Robust Rank Canonical Correlation Analysis for Multivariate Survival Data

Di He[1] , Yong Zhou[2] and Hui Zou[3]

[1] *Nanjing University,* [2] *East China Normal University*

and [3] *University of Minnesota*

Supplementary Material

The Supplementary Material contains the simulation results, theories and proofs for Spearman rank canonical correlation analysis approach.

S1 Simulation results for Spearman rank canonical correlation analysis approach

We also apply the RRCCA approach with Spearman's rank correlation in our simulation studies. It is defined in the Conclusion section in the paper and (S2.1) and (S2.2) below. For estimation accuracy of subsection 5.1 in the paper, to compare the performance, we adopt the same criterion for the canonical vector and a similar one for canonical correlation, that is,

$$MSRE(\widehat{r}_j^c) = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{(\widehat{r}_j^c)^{(m)} - r_j^c}{r_j^c}\right\}^2,$$

where r_j^c is the jth population version canonical correlation for the Spearman rank canonical correlation analysis and $(\widehat{r}_j^c)^{(m)}$ is the corresponding estimate computed from the mth sample.

We discuss how to compute the population version quantities for the Spearman's version under different sampling distributions. For (i) normal distribution or (iii) contaminated normal case, by (2.2) and using the well known relation $r = 6\pi^{-1}\arcsin(\rho/2)$ for the bivariate normal distribution, we can obtain that r_j^c is $6\pi^{-1}\arcsin(\rho_j/2)$. Under (ii) multivariate t distribution, however, as shown by Hult and Lindskog (2002), this relation does not hold for Spearman's rank correlation. Heinen and Valdesogo (2020) derived an expression for the Spearman's rank correlation of the bivariate Student t

distribution in terms of an integral and we implement it numerically to get r_j^c. For (iv) lognormal distribution, due to the invariance property against monotonic transformations of rank correlation, r_j^c is the same as that under normal distribution.

The results of estimation accuracy are presented in Table 6 and 7, and the results of empirical power of subsection 5.2 are shown in Table 8 and 9. Compared with the tables in the paper, we see that this RRCCA version performs better than the CCA-IPCW approach but is not as good as the Kendall's τ version.

S2 Theories and proofs for Spearman rank canonical correlation analysis approach

As defined in the Conclusion section in the paper, the unbiased Spearman's rank correlations $\widehat{r}_{X_kY_l}$, $\widehat{r}_{X_kX_k'}$, and $\widehat{r}_{Y_lY_l'}$ form the sample version Spearman's rank correlation matrices

$$\widehat{\mathbf{\Gamma}}_{\mathbf{XX}} = (\widehat{r}_{X_kX_{k'}})_{p\times p}, \quad \widehat{\mathbf{\Gamma}}_{\mathbf{XY}} = (\widehat{r}_{X_kY_l})_{p\times q}, \text{ and } \widehat{\mathbf{\Gamma}}_{\mathbf{YY}} = (\widehat{r}_{Y_lY_{l'}})_{q\times q} \quad \text{(S2.1)}$$

and RRCCA can be performed by solving the eigenvalues and eigenvectors of $\widehat{\mathbf{\Gamma}}_{\mathbf{XX}}^{-1}\widehat{\mathbf{\Gamma}}_{\mathbf{XY}}\widehat{\mathbf{\Gamma}}_{\mathbf{YY}}^{-1}\widehat{\mathbf{\Gamma}}_{\mathbf{YX}}$. For two random variables $\widetilde{U}$ and $\widetilde{V}$ from a joint distribution, let $(\widetilde{U}_1, \widetilde{V}_1), (\widetilde{U}_2, \widetilde{V}_2)$ and $(\widetilde{U}_3, \widetilde{V}_3)$ be three independent realiza-

Table 6: The mean squared error (MSE) of estimating the direction of the jth canonical vector

