Overall Rank Uncertainty for Correlated Populations

Shaine Rosewel Paralis Matala

A Thesis proposal submitted to UP Diliman School of Statistics

In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Statistics

School of Statistics
University of the Philippines
Month 2025

Contents

1 Introduction

Ranks are commonly of interest because they allow readers to compare populations based on estimates of interest. For example, top universities across the globe may be identified based on their institutional performance indicator, states may receive appropriate intervention according to their relative rank based on average travel times to work, and senatorial candidates who are likely to be granted a seat in the office can be reported by public opinion polling bodies prior to elections. Since ranks are computed from estimates rather than from their true, unknown values, it is implicit that their overall uncertainty—expressed through joint confidence intervals—should also be quantified. Individually, these intervals provide information on the possible range of each rank while collectively, they facilitate comparing all ranks simultaneously rather than reporting them in isolation.

Several studies have addressed this concern through different techniques. Some approaches like those by Klein et al. (2020), Mohamad et al. (2019), Mogstad et al. (2024), Andersson et al. (1998), and Lyhagen & Ahlgren (2020), relied solely on the estimates and their standard errors, constructing joint confidence intervals either for the estimated quantities or directly for the ranks themselves. Others incorporate model-based uncertainty to account for dependencies inherent in the data structure. This includes the works of Goldstein & Spiegelhalter (1996), who utilized conditioning through multilevel models where the ranked quantities are treated as residual effects. Hall & Miller (2009), on the other hand, developed a bootstrap algorithm that allows for the assumption of independence despite its potential violation.

Assuming independence when constructing joint confidence regions for estimators that are, in fact, correlated may lead to overly conservative and thus wider intervals, implying greater uncertainty—contrary to what is desired. It is therefore important to account for potential dependencies, as demonstrated by Goldstein & Spiegelhalter (1996). However, in their case, the estimators were treated as latent variables. In contrast, the present work focuses on estimators that are observable or directly measurable from the data.

A potential alternative approach, which to our knowledge has not yet been explored, is to allow for a certain degree of correlation and develop an algorithm capable of handling such dependencies while maintaining coverage close to the nominal level and producing relatively narrow joint confidence intervals. The proposed methodology uses only the observed estimators and their corresponding standard errors. Although it also employs a parametric bootstrap—commonly used to estimate overall rank uncertainties (e.g., Mohamad et al. (2019), Mogstad et al. (2024), Andersson et al. (1998), Lyhagen & Ahlgren (2020))—our implementation differs from these existing approaches.

1.1 Objective

This research builds upon Klein et al. (2020)'s methodology by extending the set of joint confidence intervals used to capture uncertainty in overall rankings. In particular, it intends to:

- Construct joint confidence intervals that utilize parametric bootstrap to obtain a narrow overall uncertainty for ranks.
- Establish joint confidence intervals for cases when ranks are assumed to be correlated.
- Evaluate the performance of the proposed approaches under different standard deviations, correlation structures, and dimensionalities.

1.2 Significance

In order to obtain joint confidence sets for overall ranks, Klein et al. (2020) requires estimating confidence intervals for the unknown parameters, with a joint coverage probability of at least $1 - \alpha$. Their goal is to produce confidence intervals that collectively produce a small difference between the upper and the lower bound to yield tighter joint uncertainty for ranks. In the same paper, they considered the set of familiar $\hat{\theta} \pm z_{\alpha/2} + SE_k$ individual confidence intervals, assuming an independently distributed $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$. This approach, while simple, disregards the idea that θ_ks may be correlated. In some cases, assuming independence in the presence of dependence leads to conservative confidence intervals resulting in wider intervals which imply a higher uncertainty in overall ranks.

For instance, in the case of ranking senatorial candidates in the Philippines, this assumption is limiting as it treats vote shares as statistically independent across contenders. Although senators are elected using Multiple Non-transferable Vote system (MNTV)—where candidates are voted for individually regardless of partisan membership and alliances (Ravanilla & Hicken (2023))—David & Legara (2015) demonstrated that candidate with name-recall advantage, such as media celebrities, incumbents, and members of dynastic families, received majority of the votes in the 2010 senatorial elections. In that year, media personalities Bong Revilla and Jinggoy Estrada secured the top spots. A similar pattern was observed in 2019, when Cynthia Villar and Grace Poe, both with prominent surnames, garnered the most votes; and again in 2022, when media figures Robin Padilla and Ramon Tulfo ranked among the top three. They also added that in weak-party systems, candidates who belong to the same political alliance or ticket commonly co-occur in ballots and hence perform with similarity.

Klein et al. (2020) also noted that although it is difficult to make generalizations about strong relationships between travel times to work, certain patterns are apparent. States with large unpopulated land areas and relatively few high-density population centers tend to report shorter travel times. In contrast, longer travel times are typically observed in highly urbanized states with large populations and high population densities. Geographic location also appears to play a role—for instance, many states with shorter travel times are located in the Mountain and Central regions, whereas majority of those with longer travel times are concentrated along the East Coast. These observations suggest the presence of potential spatial structures. Accounting for these patterns allows for a more realistic assessment of uncertainty in the estimated ranks for similar use cases.

1.3 Scope and Limitations

This study presents alternative approaches for constructing joint confidence regions to quantify uncertainty in overall ranks, building upon the main results of Klein et al. (2020). Specifically, it demonstrates the application of the parametric bootstrap and the incorporation of correlation among the populations being ranked, with an emphasis on maintaining sufficient coverage and tightness in the resulting overall uncertainty. However, several limitations must be acknowledged. First, the framework assumes that the data are generated from a multivariate normal distribution. Second, a set of correlation structures is examined to illustrate how different dependence assumptions may affect the resulting joint confidence sets; these structures are assumed rather than empirically estimated. Collectively, these limitations suggest that the findings should be regarded primarily as methodological illustrations.

