Efficient Algorithms for Bayesian Network Parameter Learning from Incomplete Data

SUPPLEMENTARY MATERIAL

A FACTORED DELETION FOR MAR

We now give a more detailed derivation of the factored deletion algorithm for MAR data. Let the query of interest be $\Pr(\mathbf{Y})$, and let $\mathbf{X}'_o = \mathbf{X}_m \setminus \mathbf{Y}_m$ and $\mathbf{Z}^i_m = \{Y^j_m | i \leq j \leq n\}$. We can then factorize the estimation of $\Pr(\mathbf{Y})$ as follows.

$$\begin{aligned}
\Pr(\mathbf{Y}) &= \sum_{\mathbf{X}'_o} \Pr(\mathbf{Y}_m, \mathbf{Y}_o, \mathbf{X}'_o) \\
&= \sum_{\mathbf{X}'_o} \Pr(\mathbf{Y}_o, \mathbf{X}'_o) \Pr(\mathbf{Y}_m | \mathbf{Y}_o, \mathbf{X}'_o) \\
&= \sum_{\mathbf{X}'_o} \Pr(\mathbf{X}_o) \Pr(\mathbf{Y}_m | \mathbf{X}_o) \\
&= \sum_{\mathbf{X}'_o} \Pr(\mathbf{X}_o) \prod_{i=1}^{n} \Pr\left(Y^i_m | \mathbf{Z}^{i+1}_m, \mathbf{X}_o\right) \\
&= \sum_{\mathbf{X}'_o} \Pr(\mathbf{X}_o) \prod_{i=1}^{n} \Pr\left(Y^i_m | \mathbf{Z}^{i+1}_m, \mathbf{X}_o, \mathbf{R}_{\mathbf{Z}^i_m} = \mathsf{ob}\right)
\end{aligned}$$

The last step makes use of the MAR assumption. This leads us to the following algorithm, based on the data distribution $\Pr_{\mathcal{D}}$, and the fully-observed proxy variables $Y^{i,\star}_m$ and $\mathbf{Z}^{i+1,\star}_m$.

$$\begin{aligned}
&\Pr(\mathbf{Y}) \\
&\approx \sum_{\mathbf{X}'_o} \Pr_{\mathcal{D}}(\mathbf{X}_o) \prod_{i=1}^{n} \Pr_{\mathcal{D}}\left(Y^{i,\star}_m | \mathbf{Z}^{i+1,\star}_m, \mathbf{X}_o, \mathbf{R}_{\mathbf{Z}^i_m} = \mathsf{ob}\right)
\end{aligned}$$

B EXTENDED EMPIRICAL EVALUATION: MCAR

Table 5 shows additional results for the classical `alarm` Bayesian network, from Section 4.1.

Table 5: `alarm` network with MCAR data

Size	EM-1-JT	EM-10-JT	D-MCAR	F-MCAR	D-MAR	F-MAR
Runtime [s]						
10^2	2	6	0	0	0	0
10^3	6	50	0	0	0	0
10^4	69	-	0	1	0	1
10^5	-	-	1	9	4	13
10^6	-	-	11	92	29	124
Test Set Log-Likelihood						
10^2	-12.18	-12.18	-12.85	-12.33	-12.82	-12.32
10^3	-10.41	-10.41	-10.73	-10.55	-10.69	-10.55
10^4	-10.00	-	-10.07	-10.04	-10.07	-10.05
10^5	-	-	-9.98	-9.98	-9.99	-9.98
10^6	-	-	-9.96	-9.96	-9.97	-9.97
Kullback-Leibler Divergence						
10^2	2.381	2.381	3.037	2.525	3.010	2.515
10^3	0.365	0.365	0.688	0.502	0.659	0.502
10^4	0.046	-	0.113	0.084	0.121	0.093
10^5	-	-	0.016	0.013	0.024	0.021
10^6	-	-	0.002	0.002	0.006	0.008

C EXTENDED EMPIRICAL EVALUATION: MAR

In this Appendix, we expand on the empirical results of Section 4 w.r.t. learning from MAR data. Here, we provide additional empirical results on standard real-world networks where inference is challenging, as originally highlighted in Table 3.

