Numerical Study of Adaptive Spectral Density Estimator for Censored Data

Supplementary Material for the paper

"*Efficient Nonparametric Spectral Density Estimation with Randomly Censored Observations*"

by Sam Efromovich and Jiaju Wu

The Supplementary Materials contain results of numerical analysis of the proposed data-driven spectral density estimator $\tilde{g}(\lambda, \hat{S}^C)$. Several experiments are considered in turn. In what follows we continue to use terminology and notations of the paper.

In *Experiment 1* the underlying time series $\{X_t\}$ is ARMA(1,1) model $X_t - 0.5X_{t-1} = 0.5(Z_t + 0.4Z_{t-1})$ with standard Gaussian $Z_t \sim N(0, 1)$. Note that the corresponding spectral density belongs to class (3) as it is explained in the paragraph below line (3). Further, the distribution of X_t is normal with variance 0.52 (Efromovich 1999, Sect. 5.2). Censoring variable C has a normal distribution which is either $N(2.1, 2)$, or $N(0.8, 2)$, or $N(0, 2)$. Accordingly, Assumption 2 holds because X_t has lighter exponential tails than $[S^C]^3$. The used censoring distributions create average censoring rates of 9%, 30%, and 50%, respectively, and we will refer to these rates as light, medium and heavy. Considered sample sizes are $n = 100$, 300, and 500. For each censoring distribution and sample size, we simulate a realization $\{(X_t, C_t), t = 1, 2, \ldots, n\}$ and repeat the simulation 1000 times. Then for each realization we calculate a corresponding censored realization $\{(V_t, \Delta_t), t = 1, 2, \ldots, n\}$. For each simulation four estimates are calculated. The first one is calculated by the Oracle-estimator based on an underlying realization $\{X_t, t = 1, 2, \ldots, n\}$, here the software of Efromovich (1999) is used. The second one is the proposed E-estimator which uses censored data $\{V_t, \Delta_t, t = 1, 2, \ldots, n\}$. The third estimator is a Naive one that simply treats V_t as X_t and ignores censoring. According to the Introduction's literature review, the Naive estimator is one of the most widely used. The fourth estimator is again a special oracle. It is the E-estimator that uses an underlying censoring survival function in place of its estimate, and we refer to it as EO-estimator. This oracle will allow us to evaluate performance of the estimator (17) of the survival function S^C. After a particular simulation is done, we calculate empirical ISEs (integrated squared errors) of the four estimates using a greed of 500 points on the interval $[0, \pi]$. The ISEs are denoted as ISEO, ISEE, ISEN and ISEEO for the Oracle-estimate, E-estimate, Naive estimate and EO-estimate, respectively. Then for each simulation ratios ISEE/ISEN, ISEE/ISEO and ISEE/ISEEO are calculated. The first ratio compares E-estimator with Naive-estimator, the second ratio compares E-estimator with Oracle estimator, and the third ratio tells us how well the proposed estimator (17) of the survival function replaces an underlying survival function of a censoring variable in the E-estimator. Table 1 shows us sample medians of 1000 ratios and corresponding mean absolute deviations (mad).

Table 1: Sample medians (mad) of ISE's ratios for Experiment 1

n	Censoring	ISEE/ISEN	ISEO/ISEE	ISEEO/ISEE
100	Light	0.98 (0.06)	0.73 (0.11)	0.77 (0.08)
	Medium	0.78 (0.07)	0.66 (0.12)	0.79 (0.09)
	Heavy	0.74 (0.07)	0.51 (0.12)	0.82 (0.10)
300	Light	0.97 (0.05)	0.84 (0.10)	0.86 (0.05)
	Medium	0.49 (0.07)	0.55 (0.13)	0.89 (0.05)
	Heavy	0.41 (0.08)	0.30 (0.13)	0.93 (0.04)
500	Light	0.89 (0.08)	0.93(0.06)	0.96 (0.02)
	Medium	0.38 (0.07)	0.54 (0.09)	0.98 (0.01)
	Heavy	0.35 (0.06)	0.26(0.09)	0.98(0.01)

Let us look at the results. We begin with the Naive estimator which is currently a staple in statistical analysis of censored time series. Corresponding results are shown in the third column. The Naive estimator performs very well for the light censoring and smaller sample sizes, and due to its simplicity it may be recommended for these cases. In other words, a light censoring can be ignored for small samples, and this recommendation coincides with the literature cited in the Introduction. At the same time, note that with increased sample size the relative efficiency of E-estimator with respect to Naive estimator increases. The outcome changes dramatically for higher censoring rates. Here E-estimator clearly dominates Naive estimator, and there is no doubt that ignoring censoring is not a feasible statistical methodology for high rate censoring.

