and $D^{0}_{k\ell} \ge 0$. The non-positiveness of the derivative and the fact that $D^{\delta}_{ij} > D^{0}_{k\ell}$ implies that

$$\begin{aligned}
\frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi(D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell})} &\le \frac{1-\Phi(D^{0}_{k\ell})}{1-\Phi(D^{0}_{k\ell}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell})} \\
&\le \frac{1-\Phi(D^{0}_{k\ell})}{1-\Phi(0)} \\
&\le 2(1-\Phi(D^{0}_{I^{0}J^{0}})).
\end{aligned}$$

where we have that $D^{0}_{k\ell} - \frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell} \le 0$ because $\frac{1}{\rho_{ij,k\ell}} \ge 1$ and $D^{0}_{k\ell} \ge 0$.

If $\delta \geq 0$, we have that $D^{\delta}_{I^{\delta}J^{\delta}} \leq D^{\delta}_{k\ell} \leq D^{0}_{k\ell}$ for all $k \leq K$ and $\ell > K$. Thus $D^{\delta}_{I^{\delta}J^{\delta}} \leq D^{0}_{I^{0}J^{0}}$ and we can combine our earlier two cases:

$$\begin{aligned}
&2(1-\Phi(D^{0}_{I^{0}J^{0}})) \leq 2(1-\Phi(D^{\delta}_{I^{\delta}J^{\delta}})) \\
&\implies \max_{\substack{k\leq K,\, \ell>K:\\ \rho_{ij,k\ell}>0}} \frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi\left(D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)} \leq 2(1-\Phi(D^{\delta}_{I^{\delta}J^{\delta}})).
\end{aligned}$$

On the other hand, if $\delta < 0$, then $D^{0}_{I^{0}J^{0}} \leq D^{0}_{k\ell} \leq D^{\delta}_{k\ell}$ for all $k \leq K$ and $\ell > K$. Thus $D^{0}_{I^{0}J^{0}} \leq D^{\delta}_{I^{\delta}J^{\delta}}$ and we can again combine our earlier two cases:

$$\begin{aligned}
&2(1-\Phi(D^{\delta}_{I^{\delta}J^{\delta}})) \leq 2(1-\Phi(D^{0}_{I^{0}J^{0}})) \\
&\implies \max_{\substack{k\leq K,\, \ell>K:\\ \rho_{ij,k\ell}>0}} \frac{1-\Phi(D^{\delta}_{ij})}{1-\Phi\left(D^{\delta}_{ij}-\frac{1}{\rho_{ij,k\ell}}D^{0}_{k\ell}\right)} \leq 2(1-\Phi(D^{0}_{I^{0}J^{0}})).
\end{aligned}$$

This is sufficient to imply the result.

2.3 Proof of Corollary 1

For this part of the argument, we fix $\delta = 0$ and prove exactly the statement in Corollary 1. To show Corollary 1, we will show that, under the specified conditions, the left-hand side of (3) is exactly equal to $2(1-\Phi(D_{IJ}))$, where I and J are from Theorem 2's statement and identical to I^0 and J^0 from the previous proof.

If Σ is diagonal or an equicorrelation matrix, then it is straightforward to check for $i, k \neq j, \ell$ that $\rho_{ij,k\ell} \geq 0$. This simplifies our procedure considerably. Again, without loss of generality, we perform our analysis conditional on $S = \{1, \ldots, K\}$. Because, $\delta = 0$ and $\rho_{ij,k\ell} \geq 0$ for $i, k \leq K$ and $j, \ell > K$, the condition (3) is satisfied when

$$\max_{i\leq K, j>K} \frac{1-\Phi(D_{ij})}{1-\Phi\left(\max\limits_{\substack{k\leq K,\, \ell>K:\\ \rho_{ij,k\ell}>0}} D_{ij}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)} \leq \alpha.$$

