Supporting Information

for

Non-parametric Estimation of Spearman's Rank Correlation with Bivariate Survival Data

by

Svetlana Eden, Chun Li, Bryan Shepherd

Contents

List of Tables

List of Figures

1 Web Appendix A: Consistency of $\widehat{\rho}_S^H$ with unbounded censoring.

Lemma A.1. *Let $X = \min(T_X, C_X)$, where random variables T_X and C_X are the time to event and time to censoring, respectively. Let $\Delta_X = \mathbb{1}(T_X \leq C_X)$ and $V_X = T_X\Delta_X$. Let $t_{max,X}$ be the "almost surely" maximum of T_X such that $\Pr(T_X > t_{max,X}) = 0$ and $\Pr(T_X > t) > 0$ for any $t < t_{max,X}$, where $t_{max,X}$ can be infinity. Also assume unbounded censoring, which means that for any t such that $\Pr(T_X \geq t) > 0$, $\Pr(T_X \leq C_X|T_X \geq t) > 0$. Given n independent and identically distributed pairs $\{(T_{X,i}, C_{X,i})\}_{i=1}^n$, let $V_{X,i} = T_{X,i}\Delta_{X,i}$ and $V_{max,n,X} = max(V_{X,1}, ..., V_{X,n})$, the largest uncensored event time. Then $V_{max,n,X} \xrightarrow{p} t_{max,X}$.*

Proof. For brevity we omit the subscript X in this proof. Because of unbounded censoring and because $\Pr(T \geq t) \geq \Pr(T > t) > 0$ for any $0 < t < t_{max}$, we have $\Pr(t \leq T \leq C) = \Pr(T \leq C|T \geq t)\Pr(T \geq t) \geq \Pr(T \leq C|T \geq t)\Pr(T > t) > 0$. Then for $V = T\Delta$ we have

$$\begin{aligned}\Pr(V < t) &= \Pr(C < T) + \Pr(T < t, T \leq C)\\ &= \Pr(C < T) + \Pr(T \leq C) - \Pr(t \leq T \leq C)\\ &= 1 - \Pr(t \leq T \leq C) < 1.\end{aligned}$$

Given data $\{(T_i, C_i)\}_{i=1}^n$, we have $\Pr(V_{max,n} < t) = \{\Pr(V < t)\}^n \longrightarrow 0$ as $n \to \infty$. Because this is true for any $0 < t < t_{max}$, for any small $\varepsilon > 0$, we have $\Pr(V_{max,n} < t_{max} - \varepsilon) \longrightarrow 0$ as $n \to \infty$. In other words, for any small $\varepsilon, \delta > 0$, there exists such n^* that $\Pr(|V_{max,n} - t_{max}| > \varepsilon) < \delta$ for any $n \geq n^*$. Therefore, $V_{max,n} \xrightarrow{p} t_{max}$. □

Theorem A.1. *Under unbounded censoring, $\widehat{\rho}_S^H$ is consistent for ρ_S, where*

$$\rho_S/c_\rho = \int_0^\infty \int_0^\infty \left\{1 - S_X(x) - S_X(x^-)\right\}\left\{1 - S_Y(y) - S_Y(y^-)\right\} S(dx, dy), \tag{1}$$

$$\widehat{\rho}_S^H/\widehat{c}_\rho^H = \sum_{i^*}\sum_{j^*}\left\{1 - \widehat{S}_X^H(x_{i^*}) - \widehat{S}_X^H(x_{i^*}^-)\right\}\left\{1 - \widehat{S}_Y^H(y_{j^*}) - \widehat{S}_Y^H(y_{j^*}^-)\right\}\widehat{S}^H(dx_{i^*}, dy_{j^*}), \tag{2}$$

i^* enumerates all the events of X plus $\widehat{\tau}_X$, j^* enumerates all the events of Y plus $\widehat{\tau}_Y$, $\widehat{\tau}_X$ and $\widehat{\tau}_Y$ are points just beyond the largest observed events, $S_X^H(x)$ and $S_Y^H(y)$ are the marginal survival functions of $S^H(x, y)$ (defined in (11) in the main manuscript), and

$$\begin{aligned}c_\rho &= \left[\mathrm{Var}\left\{1 - S_X(T_X) - S_X(T_X^-)\right\}\mathrm{Var}\left\{1 - S_Y(T_Y) - S_Y(T_Y^-)\right\}\right]^{1/2},\\ \widehat{c}_\rho^H &= \left[\mathrm{Var}\left\{1 - \widehat{S}_X^H(T_X) - \widehat{S}_X^H(T_X^-)\right\}\mathrm{Var}\left\{1 - \widehat{S}_Y^H(T_Y) - \widehat{S}_Y^H(T_Y^-)\right\}\right]^{1/2}.\end{aligned}$$