			Case (a)		Case (b)		Case (c)		
n	censoring	dist	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$	$j=3$
100	10%	normal	0.10	0.13	0.31	0.33	0.16	0.30	0.28
		$t(3)$	0.09	0.13	0.32	0.34	0.15	0.31	0.30
		contaminated normal	0.10	0.13	0.31	0.33	0.15	0.29	0.28
		lognormal	0.09	0.12	0.30	0.32	0.15	0.30	0.29
	30%	normal	0.19	0.20	0.45	0.46	0.34	0.51	0.45
		$t(3)$	0.19	0.23	0.45	0.48	0.31	0.52	0.50
		contaminated normal	0.18	0.21	0.44	0.45	0.34	0.51	0.45
		lognormal	0.18	0.20	0.46	0.47	0.34	0.50	0.44
200	10%	normal	0.07	0.09	0.21	0.23	0.10	0.20	0.19
		$t(3)$	0.07	0.09	0.22	0.24	0.09	0.20	0.20
		contaminated normal	0.07	0.09	0.22	0.24	0.10	0.19	0.18
		lognormal	0.07	0.09	0.21	0.22	0.10	0.20	0.19
	30%	normal	0.11	0.13	0.35	0.36	0.21	0.33	0.31
		$t(3)$	0.12	0.16	0.33	0.37	0.21	0.36	0.36
		contaminated normal	0.12	0.14	0.32	0.34	0.21	0.36	0.34
		lognormal	0.12	0.14	0.33	0.34	0.22	0.34	0.31

Table 7: The mean squared relative error (MSRE) of estimating the magnitude of the jth canonical correlation

			Case (a)		Case (b)		Case (c)		
n	censoring	dist	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$	$j=3$
100	10%	normal	0.01	0.29	0.02	0.07	0.00	0.02	0.13
		$t(3)$	0.01	0.29	0.02	0.09	0.00	0.02	0.16
		contaminated normal	0.01	0.28	0.02	0.07	0.00	0.02	0.15
		lognormal	0.01	0.27	0.02	0.07	0.00	0.02	0.14
	30%	normal	0.02	0.38	0.04	0.20	0.01	0.04	0.28
		$t(3)$	0.02	0.37	0.05	0.25	0.01	0.05	0.31
		contaminated normal	0.02	0.37	0.04	0.19	0.01	0.04	0.28
		lognormal	0.02	0.41	0.04	0.20	0.01	0.04	0.28
200	10%	normal	0.00	0.17	0.01	0.04	0.00	0.01	0.07
		$t(3)$	0.00	0.18	0.01	0.04	0.00	0.01	0.08
		contaminated normal	0.00	0.17	0.01	0.04	0.00	0.01	0.08
		lognormal	0.00	0.16	0.01	0.04	0.00	0.01	0.08
	30%	normal	0.01	0.30	0.03	0.10	0.01	0.02	0.18
		$t(3)$	0.01	0.33	0.02	0.14	0.01	0.03	0.26
		contaminated normal	0.01	0.29	0.02	0.12	0.01	0.03	0.19
		lognormal	0.01	0.30	0.02	0.10	0.01	0.03	0.17

Table 8: Type-I error rate of the permutation test based on maximum CCA

		distribution		
n	censoring	normal	contaminated normal	lognormal
100	10%	0.045	0.051	0.052
	30%	0.051	0.066	0.056
200	10%	0.054	0.064	0.055
	30%	0.053	0.049	0.043

Table 9: Power of the permutation test based on maximum CCA

			distribution		
n	censoring	ρ	normal	contaminated normal	lognormal
100	10%	0.1	0.111	0.116	0.097
		0.3	0.712	0.705	0.692
		0.5	0.997	0.997	0.997
	30%	0.1	0.029	0.029	0.028
		0.3	0.148	0.128	0.158
		0.5	0.587	0.554	0.611
200	10%	0.1	0.188	0.203	0.166
		0.3	0.975	0.967	0.973
		0.5	1.000	1.000	1.000
	30%	0.1	0.043	0.037	0.048
		0.3	0.458	0.378	0.455
		0.5	0.978	0.947	0.973

tions without censoring. Then, the population Spearman's rank correlation is defined by $r_{UV} = \mathrm{Cov}\left\{\mathrm{sgn}(\widetilde{U}_1 - \widetilde{U}_2), \mathrm{sgn}(\widetilde{V}_1 - \widetilde{V}_3)\right\}$ and

$$\mathbf{\Gamma_{XX}} = (r_{X_k X_{k'}})_{p\times p}, \quad \mathbf{\Gamma_{XY}} = (r_{X_k Y_l})_{p\times q}, \text{ and } \mathbf{\Gamma_{YY}} = (r_{Y_l Y_{l'}})_{q\times q} \quad \text{(S2.2)}$$

are population Spearman's rank correlation matrices that (S2.1) estimates for. We also need the following condition to ensure that the rank-based correlation matrices are well-conditioned.