2 Related Literature

2.1 Rank Uncertainty

In the problem of estimating ranks of several unknown real-valued parameters $\theta_1, \ldots, \theta_K$, $\hat{\mathbf{r}} = \mathbf{r}(\hat{\theta}_1, \ldots, \hat{\theta}_K)$ is a point estimate of $\mathbf{r}\,(\theta_1, \ldots, \theta_K)$. Naturally, this should be accompanied by a measure of uncertainty. Different approaches to quantify such uncertainty have been proposed in the literature. Some of them begin with the estimated values at hand while others employed techniques to first obtain estimates, then quantify uncertainty. Among the various approaches, the work of Klein et al. (2020) is discussed in greater detail, as it closely relates to the present study.

2.2 Klein's Joint Confidence Region for Overall Ranking Uncertainty

Klein et al. (2020) does not require knowledge of the sampling design or estimation procedure for each population. Instead, they used the estimates and their standard errors to construct joint confidence regions from which rank uncertainty is derived. This uses the idea that uncertainty in the ranks is determined by uncertainty in the parameters (Mogstad et al. (2024)).

2.2.1 Calculation of Overall Rank Uncertainty

Klein et al. (2020) quantified overall rank uncertainty using estimates of respondents' average travel time to work in each of K sampled geographical areas. They defined rank for the kth population as

$$r_k = \sum_{j=1}^{K} I(\theta_j \leq \theta_k) = 1 + \sum_{j:j \neq k} I(\theta_j \leq \theta_k), \qquad \text{for } k = 1, \ldots, K \tag{2.1}$$

Since true values, $\theta_1, \ldots, \theta_K$ are unknown, they assumed that for each $k \in \{1, 2, \ldots, K\}$, there exists L_k and U_k such that

$$\theta_k \in (L_k, U_k)$$

That is, they constructed the joint confidence region of the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ using their corresponding standard errors to estimate $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_K)$, where

$$\hat{r}_k = 1 + \sum_{j:j \neq k} I(\hat{\theta}_j \leq \hat{\theta}_k), \qquad \text{for } k = 1, \ldots, K$$

The estimated overall ranking is computed from the joint confidence region using

$$\left.\begin{array}{r} I_k = \{1, 2, \ldots, K\} - \{k\}, \\ \Lambda_{Lk} = \{j \in I_k : U_j \leq L_k\}, \\ \Lambda_{Rk} = \{j \in I_k : U_k \leq L_j\}, \\ \Lambda_{Ok} = \{j \in I_k : U_j > L_k \text{ and } U_k > L_j\} = I_k - \{\Lambda_{Lk} \cup \Lambda_{Rk}\} \end{array}\right\}$$

which implies the following:

1. $j \in \Lambda_{Lk} \leftrightarrow (L_j, U_j) \cap (L_k, U_k) = \emptyset$ and (L_j, U_j) lies to the left of (L_k, U_k);
2. $j \in \Lambda_{Rk} \leftrightarrow (L_j, U_j) \cap (L_k, U_k) = \emptyset$ and (L_j, U_j) lies to the right of (L_k, U_k);
3. $j \in \Lambda_{Ok} \leftrightarrow (L_j, U_j) \cap (L_k, U_k) \neq \emptyset$
4. $\Lambda_{Lk}, \Lambda_{Rk}$, and Λ_{Ok} are mutually exclusive, and $\Lambda_{Lk} \cup \Lambda_{Rk} \cup \Lambda_{Ok} = I_k$

Hence, for each $k \in \{1, 2, \ldots, K\}$, the joint confidence region for ranks is defined as

$$r_k \in \{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\} \tag{2.2}$$

It was noted that a smaller difference between U_k and L_k leads to a smaller $|\Lambda_{Ok}|$. Collectively, for all k, this yields narrower confidence intervals for the overall ranks, which is desirable.

They assumed a conservative confidence region whose joint coverage probability is at least as large as the nominal level, $1 - \alpha$, as shown in (2.3).

$$P\left[\bigcap_{k=1}^{K} \{\theta_k \in (L_k, U_k)\}\right] \geq 1 - \alpha \tag{2.3}$$

They showed this to result in a joint confidence set for the overall ranking, shown in (2.2), that also has a joint probability of at least $1 - \alpha$.

In line with this, they presented a proof demonstrating that if $(L_1, U_1), \ldots, (L_K, U_K)$ are constructed such that the estimator $\hat{\theta}_k \in (L_k, U_k)\ \forall k \in \{1, 2, \ldots, K\}$, then the estimated ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$ lies within the joint confidence region (2.2) with probability 1.

Moreover, they explained that (2.2) contains more than one possible overall ranking unless the values of θ_k differ from each other such that $(L_k, U_k) \cap (L_{k'}, U_{k'}) = \emptyset,\ \forall\ k \neq k'$. This implies that the unique overall ranking arises only from the narrowest attainable joint confidence region and it is the estimated ranking, $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$.

2.2.2 Construction of Joint Confidence Intervals for Parameters

Klein et al. (2020) used individual confidence intervals of the form $\hat{\theta}_k \pm z_{\alpha/2} SE_k^2$, with $\hat{\theta}_k \sim N(\theta_k, SE_k)$ for $k \in \{1, 2, \ldots, K\}$, where $\theta_1, \theta_2, \ldots \theta_k$ are unknown and $SE_1, \ldots, SE_k$ are known.

