On the Use of Spearman's Rho to Measure the Stability of Feature Rankings: Supplementary material

Sarah Nogueira, Konstantinos Sechidis and Gavin Brown

School of Computer Science, University of Manchester, Manchester M13 9PL, UK
`sarah.nogueira@manchester.ac.uk`,
WWW home page: `http://www.cs.man.ac.uk/~nogueirs/`

This document is the supplementary material of [1]. We first remind the notations used in the paper in section 1 to facilitate the reading of this material. We then provide the proofs of all theorems and corollaries of the paper in section 2.

1 Notations

We shortly remind the notations of the paper.

- M is the number of bootstrap samples taken, also the number of rankings in $\mathcal{R}$.
- d is the total number of features.
- $\mathcal{R}$ is a matrix of size $M \times d$ where the i^{th} row represents the i^th ranking $\mathbf{r}_i$.
- $r_{i,f}$ is the rank of the f^{th} feature in the i^th ranking.
- $\mathbf{r}_i = (r_{i,1}, ..., r_{i,d})$ is the i^{th} ranking. A ranking is a permutation of the integers from 1 to d (we assume to tied ranks).
- $V_r = \frac{d^2-1}{12}$.
- $\hat{\Phi}(\mathcal{R})$ is the average pairwise Spearman's rho between each pair of distinct rankings. In other words, it is the average value of $\rho(\mathbf{r}_i, \mathbf{r}_j)$ for all $M(M-1)$ pairs of ranks where $i \neq j$.
- $\hat{\Phi}^{all}(\mathcal{R})$ is the average value of $\rho(\mathbf{r}_i, \mathbf{r}_j)$ for all M^2 pairs of ranks.
- X_f is the ransom variable corresponding to the rank of the f^{th} feature.
- σ_f^2 is the maximum likelihood estimator of the variance of X_f.
- s_f^2 is the unbiased sample variance of X_f ($s_f^2 = \frac{M}{M-1}\sigma_f^2$).

2 Proof of Theorems and Corollaries

2.1 Theorem 1

Theorem 1. *The stability $\hat{\Phi}$ using Spearman's ρ can be re-written as follows:*

$$\hat{\Phi}(\mathcal{R}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{V_r}, \tag{1}$$

where $V_r = \frac{d^2-1}{12}$ is a constant only depending on d.

We calculate the stability $\hat{\Phi}(\mathcal{R})$ using the average pairwise Spearman's ρ between the rankings in $\mathcal{R}$:

$$
\begin{aligned}
\hat{\Phi}(\mathcal{R}) =& \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \rho(\mathbf{r}_i, \mathbf{r}_j) \\
=& \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \rho(\mathbf{r}_i, \mathbf{r}_j) - \frac{1}{M(M-1)} \sum_{i=1}^{M} \rho(\mathbf{r}_i, \mathbf{r}_i) \\
=& \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \rho(\mathbf{r}_i, \mathbf{r}_j) - \frac{1}{M(M-1)} \sum_{i=1}^{M} 1 \\
=& \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \rho(\mathbf{r}_i, \mathbf{r}_j) - \frac{1}{M-1} \\
=& \frac{M^2}{M(M-1)} - \frac{1}{M(M-1)} \frac{6}{d(d^2-1)} \sum_{f=1}^{d} \sum_{i=1}^{M} \sum_{j=1}^{M} (r_{i,f} - r_{j,f})^2 - \frac{1}{M-1} \\
=& \frac{M}{M-1} - \frac{1}{M(M-1)} \frac{6}{d(d^2-1)} \sum_{f=1}^{d} \sum_{i=1}^{M} \sum_{j=1}^{M} (r_{i,f}^2 - 2 r_{i,f} r_{j,f} + r_{j,f}^2) - \frac{1}{M-1} \\
=& 1 - \frac{M}{M-1} \left[\frac{6}{d(d^2-1)} \sum_{f=1}^{d} \left(\frac{2}{M} \sum_{i=1}^{M} r_{i,f}^2 \right) - \frac{6}{d(d^2-1)} \left(\frac{2}{M^2} \sum_{f=1}^{d} \sum_{i=1}^{M} \sum_{j=1}^{M} r_{i,f} r_{j,f} \right) \right] \\
=& 1 - \frac{M}{M-1} \left[\frac{12}{d(d^2-1)} \sum_{f=1}^{d} \left(\frac{1}{M} \sum_{i=1}^{M} r_{i,f}^2 \right) - \frac{12}{d(d^2-1)} \sum_{f=1}^{d} \left(\frac{1}{M} \sum_{i=1}^{M} r_{i,f} \right)^2 \right] \\
=& 1 - \frac{M}{M-1} \left[\frac{1}{d} \frac{1}{V_r} \sum_{f=1}^{d} \left(\frac{1}{M} \sum_{i=1}^{M} r_{i,f}^2 \right) - \frac{1}{d} \frac{1}{V_r} \sum_{f=1}^{d} (\bar{r}_f)^2 \right] \\
=& 1 - \frac{1}{V_r} \frac{1}{d} \sum_{f=1}^{d} \frac{M}{M-1} \left[\left(\frac{1}{M} \sum_{i=1}^{M} r_{i,f}^2 \right) - (\bar{r}_f)^2 \right] \\
=& 1 - \frac{\frac{1}{d} \sum_{f=1}^{d} s_f^2}{V_r}.
\end{aligned}
\tag{2}
$$

