Journal of Machine Learning Research 17 (2016) Supplementary Material Submitted 12/14; Revised 11/15; Published 6/16

Structure Learning in Bayesian Networks of a Moderate Size by Efficient Sampling (Supplementary Material)

Ru He HRHERU@GMAIL.COM
Department of Computer Science and Department of Statistics
Iowa State University
Ames, IA 50011, USA

Jin Tian JTIAN@IASTATE.EDU
Department of Computer Science
Iowa State University
Ames, IA 50011, USA

Huaiqing Wu ISUHWU@IASTATE.EDU
Department of Statistics
Iowa State University
Ames, IA 50011, USA

Editor: Max Chickering

1. Proof of Theorem 5 (iv)

Proof that if the quantity $\Delta = \sum_{G\in\mathcal{G}} p_{\not\perp}(G|D)$, then $\Delta \cdot \hat{p}_{\not\perp}(f|D) \leq p_{\not\perp}(f|D) \leq \Delta \cdot \hat{p}_{\not\perp}(f|D) + 1 - \Delta$.

Proof

On one hand,

$$\begin{aligned}
&\Delta \cdot \hat{p}_{\not\perp}(f|D) \\
&= \frac{\sum_{G\in\mathcal{G}} p_{\not\perp}(G,D)}{p_{\not\perp}(D)} \cdot \frac{\sum_{G\in\mathcal{G}} f(G)p_{\not\perp}(G,D)}{\sum_{G\in\mathcal{G}} p_{\not\perp}(G,D)} \\
&= \frac{\sum_{G\in\mathcal{G}} f(G)p_{\not\perp}(G,D)}{p_{\not\perp}(D)} \\
&= \frac{\sum_{G} f(G)p_{\not\perp}(G,D)}{p_{\not\perp}(D)} - \frac{\sum_{G\notin\mathcal{G}} f(G)p_{\not\perp}(G,D)}{p_{\not\perp}(D)} \\
&\leq \frac{\sum_{G} f(G)p_{\not\perp}(G,D)}{p_{\not\perp}(D)} \\
&= p_{\not\perp}(f|D).
\end{aligned}$$

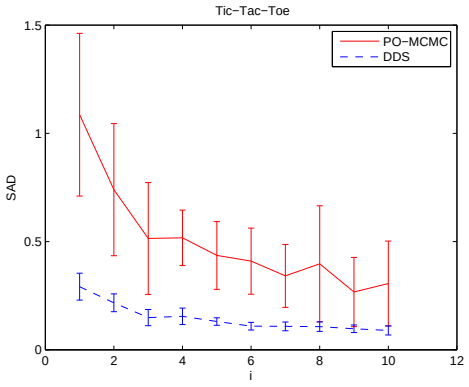

Figure 1: Plot of the SAD Performance of the PO-MCMC and the DDS for Tic-Tac-Toe

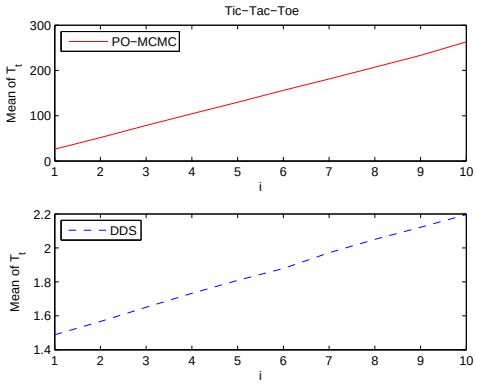

Figure 2: Plot of the Total Running Time of the PO-MCMC and the DDS for Tic-Tac-Toe

On the other hand,

$$\begin{aligned}
&\Delta \cdot \hat{p}_{\not\in}(f|D) + 1 - \Delta \\
&= \frac{\sum_{G\in\mathcal{G}} f(G) p_{\not\in}(G, D)}{p_{\not\in}(D)} + \frac{\sum_{G\notin\mathcal{G}} p_{\not\in}(G, D)}{p_{\not\in}(D)} \\
&\geq \frac{\sum_{G\in\mathcal{G}} f(G) p_{\not\in}(G, D)}{p_{\not\in}(D)} + \frac{\sum_{G\notin\mathcal{G}} f(G) p_{\not\in}(G, D)}{p_{\not\in}(D)} \\
&= \frac{\sum_{G} f(G) p_{\not\in}(G, D)}{p_{\not\in}(D)} \\
&= p_{\not\in}(f|D).
\end{aligned}$$

Combining the above two inequalities, we complete the whole proof. ■

2. Supplementary Experimental Results for the DDS

As a supplement to Section 4.1 in the main paper, this section shows more experimental results for the DDS by varying the sample size. With the same experimental settings as in Section 4.1, we performed the experiment for the data cases Tic-Tac-Toe, Wine, Child with $m = 500$, and German. By examining each figure from Figures 1, 3, 5, and 7, and the corresponding figure from Figures 2, 4, 6, and 8, we can conclude that the learning performance of the DDS with each sample size is significantly better than the performance of the PO-MCMC in each data case.

3. Supplementary Experimental Results for the IW-DDS

As a supplement to Section 4.2 in the main paper, this section shows more experimental results for the IW-DDS by varying the sample size. We performed the experiment for the data cases Wine,