Proof. The right hand side of (2) can be decomposed as a sum of four terms $A + B + C + D$, where

$$
\begin{aligned}
A &= \sum_i \sum_j \left\{1 - \widehat{S}_X(x_i) - \widehat{S}_X(x_i^-)\right\} \left\{1 - \widehat{S}_Y(y_j) - \widehat{S}_Y(y_j^-)\right\} \widehat{S}(dx_i, dy_j), \\
B &= \sum_i \left\{1 - \widehat{S}_X(x_i) - \widehat{S}_X(x_i^-)\right\} \left\{1 - \widehat{S}_Y(\widehat{\tau}_Y^-)\right\} \widehat{S}(dx_i, \widehat{\tau}_Y^-), \\
C &= \sum_j \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\} \left\{1 - \widehat{S}_Y(y_j) - \widehat{S}_Y(y_j^-)\right\} \widehat{S}(\widehat{\tau}_X^-, dy_j), \\
D &= \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\} \left\{1 - \widehat{S}_Y(\widehat{\tau}_Y^-)\right\} \widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-).
\end{aligned}
$$

Here i enumerates all the events of X, and j enumerates all the events of Y. It now suffices to show that as $n \to \infty$, A converges to the right hand side of (1), $B \xrightarrow{p} 0$, $C \xrightarrow{p} 0$, $D \xrightarrow{p} 0$, and $\widehat{c}_\rho^H \xrightarrow{p} c_\rho$.

The convergence of A to the right hand side of (1) follows from the consistency of Dabrowska's estimator (Dabrowska, 1988), the continuous mapping theorem on a metric space of functionals, and the fact that marginal and joint survival functions are from the metric space of functionals with a bounded total variation. (Details are in Section 3.9.4 of van der Vaart and Wellner (1996).)

For the convergence of B, C, and D, we note that in practice, $\widehat{S}_X(\widehat{\tau}_X^-) = \widehat{S}_X(V_{max,n,X})$, $\widehat{S}_Y(\widehat{\tau}_Y^-) = \widehat{S}_Y(V_{max,n,Y})$, and $\widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) = \widehat{S}(V_{max,n,X}, V_{max,n,Y})$. Therefore, according to Lemma A.1 and the continuous mapping theorem, $\widehat{S}_X(\widehat{\tau}_X^-) \xrightarrow{p} S_X(t_{max,X}) = 0$, $\widehat{S}_Y(\widehat{\tau}_Y^-) \xrightarrow{p} 0$, and $\widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{p} 0$, which leads to the following:

$$
\begin{aligned}
B &= \sum_i \left\{1 - \widehat{S}_X(x_i) - \widehat{S}_X(x_i^-)\right\} \left\{1 - \widehat{S}_Y(\widehat{\tau}_Y^-)\right\} \widehat{S}(dx_i, \widehat{\tau}_Y^-) \\
&\leq \sum_i \widehat{S}(dx_i, \widehat{\tau}_Y^-) = S_Y(\widehat{\tau}_Y^-) - S(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{p} 0, \\
C &= \sum_j \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\} \left\{1 - \widehat{S}_Y(y_j) - \widehat{S}_Y(y_j^-)\right\} \widehat{S}(\widehat{\tau}_X^-, dy_j) \\
&\leq \sum_j \widehat{S}(\widehat{\tau}_X^-, dy_j) = \widehat{S}_X(\widehat{\tau}_X^-) - \widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{p} 0 \\
D &= \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\} \left\{1 - \widehat{S}_Y(\widehat{\tau}_Y^-)\right\} \widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \leq \widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{p} 0.
\end{aligned}
$$

We now show that $\widehat{c}_\rho^H \xrightarrow{p} c_\rho$. Let $Z_X(x) = 1 - \widehat{S}_X^H(x) - \widehat{S}_X^H(x^-)$ and $Z_Y(y) = 1 - \widehat{S}_Y^H(y) - \widehat{S}_Y^H(y^-)$. Then $\widehat{c}_\rho^H$ is the product of the square roots of the sample variances of Z_X and Z_Y. The sample means of Z_X and Z_Y are zero for any properly defined continuous or discrete survival function (see proof of Property 8 in Li and Shepherd, 2012). Then the sample variance of Z_X is

$$
\begin{aligned}
&\sum_{i^*} \left\{1 - \widehat{S}_X^H(x_{i^*}) - \widehat{S}_X^H(x_{i^*}^-)\right\}^2 \widehat{S}_X^H(dx_{i^*}) \\
=&\sum_i \left\{1 - \widehat{S}_X(x_i) - \widehat{S}_X(x_i^-)\right\}^2 \widehat{S}_X(dx_i) + \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\}^2 \widehat{S}_X(\widehat{\tau}_X^-) \\
\xrightarrow{p}& \int \left\{1 - S_X(x) - S_X(x^-)\right\}^2 S_X(dx) + 0 = \mathrm{Var}(1 - S_X(T_X) - S_X(T_X^-)).
\end{aligned}
$$

Similarly, the sample variance of Z_Y converges to $\mathrm{Var}(1 - S_Y(T_Y) - S_Y(T_Y^-))$. Therefore, $\widehat{c}_\rho^H \xrightarrow{p} c_\rho$.

□

References:

1. Dabrowska, D. M. (1988). Kaplan–Meier estimate on the plane. The Annals of Statistics 16, 1475–1489.

2. Li, C. and Shepherd, B. E. (2012). A new residual for ordinal outcomes. Biometrika 99, 473–480.

3. van der Vaart, A. W. and Wellner, J. A. (1996). Weak Convergence and Empirical Processes. Springer.