(C3) There exists positive constants κ' and ω', such that $\min\{\lambda_{\min}(\mathbf{\Gamma_{XX}}), \lambda_{\min}(\mathbf{\Gamma_{YY}})\} \geq \kappa', \lambda^{1/2}_{\max}(\mathbf{\Gamma^{\mathrm{T}}_{XY}\Gamma_{XY}}) \leq 1/\omega'$.

The estimation consistency of Spearman rank canonical correlation analysis can be inferred from the following theorem.

Theorem 2. *Under Condition (C2) and (C3), there exists a positive constant M, where, for any $0 < \epsilon < 1$, when $n > \max\{M\epsilon^{-2}, 24p\epsilon^{-1}, 24q\epsilon^{-1}, 24(pq)^{-1/2}\epsilon^{-1}\}$, $p^2 \log p^2 n = o(n)$, $q^2 \log q^2 n = o(n)$ and $pq \log pqn = o(n)$ hold, then we have*

$$\|\widehat{\mathbf{\Gamma}}^{-1}_{\mathbf{XX}}\widehat{\mathbf{\Gamma}}_{\mathbf{XY}}\widehat{\mathbf{\Gamma}}^{-1}_{\mathbf{YY}}\widehat{\mathbf{\Gamma}}_{\mathbf{YX}} - \mathbf{\Gamma}^{-1}_{\mathbf{XX}}\mathbf{\Gamma_{XY}}\mathbf{\Gamma}^{-1}_{\mathbf{YY}}\mathbf{\Gamma_{YX}}\| = o_p(1).$$

To prove Theorem 2, we first need the sample Spearman's rank correlation for the complete data. Given a sample $(\widetilde{U}^{(i)}, \widetilde{V}^{(i)})^n_{i=1}$, $n > 3$, it is easy to show that the sample Spearman's rank correlation which is analogous to

the Pearson correlation between the rank values of two variables has the following equivalent form

$$\widetilde{r}_{UV} = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \Big\{ \sum_{j=1}^{n} I(\widetilde{U}^{(j)} > \widetilde{U}^{(i)}) \Big\} \Big\{ \sum_{m=1}^{n} I(\widetilde{V}^{(m)} > \widetilde{V}^{(i)}) \Big\} - \frac{3(n-1)}{n+1}.$$

Then, we need the following two lemmas.

Lemma 9. *For any $\epsilon > 24/n$, the Spearman correlation $\widetilde{r}_{UV}$ has the following tail bound*

$$\Pr(|\widetilde{r}_{UV} - r_{UV}| \geq \epsilon) \leq C \exp(-Cn\epsilon^2).$$

Proof of Lemma 9. Let $\epsilon = 6/n + \delta > 6/n$ in the Lemma 2 of He et al. (2021), we have

$$\begin{aligned}
\Pr(|\widetilde{r}_{UV} - r_{UV}| \geq \epsilon) &\leq 4 \exp(-\frac{(n-3)(n+1)^2(\epsilon - 6/n)^2}{24(n-2)^2}) \\
&\leq 4 \exp(-\frac{(n-3)(\epsilon - 6/n)^2}{24}) \\
&\leq 4 \exp(-\frac{(n-3)}{24}\epsilon^2 + \frac{(n-3)}{2n}\epsilon).
\end{aligned}$$

When $(n-3)\epsilon/(2n) < (n-3)\epsilon^2/48$, that is, $\epsilon > 24/n$, we have the desired result. □

