TITLE HERE

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 Year

Contents

1 Introduction

Rankings of government units derived from sample survey data are typically published without accompanying statistical statements that quantify uncertainty in estimated overall rankings *(add here uncertainty is just expressed for each element being ranked).* While the literature on quantifying overall uncertainty remains limited, existing methods overlook the potential correlation among ranks *(Literature that this is possible).* The objective of this study is to introduce a methodology that constructs joint confidence region for the true but unknown overall ranking while accounting for the correlation among them. In line with this, we also present ways to estimate correlation in a specific application—such as estimating the dependence structure among senatorial candidates' rankings.

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 tighter 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, the work of 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. 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 the case of ranking senatorial candidates in the Philippines, this assumption is limiting as it treats vote shares as statistically independent across candidates. 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)), pre-election survey and actual results (*add references*) suggest that candidates who belong to the same political alliance often perform similarly, although not always equally well. (*add more examples and references*) For example, in the 2025 election , several candidates from both the Marcos and Duterte blocs secured seats, while others from the same groups did not. Accounting for these correlations allows for a more realistic assessment of uncertainty in the estimated ranks for similar use cases.

1.3 Scope and Limitations

2 Related Literature

2.1 Joint confidence region for an overall ranking

Klein et al. (2020) proposed an approach for quantifying overall rank uncertainty following the estimation of respondents' average travel time to work in each K sampled geographical area. In their paper, rank for the kth population is defined as

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

Meanwhile, the estimated overall ranking is computed from the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$, and expressed as $(\hat{r}_1, \ldots, \hat{r}_K)$, where

$$\hat{r}_k = 1 + \sum_{j:j\neq k} I(\hat{\theta}_j \leq \hat{\theta}_k), \qquad \text{for } k = 1, \ldots, K \tag{2.2}$$

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

$$\theta_k \in (L_k, U_k) \tag{2.3}$$

and defined the following:

$$\left.\begin{aligned} I_k &= \{1, 2, \ldots, K\} - \{k\}, \\ \Lambda_{Lk} &= \{j \in I_k : U_j \leq L_k\}, \\ \Lambda_{Rk} &= \{j \in I_k : U_k \leq L_j\}, \\ \Lambda_{Ok} &= \{j \in I_k : U_j > L_k \text{ and } U_k > L_j\} = I_k - \{\Lambda_{Lk} \cup \Lambda_{Rk}\} \end{aligned}\right\} \tag{2.4}$$

Equation 2.4 can likewise be expressed in words as follows:

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 $\Lambda_{Lk} \cup \Lambda_{Rk} \cup \Lambda_{Ok} = I_k$

The above implies that for each $k \in \{1, 2, \ldots, K\}$,

$$r_k \in \{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\} \tag{2.5}$$

Equation 2.5 demonstrates that a smaller $|\Lambda_{Ok}|$ results in smaller difference between U_k and L_k. Collectively, for all k, this yields narrower confidence intervals for the overall ranks, which is desirable.

They also assumed a conservative confidence region whose joint coverage probability is at least as large as the nominal level, $1 - \alpha$, as shown in Equation 2.6.

$$P\left[\bigcap_{k=1}^{K} \{\theta_k \in (L_k, U_k)\}\right] \geq 1 - \alpha \tag{2.6}$$

This yields the joint confidence set for the overall ranking, as defined in Equation 2.7, which they showed to have a joint probability of at least $1 - \alpha$.

$$\{(r_1, \ldots, r_K) : r_k \in \{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + | \text{ for } k = 1, 2, \ldots, K\}\} \tag{2.7}$$

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 $\hat{\theta} \in (L_k, U_k)\ \forall k \in \{1, 2, \ldots, K\}$, then the estimated ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$ lies within the joint confidence region defined in Equation 2.7 with probability 1.

They also noted that the joint confidence region in Equation 2.7 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 Joint confidence intervals for θ_ks

Klein et al. (2020) used a set familiar individual confidence intervals of the form $\hat{\theta}_k \pm z_{\alpha/2} SE_k$. They assumed that $\hat{\theta}_1, \hat{\theta}_1, \ldots, \hat{\theta}_K$ are independently distributed, with $\hat{\theta}_k \sim N(\theta_k, SE_k)$ for $k = 1, 2, \ldots, K$, where $\theta_1, \theta_2, \ldots \theta_k$ are unknown and $SE_1, SE_2, \ldots, SE_k$ are known. It was noted that $SE_k = \frac{MOE_k}{z_{\alpha/2}}$. Two approaches were considered for constructing the joint confidence intervals: the Bonferroni correction and the independence assumption.