2 Web Appendix B: Code and data availability

We used the folowing R libraries: `SurvCorr`, `lcopula`, and `cubature`. Complete simulation and analysis code is posted at `https://biostat.app.vumc.org/ArchivedAnalyses`. We have implemented our methods in the R package `survSpearman`.

References:

1. Belzile, L. and Genest, C. (2017). *lcopula: Liouville Copulas.* R package version 1.0, `https://CRAN.R-project.org/package=lcopula` (accessed July 9, 2019).

2. Eden, S. K. and Li, C. and Shepherd, B. E. (2021). *survSpearman: Nonparametric Spearman's Correlation for Survival Data.* R package version 1.0.0, `https://CRAN.R-project.org/package=survSpearman` (accessed January 5, 2021).

3. Narasimhan, B. and Johnson, S. G. (2017). *cubature: Adaptive Multivariate Integration over Hypercubes.* R package version 1.3-8, `https://CRAN.R-project.org/package=cubature` (accessed July 9, 2019).

4. Ploner, M., Kaider, A., and Heinze, G. (2015). *SurvCorr: Correlation of Bivariate Survival Times.* R package version 1.0, `https://CRAN.R-project.org/package=SurvCorr` (accessed July 9, 2019).

3 Web Tables

Web Table 1: Bias [RMSE] of $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ as estimates of the overall Spearman's correlation, ρ_S, under unbounded censoring.

N	Censoring Scenario	Percent Censored	Method	Indep $\rho_S = 0$	Clayton $\rho_S = 0.2$	Frank $\rho_S = 0.2$	Frank $\rho_S = -0.2$
100	No Censoring		ρ_S^H	-0.001 [0.103]	-0.001 [0.100]	-0.005 [0.099]	0.002 [0.100]
			ρ_{IMI}	0.003 [0.102]	0.002 [0.095]	-0.018 [0.095]	0.008 [0.097]
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.001 [0.112]	-0.001 [0.108]	-0.004 [0.110]	0.003 [0.109]
			ρ_{IMI}	0.001 [0.115]	0.032 [0.118]	-0.002 [0.107]	0.000 [0.106]
		70%	ρ_S^H	0.005 [0.217]	-0.021 [0.216]	-0.006 [0.219]	0.015 [0.215]
			ρ_{IMI}	-0.004 [0.159]	0.091 [0.176]	-0.002 [0.153]	0.017 [0.157]
	$C_X \perp C_Y$,	(30%, 30%)	ρ_S^H	0.003 [0.121]	-0.004 [0.116]	-0.002 [0.116]	0.003 [0.117]
			ρ_{IMI}	0.004 [0.117]	0.036 [0.118]	-0.004 [0.111]	0.005 [0.110]
		(30%, 70%)	ρ_S^H	0.001 [0.201]	0.001 [0.191]	-0.016 [0.193]	0.012 [0.195]
			ρ_{IMI}	-0.008 [0.142]	0.065 [0.156]	-0.004 [0.138]	0.008 [0.139]
		(70%, 70%)	ρ_S^H	-0.010 [0.261]	-0.016 [0.271]	-0.036 [0.278]	0.022 [0.260]
			ρ_{IMI}	-0.007 [0.174][1]	0.100 [0.197]	-0.018 [0.169][1]	0.019 [0.179][2]
200	No Censoring		ρ_S^H	0.000 [0.068]	0.000 [0.068]	-0.003 [0.067]	0.001 [0.065]
			ρ_{IMI}	0.001 [0.069]	0.012 [0.069]	-0.012 [0.069]	0.008 [0.067]
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.002 [0.078]	0.000 [0.079]	-0.003 [0.079]	0.002 [0.076]
			ρ_{IMI}	-0.005 [0.083]	0.034 [0.083]	-0.005 [0.076]	0.003 [0.073]
		70%	ρ_S^H	0.003 [0.164]	-0.008 [0.159]	0.001 [0.156]	0.008 [0.158]
			ρ_{IMI}	0.005 [0.120]	0.094 [0.143]	-0.011 [0.110]	0.010 [0.114]
	$C_X \perp C_Y$,	(30%, 30%)	ρ_S^H	-0.003 [0.084]	-0.001 [0.081]	-0.001 [0.082]	-0.002 [0.081]
			ρ_{IMI}	0.001 [0.082]	0.040 [0.087]	-0.006 [0.077]	0.004 [0.080]
		(30%, 70%)	ρ_S^H	0.003 [0.148]	0.004 [0.139]	-0.004 [0.145]	0.002 [0.148]
			ρ_{IMI}	-0.002 [0.102]	0.081 [0.126]	-0.010 [0.097]	0.014 [0.100]
		(70%, 70%)	ρ_S^H	0.007 [0.231]	0.004 [0.209]	-0.010 [0.217]	0.009 [0.214]
			ρ_{IMI}	0.003 [0.126]	0.115 [0.167]	-0.012 [0.118]	0.015 [0.124]