Now let us look at the fourth column that allows us to understand how the E-estimator performs with respect to the Oracle-estimator that knows underlying time series $\{X_t\}$. In other words, here we may observe the effect of censoring on spectral density estimation. There are two conclusions to make. The former is that for the light censoring relative performance of the E-estimator improves as the sample size increases. The latter is that for higher rates of censoring the E-estimator simply cannot compete with the Oracle who knows the underlying uncensored time series.

Interesting outcomes are shown in the fifth column where the effect of estimate $\hat{S}^C$ on the spectral estimation is analyzed. Two outcomes to point upon are as follows. The former is well expected and is supported by the theory - the estimator improves as the sample size increases. The latter is more intriguing - the estimator improves for higher rates of censoring. The explanation is that $\hat{S}^C$ is based on censored observations, see (17), and hence it improves as the number of censored observations increases.

In Experiment 1 we fixed an underlying spectral density and then repeated the same simulation many times. A natural question is how the estimator will perform for different underlying spectral densities. There are two ways to answer this question. The former is to repeat several times Experiment 1 for different spectral densities. The later is each time to use a new underlying spectral density from a family of spectral densities and then look at outcomes for that family of spectral densities. Here we are using the second approach that is utilized as follows.

In Experiment 2 we are simulating an underlying time series $\{X_t\}$ from a Gaussian ARMA(1, 1) model with standard deviation 0.5, only now for each simulation an underlying AR and MA coefficients (a, b) are chosen uniformly from interval $[0.2, 0.8]$. Three different censoring distributions are $N(2, 2)$, $N(0.8, 2)$, $N(0, 2)$, and they impose light, medium, and heavy censoring. Results are exhibited in Table 2. They are similar to Experiment 1, and this indicates robustness of the E-estimator toward an underlying spectral density. Below the interested reader can find figures for particular simulations that are explained in the captions. Each figure also exhibits the number m of uncensored observations.

Table 2: Sample medians (mad) of ISE's ratios for Experiment 2

n	Censoring	ISEE/ISEN	ISEO/ISEE	ISEEO/ISEE
100	Light	0.95 (0.09)	0.71 (0.14)	0.78 (0.09)
	Medium	0.77 (0.09)	0.67 (0.15)	0.77 (0.12)
	Heavy	0.70 (0.10)	0.55 (0.14)	0.79 (0.12)
300	Light	0.92 (0.08)	0.83 (0.14)	0.83 (0.08)
	Medium	0.48 (0.09)	0.54 (0.15)	0.85 (0.09)
	Heavy	0.40 (0.09)	0.32 (0.14)	0.91 (0.07)
500	Light	0.85 (0.10)	0.94(0.10)	0.94 (0.08)
	Medium	0.37 (0.09)	0.57 (0.11)	0.97 (0.05)
	Heavy	0.36 (0.09)	0.28(0.09)	0.97(0.05)

Now let is check robustness of the E-estimator toward censoring distributions. We continue to use the approach of Experiment 2, only now in the new Experiment 3 censoring variable C has either Laplace(4.3, 2), or Laplace(1.5, 2), or Laplace(0, 2) distribution. The three censoring distributions impose light, medium, and heavy censoring on the underlying X_t. A Laplace distribution has a classical light (exponentially decreasing) tail, but still it is dramatically "heavier" than of a Normal distribution. Accordingly the Assumption 2 holds. As we see, for classical Gaussian time series we have a rather wide class of censoring distributions that satisfy Assumption 2.