We still have the bound

$$\max_{i\leq K, j>K} \frac{1-\Phi(D_{ij})}{1-\Phi\left(\max\limits_{\substack{k\leq K,\, \ell>K:\\ \rho_{ij,k\ell}>0}} D_{ij}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)} \leq 2(1-\Phi(D_{IJ})),$$

from proof of Theorem 2. It is also the case that

$$\begin{aligned}
\max_{i\leq K, j>K} \frac{1-\Phi(D_{ij})}{1-\Phi\left(\max\limits_{\substack{k\leq K,\, \ell>K:\\ \rho_{ij,k\ell}>0}} D_{ij}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)} &\geq \frac{1-\Phi(D_{IJ})}{1-\Phi\left(\max\limits_{\substack{k\leq K,\, \ell>K:\\ \rho_{ij,k\ell}>0}} D_{IJ}-\frac{1}{\rho_{IJ,k\ell}}D_{k\ell}\right)} \\
&\geq \frac{1-\Phi(D_{IJ})}{1-\Phi(D_{IJ}-\frac{1}{\rho_{IJ,IJ}}D_{IJ})} \\
&= 2(1-\Phi(D_{IJ})),
\end{aligned}$$

so the two expressions are in fact equal. Thus, when $\delta = 0$ and we are in the specified settings, the procedure from Theorem 1 rejects exactly when

$$1-\Phi(D_{IJ}) \leq \alpha/2,$$

as desired.

Acknowledgements

The author would like to thank John Cherian, Tim Morrison, Yash Nair, and James Yang for helpful discussions.

References

Árpád Baricz. Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340(2):1362–1370, 2008. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2007.09.063. URL https://www.sciencedirect.com/science/article/pii/S0022247X07011730.

Robert E. Bechhofer. A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with known Variances. *The Annals of Mathematical Statistics*, 25(1):16 – 39, 1954. doi: 10.1214/aoms/1177728845. URL https://doi.org/10.1214/aoms/1177728845.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982. ISSN 00401706. URL http://www.jstor.org/stable/1267823.

Eve Bofinger. Multiple comparisons and selection. *Australian Journal of Statistics*, 25(2):198–207, 1983. doi: https://doi.org/10.1111/j.1467-842X.1983.tb00373.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1983.tb00373.x.

Eve Bofinger. Multiple comparisons and type iii errors. *Journal of the American Statistical Association*, 80 (390):433–437, 1985. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2287910.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Shuenn-Ren Cheng and S Panchapakesan. Is the selected population the best?—location and scale parameter cases. *Communications in Statistics—Theory and Methods*, 38(10):1553–1560, 2009.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

M Mahamunulu Desu. A selection problem. *The Annals of Mathematical Statistics*, 41(5):1596–1603, 1970.

Vaclav Fabian. On Multiple Decision Methods for Ranking Population Means. *The Annals of Mathematical Statistics*, 33(1):248 – 254, 1962. doi: 10.1214/aoms/1177704728. URL https://doi.org/10.1214/aoms/1177704728.

Jeremy Goldwasser, Will Fithian, and Giles Hooker. Gaussian rank verification, 2025. URL https://arxiv.org/abs/2501.14142.

Shanti S Gupta. On some multiple decision (selection and ranking) rules. *Technometrics*, 7(2):225–245, 1965.

Shanti Swarup Gupta. *On a Decision Rule for a Problem in Ranking Means*. Ph.d. dissertation, University of North Carolina at Chapel Hill, 1956.

Sam Gutmann. Tests uniformly more powerful than uniformly most powerful monotone tests. *Journal of Statistical Planning and Inference*, 17:279–292, 1987. ISSN 0378-3758. doi: https://doi.org/10.1016/0378-3758(87)90120-0. URL https://www.sciencedirect.com/science/article/pii/0378375887901200.

Sam Gutmann and Zakhar Maymin. Is the selected population the best? *The Annals of Statistics*, pages 456–461, 1987.

Laurens Haan and Ana Ferreira. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

Jason C. Hsu. Simultaneous Confidence Intervals for all Distances from the "Best". *The Annals of Statistics*, 9(5):1026 – 1034, 1981. doi: 10.1214/aos/1176345582. URL https://doi.org/10.1214/aos/1176345582.

Jason C. Hsu. Constrained simultaneous confidence intervals for multiple comparisons with the best. *The Annals of Statistics*, 12(3):1136–1144, 1984. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240990.

Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47 (2):758 – 782, 2019. doi: 10.1214/17-AOS1634. URL https://doi.org/10.1214/17-AOS1634.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

Zakhar Maymin and Sam Gutmann. Testing retrospective hypotheses. *Canadian Journal of Statistics*, 20 (3):335–345, 1992.