The Bonferroni correction results in a conservative joint coverage for $\theta_1, \theta_1, \ldots, \theta_K$ of at least $1 - \alpha$. Intervals are as defined in Equation 2.8.

$$\left(\hat{\theta}_k - z_{(\theta/K)/2} SE_k,\ \hat{\theta}_k + z_{(\theta/K)/2} SE_k\right), \qquad \text{for } k = 1, 2, \ldots, K \tag{2.8}$$

In contrast, for the independence assumption, intervals that simultaneously yield a coverage equal to $1 - \alpha$ is given by Equation 2.9,

$$\left(\hat{\theta}_k - z_{\gamma/2} SE_k,\ \hat{\theta}_k + z_{\gamma/2} SE_k\right), \qquad \text{for } k = 1, 2, \ldots, K \tag{2.9}$$

where $\gamma = 1 - (1-\alpha)^{\frac{1}{K}}$.

2.3 Correlation structures

(literature on importance for accounting for correlation)

https://mgimond.github.io/Spatial/spatial-autocorrelation.html

https://cran.r-project.org/web/packages/simstudy/vignettes/corelationmat.html

3 Methodology

This section introduces the proposed methodologies to obtain confidence regions for the unknown overall true ranking. It extends approaches from Klein et al. (2020): the Bonferroni correction and the independence assumption. The following cases are tackled: case when items ranked are assumed to have zero and nonzero correlation. Both approaches are based on parametric bootstrap. Sections 3.1 and 3.2 discuss the algorithms for the cases mentioned. Section 3.3 shows the algorithms used to assess the performance of the proposed approaches. This makes use of coverage and metrics to measure the tightness of the estimated confidence regions.

For sections 3.1 and 3.2, let $\theta_1, \theta_2, \ldots, \theta_K$ be the true parameter values and $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$ be the corresponding estimates.

3.1 Joint confidence intervals for $\theta_1, \ldots, \theta_K$ by using Parametric Bootstrap

The rank-based parametric bootstrap approach assumes $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$ to be independent but not identically distributed estimates, where $\hat{\theta}_k \sim N(\theta_k, \sigma_k^2)$, $k = 1, 2, \ldots, K$. σ_k^2 is assumed known. Denote the corresponding ordered values by $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(K)}$.

Algorithm 1 Computation of Joint Confidence Region using Parametric Bootstrap

1: **for** $b = 1, 2, \ldots, B$ **do**

2: Generate $\hat{\theta}^*_{bk} \sim N\left(\hat{\theta}_k, \sigma_k^2\right)$, $k = 1, 2, \ldots, K$ and let $\hat{\theta}_{b(1)}, \hat{\theta}_{b(2)}, \ldots, \hat{\theta}_{b(K)}$ be the corresponding ordered values

	$k=1$	$k=2$	$\ldots$	$k=K$
$b=1$	$\hat{\theta}^*_{1(1)}$	$\hat{\theta}^*_{1(2)}$	$\ldots$	$\hat{\theta}^*_{1(K)}$
$b=2$	$\hat{\theta}^*_{2(1)}$	$\hat{\theta}^*_{2(2)}$	$\ldots$	$\hat{\theta}^*_{2(K)}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
$b=B$	$\hat{\theta}^*_{B(1)}$	$\hat{\theta}^*_{B(2)}$	$\ldots$	$\hat{\theta}^*_{B(K)}$

3: Compute

$$\hat{\sigma}^*_{b(k)} = \sqrt{\text{kth ordered value among } \left\{\hat{\theta}^{*2}_{b1} + \sigma_1^2, \hat{\theta}^{*2}_{b2} + \sigma_2^2, \ldots, \hat{\theta}^{*2}_{bK} + \sigma_K^2\right\} - \hat{\theta}^{*2}_{(k)}}$$

4: Compute $t^*_b = \max\limits_{1 \le k \le K} \left| \frac{\hat{\theta}^*_{b(k)} - \hat{\theta}^*_k}{\sigma^*_{b(k)}} \right|$

5: **end for**

6: Compute the $(1-\alpha)$-sample quantile of $t^*_1, t^*_2, \ldots, t^*_B$, call this $\hat{t}$.