Table 3: Sample medians (mad) of ISE's ratios for Experiment 3

n	Censoring	ISEE/ISEN	ISEO/ISEE	ISEEO/ISEE
100	Light	0.81 (0.11)	0.79 (0.13)	0.78 (0.10)
	Medium	0.63 (0.12)	0.71 (0.13)	0.79 (0.11)
	Heavy	0.52 (0.13)	0.60 (0.14)	0.79 (0.11)
300	Light	0.74 (0.10)	0.84 (0.12)	0.84 (0.10)
	Medium	0.51 (0.09)	0.58 (0.13)	0.87 (0.10)
	Heavy	0.38 (0.09)	0.35 (0.11)	0.92 (0.08)
500	Light	0.68 (0.12)	0.92(0.09)	0.95 (0.09)
	Medium	0.33 (0.10)	0.62 (0.09)	0.98 (0.08)
	Heavy	0.27(0.09)	0.33(0.07)	0.99(0.08)

Let us look at the results exhibited in Table 3. The E-estimator performs very similar with respect to the oracles. What is of interest here is that the heavier (with respect to Normal) tails of Laplace censoring distribution complicate using the Naive estimator for high rate censoring. Overall, we can conclude that the E-estimator is robust and can be recommended for analysis of censored time series.

R-software with the E-estimator is available on request from the authors.

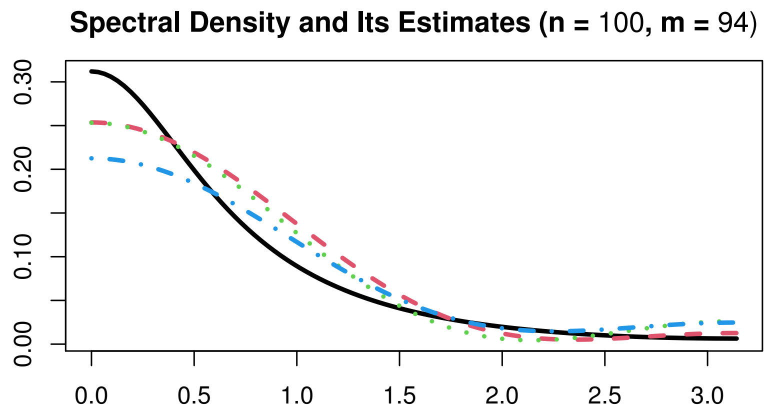

Figure 1: A realization of light censoring case with rate 6%, n = 100, AR = 0.5, MA=0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data V_t and Δ_t, and the blue dashed-dotted line is the naive estimate based on V_t.

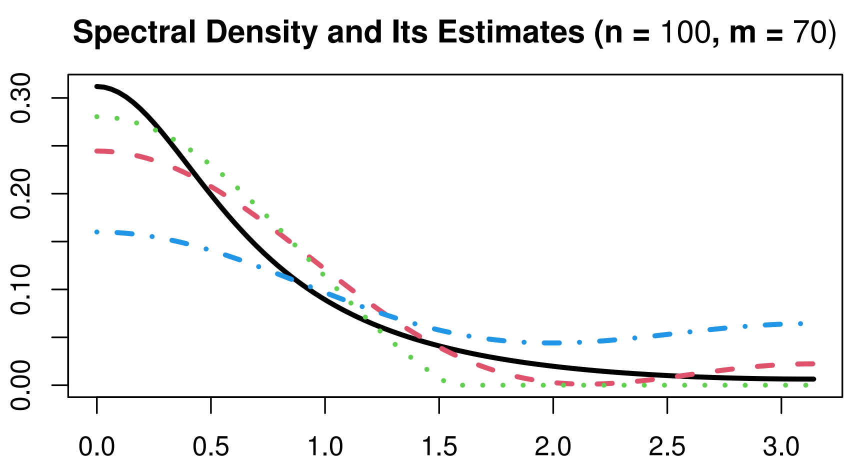

Figure 2: A realization of light censoring case with rate 30%, n = 100, AR = 0.5, MA=0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data V_t and Δ_t, and the blue dashed-dotted line is the naive estimate based on V_t.