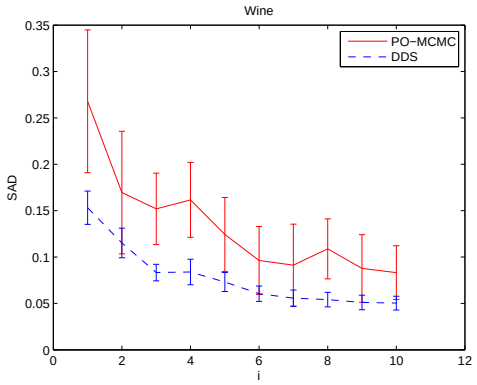

Figure 3: Plot of the SAD Performance of the PO-MCMC and the DDS for Wine

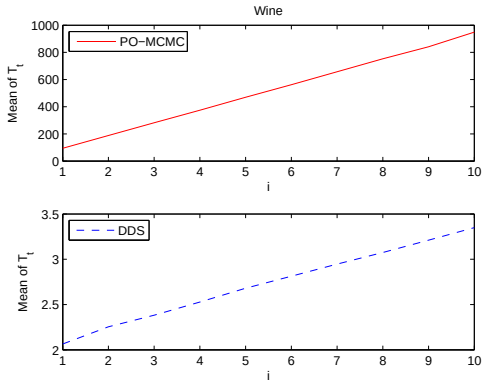

Figure 4: Plot of the Total Running Time of the PO-MCMC and the DDS for Wine

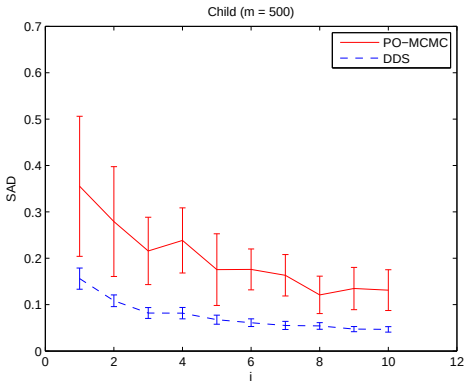

Figure 5: Plot of the SAD Performance of the PO-MCMC and the DDS for Child ($m = 500$)

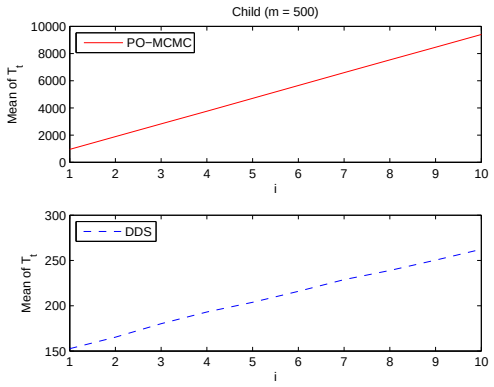

Figure 6: Plot of the Total Running Time of the PO-MCMC and the DDS for Child ($m = 500$)

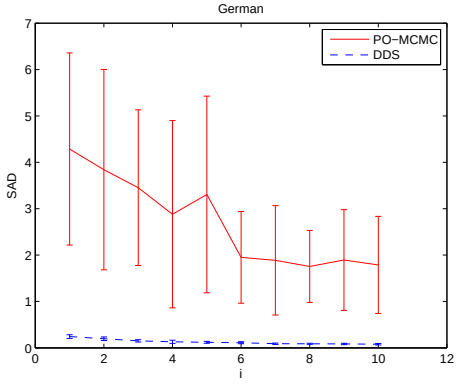

Figure 7: Plot of the SAD Performance of the PO-MCMC and the DDS for German

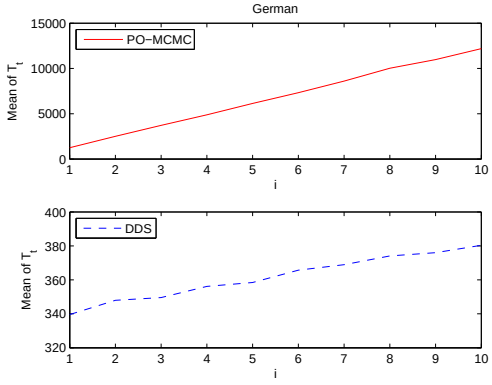

Figure 8: Plot of the Total Running Time of the PO-MCMC and the DDS for German

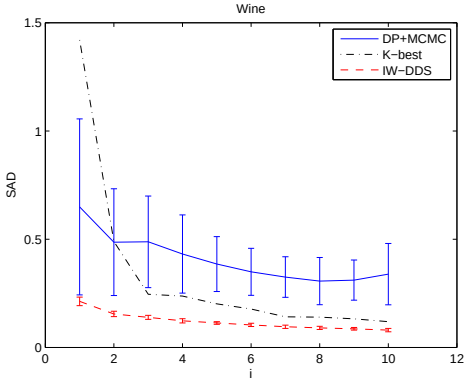

Figure 9: Plot of the SAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Wine

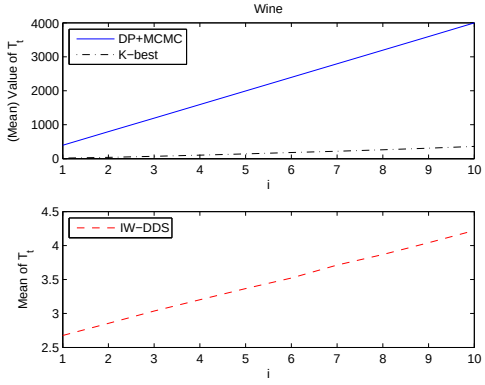

Figure 10: Plot of the Total Running Time of the DP+MCMC, the K-best, and the IW-DDS for Wine

Child with $m = 500$, and German. By examining each figure from Figures 9, 11, and 13, and the corresponding figure from Figures 10, 12, and 14, we can clearly see the advantage of the IW-DDS in the structure learning over the other two methods for each data case.

4. Memory-Saving Strategies for the DDS and the IW-DDS with a Very Large N_o

In this section, we briefly describe our memory-saving strategies for the DDS and the IW-DDS if a very large number of DAG samples are required by a user for his or her specific requirement. As the reader will see, while the memory-saving strategy for the DDS is straightforward, the memory-saving strategy for the IW-DDS is more complicated because it needs to ensure that all the duplicate DAGs among the N_o sampled DAGs are eliminated.

