

SUPPLEMENTARY MATERIAL for:

**Study of imputation procedures for nonparametric density estimation based on missing
censored lifetimes**

Sam Efromovich, Lirit Fuksman

Department of Mathematical Sciences, The University of Texas at Dallas, USA

The Supplementary Material contain color versions of the plots, results of simulations and R code.

S1. Color plots

S1.1 Liver cancer data

We analyze the liver cancer dataset which is conveniently located in plotRCS R package and contains survival status of 903 patients as well as their time to death/censoring. Rate of censoring is 35%, the indicator of censoring is available for all observations while time is missing in 6% of the cases. Out of the missing observations, 55 are uncensored and 4 are censored. The missing mechanism is MCAR. We will estimate the density on the interval $[0, 83]$ as it corresponds to the maximum time in months in the data.

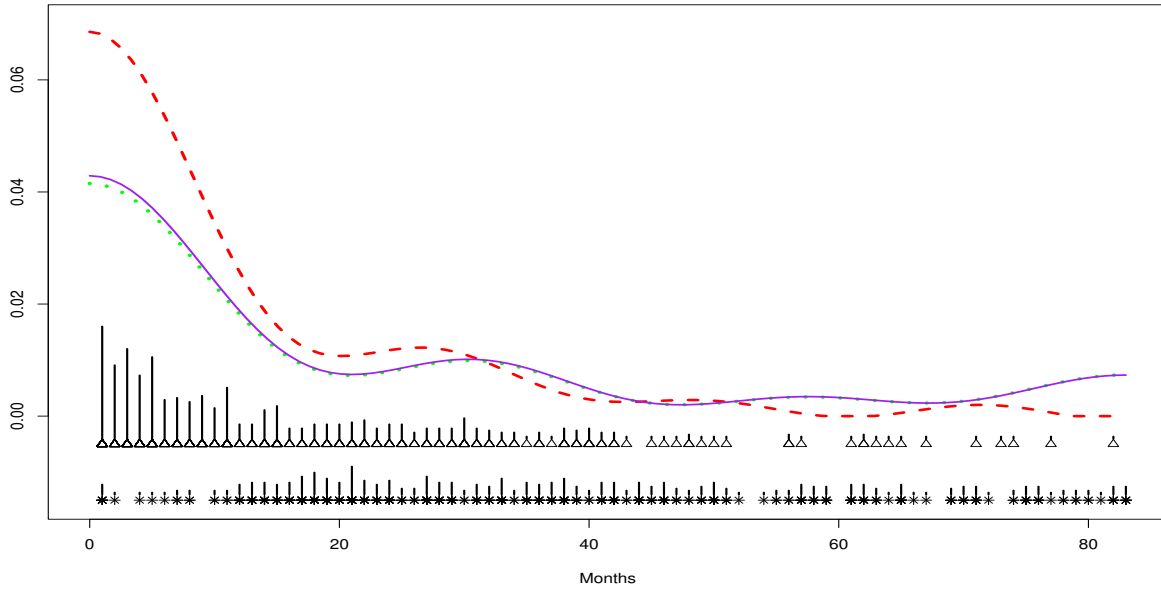


Figure 1.a: Density estimates for liver cancer data. The triangles indicate uncensored available lifetimes, the stars indicate censored available lifetimes. The vertical lines are proportional to the frequency of survival times found in the dataset, e.g. at time of 1 month there are 55 uncensored observations and 5 censored observations. The green dotted line is the MCAR E-estimate, the dashed red line is the naive estimator which ignored censoring and only involves uncensored cases, the solid purple line is the MAR E-estimate.

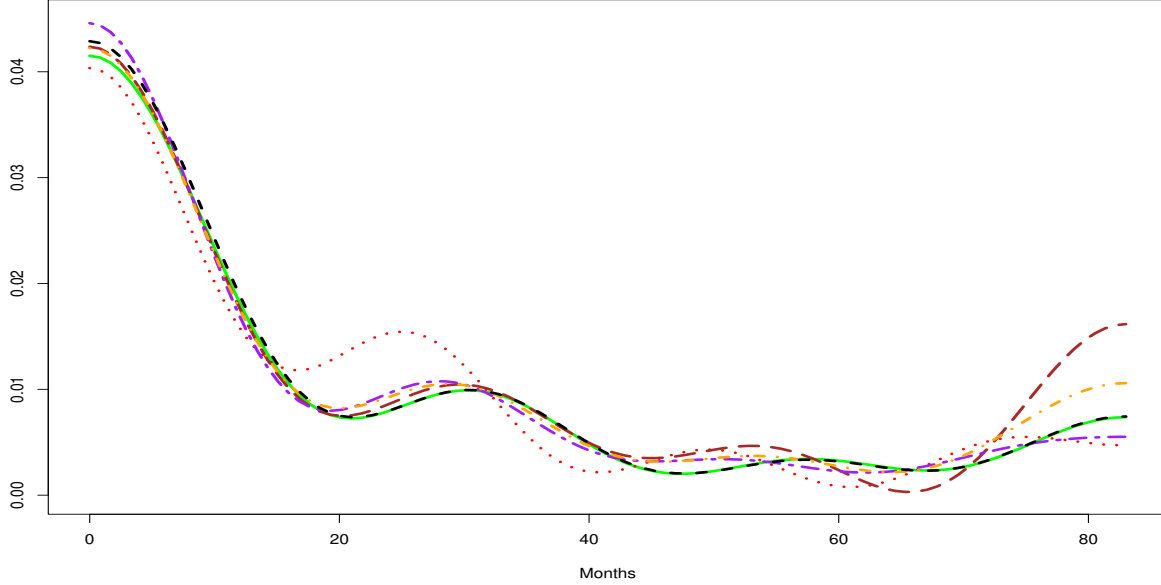


Figure 1.b: Benchmark and imputation estimates for liver cancer data. The green solid line is the MCAR E-estimate, the orange dotdashed line is multiple resampling estimator, the brown longdash line is resampling estimate, the black dashed line is multiple hotdeck estimate, the purple twodashed line is hotdeck estimate and the dotted red line is mean estimator.

S1.2 Breast cancer data

The dataset analyzed next is an example of complete missing where sample is available from a triplet $(AV, A\Delta, A)$. This dataset is a combination of two related breast cancer databases: Diagnostic Wisconsin Breast Cancer and Prognostic Wisconsin Breast Cancer which are both publicly available at UC Irvine Machine Learning Repository, see Wolberg (1992) and Mangasarian et al. (1995). Diagnostic dataset consists of breast cancer patient ID, diagnosis and multiple features of each cell nucleus. Prognostic dataset represents follow-up cases of the patients from the diagnostic dataset whose cancer was surgically removed, where each observation consists of patient ID, outcome of whether the cancer recurred or not, time to recurrence or time of last checkup and multiple measurement of the cell characteristics. We matched the two databases based on patient ID; however, not all patients from diagnostic dataset were followed-up, leading to complete missing scenario. The missing rate is 34%. Out of available observations, the censoring rate is 74.8%. The time is in months, and we estimate the density of the time to recurrence on the interval $[0,80]$ because no recurrences occur past that time.

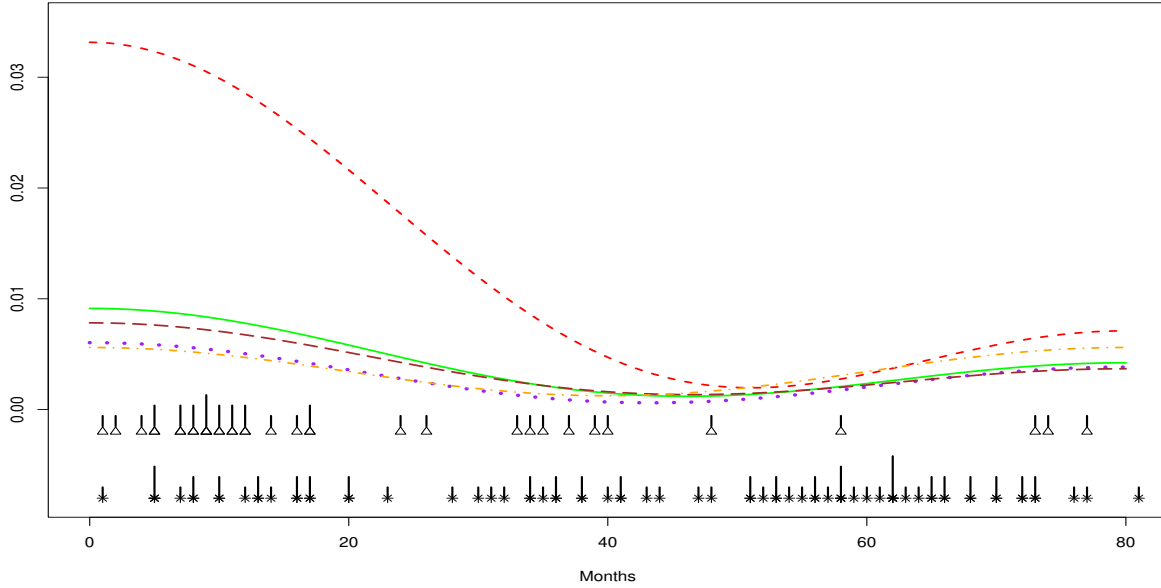


Figure 2: Density estimates for breast cancer data. The triangles indicate uncensored available lifetimes, the stars indicate censored available lifetimes. The vertical lines are proportional to the frequency of survival times found in the dataset, e.g. at time of 1 month there are 1 uncensored observations and 1 censored observations. The solid green line is the MCAR E-estimate, the dashed red line is the naive estimator which ignored censoring and only involves uncensored cases, the dotted purple line is the mean imputation estimator, the longdash brown line is resampling estimator and the dotdash orange line is the multiple resampling estimator.