John P. Mills. Table of the ratio: Area to bounding ordinate, for any portion of normal curve. *Biometrika*, 18(3/4):395–400, 1926. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2331957.

Anav Sood. Selective inference is easier with p-values, 2024. URL https://arxiv.org/abs/2411.13764.

John W Tukey. Quick and dirty methods in statistics. part ii. simple analyses for standard designs. *American Society for Quality Control*, pages 189–197, 1951.

A Losing power for certain covariance structures

Theorem 1's procedure can be more powerful than just checking the condition in Theorem 2 when for $i \neq j$ and $k \neq \ell$, correlations between pairs of differences $X_i - X_j$ and $X_k - X_\ell$ are negative. We will consider the $K = 1$ problem, and generate data

$$\begin{aligned} X_1 &= \sigma_1 Z_1 + \mu_1, \\ X_2 &= \sigma_2 Z_2 + \mu_2, \\ X_j &= -\sigma_2 Z_2 + \sigma_3 Z_j + \mu_j \text{ for } j > 2, \end{aligned}$$

where Z_i are independent standard Gaussian random variables.

If we set $\mu_1 > \mu_2 > \mu_3 = \cdots = \mu_n$ properly, then the top $K = 1$ set S will often be $\{1\}$, and the maximum in Theorem 1's condition will often be achieved by $i = 1, j = 2$. By setting $\sigma_2 > \sigma_1$ and $\sigma_2 > \sigma_3$, we can ensure that $\rho_{12,1j}$ is very negative for $j > 2$. This will mean there is a lot of benefit to running Theorem 1's full test in place of Theorem 2's simpler approach. We instantiate this problem by setting $n = 5$, $\sigma_1^2 = 1$, $\sigma_2^2 = 5$, $\sigma_3^2 = 0.1$, $\mu_1 = 5$, $\mu_2 = 3$, $\mu_3 = \mu_4 = \mu_5 = 0$. Over $B = 10000$ simulated trials run at level $\alpha = 0.1$, Theorem 1's test has an empirical power of 1.0, whereas the empirical power of Theorem 2's simpler approach is just below 0.057. Here, power is defined to be the probability of rejecting conditional on $S = \{1\}$.

B Some special cases

We document some special cases where we can better characterize the behavior of our method.

B.1 Equicorrelation

Fix $\delta \geq 0$. We will argue that if Σ is an equicorrelation matrix, i.e., $\Sigma_{ii} = \sigma^2$ and $\Sigma_{ij} = \rho\sigma^2$ when $i \neq j$, then I^δ and J^δ from the proof of Theorem 2 are the indices of the K and $(K+1)$ largest entries of X respectively.

The indices I^δ and J^δ minimize

$$D_{ij}^\delta = \frac{(X_i - X_j) - \delta}{\sqrt{2\sigma^2 - 2\rho\sigma^2}}$$

over $i \in S$ and $j \notin S$. This is clearly minimized if we take i to be the index of the smallest entry in S (i.e., the Kth largest entry) and j to be the index of the largest entry outside of S (i.e., the $(K+1)$st largest entry).

B.2 Autocorrelation

We will show if Σ is an autocorrelation matrix for an AR(1) process, i.e., $\Sigma_{ij} = \sigma^2\rho^{|i-j|}$ for some $\rho \in (-1, 1)$, then the conclusion of Corollary 1 still holds so long as $K = 1$ or $K = n - 1$.

We show the result for $K = 1$. The argument for $K = n - 1$ is analogous. When $K = 1$ the proof of Corollary 1 would go through so long as we had for $i \neq j, \ell$ that $\rho_{ij,i\ell} \geq 0$. It suffices to ensure that

$$\begin{aligned} c_{ij,i\ell} &= \mathrm{Cov}(X_i - X_j, X_i - X_\ell) \\ &= \Sigma_{ii} - \Sigma_{ij} - \Sigma_{i\ell} + \Sigma_{j\ell} \\ &= \sigma^2 - \sigma^2\rho^{|i-j|} - \sigma^2\rho^{|i-\ell|} + \sigma^2\rho^{|j-\ell|} \end{aligned}$$

is at least zero. When $\rho \geq 0$ we have that $c_{ij,i\ell} \geq \sigma^2(1-2\rho)$. When $\rho \leq 0$, we have that $c_{ij,i\ell} \geq \sigma^2(1-|\rho|-\rho^2)$. Thus, as long as $\rho \in [\frac{1-\sqrt{5}}{2}, \frac{1}{2}]$, we have $c_{ij,i\ell} \geq 0$. This is achieved whenever $|\rho| \leq 1/2$.