7: 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

$$\hat{\sigma}_{(k)} = \sqrt{\text{kth ordered value among } \left\{\hat{\theta}^2_1 + \sigma_1^2, \hat{\theta}^2_2 + \sigma_2^2, \ldots, \hat{\theta}^2_K + \sigma_K^2\right\} - \hat{\theta}^2_{(k)}}$$

3.2 Joint confidence intervals for $\theta_1, \ldots, \theta_K$ by using Nonrank-based method

The nonrank-based method assumes that $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K\right) \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. It accounts for potential correlation among items being ranked. For this case, an exchangeable correlation, $\boldsymbol{\rho}$ (See Equation 3.1.), is assumed and used in the calculation of the variance covariance matrix (See Equation 3.2.).

$$\boldsymbol{\rho} = (1-\rho)\,\mathbf{I}_K + \rho \mathbf{1}_K \mathbf{1}'_K \tag{3.1}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Delta}^{1/2} \boldsymbol{\rho} \boldsymbol{\Delta}^{1/2} \tag{3.2}$$

where $\boldsymbol{\Delta} = \text{diag}\left\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\right\}$, with known σ_k's and ρ is studied for $0.1, 0.5, 0.9$.

Algorithm 2 Computation of Joint Confidence Region using Nonrank-based Method

Let the data consist of $\hat{\theta}_1, \ldots, \hat{\theta}_K$ and suppose $\mathbf{\Sigma}$ is known

1: **for** $b = 1, 2, \ldots, B$ **do**
2: Generate $\hat{\boldsymbol{\theta}}_b^* \sim N_K\left(\hat{\boldsymbol{\theta}}, \mathbf{\Sigma}\right)$ 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
$$\mathfrak{R} = \left[\hat{\theta}_1 \pm \hat{t} \times \sigma_1\right] \times \left[\hat{\theta}_2 \pm \hat{t} \times \sigma_2\right] \times \cdots \times \left[\hat{\theta}_K \pm \hat{t} \times \sigma_K\right]$$

3.3 Evaluation

Algorithm 3 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 3.3–3.5.

$$T_1 = \frac{1}{K} \sum_{k=1}^{K} \left|\Lambda_{Ok}\right| \tag{3.3}$$

$$T_2 = \prod_{k=1}^{K} \left|\Lambda_{Ok}\right| \tag{3.4}$$

$$T_3 = 1 - \frac{K + \sum_{k=1}^{K} \left|\Lambda_{Ok}\right|}{K^2} \tag{3.5}$$

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.

Algorithm 3 Computation of Coverage Probability for Parametric Bootstrap

For given values of $\theta_1, \theta_2, \ldots, \theta_K$ and thus $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}$

1: **for** replications $= 1, 2, \ldots, 5000$ **do**
2: Generate $\hat{\theta}_k \sim N(\theta_k, \sigma_k^2)$, for $k = 1, 2, \ldots, K$
3: Compute the rectangular confidence region $\mathfrak{R}$ using Algorithm 1.
4: Check if $\left(\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}\right) \in \mathfrak{R}$ and compute T_1, T_2, and T_3
5: **end for**
6: Compute the proportion of times that the condition in step 4 is satisfied and the average of T_1, T_2, and T_3.

Algorithm 4 is similar to Algorithm 3 but computes for the coverage and average T_1, T_2, and T_3 for the nonrank-based method.

Algorithm 4 Computation of Coverage Probability for Nonrank-based Method

For given values of $\theta_1, \theta_2, \ldots, \theta_K$ and $\boldsymbol{\Sigma}$

1: **for** replications $= 1, 2, \ldots, 5000$ **do**
2: Generate $\hat{\boldsymbol{\theta}} \sim N_K(\boldsymbol{\theta}, \boldsymbol{\Sigma})$
3: Compute the rectangular confidence region $\mathfrak{R}$ using Algorithm 2.
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 step 4 is satisfied and the average of T_1, T_2, and T_3.

Bibliography

Klein, M., Wright, T., & Wieczorek, J. (2020). *A joint confidence region for an overall ranking of populations.*

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

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