S1.3 Mortality of infants data

Infant mortality in underdeveloped countries remains largely due to serious bacterial infections. The dataset we analyze is from WHO Young Infants Study Group, see discussion of the detailed methodology of data collection in The WHO Young Infants Study Group (1999a), The WHO Young Infants Study Group (1999b) and Harrell et al. (1998). The data is publicly available at Vanderbilt Biostatistics Datasets website. In brief, the data was collected over a period of 2 years from August 1991 to July 1993 in Ethio-Swedish Children's Hospital in Addis Ababa for infants up to 90 days old. For each infant patient, we observe either the number of days to death or to discharge and an indicator of whether the patient died. The rate of missing is 66% and censoring rate is 89.6%. We can assume that missing mechanism is MAR, meaning that the probability of missing a time to death or discharge depends on the outcome. Indeed, the probability of

missing the time for a dead infant is 0.08 while the probability of missing the time for a discharged infant is 0.73. The time is days, and we will estimate the density on $[0,10]$ because the vast majority of times are observed in this interval.

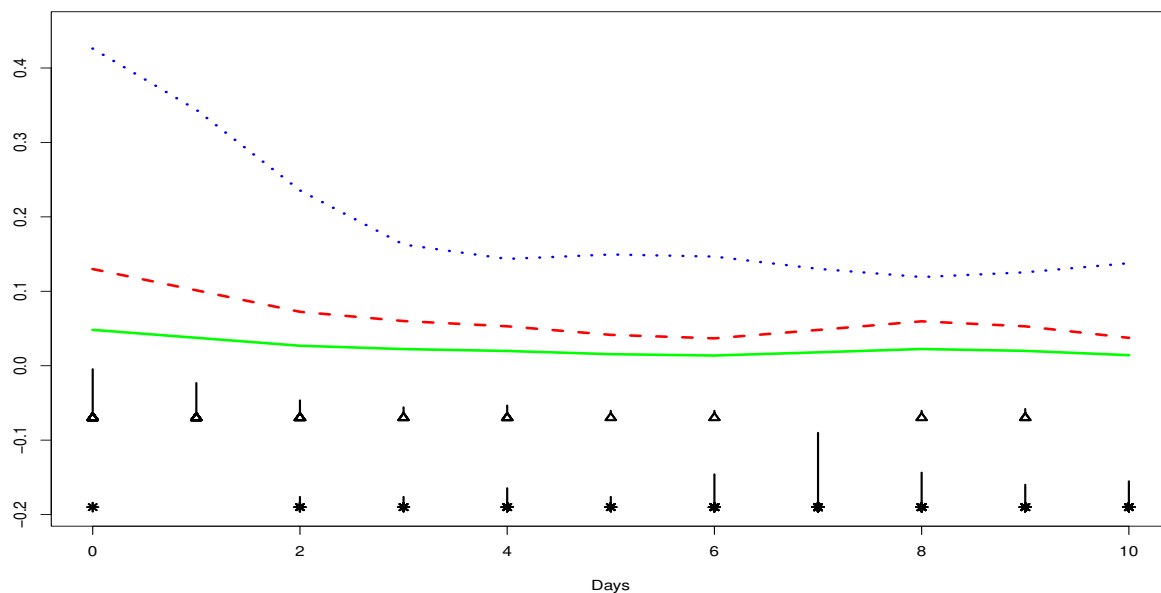


Figure 3.a: Density estimates for young infants data. The triangles indicate uncensored available lifetimes, the stars indicate censored available lifetimes in the considered interval. The vertical lines are proportional to the frequency of survival times found in the dataset, e.g. at day 0 there are 26 uncensored observations and 2 censored observations. The solid green line is the MAR E-estimate, the dashed red line is the estimator which ignored censoring and only involves uncensored cases, the dotted blue line is the naive estimate which ignored both missing and censoring.

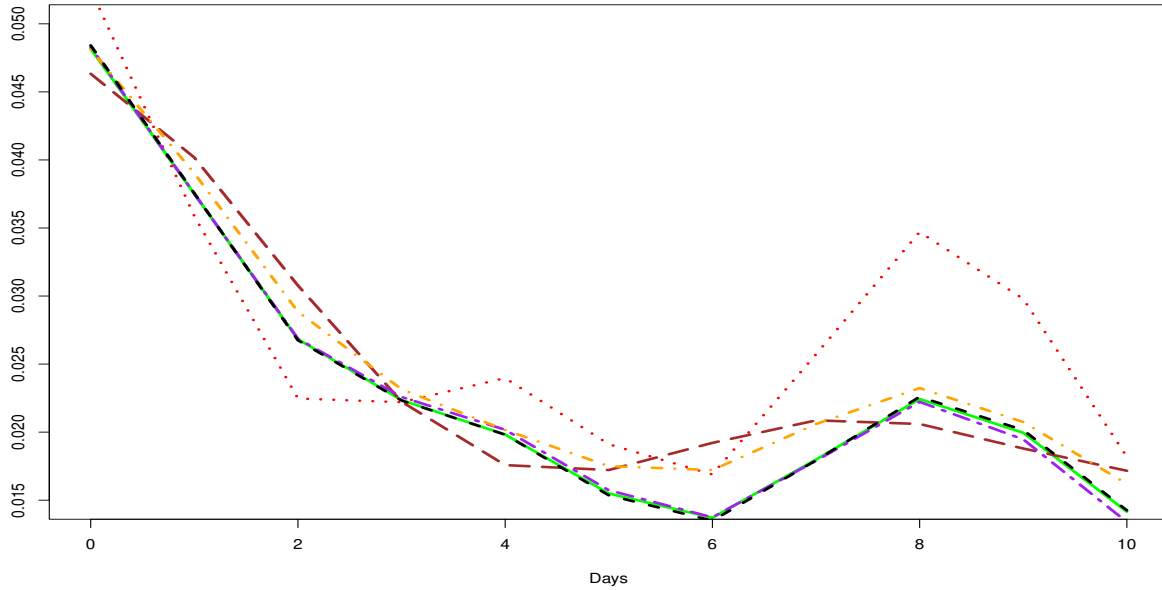


Figure 3.b: Benchmark and imputation estimates for young infants data. The green solid line is the MAR E-estimate, the orange dotdashed line is multiple resampling estimator, the brown longdash line is resampling estimate, the black dashed line is multiple hotdeck estimate, the purple twodashed line is hotdeck estimate and the dotted red line is mean estimator.

Figure 3.b shows a particular result of imputation estimators. In terms of closeness to the Benchmark, multiple hot deck performs the best, followed by hotdeck, resampling, multiple resampling and mean imputation. We reach the same conclusion when repeating the imputations for 1000 times and comparing the AISE of each imputation method, treating the Benchmark as the “true” density.

S1.4 Lung cancer data

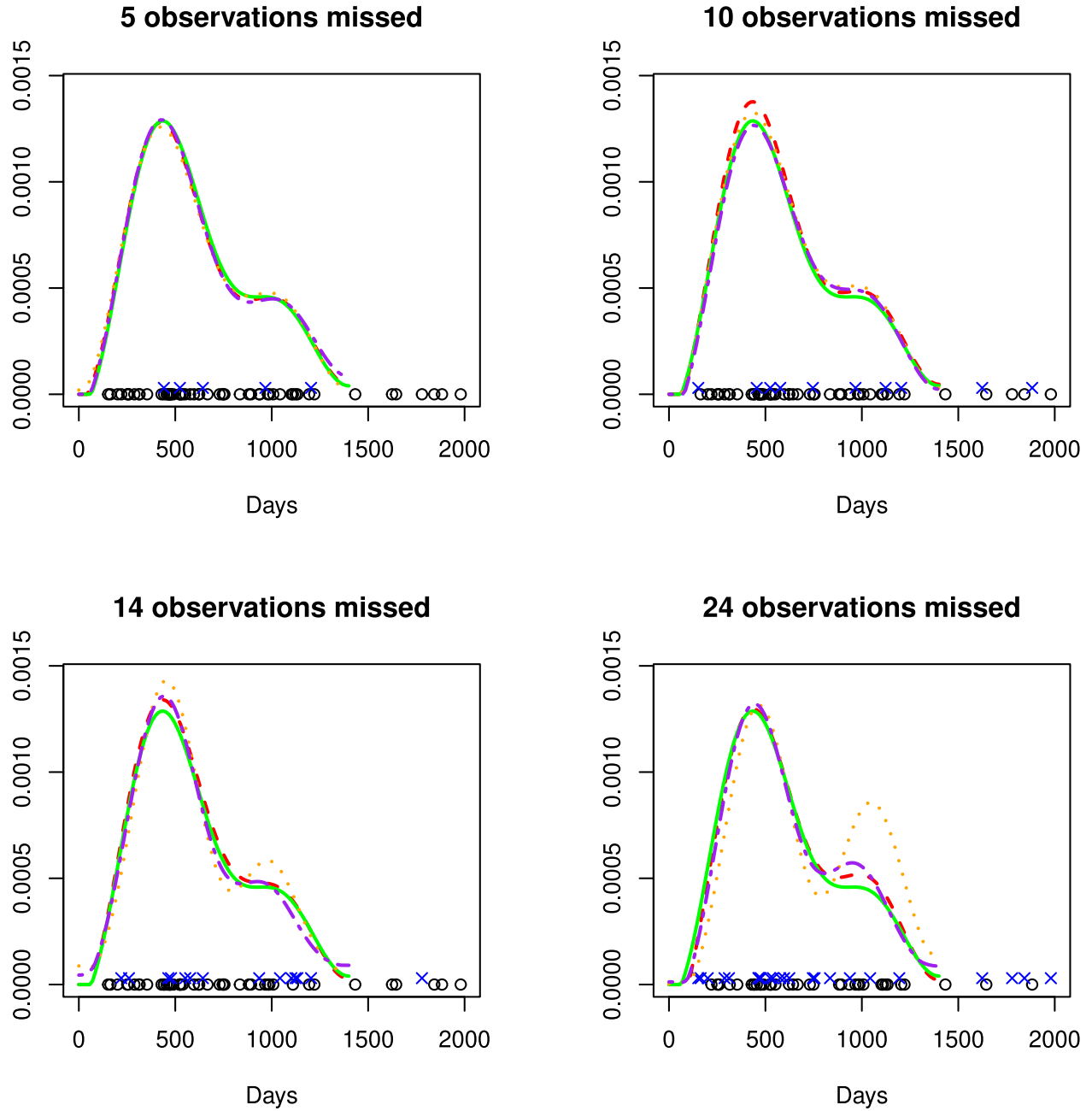


Figure 5: Density estimates for Arm A lung cancer data with various degrees of missingness. For each missing scenario, the black circles and blue crosses show available and not available censored lifetimes. The green solid line is the E-estimate. The red dashed line is the Benchmark. The orange dotted line and the purple dot-dashed line are multiple resampling and multiple hot deck imputation estimates, respectively.