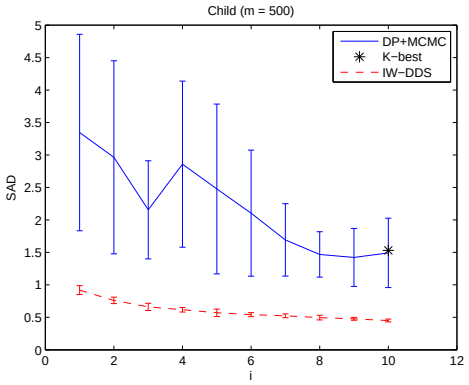

Figure 11: Plot of the SAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 500$)

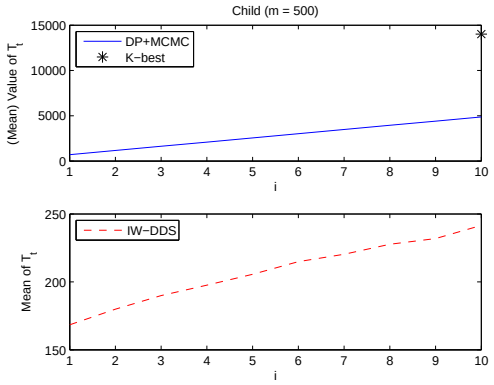

Figure 12: Plot of the Total Running Time of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 500$)

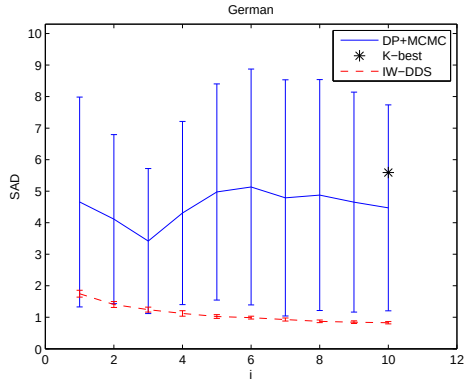

Figure 13: Plot of the SAD Performance of the DP+MCMC, the K-best, and the IW-DDS for German

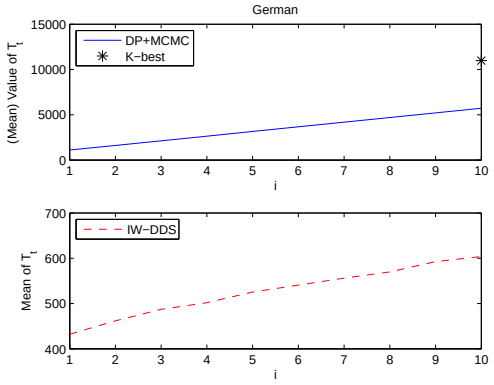

Figure 14: Plot of the Total Running Time of the DP+MCMC, the K-best, and the IW-DDS for German

As discussed in Section 3.2 in the main paper, the overall memory cost of the DDS algorithm is $O(n2^n + n^2N_o)$: Step 1 of the DDS has $O(n2^n)$ memory cost; and Steps 2 and 3 of the DDS have $O(n^2N_o)$ memory cost [1]. Since typically 10^4 to 10^5 DAG samples are sufficient for estimating $p_{\prec}(f|D)$, the $O(n^2N_o)$ space cost coming from Steps 2 and 3 does not become an issue at all. Nevertheless, if a very large N_o (such as more than 1×10^6) is needed for the estimation because of some specific requirement of a user, some memory-saving strategy needs to be used in Steps 2 and 3 to reduce their memory cost. One simple memory-saving strategy is as follows: N_o can be replaced with a smaller value of N_{bl} in the DDS algorithm, where N_{bl} is the size of a sample block, and then Steps 2 and 3 can be repeated $\lceil N_o/N_{bl} \rceil$ times. This will not change the properties of the estimator coming from the DDS algorithm but can reduce the overall memory requirement of the DDS to $O(n2^n + n^2N_{bl})$. Note that for the performance of our time-saving strategy for the DAG sampling step (described in Section 3.2.1 of the main paper), a large N_{bl} is actually preferred. Thus, N_{bl} can take a value that is large but still does not lead to the memory issue for a computer. (For instance, the value of N_{bl} can be set around several millions for a computer with 2.0 to 8.0 GB memory.) In addition, note that the estimator $\hat{p}_{\prec}(f|D)$ can be constructed by Eq. (5) in the main paper on the fly when each DAG gets sampled, so that the memory of storing all the N_o sampled DAGs can be saved.

Similarly, if a very large N_o (such as more than 1×10^6) is needed in the IW-DDS to obtain $\hat{p}_{\not\prec}(f|D) = (\sum_{G\in\mathcal{G}} f(G)p_{\not\prec}(G,D))/(\sum_{G\in\mathcal{G}} p_{\not\prec}(G,D))$ or the corresponding sound interval, the following memory-saving strategy can be used to reduce the memory requirement of the IW-DDS algorithm: N_o can be replaced with a smaller value of N_{bl} in the IW-DDS algorithm, where N_{bl} is the size of a sample block, and then Steps 2 and 3 of the DDS as well as the bias-correction step can be repeated in $\lceil N_o/N_{bl} \rceil$ iterations. (Similar to the DDS algorithm, N_{bl} can take a value that is large but still does not lead to the memory issue for a computer.) In each iteration, after the bias-correction step which eliminates the duplicates among the N_{bl} DAGs is done, up to N_{bl} DAGs get sorted according to $p(D|G)$ and then are stored as a file on the hard disk. Finally, after these $\lceil N_o/N_{bl} \rceil$ iterations are finished, the elimination of the duplicate DAGs across the DAGs stored in $\lceil N_o/N_{bl} \rceil$ files is performed as follows: each time some score threshold p_{thr} can always be found (based on $p(D|G)$ of the corresponding quantile of the sorted DAGs) such that $O(N_{bl})$ DAGs whose $p(D|G)$ is not greater than p_{thr} are newly retrieved from these $\lceil N_o/N_{bl} \rceil$ files and reloaded into the memory each time. Thus, the elimination of the duplicate DAGs only needs to be performed among these DAGs in the memory so that both the denominator $(\sum_{G\in\mathcal{G}} p_{\not\prec}(G,D))$ and the numerator $(\sum_{G\in\mathcal{G}} f(G)p_{\not\prec}(G,D))$ of $\hat{p}_{\not\prec}(f|D)$ can be updated accordingly. When all the DAGs have been retrieved from these $\lceil N_o/N_{bl} \rceil$ files, $\hat{p}_{\not\prec}(f|D)$ is obtained and its denominator $\sum_{G\in\mathcal{G}} p_{\not\prec}(G,D)$ can also be used to obtain Δ. Using this strategy, the expected time cost of the IW-DDS becomes $O(n^{k+1}C(m)+kn2^n+n^2N_o+n^{k+1}N_o+N_{bl}\log(N_{bl})\lceil N_o/N_{bl} \rceil+C_{w,r}(n)N_o+n^2N_o) = O(n^{k+1}C(m) + kn2^n + n^2N_o + n^{k+1}N_o + \log(N_{bl})N_o + C_{w,r}(n)N_o)$, where $C_{w,r}(n)$ is the time cost that a DAG of n nodes is written to the hard disk and then is reloaded from the hard disk to the memory. The corresponding memory requirement is $O(n2^n + n^2N_{bl})$, with the addition of the $O(n^2N_o)$ space in the hard disk required to record $O(N_o)$ DAGs among the $\lceil N_o/N_{bl} \rceil$ files.