[1] In one out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

[2] In four out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

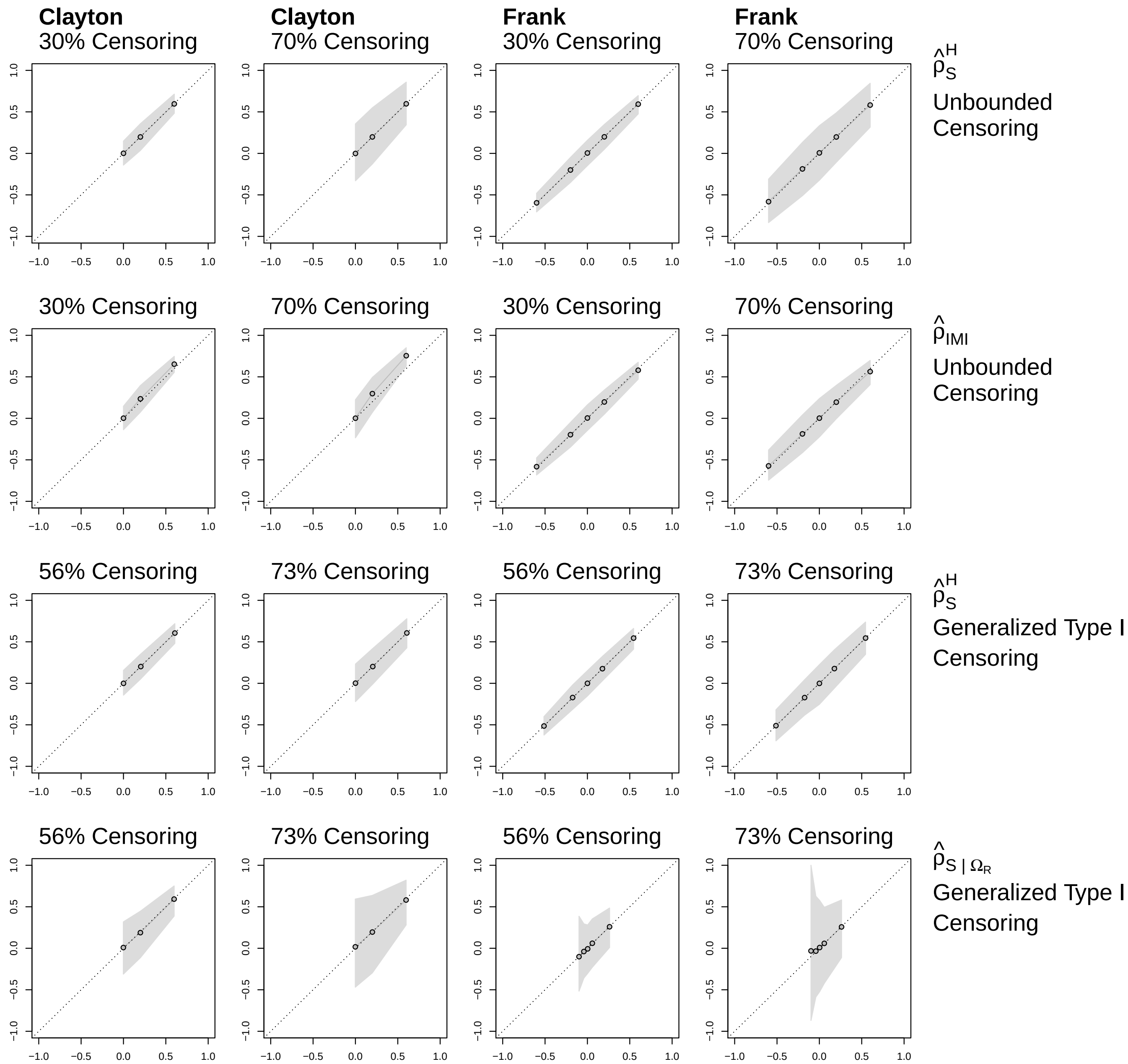
Web Table 2: Power and type I error rate for the overall Spearman's correlation, ρ_S, measured by $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ under unbounded censoring.

N	Censoring Scenario	Percent Censored	Method	Indep $\rho_S = 0$	Clayton $\rho_S = 0.2$	Frank $\rho_S = 0.2$	Frank $\rho_S = -0.2$
100	No Censoring		ρ_S^H	0.056	0.497	0.491	0.497
			ρ_{IMI}	0.055	0.532	0.447	0.488
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.041	0.393	0.390	0.396
			ρ_{IMI}	0.042	0.533	0.397	0.415
		70%	ρ_S^H	0.031	0.137	0.152	0.130
			ρ_{IMI}	0.004	0.201	0.075	0.028
	$C_X \perp C_Y$,	$(30\%, 30\%)$	ρ_S^H	0.047	0.361	0.366	0.360
			ρ_{IMI}	0.040	0.516	0.362	0.350
		$(30\%, 70\%)$	ρ_S^H	0.048	0.166	0.162	0.170
			ρ_{IMI}	0.013	0.311	0.132	0.129
		$(70\%, 70\%)$	ρ_S^H	0.032	0.101	0.082	0.090
			ρ_{IMI}	0.001[1]	0.132	0.017[1]	0.010[2]
200	No Censoring		ρ_S^H	0.042	0.802	0.803	0.819
			ρ_{IMI}	0.039	0.858	0.756	0.792
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.034	0.693	0.685	0.704
			ρ_{IMI}	0.054	0.828	0.680	0.691
		70%	ρ_S^H	0.034	0.239	0.237	0.233
			ρ_{IMI}	0.010	0.497	0.149	0.112
	$C_X \perp C_Y$,	$(30\%, 30\%)$	ρ_S^H	0.043	0.666	0.651	0.675
			ρ_{IMI}	0.043	0.837	0.657	0.630
		$(30\%, 70\%)$	ρ_S^H	0.038	0.314	0.299	0.307
			ρ_{IMI}	0.018	0.657	0.291	0.283
		$(70\%, 70\%)$	ρ_S^H	0.038	0.145	0.134	0.138
			ρ_{IMI}	0.001	0.391	0.071	0.037