S2. Simulation results

S2.1 MCAR simulations

In the paper results for two particular simulations with T having the Uniform and the Bimodal distributions of Efromovich (2018) and C being Uniform(0,2) are presented in Tables 3 and 4. Here the interested reader can find more tables as well as figures with outcomes for different underlying models. For T the underlying distributions are the Uniform, Normal, Bimodal and Strata, defined in Efromovich (2018) and presenting different patterns of smoothness. Distributions of the censoring lifetime C are Uniform(0,2), Uniform(0,1) and Exponential with rate 1/1.5. The C generated from Uniform(0,2) yields approximately 25% of censored observations. The exponential C yields about 30% of censored observations. The C generating from Uniform(0,1) yields the highest rate of censoring when about half of observations are censored.

Conclusions from the simulations are the same as those presented in the paper.

Table 1: Simulation results for Uniform when $C \sim U(0,2)$

$C \sim U(0,2)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.58	0.60	0.51	0.70	0.76
	$\omega = 0.9$	0.41	0.66	0.52	0.92	0.69
	$\omega = 0.85$	0.22	0.50	0.53	0.80	0.73
	$\omega = 0.8$	0.19	0.59	0.61	0.78	0.74
	$\omega = 0.75$	0.12	0.69	0.40	0.81	0.56
	$\omega = 0.7$	0.08	0.44	0.32	0.60	0.48
n=100	$\omega = 0.95$	0.66	0.85	0.84	1.08	1.00
	$\omega = 0.9$	0.36	0.82	0.67	1.08	0.84
	$\omega = 0.85$	0.16	0.64	0.47	0.88	0.67
	$\omega = 0.8$	0.10	0.63	0.46	0.88	0.71
	$\omega = 0.75$	0.07	0.53	0.46	0.89	0.67
	$\omega = 0.7$	0.05	0.61	0.42	0.88	0.63
n=200	$\omega = 0.95$	0.33	0.70	0.63	0.99	0.84
	$\omega = 0.9$	0.14	0.68	0.51	0.89	0.78
	$\omega = 0.85$	0.07	0.62	0.51	0.90	0.74
	$\omega = 0.8$	0.04	0.58	0.42	0.80	0.70
	$\omega = 0.75$	0.03	0.49	0.41	0.78	0.62
	$\omega = 0.7$	0.02	0.44	0.31	0.64	0.49
n=1000	$\omega = 0.95$	0.09	0.93	0.84	1.00	0.87
	$\omega = 0.9$	0.02	0.67	0.50	0.86	0.66
	$\omega = 0.85$	0.01	0.42	0.59	0.70	0.66
	$\omega = 0.8$	0.01	0.46	0.49	0.74	0.63
	$\omega = 0.75$	0.00	0.38	0.31	0.59	0.52
	$\omega = 0.7$	0.00	0.43	0.44	0.67	0.56

Table 2: Simulation results for Uniform when $C \sim \exp(\text{rate}=1/1.5)$

$C \sim \exp(\text{rate}=1/1.5)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.74	0.92	0.94	0.99	1.00
	$\omega = 0.9$	0.49	0.75	0.63	1.00	0.91
	$\omega = 0.85$	0.28	0.62	0.73	0.81	0.78
	$\omega = 0.8$	0.17	0.57	0.46	0.81	0.67
	$\omega = 0.75$	0.11	0.53	0.44	0.77	0.58
	$\omega = 0.7$	0.11	0.41	0.39	0.70	0.54
n=100	$\omega = 0.95$	0.73	0.94	0.85	1.06	1.00
	$\omega = 0.9$	0.39	0.70	0.67	0.88	0.80
	$\omega = 0.85$	0.19	0.69	0.63	0.88	0.78
	$\omega = 0.8$	0.10	0.55	0.48	0.91	0.68
	$\omega = 0.75$	0.06	0.57	0.44	0.87	0.62
	$\omega = 0.7$	0.04	0.48	0.38	0.74	0.53
n=200	$\omega = 0.95$	0.44	1.02	0.68	1.19	0.98
	$\omega = 0.9$	0.17	0.67	0.59	0.91	0.84
	$\omega = 0.85$	0.06	0.42	0.43	0.66	0.65
	$\omega = 0.8$	0.05	0.61	0.47	0.89	0.71
	$\omega = 0.75$	0.03	0.47	0.44	0.74	0.55
	$\omega = 0.7$	0.02	0.49	0.39	0.71	0.56
n=1000	$\omega = 0.95$	0.11	0.78	0.70	0.92	0.85
	$\omega = 0.9$	0.03	0.76	0.72	0.87	0.88
	$\omega = 0.85$	0.01	0.50	0.46	0.74	0.63
	$\omega = 0.8$	0.01	0.45	0.41	0.69	0.61
	$\omega = 0.75$	0.00	0.26	0.23	0.37	0.36
	$\omega = 0.7$	0.00	0.28	0.29	0.44	0.39

Table 3: Simulation results for Uniform when $C \sim U(0,1)$

$C \sim U(0,1)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.88	0.96	0.88	1.11	0.93
	$\omega = 0.9$	0.87	0.99	0.81	1.16	0.95
	$\omega = 0.85$	0.75	0.95	0.82	1.27	0.92
	$\omega = 0.8$	0.53	0.79	0.61	1.18	0.78
	$\omega = 0.75$	0.40	0.89	0.62	1.20	0.76
	$\omega = 0.7$	0.35	0.89	0.57	1.23	0.73
n=100	$\omega = 0.95$	1.07	1.02	0.91	1.19	1.00
	$\omega = 0.9$	0.83	0.93	0.73	1.17	0.93
	$\omega = 0.85$	0.55	0.90	0.71	1.15	0.89
	$\omega = 0.8$	0.38	0.91	0.65	1.28	0.77
	$\omega = 0.75$	0.25	0.78	0.60	1.24	0.72
	$\omega = 0.7$	0.18	0.80	0.54	1.17	0.66
n=200	$\omega = 0.95$	0.94	0.94	0.92	1.21	1.06
	$\omega = 0.9$	0.60	1.11	0.79	1.18	0.99
	$\omega = 0.85$	0.36	0.91	0.74	1.15	0.87
	$\omega = 0.8$	0.23	0.81	0.61	1.17	0.81
	$\omega = 0.75$	0.16	0.83	0.60	1.21	0.76
	$\omega = 0.7$	0.09	0.69	0.45	1.09	0.62
n=1000	$\omega = 0.95$	0.66	0.86	0.81	1.06	0.96
	$\omega = 0.9$	0.28	1.02	0.82	1.16	0.95
	$\omega = 0.85$	0.11	0.99	0.67	1.24	0.86
	$\omega = 0.8$	0.07	1.01	0.71	1.31	0.82
	$\omega = 0.75$	0.04	0.87	0.79	1.25	0.83
	$\omega = 0.7$	0.03	0.88	0.58	1.44	0.77

Table 4: Simulation results for Normal when $C \sim U(0,2)$

$C \sim U(0,2)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.99	0.99	0.98	1.06	1.05
	$\omega = 0.9$	0.75	0.83	0.89	1.01	1.04
	$\omega = 0.85$	0.59	0.80	0.80	1.04	1.00
	$\omega = 0.8$	0.38	0.62	0.82	0.93	1.04
	$\omega = 0.75$	0.29	0.67	0.82	0.93	1.04
	$\omega = 0.7$	0.22	0.58	0.72	0.77	0.92
n=100	$\omega = 0.95$	0.90	0.94	0.93	1.05	1.01
	$\omega = 0.9$	0.71	0.87	0.86	1.10	1.09
	$\omega = 0.85$	0.51	0.96	0.95	1.08	1.05
	$\omega = 0.8$	0.32	0.85	0.89	1.01	1.01
	$\omega = 0.75$	0.23	0.81	0.83	0.97	0.99
	$\omega = 0.7$	0.18	0.69	0.75	0.92	0.95
n=200	$\omega = 0.95$	0.87	0.92	0.89	0.96	0.98
	$\omega = 0.9$	0.60	0.87	0.90	0.96	1.04
	$\omega = 0.85$	0.35	0.91	0.87	0.98	1.04
	$\omega = 0.8$	0.20	0.77	0.80	0.95	1.04
	$\omega = 0.75$	0.12	0.66	0.72	0.84	1.00
	$\omega = 0.7$	0.10	0.61	0.80	0.81	1.09
n=1000	$\omega = 0.95$	0.29	0.91	0.94	0.93	1.04
	$\omega = 0.9$	0.08	0.70	0.85	0.78	1.00
	$\omega = 0.85$	0.04	0.53	0.83	0.64	0.93
	$\omega = 0.8$	0.02	0.42	0.74	0.52	0.85
	$\omega = 0.75$	0.01	0.36	0.59	0.43	0.80
	$\omega = 0.7$	0.01	0.35	0.58	0.40	0.81