2.2 Proof of Corollary 1

Corollary 1. $\hat{\Phi}(\mathcal{R})$ *is an unbiased and consistent estimator of:*

$$\Phi = 1 - \frac{\frac{1}{d} \sum_{f=1}^{d} \mathrm{Var}(X_f)}{V_r}. \tag{3}$$

We start by showing that $\hat{\Phi}$ is an unbiased estimator of Φ, i.e. $\mathbf{E}(\hat{\Phi}) = \Phi$.

$$\mathbf{E}(\hat{\Phi}) = \mathbf{E}\left(1 - \frac{\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2}{V_r}\right) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d}\mathbf{E}(\sigma_f^2)}{V_r} = \Phi, \tag{4}$$

since by definition s_f^2 is an unbiased estimator of $\mathrm{Var}(X_f)$ and by linearity of the expected value. We also have that s_f^2 is a consistent estimator of $\mathrm{Var}(X_f)$, therefore:

$$\begin{aligned} &\lim_{M\to\infty}(s_f^2) = \mathrm{Var}(X_f) \\ \Rightarrow &\lim_{M\to\infty}(\frac{1}{d}\sum_{f=1}^{d}s_f^2) = \frac{1}{d}\sum_{f=1}^{d}\mathrm{Var}(X_f) \\ \Rightarrow &\lim_{M\to\infty}(\hat{\Phi}) = \Phi. \end{aligned} \tag{5}$$

2.3 Proof of Theorem 2

Theorem 2. *$\hat{\Phi}$ is asymptotically bounded (as M goes to ∞) by 0 and 1.*

From Theorem 1, we have that:

$$\hat{\Phi}(\mathcal{R}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d}s_f^2}{V_r}.$$

By definition, we know that the unbiased sample variance s_f^2 is greater or equal to 0. Therefore, $\frac{1}{d}\sum_{f=1}^{d}s_f^2 \geq 0$ which implies that $\hat{\Phi}(\mathcal{R}) \leq 1$.