B.3 Multinomial

Let $Y \sim \text{Multinomial}(t, \pi_1, \ldots, \pi_n)$ and suppose we define $\hat{\pi} = Y/t$ and use a Gaussian approximation $\hat{\pi} \sim N(\pi, \hat{\pi}/t - \hat{\pi}\hat{\pi}^\top/t)$. Consider applying our method to the observation $\hat{\pi}$ while using the covariance $\Sigma = \hat{\pi}/t - \hat{\pi}\hat{\pi}^\top/t$. We will show that the conclusion of Corollary 1 still holds so long as $K = 1$.

Without loss of generality, suppose that $\hat{\pi}$ is in sorted order, so $\hat{\pi}_1 \geq \cdots \geq \hat{\pi}_n$. The proof of Corollary 1 would go through so long as we had for $j, \ell > 1$ that $\rho_{1j,1\ell} \geq 0$. It suffices to see then that

$$\begin{aligned} c_{ij,i\ell} &= \mathrm{Cov}(X_i - X_j, X_i - X_\ell) \\ &= \Sigma_{ii} - \Sigma_{ij} - \Sigma_{i\ell} + \Sigma_{j\ell} \\ &= \frac{1}{t}(\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_1\hat{\pi}_j + \hat{\pi}_1\hat{\pi}_\ell - \hat{\pi}_j\hat{\pi}_\ell) \\ &= \frac{1}{t}(\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_1\hat{\pi}_j + (\hat{\pi}_1 - \hat{\pi}_j)\hat{\pi}_\ell) \\ &\geq 0 \end{aligned}$$

Fixing $\delta \geq 0$, we show in this same setting that, when $K = 1$, the indices I^δ and J^δ from the proof of Theorem 2 are always the indices of the largest and second largest entries of $\hat{\pi}$ respectively. Because $K = 1$ we know that $I^\delta = 1$ will be the index of the largest entry. The index J^δ must then minimize

$$D_{1j}^\delta = \frac{\sqrt{t}[(\hat{\pi}_1 - \hat{\pi}_j) - \delta]}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_j(1-\hat{\pi}_j) + 2\hat{\pi}_1\hat{\pi}_j}} = \frac{\sqrt{t}[(\hat{\pi}_1 - \hat{\pi}_j) - \delta]}{\sqrt{\hat{\pi}_1 + \hat{\pi}_j - (\hat{\pi}_1 - \hat{\pi}_j)^2}}$$

over $j > 1$. It is clear that setting $J^\delta = 2$ will minimize the numerator of D_{1j}^δ and also maximize its denominator, which suffices to establish our claim.

C Simultaneous approach

For the sake of comparison, we derive a simultaneous inference approach for the problem of drawing the inference $\min_{i\in S}\mu_i > \max_{j\notin S}\mu_j$. We base our approach off of that in Bofinger [1983, 1985] and Hsu [1981], which assume an isotropic covariance and derive an approach that dominates Tukey's HSD test. To handle the case with general covariances, we will have to come up with a slightly different procedure than what currently exists in the literature. It is not tractable, but it still helps inform us of the limits of simultaneous inference.

Let $Z = X - \mu$ be a centered version of X, so that $Z \sim N(0, \Sigma)$, and Π denote the set of permutations over n elements. Defining

$$q_{1-\alpha} = \max_{\pi\in\Pi} \text{Quantile}\left(1-\alpha, \max\left\{\frac{Z_{\pi^{-1}(K)} - Z_{\pi^{-1}(K+1)}}{v_{\pi^{-1}(K),\pi^{-1}(K+1)}}, \max_{\substack{i\geq K+1,\\ j\leq K}} \frac{Z_{\pi^{-1}(i)} - Z_{\pi^{-1}(j)}}{v_{\pi^{-1}(i),\pi^{-1}(j)}}\right\}\right), \tag{7}$$

we argue that we can safely draw the inference $\min_{i\in S}\mu_i > \max_{j\notin S}\mu_j$ whenever

$$X_I \geq X_J + v_{IJ} q_{1-\alpha},$$

where I and J are as defined in the statement of Theorem 2.