The first one can be traced from Theorem 1 of Šidák (1967) which states that for a vector of random variables of dimension K, $\mathbf{X} = (X_1, X_2, \ldots, X_K)$, with $\mathbf{X} \sim N_K(\mathbf{0}, \Sigma)$ and having an arbitrary correlation matrix $\mathbf{R} = \{\rho_{jk}\}_{j,k=1}^{K}$,

$$\begin{aligned} &P(|X_1| \leq c_1, \, \ldots, |X_K| \leq c_K) \geq \\ &\qquad\qquad P(|X_1| \leq c_1) \times P(|X_2| \leq c_2, \ldots, |X_K| \leq c_K), \\ &\qquad\qquad \text{for any positive numbers } c_1, c_2, \ldots, c_K \end{aligned}$$

Under Theorem 1 and by induction,

$$P(|X_1| \leq c_1, \, \ldots, |X_K| \leq c_K) \geq \prod_{k=1}^{K} P\left(|X_k| \leq c_k\right) \tag{2.4}$$

That is, the smallest confidence level that can be attained will always be $1 - \alpha$ and in cases of presence of dependence when independence is assumed, coverage will always be more than $1 - \alpha$.

For the simultaneous confidence intervals used by Klein, Šidák (1967) considered a random sample of n vectors of $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iK})'$, $i = 1, \ldots, n$ where $Y_{ik} \sim N(\mu_k, \sigma_k^2)$ with unknown μ_k and known σ_k^2 and stated that

$$X_k = \frac{\left(\hat{\theta}_k - \mu_k\right)}{\sigma_k \big/ \sqrt{n}} \sim N(0, 1), \quad k = 1, \ldots, K \tag{2.5}$$
$$\text{where } \hat{\theta}_k = \bar{Y}_k = n^{-1} \sum_{i=1}^{n} Y_{ik}$$

satisfies the requirements of Theorem 1. Hence, when constructing a simultaneous confidence interval for $\theta_k = \mu_k$, $\forall k \in \{1, 2, \ldots, K\}$ with $(1 - \alpha)$ confidence level, it follows from (2.4) and (2.5) that,

$$\begin{aligned} \prod_{k=1}^{K} P\left(|X_k| \leq c_k\right) &= \prod_{k=1}^{K} P\left(\hat{\theta}_k - c_k \cdot \frac{\sigma}{\sqrt{n}} \leq \theta_k \leq \hat{\theta}_k + c_k \cdot \frac{\sigma}{\sqrt{n}}\right) \\ &= \prod_{k=1}^{K} P\left(\hat{\theta}_k - c_k \cdot SE_k \leq \theta_k \leq \hat{\theta}_k + c_k \cdot SE_k\right) \\ &= 1 - \alpha \end{aligned}$$

As a result, this will always simultaneously yield a confidence level for $\left(\hat{\theta}_k \pm c_k \cdot SE_k\right)$ that is least as large as $(1-\alpha)$—being equal when independence holds and larger than $(1-\alpha)$ when dependence is actually present.

For the choice of c_k, Šidák advised to assume independence with $c_1 = \cdots = c_K = c_\gamma$ where γ is the individual significance level so that

$$\prod_{k=1}^{K} P\left(|X_k| \leq c_k\right) = \prod_{k=1}^{K} (1-\gamma) = (1-\gamma)^K = 1-\alpha$$

and deriving γ returns $1-(1-\alpha)^{1/K}$. Under this condition, the two-sided $100(1-\alpha)\%$ confidence interval for the parameter $\theta_k = \mu_k$ is simultaneously given for each $k \in \{1,\ldots,K\}$ by

$$\left(\hat{\theta}_k - z_{\gamma/2} SE_k,\ \hat{\theta}_k + z_{\gamma/2} SE_k\right), \qquad \text{for } k \in \{1,2,\ldots,K\} \qquad (2.6)$$
$$\text{where } z_{\gamma/2} = \Phi^{-1}\left(1-\frac{\gamma}{2}\right)$$

Šidák also suggested the use of Bonferroni inqeuality for the case when variances are unknown and unequal.This was demonstrated by Dunn (1958) as follows:

$$P(|X_1| \leq c_1,\ \ldots, |X_K| \leq c_K) \geq 1 - 2K\left[1-\Phi(c_\alpha)\right] = 1-\alpha$$

where solving for $c_\alpha = z_{\frac{\alpha}{2K}}$ gives $\Phi^{-1}\left(1-\frac{\alpha}{2K}\right)$ resulting in a conservative joint coverage for $\theta_1,\ldots,\theta_K$ of at least $1-\alpha$. The corresponding two-sided $100(1-\alpha)\%$ confidence intervals are as defined in (2.7).

$$\left(\hat{\theta}_k - z_{(\alpha/K)/2} SE_k,\ \hat{\theta}_k + z_{(\alpha/K)/2} SE_k\right), \qquad \text{for } k = 1,2,\ldots,K \qquad (2.7)$$

Klein used (2.6) and (2.7) to come up with the joint confidence region for $\hat{\theta}_1,\ldots,\hat{\theta}_K$. These became his basis to form the joint confidence set for the ranks, as explained in Section 2.2.1.

Since the subsequent discussions rely on the sampling variability of the estimated means rather than the population dispersion, we use σ_k (instead of SE_k) to denote the standard error of $\hat{\theta}_k$.

2.3 Alternative Approaches for Ranking Uncertainty

While Klein's approach provides one framework for constructing joint confidence regions for ranks, several other studies have explored related problems using different formulations or assumptions. These alternative methods vary in whether they account for dependence structures, rely on model-based estimation, or use resampling techniques such as the bootstrap.