We consider two settings of generating MAR data, as in Section 4. In the *first setting*, the missing data mechanisms were generated with $m = 0.3$, $p = 2$, and a Beta distribution with shape parameters 1.0 and 0.5. In the second setting, we have $m = 0.9$, $p = 2$, and a Beta distribution with shape parameters 0.5 (as in Section 4.3). We consider three time limits, of 1 minute, 5 minutes, and 25 minutes. For all combinations of these setting, test set log-likelihoods are shown in Table 3, and in Tables 6 to 9.

We repeat the observations from the main paper (cf. Section 4). The EM-JT learner, which performs exact inference, does not scale well to these networks. This problem is mitigated by EM-BP, which performs *approximate* inference, yet we find that it also has difficulties scaling (dashed entries indicate that EM-JT and EM-BP did not finish 1 iteration of EM). In contrast, F-MAR, and particularly D-MAR, can scale to much larger datasets. As for accuracy, the F-MAR method typically obtains the best likelihoods (in bold) for larger datasets, although EM-BP can perform better on small datasets. We further evaluated D-MCAR and F-MCAR, although they are not in general consistent

Table 6: Log-likelihoods of large networks learned from MAR data (1 min. time limit, 1st setting).

Size		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR
10^2	Grid 90-20-1	-	-62.38	-64.15	-50.78	-63.51	**-50.24**	Water	-	-19.50	-20.51	-19.37	-20.41	**-19.35**
10^3		-	-79.75	-38.96	-32.77	-38.26	**-32.44**		-	-16.11	-16.26	-15.27	-16.09	**-15.23**
10^4		-	-	-30.65	-28.61	-30.05	**-28.34**		-	-	-15.03	-14.22	-14.86	**-14.14**
10^5		-	-	-	-	-	-		-	-	**-14.30**	-	-	-
10^2	Munin 1	-	-98.95	-103.59	-98.68	-103.54	**-98.49**	Barley	-	**-85.33**	-85.84	-85.68	-86.13	-85.75
10^3		-	-79.83	-70.49	-67.27	-69.78	**-66.97**		-	-	-67.70	-67.18	-67.67	**-67.13**
10^4		-	-	-59.25	**-57.11**	-	-		-	-	**-54.93**	-	-	-

Table 7: Log-likelihoods of large networks learned from MAR data (5 min. time limit, 1st setting).

Size		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR
10^2	Grid 90-20-1	-	-56.23	-63.34	-50.55	-62.38	**-50.06**	Water	-18.84	**-18.06**	-21.23	-19.61	-21.07	-19.57
10^3		-	-55.04	-39.89	-33.34	-39.09	**-33.01**		-	**-14.99**	-16.47	-15.33	-16.24	-15.26
10^4		-	-98.20	-30.46	-27.26	-29.73	**-26.98**		-	-17.39	-15.59	-14.52	-15.26	**-14.43**
10^5		-	-	-28.63	**-26.06**	-27.89	-		-	-	**-15.22**	-	-	-
10^6		-	-	-	-	-	-		-	-	**-15.09**	-	-	-
10^2	Munin 1	-	**-96.51**	-102.51	-98.21	-102.40	-97.95	Barley	-	**-85.59**	-85.70	-85.60	-85.99	-85.66
10^3		-	-68.04	-67.82	-65.49	-67.21	**-65.22**		-	-67.07	-67.58	-66.97	-67.53	**-66.91**
10^4		-	-95.01	-57.68	-56.00	-57.05	**-55.79**		-	-	-55.04	**-54.33**	-54.78	-
10^5		-	-	**-54.30**	-	-	-		-	-	-	-	-	-

Table 8: Log-likelihoods of large networks learned from MAR data (25 min. time limit, 1st setting).

Size		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR
10^2	Grid 90-20-1	-	**-47.66**	-59.84	-48.34	-59.39	-47.88	Water	-21.30	**-18.66**	-21.58	-19.87	-21.36	-19.83
10^3		-	-46.53	-37.29	-31.60	-36.76	**-31.28**		-17.67	-17.10	-18.64	-15.95	-18.27	**-15.86**
10^4		-	-62.98	-28.74	-26.71	-28.26	**-26.45**		-	-14.83	-16.71	-14.58	-16.30	**-14.44**
10^5		-	-	-25.88	-24.97	-25.43	**-24.75**		-	-18.78	-16.31	-14.38	-15.62	**-14.08**
10^6		-	-	-25.27	-	**-24.78**	-		-	-	**-15.25**	-	-	-
10^7		-	-	-	-	-	-		-	-	**-15.13**	-	-	-
10^2	Munin 1	-	**-90.79**	-98.57	-94.50	-98.48	-94.28	Barley	**-85.11**	-85.53	-86.00	-85.74	-86.24	-85.80
10^3		-	**-60.71**	-66.06	-63.95	-65.45	-63.67		-	**-65.96**	-67.88	-67.23	-67.79	-67.15
10^4		-	-60.35	-56.57	-55.38	-55.95	**-55.16**		-	-57.21	-55.34	-54.56	-55.05	**-54.43**
10^5		-	-	-54.29	**-53.38**	-53.67	-		-	-	**-51.09**	-	-	-