Table 5: Simulation results for Normal when $C \sim \exp(\text{rate}=1/1.5)$

$C \sim \exp(\text{rate}=1/1.5)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.92	0.94	0.96	1.05	1.07
	$\omega = 0.9$	0.72	0.87	0.94	1.09	1.11
	$\omega = 0.85$	0.58	0.79	0.95	1.04	1.09
	$\omega = 0.8$	0.40	0.78	0.80	1.04	1.07
	$\omega = 0.75$	0.30	0.78	0.72	1.02	1.03
	$\omega = 0.7$	0.23	0.71	0.61	0.90	0.92
n=100	$\omega = 0.95$	0.91	0.95	0.97	1.00	1.06
	$\omega = 0.9$	0.74	0.99	0.89	1.05	1.07
	$\omega = 0.85$	0.50	0.82	0.92	0.99	1.08
	$\omega = 0.8$	0.39	0.88	0.87	1.02	1.06
	$\omega = 0.75$	0.27	0.76	0.82	0.91	1.03
	$\omega = 0.7$	0.19	0.65	0.67	0.88	0.92
n=200	$\omega = 0.95$	0.95	1.03	1.01	1.02	1.04
	$\omega = 0.9$	0.58	0.82	0.88	0.96	0.97
	$\omega = 0.85$	0.30	0.77	0.77	0.86	1.01
	$\omega = 0.8$	0.19	0.60	0.77	0.83	0.96
	$\omega = 0.75$	0.13	0.61	0.71	0.80	0.90
	$\omega = 0.7$	0.11	0.67	0.77	0.81	1.00
n=1000	$\omega = 0.95$	0.28	0.86	0.88	0.97	1.00
	$\omega = 0.9$	0.09	0.67	0.83	0.84	0.96
	$\omega = 0.85$	0.04	0.56	0.72	0.61	0.91
	$\omega = 0.8$	0.02	0.41	0.67	0.48	0.92
	$\omega = 0.75$	0.02	0.38	0.67	0.42	0.87
	$\omega = 0.7$	0.01	0.30	0.62	0.35	0.84

Table 6: Simulation results for Normal when $C \sim U(0,1)$

$C \sim U(0,1)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.97	1.00	0.98	1.05	1.03
	$\omega = 0.9$	0.88	0.96	0.88	1.03	0.99
	$\omega = 0.85$	0.79	1.02	0.93	1.12	1.01
	$\omega = 0.8$	0.58	0.91	0.80	1.03	0.93
	$\omega = 0.75$	0.49	0.93	0.67	1.28	0.91
	$\omega = 0.7$	0.40	0.86	0.64	1.02	0.80
n=100	$\omega = 0.95$	1.03	0.96	0.97	1.06	1.06
	$\omega = 0.9$	0.86	0.93	0.93	1.15	1.04
	$\omega = 0.85$	0.64	1.03	0.82	1.23	1.02
	$\omega = 0.8$	0.43	0.91	0.76	1.13	0.97
	$\omega = 0.75$	0.29	0.73	0.75	0.96	0.92
	$\omega = 0.7$	0.23	0.73	0.79	0.96	0.92
n=200	$\omega = 0.95$	1.00	0.96	1.01	1.12	1.12
	$\omega = 0.9$	0.69	0.94	1.09	1.07	1.19
	$\omega = 0.85$	0.45	0.84	0.95	0.96	1.09
	$\omega = 0.8$	0.27	0.72	0.83	0.86	1.01
	$\omega = 0.75$	0.18	0.76	0.79	0.86	0.99
	$\omega = 0.7$	0.12	0.58	0.72	0.74	0.98
n=1000	$\omega = 0.95$	0.54	0.85	0.86	0.95	0.94
	$\omega = 0.9$	0.15	0.71	0.80	0.76	0.94
	$\omega = 0.85$	0.06	0.49	0.80	0.56	0.90
	$\omega = 0.8$	0.04	0.40	0.68	0.41	0.90
	$\omega = 0.75$	0.03	0.33	0.69	0.37	0.91
	$\omega = 0.7$	0.02	0.27	0.65	0.32	0.88

Table 7: Simulation results for Bimodal when $C \sim U(0,2)$

$C \sim U(0,2)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.96	0.98	1.00	1.04	1.07
	$\omega = 0.9$	0.93	0.96	0.96	1.06	1.06
	$\omega = 0.85$	0.84	0.97	0.96	1.08	1.08
	$\omega = 0.8$	0.80	0.96	0.98	1.12	1.14
	$\omega = 0.75$	0.79	0.99	1.03	1.14	1.17
	$\omega = 0.7$	0.67	0.96	0.97	1.16	1.12
n=100	$\omega = 0.95$	0.92	0.99	1.02	1.06	1.06
	$\omega = 0.9$	0.85	1.02	1.02	1.10	1.10
	$\omega = 0.85$	0.75	1.01	1.05	1.14	1.14
	$\omega = 0.8$	0.69	1.01	0.98	1.17	1.15
	$\omega = 0.75$	0.61	1.04	1.06	1.19	1.24
	$\omega = 0.7$	0.49	1.06	1.06	1.25	1.26
n=200	$\omega = 0.95$	0.75	1.01	1.01	1.06	1.08
	$\omega = 0.9$	0.56	0.99	1.00	1.07	1.10
	$\omega = 0.85$	0.50	1.05	1.06	1.13	1.21
	$\omega = 0.8$	0.45	1.03	1.07	1.12	1.24
	$\omega = 0.75$	0.36	1.00	1.10	1.14	1.26
	$\omega = 0.7$	0.26	1.02	1.17	1.17	1.30
n=1000	$\omega = 0.95$	0.52	0.96	0.97	1.01	1.02
	$\omega = 0.9$	0.17	0.95	0.96	1.01	1.02
	$\omega = 0.85$	0.09	0.86	0.90	0.94	1.01
	$\omega = 0.8$	0.07	0.86	1.00	0.90	1.04
	$\omega = 0.75$	0.05	0.81	0.95	0.92	1.13
	$\omega = 0.7$	0.03	0.77	0.95	0.87	1.11

Table 8: Simulation results for Bimodal when $C \sim \exp(\text{rate}=1/1.5)$

$C \sim \exp(\text{rate}=1/1.5)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.95	0.99	0.97	1.04	1.06
	$\omega = 0.9$	0.97	1.00	1.02	1.08	1.12
	$\omega = 0.85$	0.94	1.01	1.01	1.11	1.14
	$\omega = 0.8$	0.86	1.05	1.00	1.14	1.14
	$\omega = 0.75$	0.77	1.00	0.99	1.13	1.15
	$\omega = 0.7$	0.63	1.00	1.01	1.17	1.16
n=100	$\omega = 0.95$	0.91	1.01	1.02	1.10	1.09
	$\omega = 0.9$	0.83	1.03	1.08	1.14	1.14
	$\omega = 0.85$	0.76	1.03	1.08	1.16	1.19
	$\omega = 0.8$	0.64	1.06	1.07	1.20	1.24
	$\omega = 0.75$	0.56	1.03	1.06	1.24	1.23
	$\omega = 0.7$	0.47	1.05	1.05	1.28	1.29
n=200	$\omega = 0.95$	0.79	1.03	1.06	1.08	1.09
	$\omega = 0.9$	0.62	1.01	1.02	1.09	1.12
	$\omega = 0.85$	0.50	0.97	1.02	1.09	1.16
	$\omega = 0.8$	0.46	0.98	1.10	1.11	1.22
	$\omega = 0.75$	0.40	1.07	1.06	1.18	1.25
	$\omega = 0.7$	0.32	1.01	1.10	1.20	1.34
n=1000	$\omega = 0.95$	0.55	1.01	0.96	1.03	1.01
	$\omega = 0.9$	0.19	0.89	0.92	0.95	0.98
	$\omega = 0.85$	0.09	0.89	0.96	0.96	1.04
	$\omega = 0.8$	0.07	0.79	0.91	0.89	1.04
	$\omega = 0.75$	0.05	0.78	0.93	0.87	1.06
	$\omega = 0.7$	0.04	0.77	0.92	0.82	1.09

Table 9: Simulation results for Bimodal when $C \sim U(0,1)$

$C \sim U(0,1)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	1.01	0.99	0.97	1.04	1.02
	$\omega = 0.9$	0.97	0.96	0.93	1.02	1.01
	$\omega = 0.85$	0.94	1.00	0.91	1.06	1.00
	$\omega = 0.8$	0.93	0.96	0.90	1.09	1.02
	$\omega = 0.75$	0.84	1.03	0.93	1.15	1.03
	$\omega = 0.7$	0.77	0.97	0.92	1.14	1.07
n=100	$\omega = 0.95$	0.97	1.00	1.00	1.07	1.07
	$\omega = 0.9$	0.96	1.02	1.03	1.13	1.12
	$\omega = 0.85$	0.91	1.01	1.01	1.14	1.12
	$\omega = 0.8$	0.85	1.01	1.01	1.18	1.18
	$\omega = 0.75$	0.78	1.01	1.00	1.23	1.17
	$\omega = 0.7$	0.71	1.04	1.02	1.24	1.22
n=200	$\omega = 0.95$	0.92	0.99	1.03	1.07	1.07
	$\omega = 0.9$	0.89	1.06	1.05	1.14	1.14
	$\omega = 0.85$	0.83	1.00	1.06	1.13	1.19
	$\omega = 0.8$	0.80	1.06	1.10	1.22	1.25
	$\omega = 0.75$	0.74	1.05	1.07	1.26	1.27
	$\omega = 0.7$	0.64	1.07	1.15	1.26	1.34
n=1000	$\omega = 0.95$	0.80	0.98	0.99	1.00	1.02
	$\omega = 0.9$	0.47	0.91	0.95	1.00	1.07
	$\omega = 0.85$	0.27	0.84	1.00	0.98	1.15
	$\omega = 0.8$	0.21	0.88	1.08	0.98	1.22
	$\omega = 0.75$	0.17	0.86	1.08	0.95	1.29
	$\omega = 0.7$	0.12	0.73	1.07	0.88	1.28