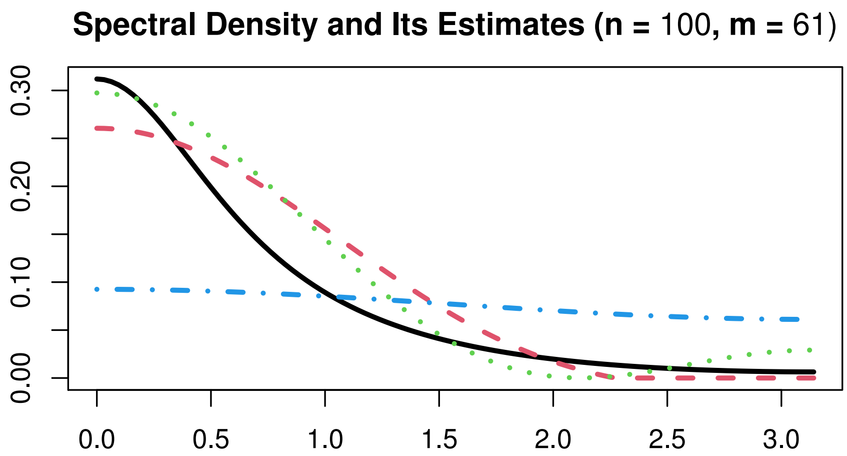

Figure 3: A realization of heavy censoring case with rate 39%, n = 100, AR = 0.5, MA=0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data V_t and Δ_t, and the blue dashed-dotted line is the naive estimate based on V_t.

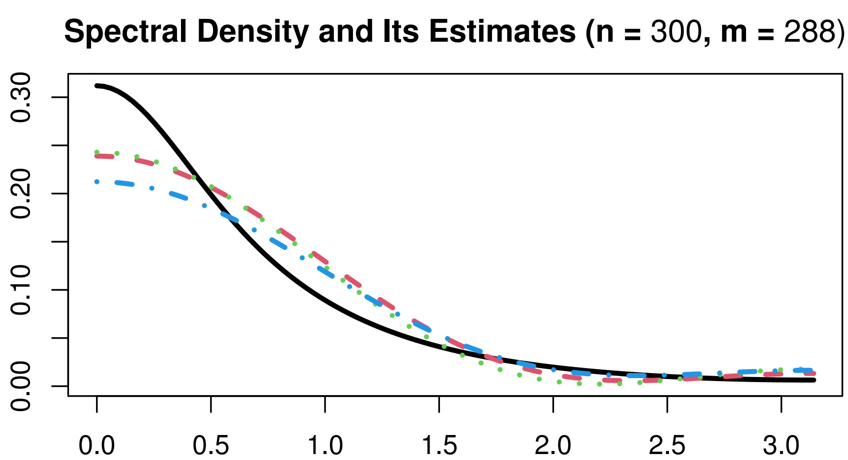

Figure 4: A realization of light censoring case with rate 4%, n = 300, AR = 0.5, MA = 0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data V_t and Δ_t, and the blue dashed-dotted line is the naive estimate based on V_t.

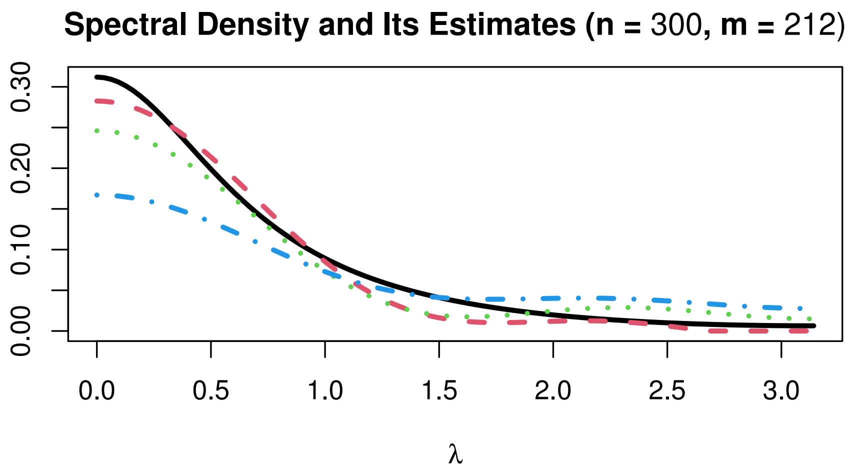

Figure 5: A realization of medium censoring case with rate 29%, n = 300, AR = 0.5, MA=0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data V_t and Δ_t, and the blue dashed-dotted line is the naive estimate based on V_t.