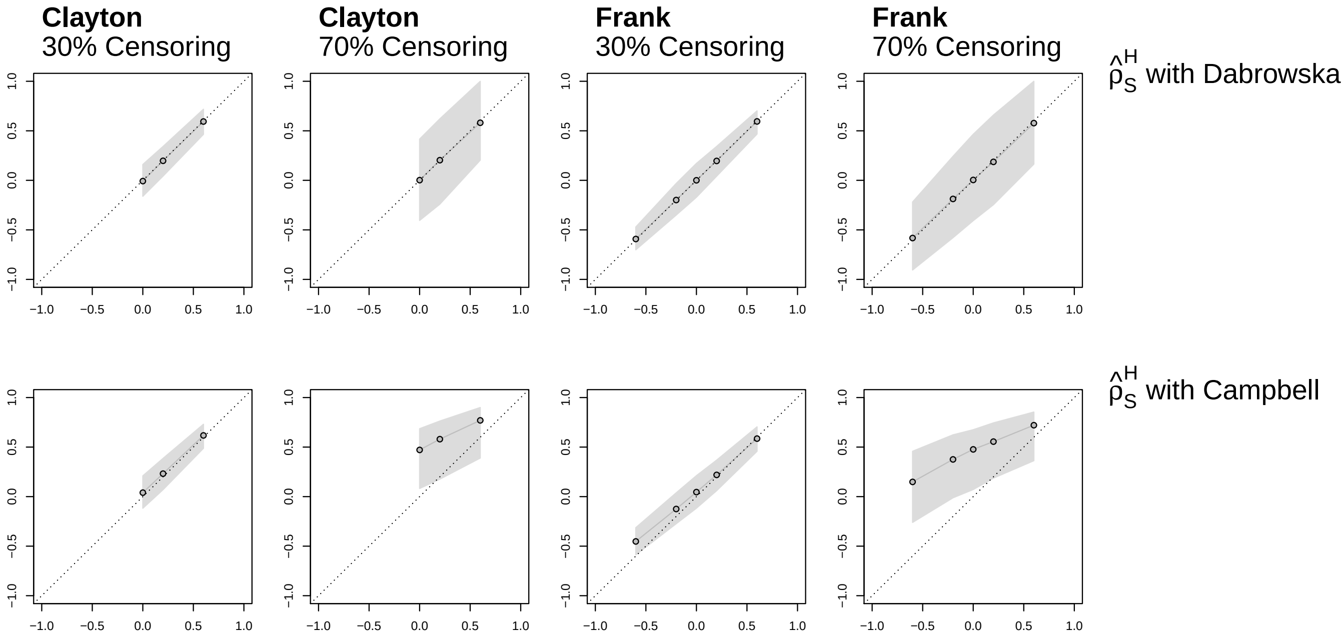
[1] In one out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