Table 10: Simulation results for Strata when $C \sim U(0,2)$

$C \sim U(0,2)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.87	0.99	0.99	1.02	1.02
	$\omega = 0.9$	0.79	0.94	0.97	1.01	1.02
	$\omega = 0.85$	0.71	0.93	0.99	0.98	1.03
	$\omega = 0.8$	0.60	0.92	1.01	1.01	1.06
	$\omega = 0.75$	0.49	0.97	0.99	1.05	1.08
	$\omega = 0.7$	0.39	0.97	1.00	1.06	1.09
n=100	$\omega = 0.95$	0.82	1.01	1.03	1.05	1.07
	$\omega = 0.9$	0.68	0.99	1.06	1.07	1.10
	$\omega = 0.85$	0.60	1.01	1.05	1.08	1.11
	$\omega = 0.8$	0.52	0.99	1.02	1.08	1.14
	$\omega = 0.75$	0.39	1.00	1.04	1.10	1.14
	$\omega = 0.7$	0.28	1.04	1.10	1.11	1.21
n=200	$\omega = 0.95$	0.67	1.00	0.97	1.00	1.02
	$\omega = 0.9$	0.49	0.96	1.02	1.00	1.05
	$\omega = 0.85$	0.39	0.92	1.01	0.97	1.05
	$\omega = 0.8$	0.33	0.91	1.05	0.97	1.09
	$\omega = 0.75$	0.23	0.91	1.02	0.97	1.13
	$\omega = 0.7$	0.16	0.94	1.00	1.00	1.16
n=1000	$\omega = 0.95$	0.65	0.96	1.00	0.97	1.00
	$\omega = 0.9$	0.30	0.93	0.99	0.93	1.01
	$\omega = 0.85$	0.20	0.89	0.99	0.90	1.01
	$\omega = 0.8$	0.13	0.85	1.00	0.87	1.01
	$\omega = 0.75$	0.08	0.82	0.98	0.83	1.00
	$\omega = 0.7$	0.06	0.77	0.97	0.79	1.00

Table 11: Simulation results for Strata when $C \sim \exp(\text{rate}=1/1.5)$

$C \sim \exp(\text{rate}=1/1.5)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	0.92	1.01	1.00	1.02	1.03
	$\omega = 0.9$	0.85	0.96	1.02	1.01	1.04
	$\omega = 0.85$	0.79	0.95	0.98	1.01	1.04
	$\omega = 0.8$	0.69	0.96	1.00	1.02	1.04
	$\omega = 0.75$	0.56	0.95	1.01	1.03	1.05
	$\omega = 0.7$	0.43	1.01	1.00	1.07	1.08
n=100	$\omega = 0.95$	0.79	1.00	1.01	1.03	1.06
	$\omega = 0.9$	0.67	1.01	1.01	1.06	1.08
	$\omega = 0.85$	0.60	0.99	1.04	1.06	1.13
	$\omega = 0.8$	0.50	0.96	1.05	1.05	1.14
	$\omega = 0.75$	0.39	0.97	1.05	1.08	1.17
	$\omega = 0.7$	0.28	0.97	1.00	1.11	1.21
n=200	$\omega = 0.95$	0.67	0.97	1.01	1.01	1.03
	$\omega = 0.9$	0.51	0.93	1.01	1.00	1.05
	$\omega = 0.85$	0.43	0.95	1.02	0.98	1.06
	$\omega = 0.8$	0.34	0.95	1.00	0.99	1.08
	$\omega = 0.75$	0.25	0.94	1.03	0.98	1.11
	$\omega = 0.7$	0.18	0.87	1.07	0.99	1.16
n=1000	$\omega = 0.95$	0.65	0.96	1.00	0.96	1.00
	$\omega = 0.9$	0.32	0.91	0.99	0.93	1.00
	$\omega = 0.85$	0.21	0.87	0.99	0.89	1.01
	$\omega = 0.8$	0.13	0.82	0.99	0.85	1.01
	$\omega = 0.75$	0.09	0.79	0.97	0.81	1.00
	$\omega = 0.7$	0.06	0.75	0.95	0.77	1.01

Table 12: Simulation results for Strata when $C \sim U(0,1)$

$C \sim U(0,1)$		AISEB/AISEM	AISEB/AISER	AISEB/AISEHD	AISEB/AISEMR	AISEB/AISEMHD
n=50	$\omega = 0.95$	1.00	1.00	1.00	1.04	1.04
	$\omega = 0.9$	0.98	0.95	0.94	1.04	1.02
	$\omega = 0.85$	0.95	1.00	0.96	1.06	1.05
	$\omega = 0.8$	0.91	0.92	0.96	1.05	1.05
	$\omega = 0.75$	0.88	0.96	0.96	1.15	1.07
	$\omega = 0.7$	0.77	1.01	1.03	1.17	1.14
n=100	$\omega = 0.95$	0.93	0.98	1.02	1.05	1.06
	$\omega = 0.9$	0.87	0.99	1.04	1.08	1.10
	$\omega = 0.85$	0.86	1.01	1.04	1.09	1.15
	$\omega = 0.8$	0.82	1.02	1.07	1.11	1.16
	$\omega = 0.75$	0.76	0.99	1.07	1.11	1.18
	$\omega = 0.7$	0.62	0.95	1.05	1.10	1.20
n=200	$\omega = 0.95$	0.82	0.99	1.01	1.03	1.06
	$\omega = 0.9$	0.69	0.95	1.06	1.03	1.10
	$\omega = 0.85$	0.62	0.95	1.09	1.05	1.14
	$\omega = 0.8$	0.60	0.94	1.07	1.02	1.16
	$\omega = 0.75$	0.56	0.92	1.06	1.03	1.19
	$\omega = 0.7$	0.46	0.95	1.06	1.04	1.23
n=1000	$\omega = 0.95$	0.72	0.94	0.99	0.95	1.00
	$\omega = 0.9$	0.37	0.86	0.97	0.89	1.01
	$\omega = 0.85$	0.23	0.80	0.99	0.82	1.00
	$\omega = 0.8$	0.19	0.75	0.99	0.77	1.02
	$\omega = 0.75$	0.15	0.68	0.99	0.72	1.03
	$\omega = 0.7$	0.11	0.63	1.00	0.67	1.04

The figures on the following pages are a graphical summary of the ratios of average ISEs presented in the tables. Each figure represents a different censoring distribution, while the four diagrams in each figure correspond to the four underlying densities f^T . In each diagram the vertical dashed lines separate experiments with different sample sizes and the corresponding integers 1,2,3 and 4 indicate the underlying sample size for the experiments in the group (1 indicates $n = 50$, 2 indicates $n = 100$, 3 indicates $n = 200$, 4 indicates $n = 1000$). Further, for each sample size the tick marks indicate the availability likelihood $w = 0.95, 0.9, 0.85, 0.8, 0.75, 0.7$, respectively. The letters A, B, C, D, E indicate AISEB/AISEM, AISEB/AISER, AISEB/AISEHD, AISEB/AISEMR and AISEB/AISEMHD respectively.

These figures shed an interesting light on the empirical results.

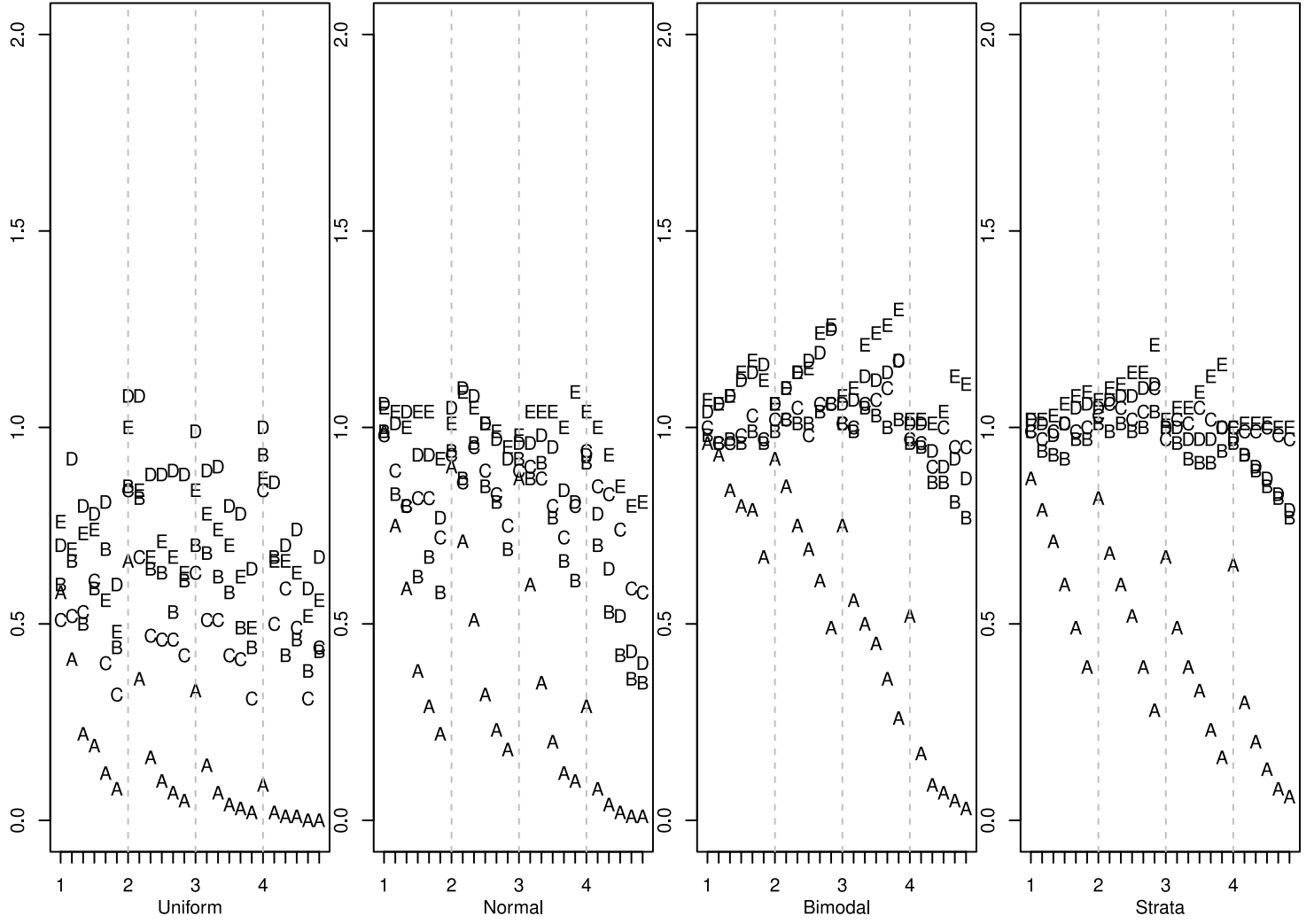


Figure 6: Graphical summary of ratios of average ISEs when censoring is Uniform(0,2)