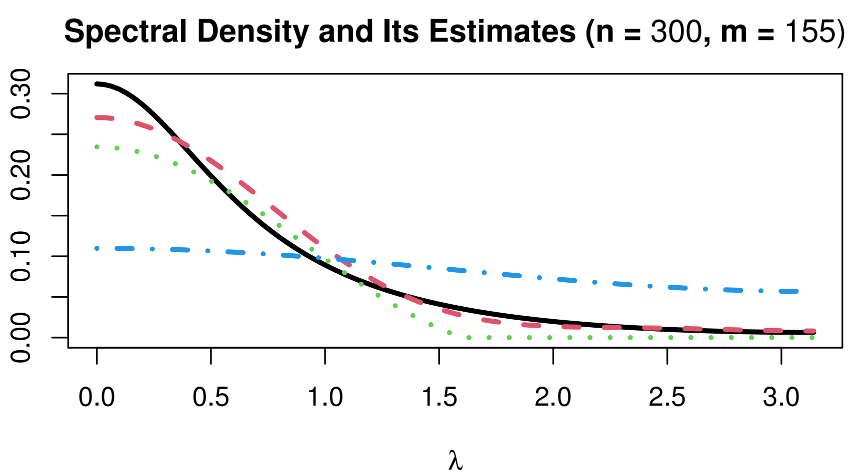

Figure 6: A realization of heavy censoring case with rate 49%, n = 300, AR = 0.5, MA=0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data V_t and Δ_t, and the blue dashed-dotted line is the naive estimate based on V_t.

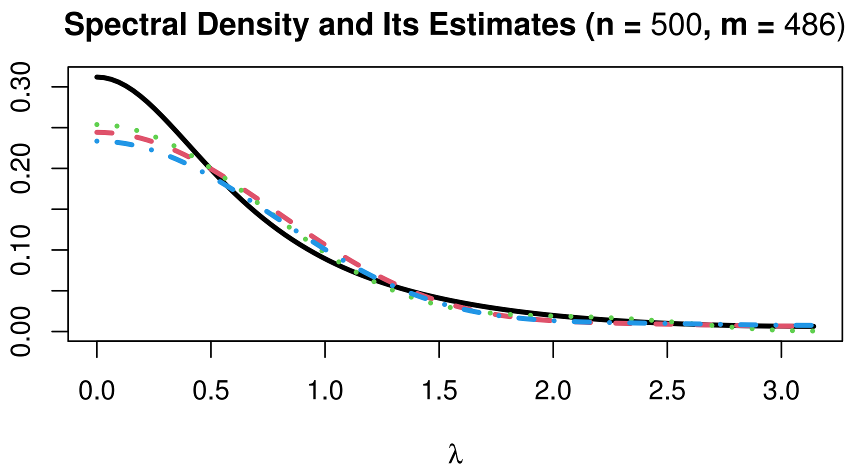

Figure 7: A realization of light censoring case with rate 3%, n = 500, AR = 0.5, MA = 0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data, and the blue dashed-dotted line is the naive estimate based on V_t.

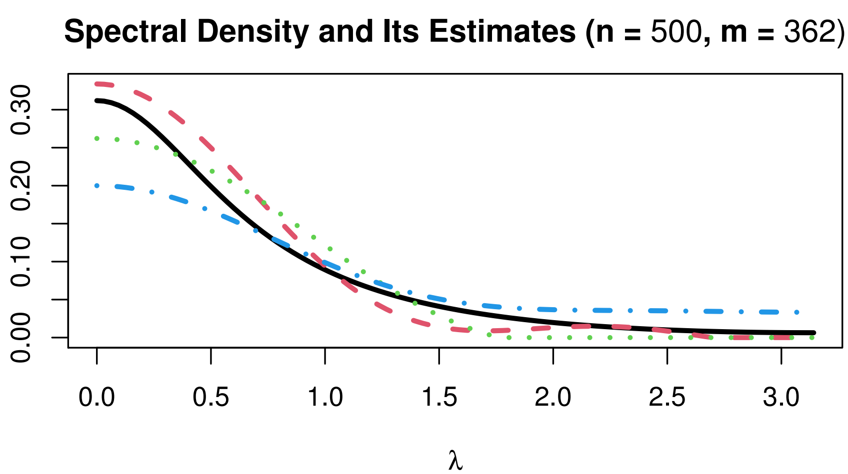

Figure 8: A realization of medium censoring case with rate 27.6%, n = 500, AR = 0.5, MA = 0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data, and the blue dashed-dotted line is the naive estimate based on V_t.