We demonstrate the performance of our memory-saving strategy for the IW-DDS based on the data case Insur19 with $m = 200$ in Figures 15 and 16. In our experiment we fixed $N_{bl} = 2 \times 10^6$

1. As described in the main paper, since Step 1 of the DDS limits our application to Bayesian networks with up to around $n = 25$ variables, the actual memory requirement of Steps 2 and 3 of the DDS is to hold $O(N_o) \times n$ (32-bit) integers in the memory when vector representations for orders and DAGs are used.

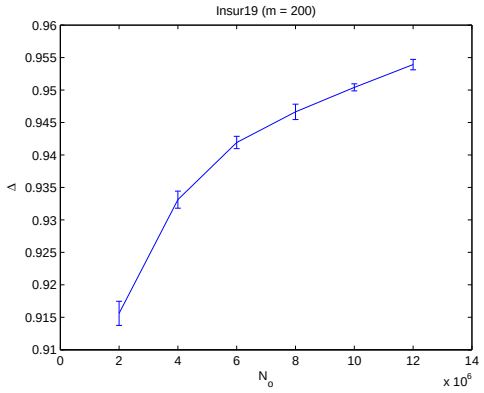

Figure 15: Plot of Δ versus N_o for Insur19 ($m = 200$)

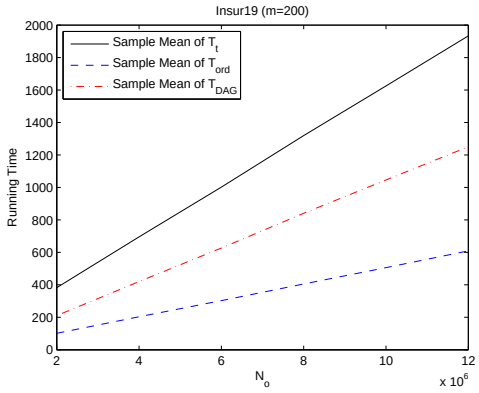

Figure 16: Plot of the Running Time versus N_o for Insur19 ($m = 200$)

as the size of the sample block. We increased N_o from 2×10^6 to 1.2×10^7 with each increment 2×10^6 and showed the corresponding change of Δ and the running time in Figures 15 and 16. (We performed 20 independent runs for the data case to get the results.) By temporarily storing the sampled DAGs in the hard disk, our IW-DDS (equipped with our memory-saving strategy) is shown to be able to efficiently sample $N_o = 1.2 \times 10^7$ DAGs so that the resulting mean of Δ can reach 95.39% with the time cost $\hat{\mu}(T_t) = 1,933.44$ seconds.

5. Boxplots for Comparing the PO-MCMC, the DOS, and the DDS in Terms of MAD

In this section we use boxplots to re-demonstrate the comparison of the PO-MCMC, the DOS, and the DDS for all the 33 data cases in Table 1 in the main paper. Note that in these boxplots the criterion MAD (= SAD$/(n(n-1))$) is used instead of SAD. Though everything except the scale of the Y axis is the same between a boxplot using SAD and a boxplot using MAD, the comparison among data sets with different values of n can be directly made by using MAD. These boxplots (Figures 17 to 49) clearly illustrate the advantage of our DDS method over the PO-MCMC method, as described in Section 4.1 in the main paper.

6. Boxplots for Comparing the DP+MCMC, the K-best, and the IW-DDS in Terms of MAD

In this section we use boxplots to re-demonstrate the comparison of the DP+MCMC, the K-best, and the IW-DDS in terms of MAD for all the 33 data cases in Table 3 in the main paper. These boxplots (Figures 50 to 82) clearly illustrate the advantage of our IW-DDS algorithm, as described in Section 4.2 in the main paper.