[2] In four out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

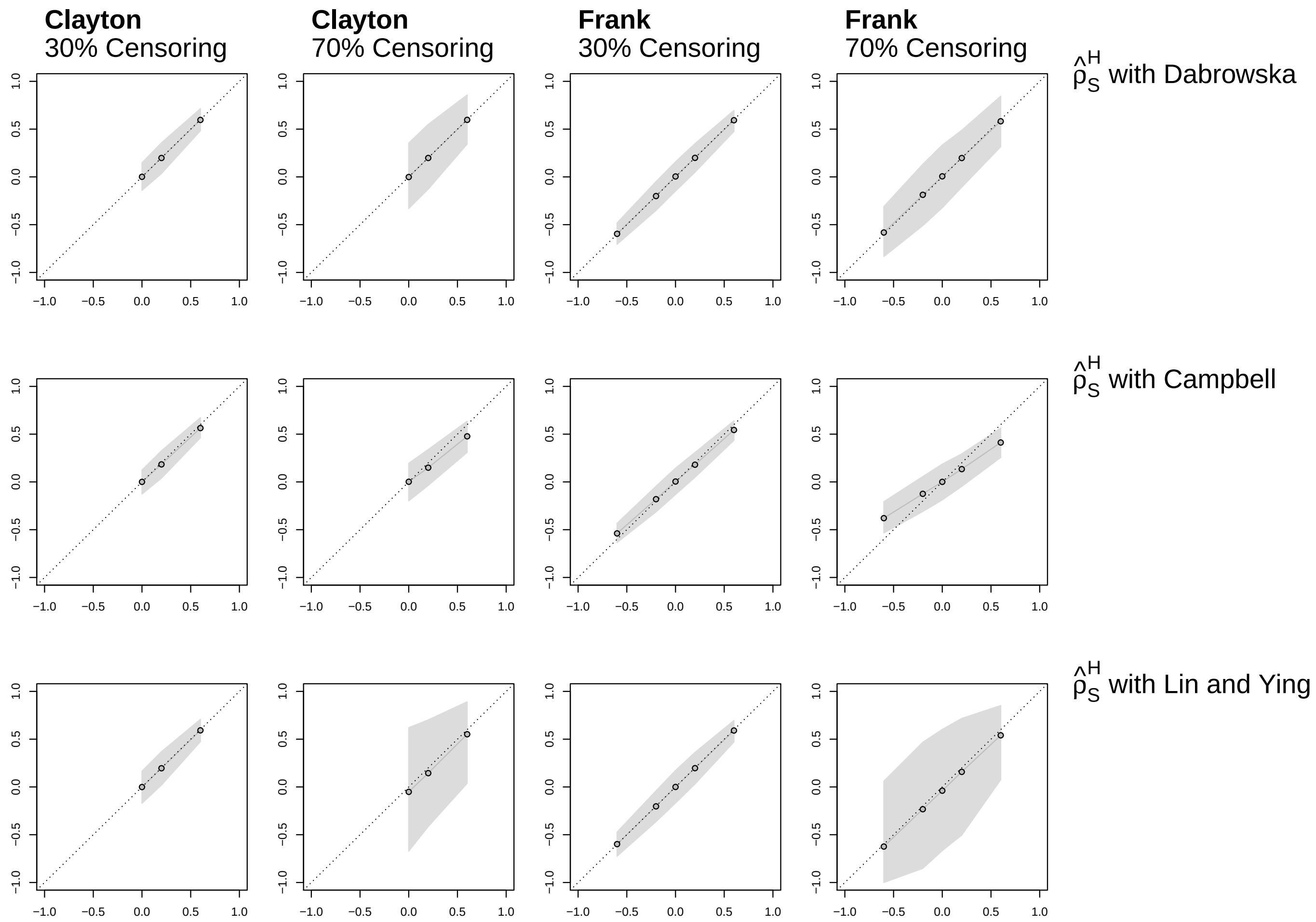
4 Web Figures



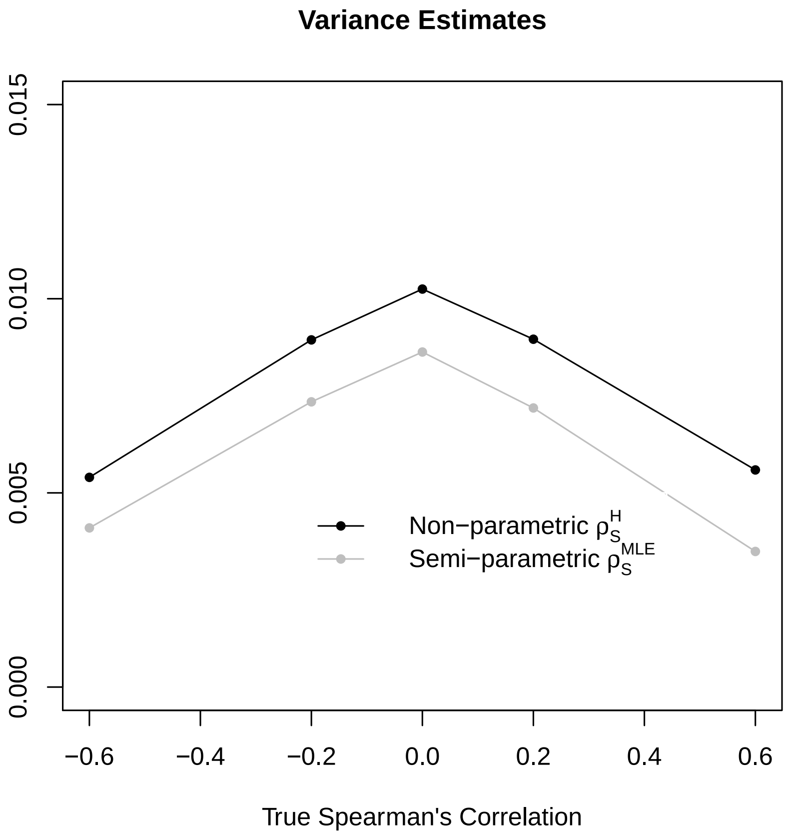
Web Figure 1: Point estimates (x-axis) vs population parameters (y-axis) under different univariate censoring scenarios. The top and second rows are $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ as estimators of the overall Spearman's correlation, ρ_S. The third row is $\widehat{\rho}_S^H$ as an estimator of ρ_S^H. The bottom row is $\widehat{\rho}_{S|\Omega_R}$ as an estimator of $\rho_{S|\Omega_R}$. The columns represent Clayton's and Frank's copulas. The population parameters for Clayton's family are 0, 0.2, and 0.6 for all estimates. For Frank's family, the population parameters of ρ_S are -0.6, -0.2, 0.2, and 0.6; the population parameters of ρ_S^H are -0.512, -0.173, 0.180, and 0.545; the population parameters of $\rho_{S|\Omega_R}$ are -0.098, -0.042, 0.058, and 0.261. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles. For generalized type I censoring, the restricted region, Ω_R, is defined by the median survival times.