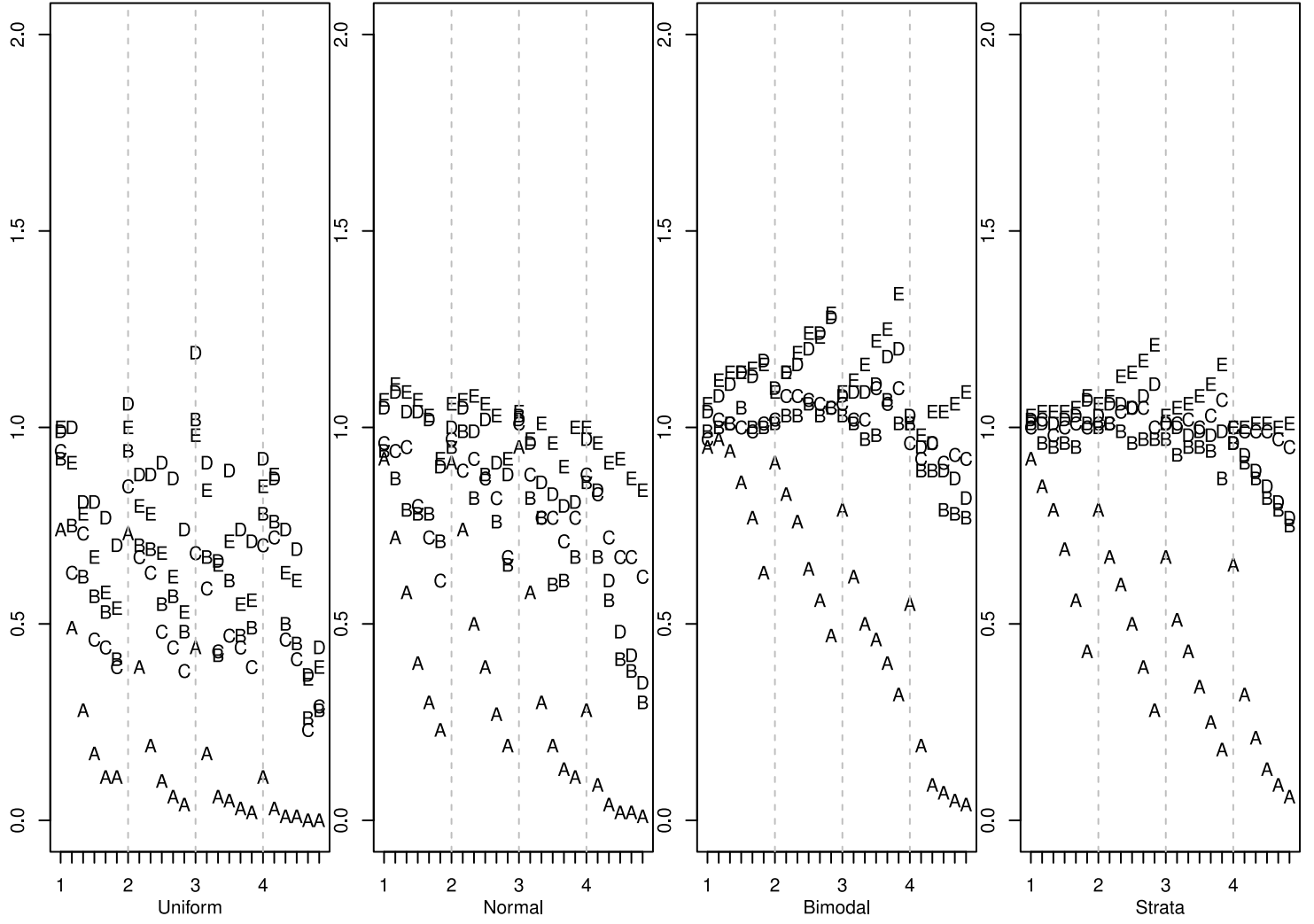


Figure 7: Graphical summary of ratios of average ISEs when censoring is Exponential(rate=1/1.5)

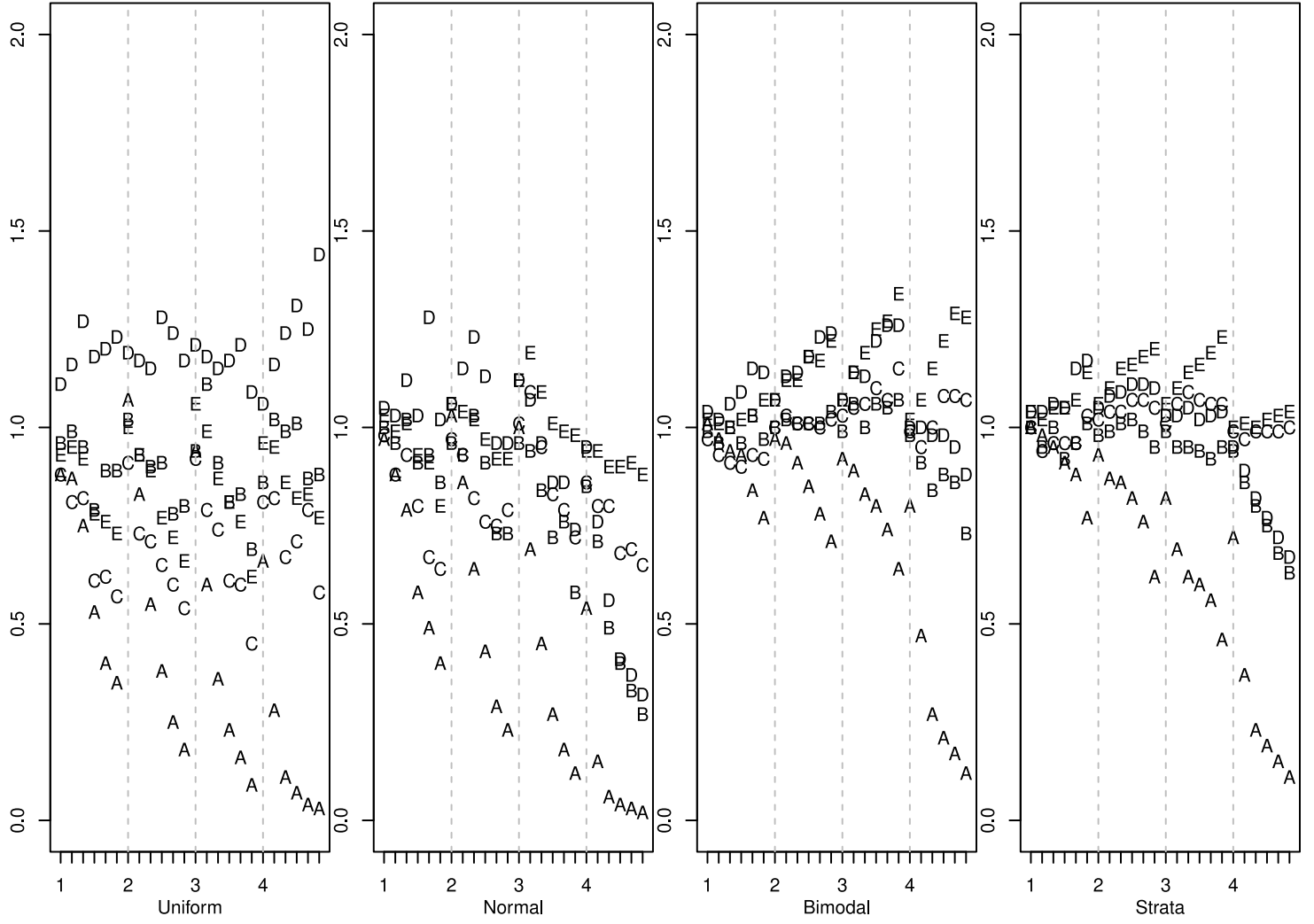


Figure 8: Graphical summary of ratios of average ISEs when censoring is Uniform(0,1)