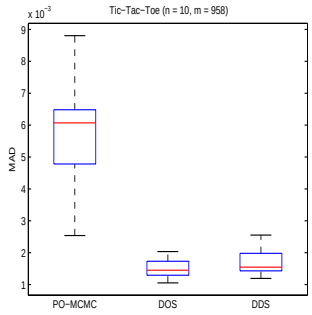

Figure 17: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Tic-Tac-Toe

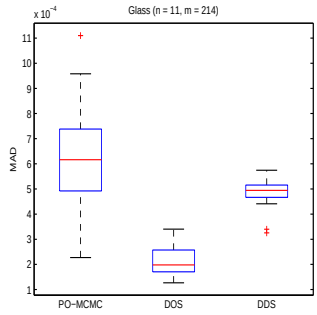

Figure 18: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Glass

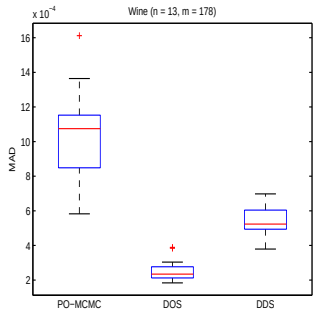

Figure 19: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Wine

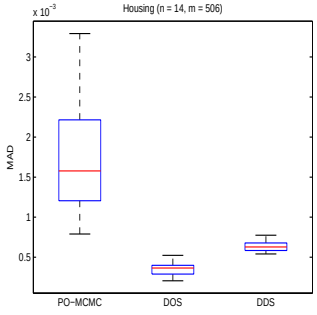

Figure 20: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Housing

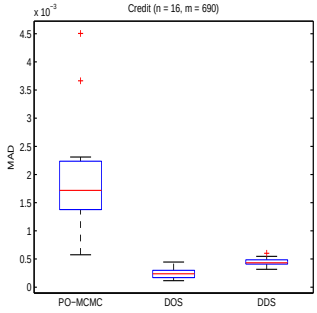

Figure 21: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Credit

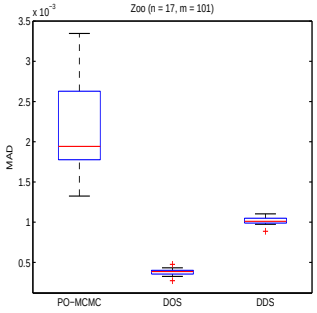

Figure 22: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Zoo

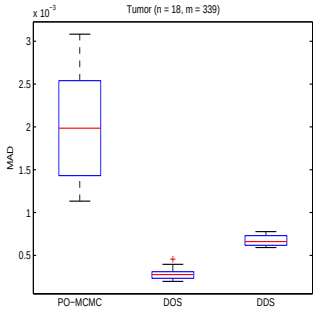

Figure 23: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Tumor

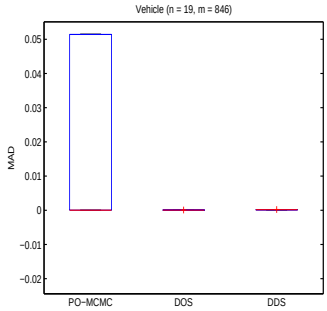

Figure 24: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Vehicle

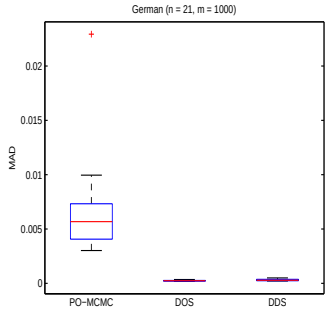

Figure 25: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for German

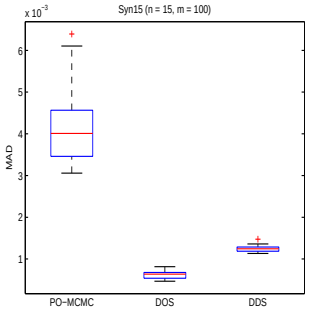

Figure 26: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Syn15 ($m = 100$)

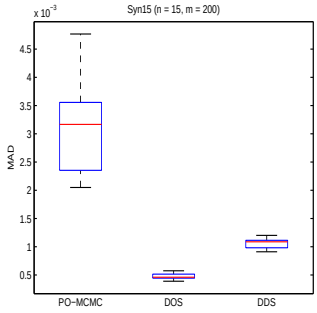

Figure 27: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Syn15 ($m = 200$)

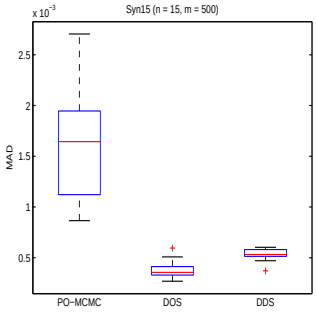

Figure 28: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Syn15 ($m = 500$)

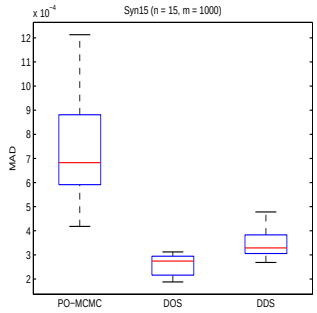

Figure 29: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Syn15 ($m = 1000$)

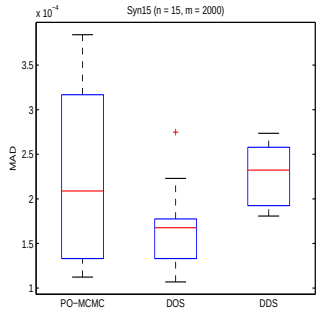

Figure 30: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Syn15 ($m = 2000$)

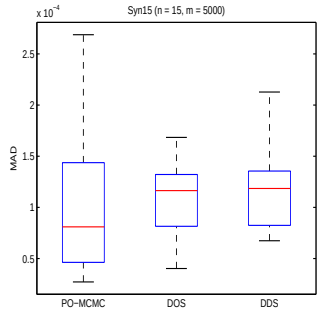

Figure 31: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Syn15 ($m = 5000$)

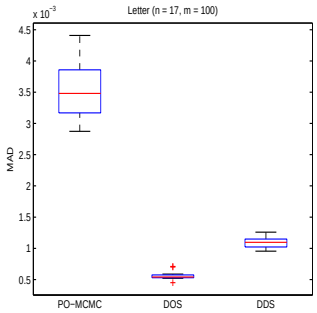

Figure 32: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Letter ($m = 100$)

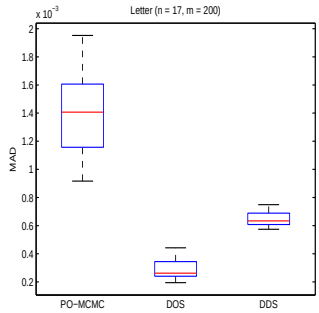

Figure 33: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Letter ($m = 200$)

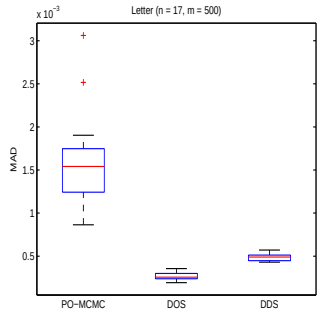

Figure 34: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Letter ($m = 500$)

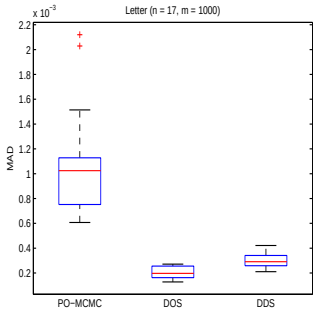

Figure 35: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Letter ($m = 1000$)

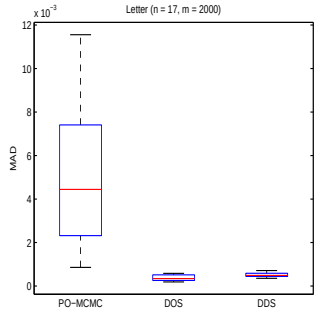

Figure 36: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Letter ($m = 2000$)

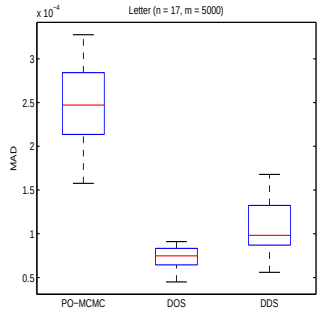

Figure 37: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Letter ($m = 5000$)