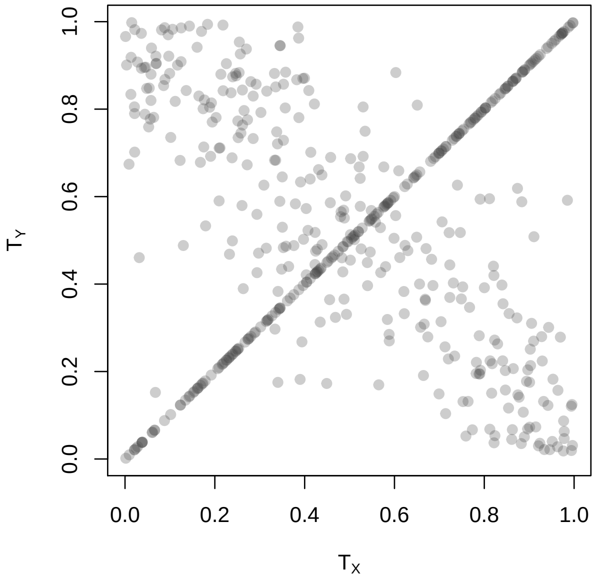
Web Figure 2: Performance of $\widehat{\rho}_S^H$ with survival surface estimators of Dabrowska (1988) (top row) and Campbell (1981) (bottom row) under bivariate unbounded censoring. The columns represent Clayton's and Frank's copulas under moderate and heavy censoring. The x-axis is the true overall Spearman's correlation; the y-axis is an estimate. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles.



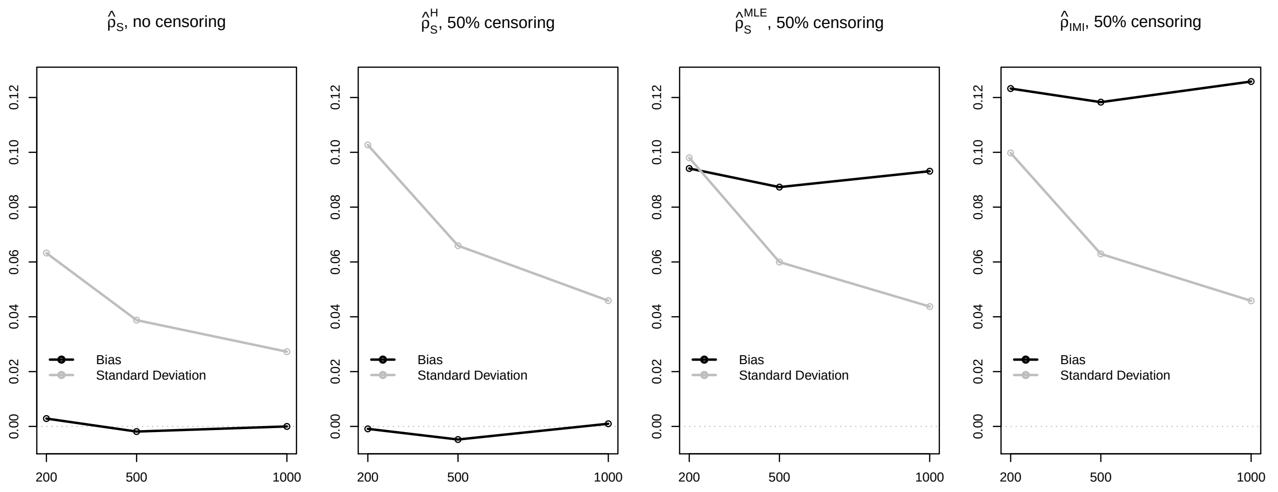
Web Figure 3: Performance of $\widehat{\rho}_S^H$ with survival surface estimators of Dabrowska (1988) (top row), Campbell (1981) (middle row), and Lin and Ying (1993) (bottom row) under univariate unbounded censoring. The columns represent Clayton's and Frank's copulas under moderate and heavy censoring. The x-axis is the true overall Spearman's correlation; the y-axis is an estimate. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles.



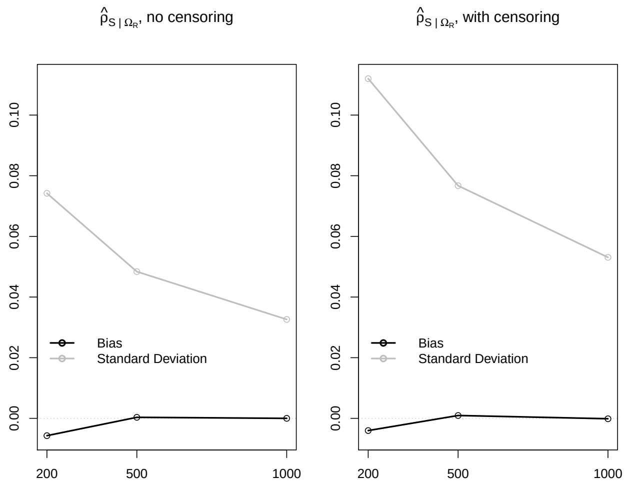
Web Figure 4: Efficiency of $\widehat{\rho}_S^H$ vs $\widehat{\rho}_S^{MLE}$, the maximum likelihood estimator of ρ_S assuming Frank's copula dependency structure. The black and gray lines are the variances of $\widehat{\rho}_S^H$ vs $\widehat{\rho}_S^{MLE}$ respectively. The data are simulated 1000 times with 200 pairs generated from Frank's copula family; the univariate unbounded censoring at 50% is applied. The relative efficiency $\text{Var}(\widehat{\rho}_S^H)/\text{Var}(\widehat{\rho}_S^{MLE})$ ranged from 1.19 (for $\rho_S = 0$) to 1.60 (for $\rho_S = 0.6$).