2.3.1 Calculated Measure

Other studies with similar concern include that of Andersson et al. (1998) who suggested the use of a statistic C which quantifies the number of positions ranked populations would on average change their order due to random variation. They calculated the measure using a bootstrap approach. Since they worked on risk ratios p_k of K units, they drew B bootstrap proportions $\hat{p}_k^*$ from (2.8).

$$\hat{p}_{bk}^* \sim N\left(\hat{p}_k, \frac{\hat{p}_k(1-\hat{p}_k)}{n_k}\right), \quad k = 1, \dots, K;\ b = 1, \dots, B \tag{2.8}$$

For each bootstrap iteration b, they sorted $\hat{p}_{bk}^*$ to get the corresponding rank r_{bk}^* and calculated the difference d_{bk} between the original and bootstrap rank as $|\hat{r}_k - \hat{r}_{bk}^*|$ to obtain the expected change $\bar{d}_k$ in the ranking for unit k. $\bar{d}_k$ is in turn obtained by taking the average of d_{bk} across the bootstrap samples. Finally, the overall measure C is calculated as the average of $\bar{d}_k$ across all K units.

2.3.2 Pairwise Difference

Mohamad et al. (2019), requiring only the mean estimates and their corresponding standard errors similar to Klein, applied Tukey's Honest Significant Difference (HSD) to test $H_0 : \theta_j - \theta_k = 0$ for all $j \neq k \in 1, ..., K$ at level α. Tukey's HSD is typically used to provide simultaneous confidence statements about the differences between the means while controlling the family-wise error rate (FWER). This allowed them to come up with a joint confidence set for ranks expressed as

$$\left(1 + \#\left\{j : \frac{\hat{\theta}_j - \hat{\theta}_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} > q_{1-\alpha}\right\},\ K - \#\left\{j : \frac{\hat{\theta}_j - \hat{\theta}_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} < -q_{1-\alpha}\right\}\right), \tag{2.9}$$
$$\text{for } k = 1, 2, \dots, K$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of of the distribution of the studentized range,

$$\max_{j,k=1,\dots K} \frac{|\theta_j - \theta_k|}{\sqrt{\sigma_j^2 + \sigma_k^2}}$$

In (2.9), $\#\{\cdot\}$ counts the number of pairwise hypotheses that are rejected according to Tukey's HSD, which determines the lower and upper bounds of the confidence interval for the rank of $\hat{\theta}_k$.

They showed this to yield uniformly narrower intervals than that of Klein's, for the case when σ's are equal. It also has a simultaneous coverage of at least $1-\alpha$ and exactly $1-\alpha$ when all true performances are equal. However, their approach tends to be overly conservative, showing coverage levels between 0.996 and 1.0 at a 0.90 nominal level in simulations, when performances differ (i.e., there are no ties). They also demonstrated that as the true performance differences increase from 0 to 0.5, the coverage quickly increases from the nominal level to 1. As a remedy, they proposed a rescaling technique that brings the coverage closer to the nominal level, though it remains conservative (e.g., from 1.0 to 0.978, from 0.998 to 0.961—at 0.90 confidence level).

Mogstad et al. (2024) presented another technique that closely resembles the procedure by Klein and Mohamad. However, they defined ranks in the opposite way (i.e., larger rank value for lower estimate). They constructed the rectangular confidence region in (2.10), from the pairwise differences of estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ and an estimator of the variance of $\hat{\theta}_j - \hat{\theta}_k$, $\sqrt{\sigma_j + \sigma_k}$. $\hat{\theta}_1, \ldots, \hat{\theta}_K$ need not be independent. P_k is the distribution from which $\hat{\theta}_k$ is estimated; $\hat{P}_k$ denotes the estimate of P_k.

$$\begin{gathered} C(1-\alpha, S) = \prod_{(j,k)\in S} \left[\hat{\theta}_j - \hat{\theta}_k \pm \sqrt{\sigma_j + \sigma_k}\, L^{-1}(1-\alpha, S, \hat{P})\right], \\ S \subseteq \{(j,k) \in K \times K : j \neq k\} \end{gathered} \tag{2.10}$$

They added that if the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are jointly asymptotically normally distributed, then the quantiles $L^{-1}(1-\alpha, S, \hat{P})$, can be computed from the limiting distributions of the max-statistics shown in (2.11), through resampling methods. In particular, they obtained their $L^{-1}(1-\alpha, S, \hat{P})$ by repeatedly drawing K standard normal variates, recording the maximum for each draw, and taking the relevant quantile of these maxima.

$$L(x, S, P) = P\left\{\max_{(j,k)\in S} \frac{|\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)|}{\sqrt{\sigma_j + \sigma_k}} \leq x\right\}, \quad \Delta_{j,k} = \theta_j - \theta_k \tag{2.11}$$

When the population distribution P_k for $k \in \{1, \ldots, K\}$ is a set of distributions on $\mathbb{R}^K$ satisfying uniform integrability, using bootstrap leads to confidence sets that satisfy (2.12) when $\boldsymbol{\theta}$ is the population mean and $\hat{\boldsymbol{\theta}}$ is the sample mean.

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left\{\Delta_S(P) \in C(1-\alpha, S)\right\} \geq 1-\alpha \tag{2.12}$$

(2.10) is used to produce the overall uncertainty for ranks. For each k, they counted the number of intervals strictly greater than zero ($|N_k^+|$) and strictly less than zero ($|N_k^-|$). By *strictly*, they meant that the entire confidence interval does not cover zero. Their initial joint confidence set for ranks with at least $(1-\alpha)$ coverage is given by

$$R^{\text{joint}} = \left\{|N_k^-| + 1, \ldots, K - |N_k^+|\right\}, \qquad k \in \{1, \ldots, K\} \tag{2.13}$$

They further refined their approach through a stepwise multiple hypothesis testing algorithm that controls the mixed directional family-wise error rate (mdFWER) by counting the number of rejections for the null hypothesis $\Delta_{j,k}(P) = 0$ in favor of it being strictly smaller or larger than zero, rather than merely not equal to zero. They used these counts in place of $|N_k^-|$ and $|N_k^+|$ in (2.13) (see Algorithm 3.2 in Mogstad et al. (2024)). Under certain conditions (see Theorem 3.4 in Mogstad et al. (2024)),

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left\{r(P) \in R^{\text{joint}}\right\} \geq 1-\alpha \tag{2.14}$$

It should be noted that (2.12) and (2.14), only asymptotically controls the coverage probability (Bazylik et al. (2025)). Through simulations, they studied the finite sample performance of their proposed method and showed that the presence of ties led to a coverage closer to nominal level but the confidence interval for a given k contains all or almost all possible values of the rank. On the other hand, when there are no ties, the coverage frequency is close to one and the expected length is small. This is because in the absence of ties, the confidence sets for the differences all exclude zero and lie on the correct side of zero with probability approaching one as the sample size increases. Consequently, the resulting confidence sets for the ranks are small as they contain only the true rank with probability approaching one as the sample size increases. They also showed that their approach generally leads to a confidence set that is narrower than that of Klein's.