Table 13: Effect of the number m of repetitions on $AISB/AISMR$

Model	n	m	$w = 0.95$	$w = 0.90$	$w = 0.85$	$w = 0.80$	$w = 0.75$	$w = 0.70$
$T \sim Unif(0, 1)$ $C \sim Unif(0, 2)$	50	5	0.70	0.92	0.80	0.78	0.81	0.60
	50	10	0.80	0.97	0.85	0.81	0.79	0.69
	50	30	0.88	0.96	0.83	0.87	0.78	0.72
	100	5	1.08	1.08	0.88	0.88	0.89	0.88
	100	10	1.06	1.08	0.97	0.88	0.98	0.95
	100	30	1.08	1.18	1.01	0.96	0.99	0.93
	200	5	0.99	0.89	0.90	0.80	0.78	0.64
	200	10	0.99	0.99	0.95	0.88	0.85	0.71
	200	30	1.03	0.99	0.99	0.92	0.92	0.74
$T \sim N(0.5, sd = 0.15)$ $C \sim Unif(0, 2)$	50	5	1.06	1.01	1.04	0.93	0.93	0.77
	50	10	1.08	1.04	1.05	0.95	0.96	0.80
	50	30	1.09	1.07	1.06	0.98	1.00	0.86
	100	5	1.05	1.10	1.08	1.01	0.97	0.92
	100	10	1.04	1.11	1.11	1.08	0.99	0.97
	100	30	1.05	1.11	1.10	1.08	1.01	1.02
	200	5	0.96	0.96	0.98	0.95	0.84	0.81
	200	10	0.96	1.02	1.01	1.00	0.87	0.86
	200	30	0.99	1.05	1.04	1.03	0.92	0.90
$T \sim Bimodal$ $C \sim Unif(0, 2)$	50	5	1.04	1.06	1.08	1.12	1.14	1.16
	50	10	1.06	1.09	1.11	1.16	1.17	1.19
	50	30	1.06	1.09	1.13	1.16	1.18	1.20
	100	5	1.06	1.10	1.14	1.17	1.19	1.25
	100	10	1.07	1.12	1.17	1.19	1.22	1.28
	100	30	1.07	1.12	1.17	1.20	1.25	1.31
	200	5	1.06	1.07	1.13	1.12	1.14	1.17
	200	10	1.06	1.09	1.16	1.16	1.19	1.21
	200	30	1.06	1.09	1.18	1.18	1.22	1.24
$T \sim Strata$ $C \sim Unif(0, 2)$	50	5	1.02	1.01	0.98	1.01	1.05	1.06
	50	10	1.02	1.01	1.00	1.01	1.05	1.06
	50	30	1.02	1.02	1.02	1.02	1.07	1.07
	100	5	1.05	1.07	1.08	1.08	1.10	1.11
	100	10	1.05	1.07	1.09	1.10	1.12	1.13
	100	30	1.06	1.07	1.08	1.11	1.12	1.15
	200	5	1.00	1.00	0.97	0.97	0.97	1.00
	200	10	1.00	1.01	0.98	0.98	0.99	1.01
	200	30	1.00	1.01	0.98	0.98	1.00	1.01
$T \sim Strata$ $C \sim Exponential(rate = 1/1.5)$	50	5	1.02	1.01	1.01	1.02	1.03	1.07
	50	10	1.02	1.02	1.02	1.01	1.02	1.09
	50	30	1.02	1.02	1.02	1.03	1.03	1.09
	100	5	1.03	1.06	1.06	1.05	1.08	1.11
	100	10	1.04	1.06	1.06	1.07	1.10	1.12
	100	30	1.05	1.06	1.08	1.08	1.12	1.14
	200	5	1.01	1.00	0.98	0.99	0.98	0.99
	200	10	1.01	1.01	0.99	1.00	0.99	1.00
	200	30	1.01	1.01	0.99	1.00	0.99	1.01

S2.2 MAR simulations

The simulations for density estimation are conducted when the random variable of interest T follows one of the four distributions with differing degrees of smoothness: Uniform, Normal, Bimodal and Strata. For each of four underlying distributions, the censoring variable C follows one of the following: Uniform(0, 2), Uniform(0, 1), or exp(rate = 1/1.5). When C is generated from Uniform(0,2), approximately 25% of the observations are censored. When C is generated from exp(rate = 1/1.5), around 30% of the observations are censored. Generating C from Uniform(0, 1) results in the highest proportion of censored observations of approximately 50%. For each combination of T and C , there are four missing scenarios considered. $\omega = 0.7*(1-\delta)+0.9*\delta$ leads to approximately 15% of observation being missed, $\omega = 0.9*(1-\delta)+0.7*\delta$ leads to 25% missing, $\omega = 0.5*(1-\delta)+0.7*\delta$ leads to 35% missing and $\omega = 0.7*(1-\delta)+0.5*\delta$ leads to 45% missing. Each combination is repeated for samples size n equals to 50, 100, 200 and 1000. For each run, integrated square error is computed. Each scenario was performed 100 times and the results are averaged to yield mean integrated square error.

In order to understand the performance of our estimator, two alternative imputation approaches were used: hot deck and multiple hot deck. In hot deck imputation, missing lifetimes are replaced by randomly chosen lifetimes from a subset of available ones with the same indicator. In multiple hot deck imputation, the above-defined hot-deck imputation is repeated m times, and then the calculated density estimates are averaged. Following van Buuren (2018), $m = 5$ is used.

The tables below summarize the simulation results for density estimation.

Table 14: Simulation results for density estimation, $T \sim U(0,1)$

	n=50		n=100		n=200		n=1000	
	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$
$C \sim U(0, 2)$								
$\omega = 0.7*(1-\delta)+0.9*\delta$	0.53	0.73	0.68	0.88	0.50	0.70	0.45	0.66
$\omega = 0.9*(1-\delta)+0.7*\delta$	0.27	0.39	0.42	0.58	0.28	0.48	0.43	0.58
$\omega = 0.5*(1-\delta)+0.7*\delta$	0.30	0.39	0.43	0.62	0.33	0.53	0.41	0.52
$\omega = 0.7*(1-\delta)+0.5*\delta$	0.13	0.18	0.23	0.32	0.25	0.39	0.27	0.39
$C \sim exp(rate = 1/1.5)$								
$\omega = 0.7*(1-\delta)+0.9*\delta$	0.69	0.96	0.70	0.85	0.55	0.78	0.64	0.82
$\omega = 0.9*(1-\delta)+0.7*\delta$	0.41	0.57	0.31	0.46	0.36	0.52	0.25	0.35
$\omega = 0.5*(1-\delta)+0.7*\delta$	0.44	0.67	0.33	0.48	0.36	0.49	0.25	0.39
$\omega = 0.7*(1-\delta)+0.5*\delta$	0.21	0.31	0.27	0.36	0.26	0.34	0.17	0.23
$C \sim U(0, 1)$								
$\omega = 0.7*(1-\delta)+0.9*\delta$	2.71	0.93	2.29	0.96	3.44	0.93	4.50	0.93
$\omega = 0.9*(1-\delta)+0.7*\delta$	1.27	0.69	1.56	0.66	1.71	0.65	2.31	0.82
$\omega = 0.5*(1-\delta)+0.7*\delta$	1.23	0.70	1.69	0.64	1.86	0.73	2.29	0.73
$\omega = 0.7*(1-\delta)+0.5*\delta$	0.74	0.49	0.68	0.42	1.18	0.54	0.64	0.36

Table 15: Simulation results for density estimation, $T \sim \text{Normal}(0, \text{sd}=0.15)$

	n=50		n=100		n=200		n=1000	
$C \sim U(0, 2)$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.91	1.06	0.88	1.10	0.85	1.05	0.86	0.98
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	0.75	0.98	0.75	0.99	0.79	1.10	0.62	0.87
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.78	0.95	0.74	0.97	0.77	1.06	0.60	0.81
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.53	0.65	0.59	0.81	0.74	0.94	0.50	0.71
$C \sim \exp(\text{rate} = 1/1.5)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.86	1.06	0.89	1.01	0.92	1.15	0.80	0.94
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	0.63	0.82	0.66	0.84	0.82	1.01	0.68	0.90
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.61	0.86	0.69	0.89	0.84	1.01	0.65	0.83
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.79	0.95	0.46	0.65	0.68	0.89	0.54	0.68
$C \sim U(0, 1)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.89	0.98	0.93	1.07	0.94	1.02	0.72	0.92
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	0.66	0.84	0.81	0.94	0.76	1.01	0.62	0.91
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.68	0.87	0.78	0.91	0.79	1.04	0.49	0.79
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.64	0.70	0.54	0.68	0.72	0.94	0.49	0.72

Table 16: Simulation results for density estimation, $T \sim \text{Bimodal}$

	n=50		n=100		n=200		n=1000	
$C \sim U(0, 2)$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.95	1.06	1.01	1.13	1.02	1.12	0.97	1.01
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	0.98	1.11	1.10	1.29	1.16	1.29	0.92	1.10
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.97	1.11	1.09	1.29	1.14	1.32	0.91	1.12
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.98	1.15	1.05	1.29	1.23	1.52	1.05	1.28
$C \sim \exp(\text{rate} = 1/1.5)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	1.01	1.08	1.06	1.14	1.04	1.12	0.92	1.00
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	1.01	1.16	1.03	1.29	1.07	1.32	0.90	1.08
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	1.02	1.16	1.04	1.29	1.08	1.26	0.94	1.06
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.89	1.04	1.20	1.38	1.13	1.38	1.09	1.30
$C \sim U(0, 1)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.96	1.04	1.02	1.12	1.03	1.15	0.97	1.06
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	0.92	1.05	1.05	1.26	1.09	1.31	1.07	1.31
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.89	1.05	1.04	1.21	1.14	1.32	1.09	1.31
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.85	0.98	0.84	1.09	1.15	1.46	1.27	1.61

Table 17: Simulation results for density estimation, $T \sim \text{Strata}$

	n=50		n=100		n=200		n=1000	
	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$
$C \sim U(0, 2)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.97	1.01	1.03	1.09	1.03	1.06	0.98	1.01
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	1.00	1.08	1.11	1.21	1.00	1.15	0.97	1.00
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	1.01	1.09	1.11	1.21	1.01	1.14	0.97	1.00
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.99	1.04	1.15	1.31	1.12	1.25	0.97	1.01
$C \sim \exp(\text{rate} = 1/1.5)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	1.01	1.04	1.01	1.09	1.03	1.05	0.98	1.00
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	0.99	1.07	1.03	1.26	1.10	1.17	0.96	1.00
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.97	1.05	0.98	1.22	1.08	1.16	0.96	1.00
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.95	1.10	1.10	1.23	1.08	1.22	0.95	1.01
$C \sim U(0, 1)$								
$\omega = 0.7 * (1 - \delta) + 0.9 * \delta$	0.97	1.05	1.02	1.12	1.06	1.12	0.97	0.99
$\omega = 0.9 * (1 - \delta) + 0.7 * \delta$	1.00	1.10	1.06	1.17	1.09	1.25	1.00	1.06
$\omega = 0.5 * (1 - \delta) + 0.7 * \delta$	0.98	1.11	1.09	1.25	1.07	1.23	0.99	1.03
$\omega = 0.7 * (1 - \delta) + 0.5 * \delta$	0.97	1.13	0.96	1.18	1.20	1.36	1.12	1.19

When T is from Uniform distribution, our Benchmark estimator generally outperforms the imputations ones and as missing becomes heavier, its performance becomes progressively better. Hot deck performs surprisingly well when $C \sim U(0, 1)$ but otherwise multiple hot deck is better than hot deck as expected but worse than the Benchmark. Similar remarks can be said for scenarios when T is from Normal distribution. While multiple hot deck performs better than hot deck, Benchmark estimator is a clear winner for all censoring and missing scenarios, especially at $n = 1000$. When T comes from Bimodal or Strata, our Benchmark estimator is comparable to multiple hot deck one. In fact, the rougher the density the better relative performance of imputation becomes. When T is Bimodal, multiple hot deck outperforms for heavier censoring cases but comparable to our Benchmark estimator for all other scenarios. When T is from Strata, the Benchmark outperforms hot deck and is comparable to multiple hot deck.