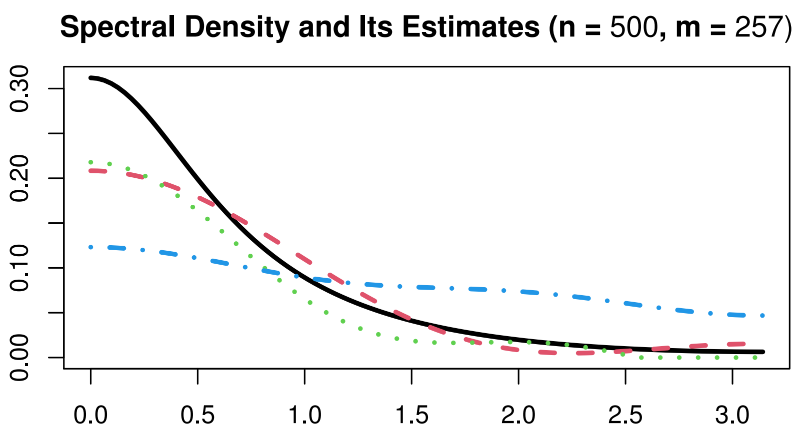

Figure 9: A realization of heavy censoring case with rate 49%, n = 500, AR = 0.5, MA = 0.4. The black solid line is the underlying spectral density; the red dashed line is the Oracle estimate based on X_t, the green dotted line is the E-estimate based on the censored data, and the blue dashed-dotted line is the naive estimate based on V_t.

n	Censoring	MISEO	MISEE	MISEN	Ratio1	Ratio2
100	Light	0.00051	0.00060	0.00063	0.9524	1.1764
	Medium	0.00052	0.00107	0.00120	0.8916	2.0577
	Heavy	0.00055	0.00262	0.00299	0.8673	4.7636
300	Light	0.00021	0.00023	0.00026	0.8846	1.0952
	Medium	0.00022	0.00041	0.00077	0.5324	1.8626
	Heavy	0.00020	0.00077	0.00206	0.3737	3.8500
500	Light	0.00014	0.00016	0.00019	0.8421	1.1428
	Medium	0.00014	0.00028	0.00070	0.4000	2.0000
	Heavy	0.00014	0.00057	0.00196	0.2908	4.0700

Table 4: Empirical MISEs and ratios for ARMA(1,1) with Normal censoring, AR = 0.3 and MA = 0.4.

n	Censoring	MISEO	MISEE	MISEN	Ratio1	Ratio2
100	Light	0.0134	0.0135	0.0149	0.9020	1.0066
	Medium	0.0140	0.0211	0.0261	0.8067	1.5111
	heavy	0.0132	0.0253	0.0388	0.6527	1.9137
300	Light	0.0056	0.0056	0.0065	0.8648	1.0072
	Medium	0.0053	0.0082	0.0162	0.5024	1.5348
	heavy	0.0056	0.0159	0.0334	0.4762	2.8530
500	Light	0.0038	0.0038	0.0047	0.8085	1.0159
	Medium	0.0038	0.0057	0.0143	0.3994	1.5040
	Heavy	0.0039	0.0129	0.0316	0.4062	3.3290

Table 5: Empirical MISEs and ratios for ARMA(1,1) with Normal censoring, AR = 0.7 and MA = 0.4.

n	Censoring	MISEO	MISEE	MISEN	Ratio1	Ratio2
100	Light	0.0235	0.0256	0.0365	0.7027	1.0909
	Medium	0.0237	0.0318	0.1128	0.2815	1.3426
	Heavy	0.0234	0.0440	0.2369	0.1861	1.8858
300	Light	0.0111	0.0118	0.0192	0.6156	1.0639
	Medium	0.0109	0.0138	0.0931	0.1487	1.2712
	Heavy	0.0126	0.0242	0.2281	0.1061	1.9099
500	Light	0.0071	0.0077	0.0147	0.5238	1.0830
	Medium	0.0068	0.0089	0.0806	0.1109	1.3094
	Heavy	0.0070	0.0131	0.2140	0.0611	1.8780

Table 6: Empirical MISEs and ratios for ARMA(1,1) with Laplace censoring, random AR and MA.