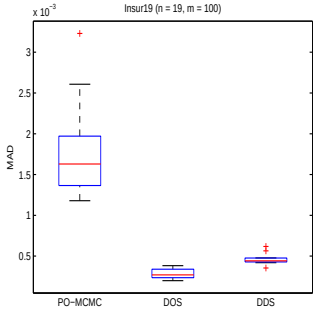

Figure 38: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Insur19 ($m = 100$)

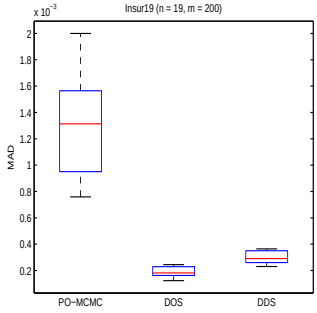

Figure 39: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Insur19 ($m = 200$)

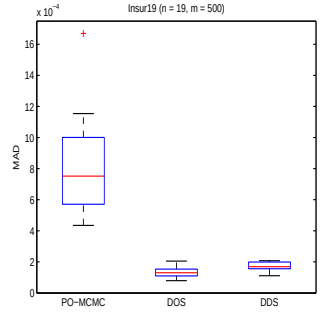

Figure 40: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Insur19 ($m = 500$)

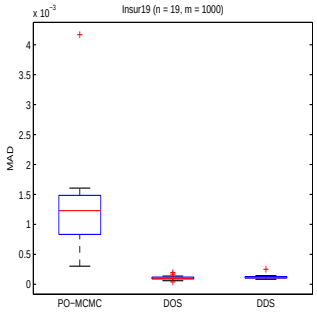

Figure 41: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Insur19 ($m = 1000$)

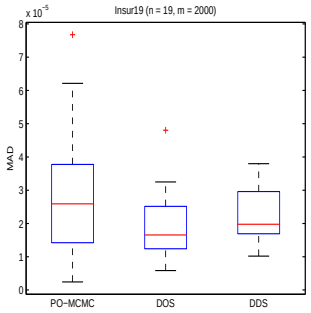

Figure 42: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Insur19 ($m = 2000$)

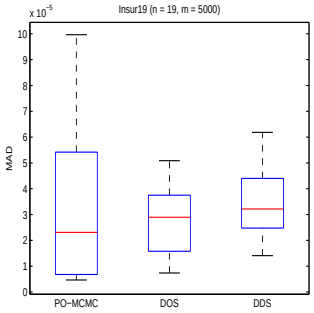

Figure 43: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Insur19 ($m = 5000$)

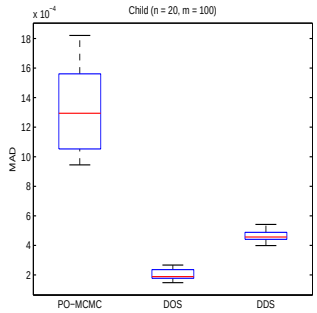

Figure 44: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Child ($m = 100$)

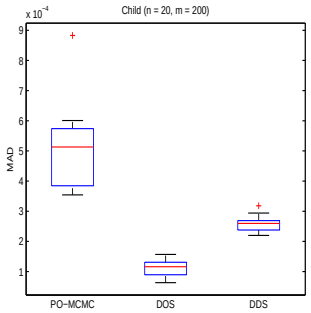

Figure 45: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Child ($m = 200$)

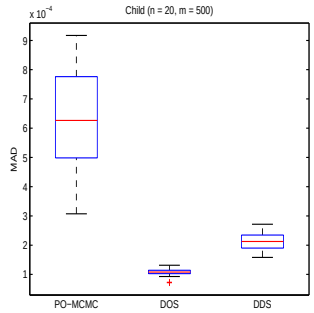

Figure 46: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Child ($m = 500$)

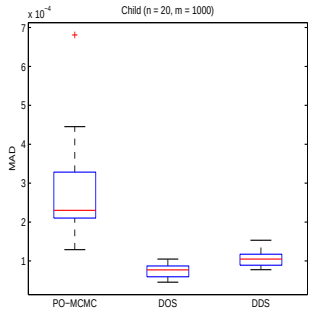

Figure 47: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Child ($m = 1000$)

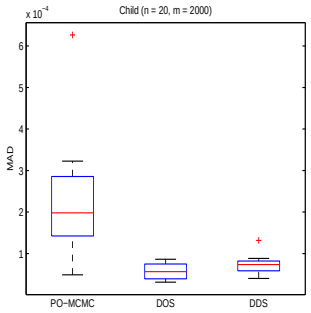

Figure 48: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Child ($m = 2000$)

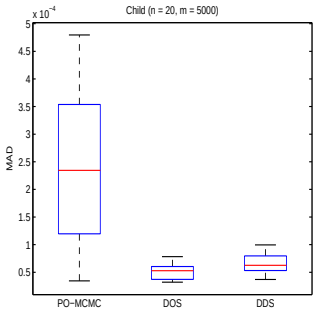

Figure 49: Boxplot of the MAD Performance of the PO-MCMC, the DOS, and the DDS for Child ($m = 5000$)