To prove that $\hat{\Phi}(\mathcal{R})$ is asymptotically positive, we will show that is can be re-written as follows:

$$\hat{\Phi}(\mathcal{R}) = \frac{1}{M(M-1)}\frac{1}{d^2}\sum_{f<f'}\underbrace{\left[\sum_{i=1}^{M}\frac{r_{i,f}-r_{i,f'}}{\sqrt{V_r}}\right]^2}_{\geq 0} - \underbrace{\frac{1}{M-1}}_{\underset{M\to+\infty}{\to}0}$$

which gives us that $\underset{M\to+\infty}{lim}[\hat{\Phi}(\mathcal{R})] \geq 0$. Indeed, we have:

$$\frac{1}{M(M-1)}\frac{1}{d^2}\sum_{f<f'}\left[\sum_{i=1}^{M}\frac{r_{i,f}-r_{i,f'}}{\sqrt{V_r}}\right]^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d^2}\sum_{f<f'}\left[\sum_{i=1}^{M}\frac{r_{i,f}}{\sqrt{V_r}}-\sum_{i=1}^{M}\frac{r_{i,f'}}{\sqrt{V_r}}\right]^2$$

$$=\frac{1}{M(M-1)}\left[\frac{1}{d}\sum_{f=1}^{d}\left(\sum_{i=1}^{M}\frac{r_{i,f}}{\sqrt{V_r}}\right)^2-\left(\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\frac{r_{i,f}}{\sqrt{V_r}}\right)^2\right]$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{r_{i,f}r_{j,f}}{V_r}-\frac{1}{M(M-1)}\left(\frac{1}{d}\sum_{i=1}^{M}\frac{1}{\sqrt{V_r}}\left(\sum_{f=1}^{d}r_{i,f}\right)\right)^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{r_{i,f}r_{j,f}}{V_r}-\frac{1}{M(M-1)}\left(\frac{1}{d}\sum_{i=1}^{M}\frac{1}{\sqrt{V_r}}\frac{d(d+1)}{2}\right)^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{r_{i,f}r_{j,f}}{V_r}-\frac{1}{M(M-1)}\left(\frac{M}{\sqrt{V_r}}\frac{(d+1)}{2}\right)^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{r_{i,f}r_{j,f}}{V_r}-\frac{M}{M-1}\frac{1}{V_r}\frac{(d+1)^2}{4}$$

$$=\frac{1}{M(M-1)}\frac{6}{d(d^2-1)}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}2r_{i,f}r_{j,f}-\frac{M}{M-1}\frac{1}{V_r}\frac{(d+1)^2}{4}$$

$$=\hat{\Phi}(\mathcal{R})-1+\frac{1}{M-1}\frac{12}{d(d^2-1)}\sum_{f=1}^{d}\sum_{i=1}^{M}r_{i,f}^2-\frac{M}{M-1}\frac{1}{V_r}\frac{(d+1)^2}{4}\quad \text{using Equation (2)}$$

$$=\hat{\Phi}(\mathcal{R})-1+\frac{1}{M-1}\frac{12}{d(d^2-1)}\sum_{i=1}^{M}\frac{d(d+1)(2d+1)}{6}-\frac{M}{M-1}\frac{1}{V_r}\frac{(d+1)^2}{4}$$

$$=\hat{\Phi}(\mathcal{R})-1+\frac{M}{M-1}\frac{1}{V_r}\frac{(d+1)(2d+1)}{6}-\frac{M}{M-1}\frac{1}{V_r}\frac{(d+1)^2}{4}$$

$$=\hat{\Phi}(\mathcal{R})-1+\frac{M}{M-1}\frac{1}{V_r}\left[\frac{(d+1)(2d+1)}{6}-\frac{(d+1)^2}{4}\right]$$

$$=\hat{\Phi}(\mathcal{R})-1+\frac{M}{M-1}$$

$$=\hat{\Phi}(\mathcal{R})+\frac{1}{M-1}.$$

2.4 Proof of Theorem 3

Theorem 3 (Correction For Chance). *$\hat{\Phi}$ is corrected by chance which means that its expected value is constant and equal to 0 when the FR is random (i.e. when all rankings/permutations have equal probability).*