Table 9: Log-likelihoods of large networks learned from MAR data (1 min. time limit, 2nd setting).

Size		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR		EM-JT	EM-BP	D-MCAR	F-MCAR	D-MAR	F-MAR
10^2	Grid 90-20-1	-	-62.25	-80.10	-56.59	-79.93	**-56.07**	Water	-	**-20.15**	-26.40	-22.85	-26.24	-22.88
10^3		-	-129.38	-38.74	-29.88	-38.51	**-29.70**		-	-17.76	-20.45	-17.80	-20.32	**-17.64**
10^4		-	-	-27.83	-24.30	-27.25	**-23.97**		-	-	-17.59	-15.40	-17.28	**-15.29**
10^5		-	-	-	-	-	-		-	-	**-15.38**	-	-	-
10^2	Munin 1	-	**-99.49**	-111.95	-104.07	-111.72	-103.10	Barley	-	-89.16	-89.63	-89.13	-89.66	**-88.99**
10^3		-	-99.56	-70.32	-66.08	-69.76	**-65.57**		-	-	-71.76	**-70.50**	-71.74	-
10^4		-	-	-56.25	**-54.36**	-	-		-	-	**-56.59**	-	-	-

for MAR data, and find that they scale even further, and can also produce relatively good estimates (in terms of likelihood).

D EXAMPLE: DATA EXPLOITATION BY CLOSED-FORM ESTIMATORS

This appendix demonstrates with an example how each learning algorithm exploits varied subsets of data to estimate marginal probability distributions, given the manifest (or data) distribution in Table 10 which consists of four variables, $\{X, Y, Z, W\}$ such that $\{X, Y\} \in \mathbf{X}_m$ and $\{Z, W\} \in \mathbf{X}_o$.

We will begin by examining the data usage by deletion algorithms while estimating $\Pr(x, w)$ under the MCAR assumption. All three deletion algorithms, namely listwise deletion, direct deletion and factored deletion guarantee consistent estimates when data are MCAR. Among these algorithms, listwise deletion utilizes the least amount of data (4 distinct tuples out of 36 available tuples, as shown in table 11) to compute $\Pr(xw)$ whereas factored deletion employs two thirds of the tuples (24 distinct tuples out of 36 available tuples as shown in table 11) for estimating $\Pr(xw)$.

Under MAR, no guarantees are available for listwise deletion. However the three algorithms, namely direct deletion, factored deletion and informed deletion, guarantee consistent estimates. While estimating $\Pr(x, y)$, all the three algorithms utilize every tuple in the manifest distribution at least once (see Table 12). Compared to the direct deletion algorithm, the factored deletion algorithm utilizes more data while computing $\Pr(x, y)$ since it has multiple factorizations with more than two factors in each of them; this allows more data to be used while computing each factor (see Table 11). In contrast to both direct and factored deletion, the informed deletion algorithm yields an estimator that involves factors with fewer elements in them ($\Pr(w)$ vs. $\Pr(zw)$) and hence can be computed using more data ($\Pr(w = 0)$ uses 18 tuples compared to $\Pr(z = 0, w = 0)$ that uses 9 tuples).

Precise information regarding the missingness process is required for estimation when dataset falls under the MNAR category. In particular, only algorithms that consult the missingness graph can answer questions about the estimability of queries.

Table 10: Manifest (Data) Distribution with $\{X, Y\} \in \mathbf{X}_m$ and $\{Z, W\} \in \mathbf{X}_o$.