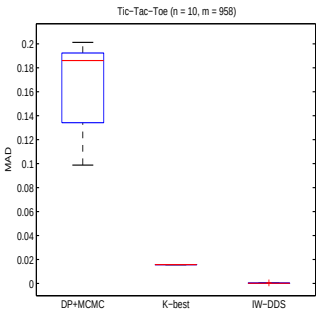

Figure 50: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Tic-Tac-Toe

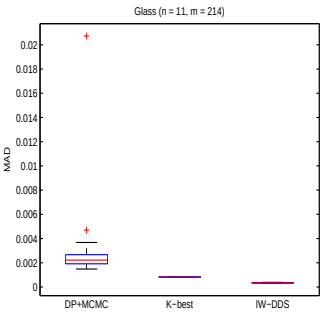

Figure 51: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Glass

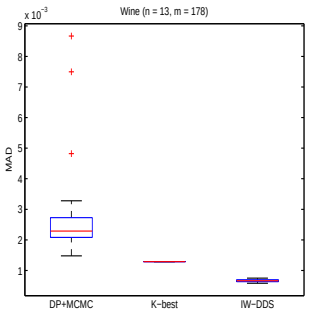

Figure 52: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Wine

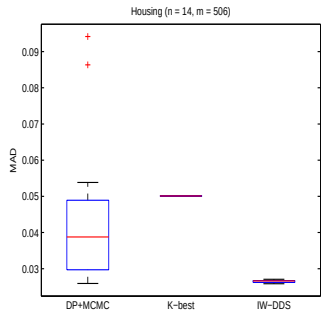

Figure 53: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Housing

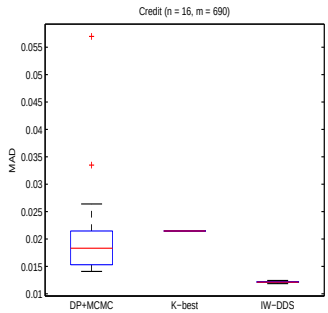

Figure 54: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Credit

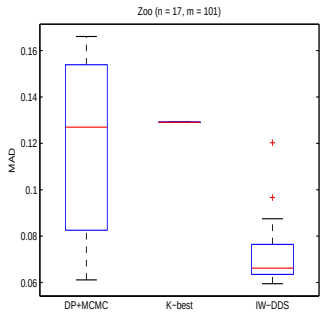

Figure 55: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Zoo

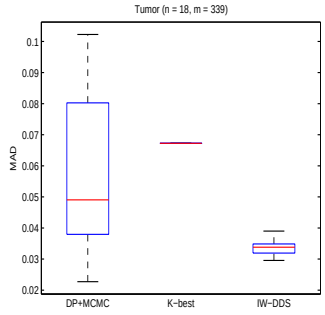

Figure 56: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Tumor

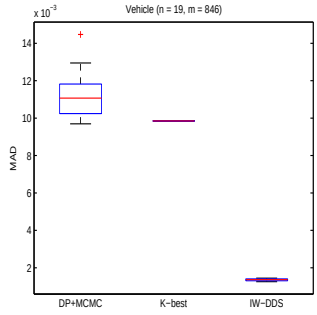

Figure 57: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Vehicle

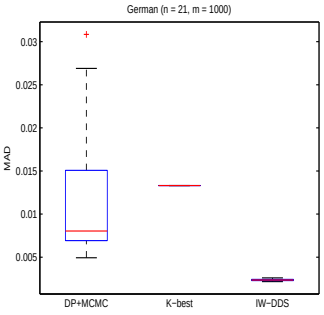

Figure 58: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for German

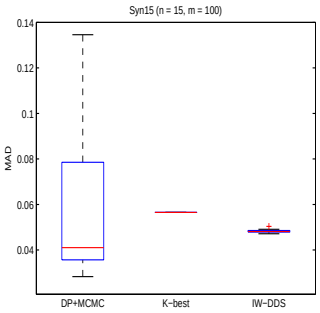

Figure 59: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Syn15 ($m = 100$)

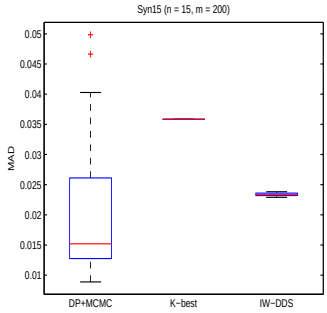

Figure 60: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Syn15 ($m = 200$)

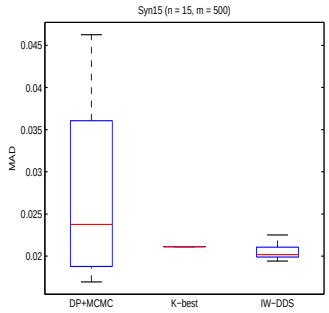

Figure 61: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Syn15 ($m = 500$)

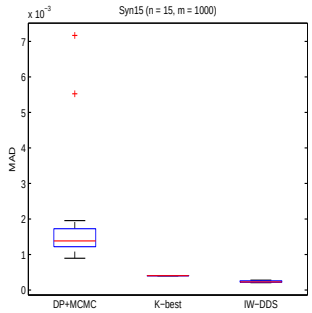

Figure 62: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Syn15 ($m = 1000$)

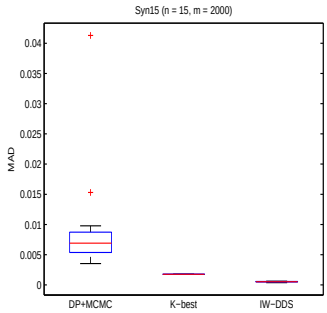

Figure 63: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Syn15 ($m = 2000$)

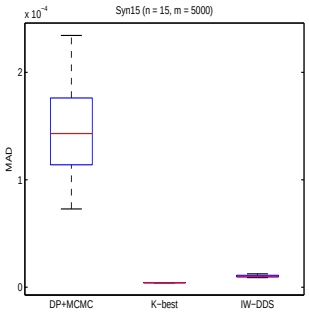

Figure 64: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Syn15 ($m = 5000$)

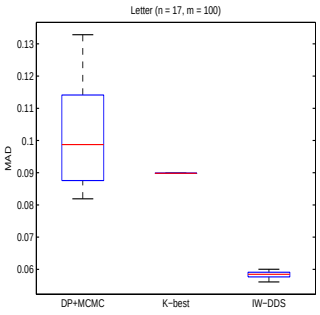

Figure 65: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Letter ($m = 100$)

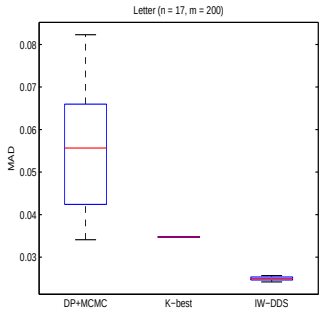

Figure 66: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Letter ($m = 200$)

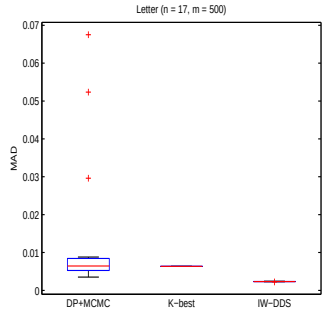

Figure 67: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Letter ($m = 500$)