Define I' to be the largest index in S (corresponding to the smallest mean) and J' to be the smallest index not in S (corresponding to the largest mean). A false rejection happens exactly when $X_I > X_J + v_{IJ}q_{1-\alpha}$ and also $\mu_{I'} - \mu_{J'} \leq 0$. Note that $I' \geq K$, and $I' = K \implies J' = K+1$ and $I' > K \implies J' \leq K$. With this in mind, we can bound

$$\begin{aligned}
P(\text{false rejection}) &= P\left(\mu_{I'} - \mu_{J'} \leq 0, \frac{X_I - X_J}{v_{IJ}} - q_{1-\alpha} \geq 0\right) \\
&\leq P\left(\frac{\mu_{I'} - \mu_{J'}}{v_{I'J'}} \leq 0, \frac{X_{I'} - X_{J'}}{v_{I'J'}} - q_{1-\alpha} \geq 0\right) \\
&\leq P\left(\frac{\mu_{I'} - \mu_{J'}}{v_{I'J'}} \leq \frac{X_{I'} - X_{J'}}{v_{I'J'}} - q_{1-\alpha}\right) \\
&\leq P\left(q_{1-\alpha} \leq \frac{X_{I'} - \mu_{I'}}{v_{I'J'}} - \frac{X_{J'} - \mu_{J'}}{v_{I'J'}}\right) \\
&\leq P\left(q_{1-\alpha} \leq \frac{Z_{I'} - Z_{J'}}{v_{I'J'}}\right) \\
&\leq P\left(q_{1-\alpha} \leq \max\left\{\frac{Z_K - Z_{K+1}}{v_{K,K+1}}, \max_{\substack{i\geq K+1,\\ j\leq K}} \frac{Z_i - Z_j}{v_{ij}}\right\}\right) \\
&\leq \alpha
\end{aligned}$$

where the last inequality follows from the definition of $q_{1-\alpha}$.

Having proven the validity of a simultaneous method, we make two points. First, we could not use the $1-\alpha$ quantile of

$$\max\left\{\frac{Z_K - Z_{K+1}}{v_{K,K+1}}, \max_{\substack{i\geq K+1,\\ j\leq K}} \frac{Z_i - Z_j}{v_{ij}}\right\}$$

in our procedure because, in the non-isotropic covariance case, computing this quantile requires us to know how to order the samples by their means. Second, $q_{1-\alpha}$ is certainly at most

$$h_{1-\alpha} = \text{Quantile}\left(1-\alpha, \max_{i\neq j} \frac{|Z_i - Z_j|}{v_{ij}}\right), \tag{8}$$

which justifies the validity of the Tukey's HSD variant we proposed in the main text.

Now, let's compare Theorem 1's test to this simultaneous approach. It is easy to see that $q_{1-\alpha} \geq z_{1-\alpha/2}$, where the inequality is strict whenever $n > 2$. Thus Theorem 1's test has a rejection region that matches the simultaneous approach's rejection region when $n = 2$, and is a strict superset of it when $n > 2$.

In the case that the X_i are independent, we can more explicitly quantify the difference in the two rejection regions. Suppose that n is arbitrarily large and, without loss of generality, that $K < n/2$ (the case that $K \geq n/2$ can be handled with an identical argument). Let π be the permutation such that $\pi^{-1}(K)$ satisfies $\Sigma_{\pi^{-1}(K),\pi^{-1}(K)} \leq \Sigma_{m,m}$ for all m. Then,