2.3.3 Accounting for Data Dependencies

Some approaches explicitly accounted for dependencies in the data. Goldstein & Spiegelhalter (1996) used multilevel models, in the context of ranking education and health institutions (e.g., schools, hospitals, medical practitioners, etc.), to address the hierarchical nature of data structures associated with institutional performance. Rank uncertainty was presented through a visualization in which non-overlap of confidence intervals conveyed a significant difference between compared institutions (Goldstein & Healy (1995)). In an alternative approach, along with institution effect estimation through Gibbs sampling, the rank was obtained for each iteration. Their example illustrated that while the multilevel model made individual estimates more accurate, it also had the effect of making the ranks even more uncertain.

Zhang et al. (2013) analyzed U.S. age-adjusted cancer incidence and mortality rates across states and counties by computing individual and overall simultaneous confidence intervals for age-adjusted health index using the Monte Carlo method. Because many health conditions are age-dependent, they used age-adjusted rates to minimize the confounding effect of age differences when comparing different population groups. They also extended their method to handle cases where only the adjusted rates and confidence intervals are available, aligning it more closely with the approach of Klein et al. (2020). Mohamad et al. (2019) showed their technique to result in joint confidence sets with very low coverage probabilities and which are only able to reach the nominal level when differences among the means are large enough.

Hall & Miller (2009) mentioned that in some use cases such as institutions ranking, dependencies can be accommodated through conditioning, similar to the above approaches. However, in genomics where data on expression levels of different genes from the same individual are generally not independent, they suggested using an "independent component" version of the bootstrap on the sample, where m-out-of-n bootstrap (m < n) is applied as though the ranked variables were statistically independent. They showed this to perform at its best when a reasonable level of correlation is present among the variables.

Bazylik et al. (2025), in their recent study, tackled the ranking of political candidates or parties using the estimated share of support each one receives in surveys. They used the multinomial distribution to develop confidence sets for finite samples and explored bootstrap in the case of approximately large samples. They addressed the dependence attributed to the success probabilities of different categories by using their proposed bootstrap algorithm. Their simulations showed that bootstrap-based confidence sets may have coverage probability below the nominal level despite them being excessively wide. In contrast, the finite-sample confidence sets have coverage probability at least as large as expected and may even be relatively shorter.

3 Methodology

This section introduces the proposed methodologies to obtain joint confidence intervals that can later be used to quantify uncertainty for the unknown overall true ranking using Klein's main result in Section 2.2.1. It adds approaches, on top of the Bonferroni correction and independence assumption in Section 2.2.2, by addressing the case when estimates being ranked are assumed correlated to certain degrees. Section 3.1 lists the algorithms employed to compute the joint confidence regions. These include a non-rank and rank-based methods. It also has a subsection that discusses correlation structures suitable to the intended use cases. The calculated joint confidence regions are then assessed on the basis of coverage and metrics that measure the tightness of estimated confidence regions. These are tackled in Section 3.2.

3.1 Parametric bootstrap approaches for constructing joint confidence intervals for correlated $\theta_1, \dots, \theta_K$

The proposed approaches only utilizes $\hat{\theta}_1, \dots, \hat{\theta}_K$ and their corresponding standard errors, $\sigma_1, \dots, \sigma_K$; knowledge of the sampling design and estimation methodology for each population is not required. These are constructed to account for assumed correlation among items being ranked. Hence, various correlation structures $\mathbf{R}$ are listed in section 3.1.3, to be later examined in a simulation study. The correlation matrix is used in the calculation of the covariance matrix as show in (3.1).

$$\boldsymbol{\Sigma} = \boldsymbol{\Delta}^{1/2}\mathbf{R}\boldsymbol{\Delta}^{1/2}; \quad \boldsymbol{\Delta} = \text{diag}\left\{\sigma_1^2, \sigma_2^2, \dots, \sigma_K^2\right\} \tag{3.1}$$

with assumed $\mathbf{R}$. This form of $\boldsymbol{\Sigma}$ will be used in Sections 3.1.1 and 3.1.2.

We primarily use parametric bootstrap to approximate the quantile that will be used to construct the confidence intervals while controlling the FWER to be around the nominal level. The design closely parallels that of Andersson et al. (1998) and Leyland & Langford (Goldstein & Spiegelhalter (1996)), who generated bootstrap samples from a normal distribution by applying the plug-in principle (Efron & Tibshirani (1993)) of using the observed estimator and its corresponding standard error as parameters. In our case however, correlation is assumed. Hence, we sample from the multivariate normal distribution, with the vector of estimates, $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K\right)'$, as mean and $\boldsymbol{\Sigma}$ as defined in (3.1).

Common across the proposed procedure is calculating $\left|\frac{\hat{\theta}_{bk}^* - \hat{\theta}_k}{\sigma_k}\right|$ and taking its maximum across $k \in \{1, \ldots, K\}$. This step keeps the coverage of the rectangular confidence region approximately equal to the nominal level. The idea is also conceptually similar to that of Mogstad et al. (2024) who used resampling to obtain the quantile as described in Section 2.3.2.