First of all, let us prove that $V_r = \frac{d^2-1}{12}$ is the variance of X_f when the feature ranker (FR) is random. Let us assume X_f is the rank of the f^{th} feature by a random FR. By definition we have that:

$$Var(X_f) = \mathbf{E}(X_f^2) - (\mathbf{E}(X_f))^2 \tag{6}$$

Let us calculate $\mathbf{E}(X_f)$:

$$\mathbf{E}(X_f) = \sum_{i=1}^{d} i \times \mathbf{P}(X_f = i), \tag{7}$$

where $\mathbf{P}(X_f = i)$ is the probability that the rank of the f^{th} feature is equal to i. Since the feature ranker is random, all ranks are equiprobable, therefore $\mathbf{P}(X_f = i) = \frac{(d-1)!}{d!} = \frac{1}{d}$ since there are $d!$ permutations of the natural numbers from 1 to d and that in $(d-1)!$ of them, the f^{th} feature has a rank equal to i. Replacing this in Equation (8), we get that:

$$\mathbf{E}(X_f) = \sum_{i=1}^{d} i\frac{1}{d} = \frac{1}{d}\sum_{i=1}^{d} i = \frac{1}{d}\frac{d(d+1)}{2} = \frac{d+1}{2}. \tag{8}$$

Now, let us do the same type of calculation for $\mathbf{E}(X_f^2)$:

$$\mathbf{E}(X_f^2) = \sum_{i=1}^{d} i^2 \times \mathbf{P}(X_f = i) = \frac{1}{d}\sum_{i=1}^{d} i^2 = \frac{(d+1)(2d+1)}{6}. \tag{9}$$

Now using the results of equations (8) and (9) in Equation (10), we get that:

$$Var(X_f) = \frac{(d+1)(2d+1)}{6} - \left(\frac{d+1}{2}\right)^2 = \frac{d^2-1}{12}. \tag{10}$$

Therefore, we get that $V_r = \frac{d^2-1}{12}$.
Using Equation (4) that $\mathbf{E}(\hat{\Phi}) = \Phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d}\mathrm{Var}(X_f)}{V_r}$. Since $\mathrm{Var}(X_f) = V_r$ when the FR is random, we get that $\frac{\frac{1}{d}\sum_{f=1}^{d}\mathrm{Var}(X_f)}{V_r} = 1$ in that case and therefore that $\mathbf{E}(\hat{\Phi}) = 0$.

2.5 Proof of Theorem 4

Theorem 4. *The average squared error of the mean rank over the d features can be decomposed into two* ***positive*** *terms as follows:*

$$\overbrace{\frac{1}{d}\sum_{f=1}^{d}(\bar{r}_f - r_f^*)^2}^{av.\ SE\ of\ the\ mean\ ranker} = \overbrace{\frac{1}{d}\sum_{f=1}^{d}\left(\frac{1}{K}\sum_{i=1}^{K}(r_{i,f} - r_f^*)^2\right)}^{av.\ MSE\ of\ the\ K\ rankers} - \overbrace{(1-\hat{\Phi}^{all})V_r}^{ambiguity\ term}, \tag{11}$$

where the ambiguity term is also equal to $\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2$ *and where* $V_r = \frac{d^2-1}{12}$*. Therefore, the error of the ensemble ranker is guaranteed to be less or equal than the one of the individual rankers on average.*

Let us first calculate the average MSE term:

$$\begin{aligned}
&\frac{1}{d}\sum_{f=1}^{d}\left(\frac{1}{K}\sum_{i=1}^{K}(r_{i,f}-r_f^*)^2\right)\\
=&\frac{1}{d}\sum_{f=1}^{d}\frac{1}{M}\sum_{i=1}^{M}(r_{i,f}-r_f^*)^2 \qquad (12)\\
=&\frac{1}{d}\sum_{f=1}^{d}\left(\frac{1}{M}\sum_{i=1}^{M}r_{i,f}^2\right)-2\bar{r}_f r_f^*+(r_f^*)^2
\end{aligned}$$