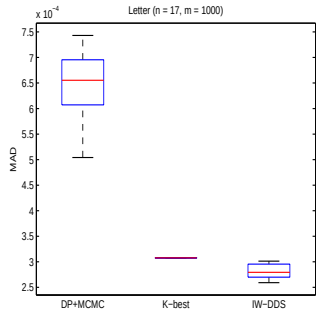

Figure 68: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Letter ($m = 1000$)

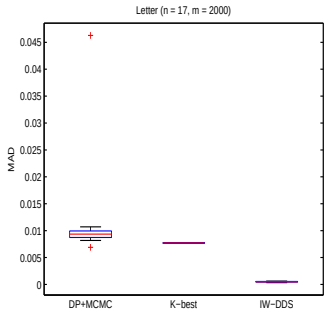

Figure 69: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Letter ($m = 2000$)

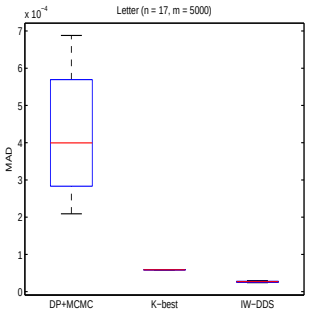

Figure 70: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Letter ($m = 5000$)

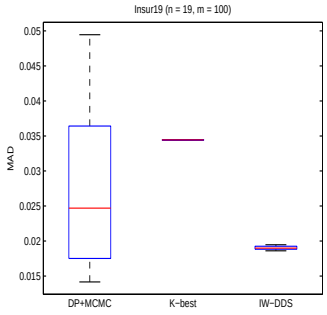

Figure 71: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Insur19 ($m = 100$)

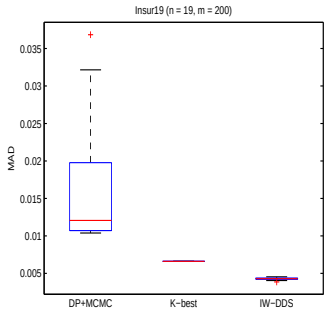

Figure 72: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Insur19 ($m = 200$)

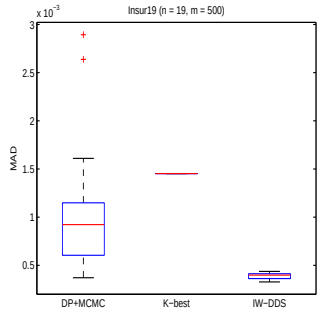

Figure 73: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Insur19 ($m = 500$)

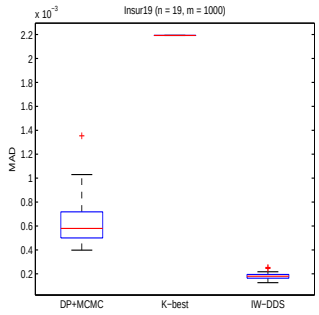

Figure 74: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Insur19 ($m = 1000$)

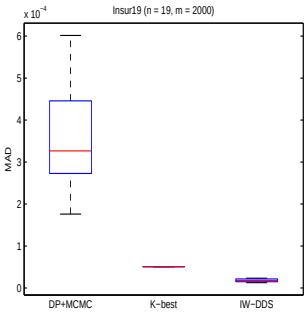

Figure 75: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Insur19 ($m = 2000$)

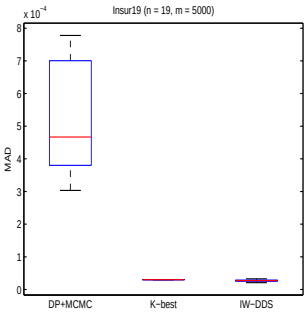

Figure 76: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Insur19 ($m = 5000$)

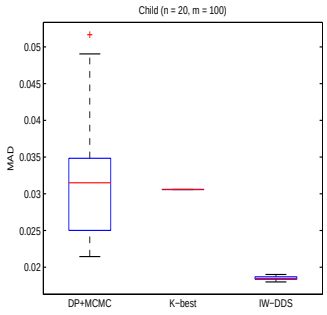

Figure 77: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 100$)

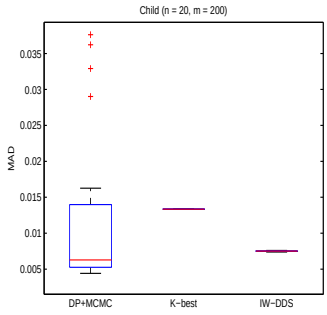

Figure 78: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 200$)

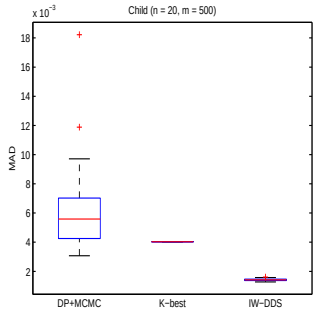

Figure 79: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 500$)

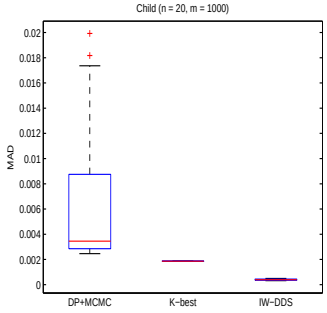

Figure 80: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 1000$)

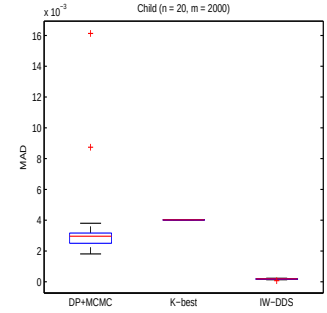

Figure 81: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 2000$)

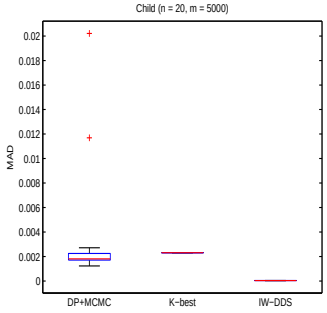

Figure 82: Boxplot of the MAD Performance of the DP+MCMC, the K-best, and the IW-DDS for Child ($m = 5000$)