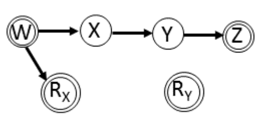
#	X	Y	W	Z	R_X	R_Y
1	0	0	0	0	ob	ob
2	0	0	0	1	ob	ob
3	0	0	1	0	ob	ob
4	0	0	1	1	ob	ob
5	0	1	0	0	ob	ob
6	0	1	0	1	ob	ob
7	0	1	1	0	ob	ob
8	0	1	1	1	ob	ob
9	1	0	0	0	ob	ob
10	1	0	0	1	ob	ob
11	1	0	1	0	ob	ob
12	1	0	1	1	ob	ob
13	1	1	0	0	ob	ob
14	1	1	0	1	ob	ob
15	1	1	1	0	ob	ob
16	1	1	1	1	ob	ob
17	0	?	0	0	ob	unob
18	0	?	0	1	ob	unob

#	X	Y	W	Z	R_X	R_Y
19	0	?	1	0	ob	unob
20	0	?	1	1	ob	unob
21	1	?	0	0	ob	unob
22	1	?	0	1	ob	unob
23	1	?	1	0	ob	unob
24	1	?	1	1	ob	unob
25	?	0	0	0	unob	ob
26	?	0	0	1	unob	ob
27	?	0	1	0	unob	ob
28	?	0	1	1	unob	ob
29	?	1	0	0	unob	ob
30	?	1	0	1	unob	ob
31	?	1	1	0	unob	ob
32	?	1	1	1	unob	ob
33	?	?	0	0	unob	unob
34	?	?	0	1	unob	unob
35	?	?	1	0	unob	unob
36	?	?	1	1	unob	unob

Table 11: Enumeration of sample # used for computing $\Pr(x, w)$ by listwise deletion, direct deletion and factored deletion algorithms under MCAR assumptions.

Algorithm	Estimator and Sample #
Listwise	$\Pr(xw) = \Pr(xw \mid R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ 11,12,15,16
Direct	$\Pr(xw) = \Pr(xw \mid R_X = \mathsf{ob})$ 11,12,15,16,23,24
Factored	$\Pr(xw) = \Pr(x \mid w, R_X = \mathsf{ob}) \Pr(w)$ 3,4,7,8,11,12,15,16,19,20,23,24,27,28,31,32,35,36 $\Pr(xw) = \Pr(w \mid x, R_X = \mathsf{ob}) \Pr(x \mid R_X = \mathsf{ob})$ 9,10,11,12,13,14,15,16,21,22,23,24

Table 12: Enumeration of sample # used for computing $\Pr(x, y)$ by direct deletion, factored deletion and informed deletion algorithms under MAR assumption.

Algorithm	Estimator and Sample #
Direct	$\Pr(xy) = \sum_{z,w} \Pr(xy \mid w, z, R_X = \mathsf{ob}, R_Y = \mathsf{ob}) \Pr(zw)$ 13, 14, 15, 16 for $\Pr(xy \mid w, z, R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ all tuples: [1,36] for $\Pr(z, w)$
Factored	$\Pr(xy) = \sum_{z,w} \Pr(x \mid w, z, y, R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ $\Pr(y \mid z, w, R_Y = \mathsf{ob}) \Pr(zw)$ 13, 14, 15, 16 for $\Pr(x \mid y, w, z, R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ 5, 6, 7, 8, 13, 14, 15, 16, 29, 30, 31, 32 for $\Pr(y \mid w, z, R_Y = \mathsf{ob})$ all tuples: [1,36] for $\Pr(z, w)$ $\Pr(xy) = \sum_{z,w} \Pr(y \mid x, w, z, R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ $\Pr(x \mid z, w, R_X = \mathsf{ob}) \Pr(zw)$ 13, 14, 15, 16 for $\Pr(y \mid x, w, z, R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ 9, 10, 11, 12, 13, 14, 15, 16, 21, 22, 23, 24 for $\Pr(x \mid w, z, R_X = \mathsf{ob})$ all tuples: [1,36] for $\Pr(z, w)$
Informed (direct) 	$\Pr(xy) = \sum_{w} \Pr(xy \mid w, R_X = \mathsf{ob}, R_Y = \mathsf{ob}) \Pr(w)$ 13, 14, 15, 16 for $\Pr(xy \mid w, R_X = \mathsf{ob}, R_Y = \mathsf{ob})$ all tuples: [1,36] for $\Pr(w)$