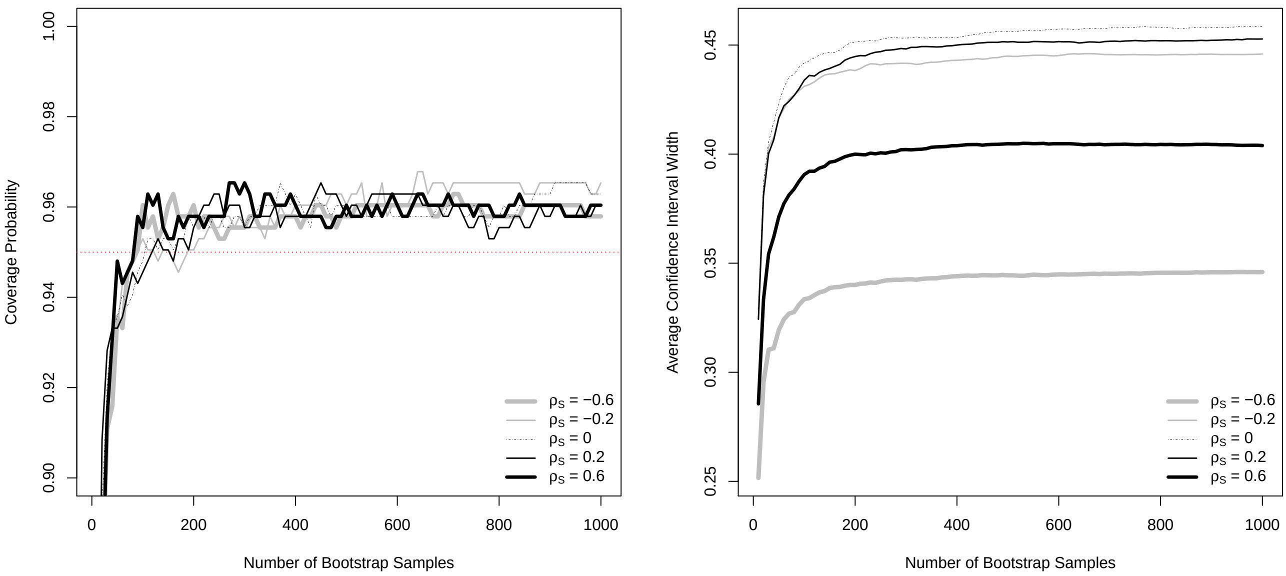
Web Figure 5: Illustration of the mixture distribution composed of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula family with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the overall Spearman's correlation being about -0.0813. T_X and T_Y are uniformly distributed.



Web Figure 6: Bias and standard deviation as functions of the sample size for $\widehat{\rho}_S$ (first panel), $\widehat{\rho}_S^H$ (second panel), $\widehat{\rho}_S^{MLE}$ (third panel), and $\widehat{\rho}_{IMI}$ (forth panel). Estimator $\widehat{\rho}_S$ is computed as Spearman's rank correlation for uncensored data. Estimators $\widehat{\rho}_S^H$, $\widehat{\rho}_S^{MLE}$, and $\widehat{\rho}_{IMI}$ are computed under 50% random unbounded censoring. The bivariate survival data are simulated as a mixture of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula family with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the overall Spearman's correlation being about -0.0813 (see Web Figure 5 for illustration).



Web Figure 7: Bias and standard deviation as functions of the sample size for estimates of Spearman's correlation within the restricted region with no censoring and therefore using standard methods (left panel) and with censoring and therefore computing $\widehat{\rho}_{S|\Omega_R}$ (right panel) as described in Section 3.1. Estimator $\widehat{\rho}_S$ is computed as Spearman's rank correlation for uncensored pairs with each event time less than the median event time. The restricted region Ω_R is defined by the median follow-up time for both times to event. The bivariate survival data are simulated as a mixture of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula family with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the true overall and restricted Spearman's correlations being about -0.081 and 0.85 respectively. Unbounded 50% censoring is applied to the entire sample. The effective proportion of uncensored events for $\widehat{\rho}_{S|\Omega_R}$ is 25% for each time to event (Web Figure 5 illustrates an uncensored sample).



Web Figure 8: Coverage probability (left panel) and average width (right panel) of the bootstrap confidence intervals for $\widehat{\rho}_S^H$ as an estimate of ρ_S under 50% unbounded censoring. The data are simulated from Frank's copula with parameters corresponding to Spearman's correlation of -0.6, -0.2, 0, 0.2, and 0.6. The sample size is 200 and the number of simulations is 1000. The 95% bootstrap confidence bounds are computed as the 0.025^{th} and 0.975^{th} percentiles.