$$\begin{aligned}
&\max\left\{\frac{Z_{\pi^{-1}(K)} - Z_{\pi^{-1}(K+1)}}{v_{\pi^{-1}(K),\pi^{-1}(K+1)}}, \max_{\substack{i \geq K+1,\\ j \leq K}} \frac{Z_{\pi^{-1}(i)} - Z_{\pi^{-1}(j)}}{v_{\pi^{-1}(i),\pi^{-1}(j)}}\right\} \\
&\geq \max_{i \geq K+1} \frac{Z_{\pi^{-1}(i)} - Z_{\pi^{-1}(K)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} \\
&\geq \max_{i \geq K+1} \frac{Z_{\pi^{-1}(i)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} - \max_{i \geq K+1} \frac{Z_{\pi^{-1}(K)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} \\
&\geq \max_{i \geq K+1} \frac{Z_{\pi^{-1}(i)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} - \max_{i \geq K+1} \frac{Z_{\pi^{-1}(K)} I(Z_{\pi^{-1}(K)} > 0)}{v_{\pi^{-1}(i),\pi^{-1}(K)}} \\
&= \max_{i \geq K+1} \frac{Z_{\pi^{-1}(i)} I(Z_{\pi^{-1}(i)} > 0)}{v_{\pi^{-1}(i),\pi^{-1}(K)}} - \frac{Z_{\pi^{-1}(K)} I(Z_{\pi^{-1}(K)} > 0)}{\sqrt{2}\sqrt{\Sigma_{\pi^{-1}(K),\pi^{-1}(K)}}} + o_p(1) \\
&\geq \max_{i \geq K+1} \frac{Z_{\pi^{-1}(i)} I(Z_{\pi^{-1}(i)} > 0)}{\sqrt{2}\sqrt{\Sigma_{\pi^{-1}(i),\pi^{-1}(i)}}} + O_p(1) \\
&\geq \frac{1}{\sqrt{2}} \max_{i \geq K+1} \frac{Z_{\pi^{-1}(i)}}{\sqrt{\Sigma_{\pi^{-1}(i),\pi^{-1}(i)}}} + O_p(1) \\
&= O(\sqrt{\log n}) + O_p(1)
\end{aligned}$$

where the last equality follows from applying standard extreme value theory results regarding the concentration of the maximum of independent standard Gaussians [Haan and Ferreira, 2006] and the fact that $n - K \geq n/2$ per our assumption. As a consequence, $q_{1-\alpha}$ must grow at least on the order of $\sqrt{\log n}$ as well. This implies that, in the independent case, the HSD quantile (8) grows at least on the order of $\sqrt{\log n}$ also.

D Getting a confidence lower bound

By inverting the test (3) for different values of δ (i.e., considering the set of δ for which we fail to reject), we get a confidence region for the gap $\min_{i \in S} \mu_i - \max_{j \neq S} \mu_j$ between the smallest mean in the selected set and the largest mean in the unselected set that is valid conditional on S. It is not immediately clear, however, that this region will result in a confidence lower bound (i.e., there is some smallest δ for which we fail to reject). We provide an argument that it does.

Appendix B.3 of Sood [2024] tells us that, because our original marginal p-values p_{ij}^{δ} in (4) come from the UMP test in a MLR family and because our selection event does not depend on the parameter δ we are testing, the selective p-values $p_{sel,ij}^{\delta}$ from (5) are non-decreasing in δ. If, for $i \in S$ and $j \notin S$ we define

$$\hat{\mu}_{ij} = \begin{cases} \infty, & p_{sel,ij}^{\delta} < \alpha \text{ for all } \delta, \\ -\infty, & p_{sel,ij}^{\delta} > \alpha \text{ for all } \delta, \\ \sup\{\delta : p_{sel,ij}^{\delta} = \alpha\} & \text{otherwise,} \end{cases}$$

then it is straightforward to argue that Theorem 1's procedure will fail to reject if and only if $\delta > \min_{i \in S, j \notin S} \hat{\mu}_{ij}$. Therefore the inverted confidence region does indeed correspond to a confidence lower bound.

Recalling I^{δ} and J^{δ} from the proof of Theorem 2, the more general result we prove in Theorem 2 implies that the following more computationally easier confidence lower bound for $\min_{i \in S} \mu_i - \max_{j \neq S} \mu_j$ is still valid conditional on S. First, check if $1 - \Phi(D^0_{I^0 J^0}) > \alpha/2$. If so, return $-\infty$. Otherwise, return the δ for which $\alpha/2 = 1 - \Phi(D^\delta_{I^\delta J^\delta})$. Noting that

$$\Phi(D^\delta_{I^\delta J^\delta}) = \min_{i \in S, j \notin S} 1 - \Phi\left(\frac{X_i - X_j - \delta}{v_{ij}}\right)$$

is the minimum of a finite number of Gaussian p-values from UMP one-sided testing, there will be some first time that this minimum equals $\alpha/2$. Keep in mind that, while easier to compute, this confidence lower bound will always be at least as large as the one that results from inverting the test (3) (this is an implication of the proof of Theorem 2).