Lemma 10. *There exists a positive constant M, for any $0 < \epsilon < 1$, when*

$n > \max\{M\epsilon^{-2}, 24\epsilon^{-1}\}$, *we have*

$$\Pr(|\widehat{r}_{X_kY_l} - r_{X_kY_l}| \geq \epsilon) \leq Cn\exp(-Cn\epsilon^2),$$

$$\Pr(|\widehat{r}_{X_kX_k'} - r_{X_kX_k'}| \geq \epsilon) \leq Cn\exp(-Cn\epsilon^2),$$

$$\Pr(|\widehat{r}_{Y_lY_l'} - r_{Y_lY_l'}| \geq \epsilon) \leq Cn\exp(-Cn\epsilon^2).$$

Proof of Lemma 10. Rewrite $\widehat{r}_{X_kY_l} = \frac{12(n-2)}{n+1}\binom{n}{3}^{-1}\sum_{i<j<m} h_2\left\{\mathbf{W}^{(i)}, \mathbf{W}^{(j)}, \mathbf{W}^{(m)}\right\} - \frac{3(n-1)}{n+1}$, where

$$\begin{aligned} h_2\left\{\mathbf{W}^{(i)}, \mathbf{W}^{(j)}, \mathbf{W}^{(m)}\right\} =& \frac{1}{3}\left\{\frac{\delta_k^{(i)}}{(\widehat{S}_k^{(i)})^2}\frac{\phi_l^{(i)}}{(\widehat{S}_l^{(i)})^2} I(X_k^{(j)} > X_k^{(i)}) I(Y_l^{(m)} > Y_l^{(i)}) \right. \\ &+ \frac{\delta_k^{(i)}}{(\widehat{S}_k^{(i)})^2}\frac{\phi_l^{(i)}}{(\widehat{S}_l^{(i)})^2} I(X_k^{(m)} > X_k^{(i)}) I(Y_l^{(j)} > Y_l^{(i)}) \\ &\left. + \frac{\delta_k^{(m)}}{(\widehat{S}_k^{(m)})^2}\frac{\phi_l^{(m)}}{(\widehat{S}_l^{(m)})^2} I(X_k^{(j)} > X_k^{(m)}) I(Y_l^{(i)} > Y_l^{(m)}) \right\} \end{aligned}$$

is the symmetric kernel of $\{\widehat{r}_{X_kY_l} + \frac{3(n-1)}{n+1}\}/\frac{12(n-2)}{n+1}$. Hence, $\{\widehat{r}_{X_kY_l} + \frac{3(n-1)}{n+1}\}/\frac{12(n-2)}{n+1}$ is a U-statistic.

Let $U_n[f] \equiv \binom{n}{3}^{-1}\sum_{i<j<m}\left\{f(x^{(i)}, x^{(j)}, x^{(m)}) + f(x^{(i)}, x^{(m)}, x^{(j)}) + f(x^{(m)}, x^{(j)}, x^{(i)})\right\}/3$ denote the empirical function for this U-statistics. Similar as the proof of Lemma 4, there exists a positive constant M', for any $0 < \epsilon < 1$, when $n > 4M'\epsilon^{-2}$, we have

$$\Pr\left\{\frac{n+1}{12(n-2)}|\widehat{r}_{X_kY_l} - \widetilde{r}_{X_kY_l}| \geq \epsilon\right\} \leq Cn\exp(-Cn\epsilon^2).$$

Since $\frac{n+1}{12(n-2)} > \frac{1}{12}$, we have

$$\Pr(|\widehat{r}_{X_kY_l} - \widetilde{r}_{X_kY_l}| \geq \epsilon) \leq Cn\exp(-Cn\epsilon^2).$$

By triangle inequality $|\widehat{r}_{X_kY_l} - r_{X_kY_l}| \leq |\widehat{r}_{X_kY_l} - \widetilde{r}_{X_kY_l}| + |\widetilde{r}_{X_kY_l} - r_{X_kY_l}|$ and Lemma 9, we have the desired conclusion. The other two inequalities can be shown in the same way. □

Proof of Theorem 2. With Lemma 10, Theorem 2 can be proved in a similarly way as that of Theorem 1. □

Bibliography

He, D., Zhou, Y., and Zou, H. (2021). On sure screening with multiple responses. *Statistica Sinica*, 31(4):1749–1777.

Heinen, A. and Valdesogo, A. (2020). Spearman rank correlation of the bivariate student t and scale mixtures of normal distributions. *Journal of Multivariate Analysis*, 179:104650.

Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3):587–608.