Now we calculate the ambiguity term as follows:

$$\begin{aligned}
&(1-\hat{\Phi}^{all})V_r\\
=&\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2\\
=&\frac{1}{d}\sum_{f=1}^{d}\left[\left(\frac{1}{M}\sum_{i=1}^{M}r_{i,f}^2\right)-\left(\frac{1}{M}\sum_{i=1}^{M}r_{i,f}\right)^2\right] \quad \text{by definition of the sample variance}\\
=&\frac{1}{d}\sum_{f=1}^{d}\left[\left(\frac{1}{M}\sum_{i=1}^{M}r_{i,f}^2\right)-(\bar{r}_f)^2\right]
\end{aligned} \qquad (13)$$

Now subtracting the ambiguity term from Equation (13) from the average MSE term from Equation (12), we get the following:

$$\begin{aligned}
&\frac{1}{d}\sum_{f=1}^{d}\left(\frac{1}{K}\sum_{i=1}^{K}(r_{i,f}-r_f^*)^2\right)-(1-\hat{\Phi}^{all})V_r\\
=&\frac{1}{d}\sum_{f=1}^{d}(r_f^*)^2-2\bar{r}_f r_f^*+(\bar{r}_f)^2 \qquad (14)\\
=&\frac{1}{d}\sum_{f=1}^{d}\left(\bar{r}_f-r_f^*\right)^2,
\end{aligned}$$

which is the average squared error term given on the left-side of the equation.

Let us know show that the average MSE term and the ambiguity term are both positive. By definition, we know that a sum of squares is positive, therefore the average MSE term is positive. Let us now prove that the ambiguity term is positive as well:

$$\begin{aligned}
&0 \leq (1-\hat{\Phi}^{all})V_r \\
&\Leftrightarrow 0 \leq 1-\hat{\Phi}^{all} \\
&\Leftrightarrow \hat{\Phi}^{all} \leq 1 \\
&\Leftrightarrow 1-\frac{\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2}{V_r} \leq 1 \\
&\Leftrightarrow \frac{\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2}{V_r} \geq 0 \\
&\Leftrightarrow \frac{1}{d}\sum_{f=1}^{d}\sigma_f^2 \geq 0
\end{aligned} \tag{15}$$

Since σ_f^2 is always positive by definition, we have that $\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2 \geq 0$ and therefore that the ambiguity term is positive.

2.6 Proof of Theorem 5

Theorem 5. *Assuming the K rankings in the ensemble are independent and identically distributed (i.i.d), the stability of the mean ranking is reduced by $\frac{1}{K}$ compared to the stability of the individual FR:*

$$\Psi = \frac{K-1}{K} + \frac{\Phi}{K}. \tag{16}$$

By definition, we have:

$$\begin{aligned}
\Psi(\bar{\mathbf{r}}) &= 1-\frac{\frac{1}{d}\sum_{f=1}^{d}\mathrm{Var}(\bar{r}_f)}{V_r} \\
&= 1-\frac{\frac{1}{d}\sum_{f=1}^{d}\mathrm{Var}(\frac{1}{K}\sum_{i=1}^{K}r_{i,f})}{V_r} \\
&= 1-\frac{\frac{1}{d}\sum_{f=1}^{d}\frac{1}{K^2}\sum_{i=1}^{K}\mathrm{Var}(r_{i,f})}{V_r} \quad \text{since } Cov(r_{i,f},r_{j,f})=0 \text{ for } i\neq j \text{ using the i.i.d. assumption} \\
&= 1-\frac{\frac{1}{d}\sum_{f=1}^{d}\frac{1}{K^2}\sum_{i=1}^{K}\mathrm{Var}(X_f)}{V_r} \\
&= 1-\frac{1}{K}\frac{\frac{1}{d}\sum_{f=1}^{d}\mathrm{Var}(X_f)}{V_r}
\end{aligned}$$

Therefore:

$$\Psi(\bar{\mathbf{r}}) = 1 - \frac{1}{K}(1 - \Phi) = 1 - \frac{1}{K}(1 - \phi) = 1 - \frac{1}{K} - \frac{\Phi}{K} = \frac{K-1}{K} - \frac{\Phi}{K}$$

References

1. Nogueira, S., Brown, G.: On the use of spearman rho to measure the stability of feature rankings. In: Under review at IbPRIA (2017)