3.1.1 Nonrank-based method

The nonrank-based method, as implied by its name, does not incorporate order statistics in the algorithm. It focuses on the minimum requirement of constructing a sampling distribution from which the $(1-\alpha)$-quantile that keeps the simultaneous coverage at the nominal level will be derived.

Algorithm 1 Computating the joint confidence region (nonrank)

```
Let the data be represented by $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K\right)'$ and suppose that $\boldsymbol{\Sigma}$ is known
1: for $b = 1, 2, \ldots, B$ do
2:     Generate $\hat{\boldsymbol{\theta}}_b^* \sim N_K\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}\right)$ and write $\hat{\boldsymbol{\theta}}_b^* = \left(\hat{\theta}_{b1}^*, \hat{\theta}_{b2}^*, \ldots, \hat{\theta}_{bK}^*\right)'$
3:     Compute
           $$t_b^* = \max_{1 \le k \le K} \left|\frac{\hat{\theta}_{bk}^* - \hat{\theta}_k}{\sigma_k}\right|$$
4: end for
5: Compute the $(1 - \alpha)$-sample quantile of $t_1^*, t_2^*, \ldots, t_B^*$, call this $\hat{t}$.
6: The joint confidence region of $\theta_1, \theta_2, \ldots, \theta_K$ is given by
           $$\mathfrak{R} = \left[\hat{\theta}_1 \pm \hat{t} \times \sigma_1\right] \times \left[\hat{\theta}_2 \pm \hat{t} \times \sigma_2\right] \times \cdots \times \left[\hat{\theta}_K \pm \hat{t} \times \sigma_K\right]$$
```

3.1.2 Rank-based methods

For the rank-based methods, order statistics are considered for the bootstrap sampled estimates. That is, for each bootstrap b, the estimates are sorted in increasing order. Similar to nonrank-based method, the data is consist of $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K\right)'$ and a known $\boldsymbol{\Sigma}$.

3.1.2.1 Asymptotic variance The asymptotic definition of variance is employed in Algorithm 2 since it is unknown for $\hat{\theta}_{(k)}$.

3.1.2.2 Variance from second-level bootstrap As an alternative to using the asymptotic variance, a second-level (or double) bootstrap can be employed to estimate the variance, as illustrated in Algorithm 3. However, this approach is computationally intensive.

Algorithm 2 Computing the joint confidence region (asymptotic variance)

```
for $b = 1, 2, \ldots, B$ do
    Generate $\hat{\boldsymbol{\theta}}_b^* = \left(\hat{\theta}_{b1}^*, \hat{\theta}_{b2}^*, \ldots, \hat{\theta}_{bK}^*\right)' \sim N_K\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}\right)$ and let $\hat{\theta}_{b(1)}^*, \hat{\theta}_{b(2)}^*, \ldots, \hat{\theta}_{b(K)}^*$ be the corresponding ordered values
    Compute
        $\hat{\sigma}_{b(k)}^* = \sqrt{\left[\text{kth ordered value among } \left\{\hat{\theta}_{b1}^{*2} + \sigma_1^2, \hat{\theta}_{b2}^{*2} + \sigma_2^2, \ldots, \hat{\theta}_{bK}^{*2} + \sigma_K^2\right\}\right] - \hat{\theta}_{(k)}^{*2}}$
    Compute $t_b^* = \max_{1 \le k \le K} \left| \frac{\hat{\theta}_{b(k)}^* - \hat{\theta}_k^*}{\hat{\sigma}_{b(k)}^*} \right|$
end for
Compute the $(1-\alpha)$-sample quantile of $t_1^*, t_2^*, \ldots, t_B^*$, call this $\hat{t}$.
The joint confidence region of $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}$ is given by
    $\mathfrak{R} = \left[\hat{\theta}_{(1)} \pm \hat{t} \times \hat{\sigma}_{(1)}\right] \times \left[\hat{\theta}_{(2)} \pm \hat{t} \times \hat{\sigma}_{(2)}\right] \times \cdots \times \left[\hat{\theta}_{(K)} \pm \hat{t} \times \hat{\sigma}_{(K)}\right]$
where $\hat{\sigma}_{(k)}$ is computed as
    $\hat{\sigma}_{(k)} = \sqrt{\text{kth ordered value among } \left\{\hat{\theta}_1^2 + \sigma_1^2, \hat{\theta}_2^2 + \sigma_2^2, \ldots, \hat{\theta}_K^2 + \sigma_K^2\right\} - \hat{\theta}_{(k)}^2}$
```