S3. Code

The main function that should be called by the user is *E_Estimate* which takes the following arguments:

V: vector of observed lifetimes.

delta: vector of indicators of censoring.

A: vector of values of availability function.

n: sample size.

a: is set to 1 because the user should first rescale the interval of estimation to $[0,1]$.

cTH: constant with preset value, controls soft thresholding procedure that selects which Fourier coefficients to keep and which should be set to zero.

knots: number of points equidistant on $[0,1]$ at which the resulting density is evaluated. The default value of knots is 100. After obtaining the density estimates at each knot, the user should rescale the resulting density onto the original interval.

cJ0, cJ1, cJM: constants with corresponding default values. User can control those to change the number of Fourier coefficients calculated.

cB: constant with preset value, controls how small the bump has to be in order to be removed.

After being called, *E_Estimate* will call the following functions:

tensor_j: generates cosine basis on $[0,1]$;

estimateCoefVar: calculates Fourier coefficients, its variance and sample-mean estimator of variance;

negden: performs bona fide projection and bump removal. This function uses *rem.bump1* and *abvec* functions from the R package Efromovich (1999).

```

E_Estimate <- function(V, delta, A, n, a = 1, cTH = 4, knots = 100,
  cJ0 = 4, cJ1 = 0.5, cJM = 6, cB = 2) {
  ### step 1: find Fourier coefficients
  Jn <- floor(cJ0 + cJ1 * log(n))
  JFourier <- cJM * Jn
  est <- estimateCoefVar(V, delta, A, n, a, JFourier) # missing of V
  coef <- est$coef
  vjn <- est$vjn

  ### step 2: find optimal empirical cutoffs J
  Msummation <- cumsum(2 * vjn - (coef ^ 2))[0:(Jn + 1)] #cutoffs up to Jn
  J <- which(Msummation == min(Msummation), arr.ind = TRUE) # optimal empirical cutoff
  thetaCutoff <- coef[0:J] # theta to be used
  thetaCutoff[thetaCutoff ^ 2 <= cTH * vjn[0:J]] <- 0 # perform thresholding

  # calculate density over [0,1]
  z1 <- seq(from = 0, to = 1, len = knots)
  if (J == 1) {
    Basis <- matrix(1, ncol = 1, nrow = length(z1))
  } else{
    Basis <- cbind(matrix(1, ncol = 1, nrow = length(z1)),
      (2 ^ (1 / 2)) * cos(outer(z1, pi * (1:(J - 1)))))
  }
  func <- Basis %*% thetaCutoff

  ### step 3: bona fide projection - calling functions from Efromovich (1999)
  denProj <- negden(func, delta = 0.01, FLAGBUMP = 1, cB = cB)
  return(denProj) # final density to return
}

tensor_j <- function(y,j){ # on [0,1]
  return(1*(j==0)+sqrt(2)*cos(pi*j*y)*1*(j>0))
}

estimateCoefVar <- function(V, delta, A, n, a = 1, J_cutoff) {
  # returns Fourier coefficients, its variance and sample-mean estimator of variance
  J <- J_cutoff - 1
  khat <- mean(A)

```

```

survival <- function(V, delta, y, n) {
  sum1 <- 0
  for (l in 1:n) {
    ind1 <- ifelse((A[l] * V[l] <= y), 1, 0)
    sum2 <- sum(A * V >= A[l] * V[l])
    sum1 <- sum1 + A[l] * (1 - delta[l]) * ind1 / sum2
  }
  return(exp(-sum1))
}

Mcoef <- rep(0, (J + 1))
Mterm <- matrix(0, ncol = (J + 1), nrow = n)
surv <- sapply(V, survival, V = V, delta = delta, n = n)

for (j in 0:J) {
  for (l in 1:n) {
    ind <- ifelse(V[l] >= 0. && V[l] <= a, 1, 0)
    tensorj <- tensor_j(V[l], j)
    Mterm[l, (j + 1)] <- (A[l] * delta[l] * ind * tensorj / (khat * surv[l]))
  }
}

Mcoef <- apply(Mterm, 2, mean) # Fourier coefficient
Mvar <- apply(Mterm, 2, var) # variance of each Fourier
vjn <- Mvar / n # variance of each Fourier coefficient
retlist <- list("coef" = Mcoef, "vjn" = vjn, "var" = Mvar)
return(retlist)
}

negden <- function(f = NA, delta = 0.01, FLAGBUMP = 1, cB = 1) {
  # finds nonnegative projection
  # FLAGBUM =1 then the program removes bumps whose int f^2 dx
  # less than cof*\int (f - f.neg)^2 dx
  flag <- 0
  f1 <- f
  k <- length(f)
  AREA <- (k / (k - 1)) * mean(f) - (f[1] + f[k]) / (2 * (k - 1))
  if (all(f >= 0)) {
    flag <- 1
  }
}

```

```

if (all(f <= 2 * delta) | (AREA <= 2 * delta)) {
  flag <- 2
}
while (flag == 0) {
  f <- f - delta
  f[f < 0] <- 0
  int <- (k / (k - 1)) * mean(f) - (f[1] + f[k]) / (2 * (k - 1))
  if (int <= AREA) {
    if (int > (10 * delta)) {
      f <- f * (AREA / int)
    }
    flag <- 1
  }
}
if (FLAGBUMP == 1) {
  AREASQ <- mean((f - f1) ^ 2) # area removed so far
  f <- rem.bump1(f = f, AREASQ = AREASQ, coef = cB)
}
if (flag == 1) {
  if (mean(f) > (10 * delta)) {
    f <- f * (AREA / mean(f))
  }
}
f[f < 0] <- 0
f
}

rem.bump1 <- function(f = NA, AREASQ = NA, coef = 1) {
  ## removes any bump with L_2 area larger than coef*AREASQ
  n <- length(f)
  vec <- abvec(f)
  if (length(vec) > 2) {
    vec <- vec[-c(1, 2)]
    k <- length(vec) / 2
    for (s in 1:k) {
      if (sum((f[vec[2 * s - 1]:vec[2 * s]]) ^ 2) / n <= coef * AREASQ) {
        f[vec[2 * s - 1]:vec[2 * s]] <- 0
      }
    }
  }
}

```

```

    }
    f
  }

abvec <- function(f = NA) {
  ## returns intervals where the function f is positive
  ### The first two elements of calculated vec are 00.
  #### This indicates that all entries of f are nonpositive.
  n <- length(f) + 1
  f <- c(f, 0)
  vec <- c(0, 0)
  if (all(f > 0)) {
    vec <- c(vec, 1, n)
  }
  else {
    seq.pos <- (1:n)[f > 0]
    seq.neg <- (1:n)[f <= 0]
    seq.neg <- seq.neg[seq.neg > 1]
    a <- 1
    while (length(seq.pos) * length(seq.neg) > 0) {
      if (f[a + 1] > 0) {
        b <- min(seq.neg) - 1
      }
      else {
        a <- min(seq.pos)
        seq.neg <- seq.neg[seq.neg > a]
        b <- min(seq.neg) - 1
      }
      vec <- c(vec, a, b)
      seq.pos <- seq.pos[seq.pos > b]
      seq.neg <- seq.neg[seq.neg > b]
      a <- b
    }
  }
  if (vec[length(vec)] == n) {
    vec[length(vec)] <- n - 1
  }
  vec
}

```

References

- Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. New York: Springer.
- Efromovich, S. (2018). *Missing and Modified Data in Nonparametric Estimation: With R Examples*. Chapman and Hall/CRC.
- Harrell, F.E., Jr., Margolis, P.A., Gove, S., Mason, K.E., Mulholland, E.K., Lehmann, D., Muhe, L., Gatchalian, S. and Eichenwald, H.F. (1998). Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants. *Statist. Med.* **17**, 909–944.
- Mangasarian, O.L., Street, W.N. and Wolberg, W.H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research* **43**, 570–577.
- The WHO Young Infants Study Group. (1999). Methodology for a multicenter study of serious infections in young infants in developing countries. (1999). *The Pediatric Infectious Disease Journal* **18**, S8–S16.
- The WHO Young Infants Study Group. (1999). Clinical prediction of serious bacterial infections in young infants in developing countries. *The Pediatric Infectious Disease Journal* **18**, S23–S31.
- van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.
- Vanderbilt Department of Biostatistics. 2023. Vanderbilt Biostatistics Datasets. <https://hbiostat.org/data/>.
- Wolberg, W. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. <https://doi.org/10.24432/C5HP4Z>.