Algorithm 3 Computing the joint confidence region (variance from double bootstrap)

```
for $b = 1, 2, \ldots, B$ do
    Generate $\hat{\boldsymbol{\theta}}_b^* = \left(\hat{\theta}_{b1}^*, \hat{\theta}_{b2}^*, \ldots, \hat{\theta}_{bK}^*\right)' \sim N_K\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}\right)$ and let $\hat{\theta}_{b(1)}^*, \hat{\theta}_{b(2)}^*, \ldots, \hat{\theta}_{b(K)}^*$ be the corresponding ordered values of $\hat{\theta}_{b1}^*, \hat{\theta}_{b2}^*, \ldots, \hat{\theta}_{bK}^*$
    for $c = 1, 2, \ldots, C$ do
        Generate $\hat{\boldsymbol{\theta}}_{bc}^{**} = \left(\hat{\theta}_{bc1}^{**}, \hat{\theta}_{bc2}^{**}, \ldots, \hat{\theta}_{bcK}^{**}\right) \sim N_K\left(\hat{\boldsymbol{\theta}}_b^*, \boldsymbol{\Sigma}\right)$ and let $\hat{\theta}_{bc(1)}^{**}, \hat{\theta}_{bc(2)}^{**}, \ldots, \hat{\theta}_{bc(K)}^{**}$ be the corresponding ordered values of $\hat{\theta}_{bc1}^{**}, \hat{\theta}_{bc2}^{**}, \ldots, \hat{\theta}_{bcK}^{**}$
        Compute $\hat{\sigma}_{b(k)}^* = \frac{\sum_{c=1}^{C}\left(\hat{\theta}_{bc(k)}^{**} - \bar{\hat{\theta}}_{b\cdot(k)}^{**}\right)^2}{C-1}, \quad \bar{\hat{\theta}}_{b\cdot(k)}^{**} = \frac{1}{C}\sum_{c=1}^{C}\hat{\theta}_{bc(k)}^{**}$
    end for
    Compute $t_b^* = \max_{1 \le k < K} \left| \frac{\hat{\theta}_{b(k)}^* - \hat{\theta}_{(k)}}{\hat{\sigma}_{b(k)}^*} \right|$
end for
Compute the $(1-\alpha)$-sample quantile of $t_1^*, t_2^*, \ldots, t_B^*$, call this $\hat{t}$.
The joint confidence region of $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}$ is
    $\mathfrak{R} = \left[\hat{\theta}_{(1)} \pm \hat{t} \times \hat{\sigma}_{(1)}\right] \times \left[\hat{\theta}_{(2)} \pm \hat{t} \times \hat{\sigma}_{(2)}\right] \times \cdots \times \left[\hat{\theta}_{(K)} \pm \hat{t} \times \hat{\sigma}_{(K)}\right]$
where $\hat{\sigma}_{(k)}$ is computed as
    $\hat{\sigma}_{(k)} = \frac{\sum_{b=1}^{B}\left(\hat{\theta}_{b(k)}^* - \bar{\hat{\theta}}_{\cdot(k)}^*\right)^2}{B-1}, \quad \bar{\hat{\theta}}_{\cdot(k)}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{b(k)}^*$
```

3.1.3 Correlation structures

This section discusses the correlation structures considered in the simulation. Since the estimation of correlation matrices is beyond the scope of this study, it is enough to assure that any assumed matrix of correlation is indeed valid a correlation matrix $\mathbf{R}$ satisfying the following:

- $\mathbf{R}$ is nonnegative definite (or positive semidefinite)
- $0 \leq |\mathbf{R}| \leq 1$
- If $|\mathbf{R}| = 1$ then $\mathbf{R} = \mathbf{I}$

For simplicity, an equicorrelation matrix in (3.2) is included. This assumes that the k variables are equally correlated, such that $\rho_{jk} = \rho$ where $\rho \in [-1, 1]$ for $j \neq k \in \{1, \ldots, K\}$.

$$\mathbf{R}_{\text{eq}} = (1-\rho)\,\mathbf{I}_K + \rho \mathbf{1}_K \mathbf{1}_K' = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}_{K \times K} \tag{3.2}$$

In a block correlation matrix $\mathbf{R}_{block}$ with G blocks, as represented by Archakova & Hansen (2020), the correlation between any two variables is determined by the block to which the two variables belong. Each diagonal block represents an equicorrelation structure within group g, denoted by

$$\mathbf{R}_{\text{eq,g}} = (1-\rho_g)\,\mathbf{I}_{n_g} + \rho_g \mathbf{1}_{n_g} \mathbf{1}_{n_g}'$$

where ρ_g is the within-block correlation and n_g is the number of variables in block g such that $\sum_{g=1}^{G} n_g = K$. The off-diagonal blocks capture between-block correlations, represented by

$$\mathbf{C}_{g'g} = \mathbf{C}_{gg'} = \rho_{gg'} \mathbf{1}_{n_g} \mathbf{1}_{n_g}' \\ \text{where } g \neq g' \in \{1, \ldots, G\}$$

Thus, the full block correlation matrix can be expressed as in (3.3).

$$\mathbf{R}_{\text{block}} = \begin{bmatrix} \mathbf{R}_{eq,1} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1G} \\ \mathbf{C}_{11} & \mathbf{R}_{eq,2} & \cdots & \mathbf{C}_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{G1} & \mathbf{C}_{G2} & \cdots & \mathbf{R}_{eq,G} \end{bmatrix}_{K \times K} \tag{3.3}$$

In the context of pre-election surveys, each block may represent correlations induced by party or ticket membership, reflecting stronger associations within parties and weaker associations between them.

Correlation structures that account for spatial proximity can be borrowed from geostatistics. This is particularly relevant in light of Klein's observation that states located within certain regions exhibit similar travel time characteristics. In such cases, spatial dependence can be modeled using a stationary (i.e., no directional dependence) Matérn correlation function, which for two locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is expressed as in (3.4).

$$\rho_{\text{matern}} = \frac{2^{1-\nu}}{\Gamma(\nu)}(\kappa \parallel \mathbf{s}_i - \mathbf{s}_j \parallel)^{\nu} K_{\nu}(\kappa \parallel \mathbf{s}_i - \mathbf{s}_j \parallel) \tag{3.4}$$

where $\|\cdot\|$ denotes the Euclidean distance and K_ν is the second kind of the modified Bessel function. It has a scale parameter $\kappa > 0$ and a smoothness parameter $\nu > 0$. ρ_{matern} reduces to the exponential correlation when $\nu = 0.5$ and to Gaussian correlation function when $\nu = \infty$. In this paper, the R package "BayesNGSP" (Turek & Risser (2022)), is used to construct the $\mathbf{R}_{\text{matern}}$.

3.2 Evaluation

Algorithm 4 is employed to estimate the coverage, which corresponds to the proportion of replications in which the true parameter values are contained within the confidence intervals for all K simultaneously. Likewise,the tightness of the joint confidence region is is assessed using three summary measures: the arithmetic mean (T_1), geometric mean (T_2), and the metric T_3 introduced by Wright (2025), as presented in Equations 3.5–3.7.

$$T_1 = \frac{1}{K}\sum_{k=1}^{K}\left|\Lambda_{Ok}\right| \tag{3.5}$$

$$T_2 = \prod_{k=1}^{K}\left|\Lambda_{Ok}\right| \tag{3.6}$$

$$T_3 = 1 - \frac{OP}{K^2} \tag{3.7}$$

In equation 3.7, $OP = K + \sum_{k=1}^{K}\left|\Lambda_{Ok}\right|$ denotes the total number of occupied positions in a joint confidence region out of the total number of positions K^2; or the sum of the differences between the upper and lower bound of the simultaneous rank intervals added by 1, for each population k. Higher values of T_1 and T_2 indicate wider confidence intervals and are therefore less desirable, whereas higher values of T_3 are preferable. T_3 can range from 0, indicating no tightness, to $\frac{K-1}{K}$, implying the confidence region only contains the estimated ranking which is likely the true ranking.

Algorithm 4 Computing the coverage probability and tightness measures

```
For given values of $\mathbf{\Sigma}$ and $\theta_1, \theta_2, \ldots, \theta_K$ (with corresponding $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}$ for rank-based methods)
1: for replications = $1, 2, \ldots, 5000$ do
2:     Generate $\hat{\boldsymbol{\theta}} \sim N_K(\boldsymbol{\theta}, \mathbf{\Sigma})$
3:     Compute the rectangular confidence region $\mathfrak{R}$ using Algorithm 1 (using Algorithm 2 and 3 for rank-based methods).
4:     Check if $(\theta_1, \theta_2, \ldots, \theta_K) \in \mathfrak{R}$ and compute $T_1, T_2$, and $T_3$.
5: end for
6: Compute the proportion of times that the condition in line 4 is satisfied and the average of $T_1, T_2$, and $T_3$.
```

3.3 Simulation study

The resulting joint confidence intervals in Section 3.1 are used as basis in constructing the joint confidence intervals for overall rank uncertainty according to Klein's main result in Section 2.2.1. These are compared with the outcomes of joint confidence intervals in Section 2.2.2 in terms of coverage and overall measures of tightness resulting from Section 3.2.

In each simulation scenario, the components of the mean vector for the multivariate normal distribution were drawn from a normal distribution with mean 23.8–corresponding to the average of the mean travel time estimates across 51 states in Klein's study—and standard deviation $sd \in \{2, 3.6, 6\}$. These settings are selected to represent varying degrees of separation among the true parameter values, thereby influencing the difficulty of maintaining simultaneously narrow confidence intervals. By intuition, wider spread among true means facilitates clearer differentiation between estimates.

The number of populations being ranked was varied as $K \in \{5, 10, 20, 30, 40, 50\}$), to examine how dimensionality affects the uncertainty of the estimated rankings. Correlation among the parameters was imposed according to the structures outlined in Section 3.1.3, enabling comparison across distinct dependency patterns and among different joint confidence region constructions. Each case is carried out with $\alpha = 0.05$.

Bibliography

Andersson, J., Carling, K., & Mattson, S. (1998). Random ranking of hospitals is unsound. *CHANCE*, *11*(3), 33–39. https://doi.org/doi:10.1080/09332480.1998.10542106

Archakova, I., & Hansen, P. R. (2020). *A canonical representation of block matrices with applications to covariance and correlation matrices.*

Bazylik, S., Mogstad, M., Romano, J. P., Shaikh, A. M., & Wilhelm, D. (2025). Simultaneous confidence regions for ranks. *Journal of Econometrics.* https://doi.org/https://doi.org/10.1016/j.jeconom.2025.106010

David, C., & Legara, E. F. (2015). *How voters combine candidates on the ballot: The case of the philippine senatorial elections.*

Dunn, O. J. (1958). *Estimation of the means of dependent variables.*

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57). Chapman & Hall/CRC.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 175–177.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 385–443.

Hall, P., & Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, *37*(6B), 3929–3959.

Klein, M., Wright, T., & Wieczorek, J. (2020). A joint confidence region for an overall ranking of populations. *Journal of the Royal Statistical Society*, 589–606.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., & Rue, H. (2019). *Advanced spatial modeling with stochastic partial differential equations using r and INLA*. Chapman & Hall/CRC Press.

Lyhagen, J., & Ahlgren, P. (2020). Uncertainty and the ranking of economics journals. *Scientometrics*, *125*, 2545–2560. https://doi.org/https://doi.org/10.1007/s11192-020-03681-5

Mogstad, M., Romano, J. P., Shaikh, A. M., & Wilhelm, D. (2024). Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies*, *91*(1), 476–518.

Mohamad, D. A., Zwet, E. W. van, & Goeman, J. J. (2019). Simultaneous confidence intervals for ranks with application to ranking institutions. *Journal of the International Biometric Society*.

Ravanilla, N., & Hicken, A. (2023). *When legislators don't bring home the pork: The case of philippine senators*.

Šidák, Z. (1967). *Rectangular confidence regions for the means of multivariate normal distributions*.

Turek, D., & Risser, M. (2022). *bayesNSGP: Bayesian analysis of non-stationary gaussian process models*. https://cran.stat.auckland.ac.nz/web/packages/BayesNSGP/BayesNSGP.pdf

Wright, T. (2025). *Optimal tightening of the KWW joint confidence region for a ranking*.

Zhang, S., Luo, J., Zhu, L., Stinchcomb, D. G., Campbell, D., Carter, G., Gilkesone, S., & Feuerc, E. J. (2013). *Confidence intervals for ranks of age-adjusted rates across states or counties*.