2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–13

Masking schemes for universal marginalisers

Divya Gautam DIVYA.GAUTAM@BABYLONHEALTH.COM
Maria Lomeli MARIA.LOMELI@BABYLONHEALTH.COM
Kostis Gourgoulias KOSTIS.GOURGOULIAS@BABYLONHEALTH.COM
Daniel H. Thompson DANIEL.THOMPSON@BABYLONHEALTH.COM
Saurabh Johri SAURABH.JOHRI@BABYLONHEALTH.COM

Abstract

We consider the effect of structure-agnostic and structure-dependent masking schemes when training a universal marginaliser (Douglas et al., 2017) in order to learn conditional distributions of the form $P(x_i|\mathbf{x_b})$, where x_i is a given random variable and $\mathbf{x_b}$ is some arbitrary subset of all random variables of the generative model of interest. In other words, we mimic the self-supervised training of a denoising autoencoder, where a dataset of unlabelled data is used as partially observed input and the neural approximator is optimised to minimise reconstruction loss. We focus on studying the underlying process of the partially observed data—how good is the neural approximator at learning all conditional distributions when the observation process at prediction time differs from the masking process during training? We compare networks trained with different masking schemes in terms of their predictive performance and generalisation properties.

1. Motivation

In automated medical diagnosis a Bayesian network (BN) can be used as a statistical model for risk factors, diseases and symptoms in order to provide medical recommendations. For a BN to be appropriate for medical diagnosis, it typically needs to capture the relationships between thousands of diseases, symptoms and risk factors, making exact inference intractable. In this work, we focus on unlabeled binary data generated from a three-layer risk-disease-symptom BN as the generative model of interest.

During a typical diagnosis, only a subset of the evidence will be available, and hence we need to be able to efficiently (implicitly) marginalise over unobserved symptoms and risk factors. To obtain estimators of the posterior marginal distributions of diseases, we follow the amortised inference paradigm (Gershman and Goodman, 2014) and train a denoising autoencoder (Vincent et al., 2010; Bengio et al., 2013) using a variety of masking schemes for the input. This results in neural approximators for arbitrary conditional probabilities (Douglas et al., 2017; Belghazi et al., 2019). In this scenario, the quantities of interest are the posterior probability of diseases given sets of evidence containing observations from risk factors and/or symptoms from each patient. These estimates can then be used for diagnostic accuracy, namely, how high the algorithm ranks the reference diagnoses on a differential diagnosis (Shwe et al., 1991b). Since the goal is to provide medically accurate recommendations, it is crucial to obtain accurate marginal estimates as well as to

© D. Gautam, M. Lomeli, K. Gourgoulias, D.H. Thompson & S. Johri.

have computationally feasible inference schemes given the exponential number of queries required.

The main contribution of this work is a study of the effect on learning conditional distributions when the pattern of observed values in the data is misspecified. We consider multiple masking distributions, some of which are independent of the underlying structure of the BN, and others which are more tailored to it.

2. Background and related work

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$, $X \in \{0,1\}^n$, denote the collection of random variables in a BN of n nodes with joint distribution denoted by $P(\mathbf{x})$. $M(\mathbf{b})$ is the masking distribution for the masking random variable $\mathbf{B} \in \{0,1\}^n$. The masks are applied to the nodes through the element-wise product: $\mathbf{x} * \mathbf{b} = (x_1 b_1, \ldots, x_n b_n)$. If $b_i = 1$, we say that the value x_i is observed, otherwise it is unobserved or masked. We will also use the notation $\mathbf{x_b}$ to denote the subset of unmasked nodes.

We use the universal marginaliser (UM) (Douglas et al., 2017), a feedforward neural network that takes masked samples $\mathbf{x} * \mathbf{b}$ from a given BN as input, and learns to predict the marginals of all the nodes in the BN conditioned on the assignment of the observed nodes in the input. It is trained using the multi-label binary cross entropy loss:

$$\min_{\theta} \left\{ \mathbb{E}_{P(\mathbf{x})} \mathbb{E}_{M(\mathbf{b})} \sum_{i=1}^{n} CE(x_i | \mathbf{x}, \mathbf{b}; \theta) \right\}, \text{where}$$
$$CE(x_i | \mathbf{x}, \mathbf{b}; \theta) = -x_i \log Q_\theta(X_i = 1 | \mathbf{x_b}) - (1 - x_i) \log(1 - Q_\theta(X_i = 1 | \mathbf{x_b})) \tag{1}$$

in the same manner as a denoising autoencoder (Vincent et al., 2010). As a consequence, the UM, denoted by Q_θ, learns to output $Q_\theta(X_i = 1 | \mathbf{x_b}) \approx P(X_i = 1 | \mathbf{x_b})$, for all $i = 1, \ldots, n$. Joint posterior distributions can be estimated using the product rule and repeated evaluations of the UM. Section B in the supplementary material contains further details about our implementation.

There exist other types of universal marginalisers that use a variety of generative probabilistic models based on other neural networks to estimate arbitrary conditional distributions. Germain et al. (2015) modify an autoencoder neural network to estimate conditional probabilities with autoregressive constraints. The authors mask the weighted connections of a standard autoencoder to convert it into a distribution estimator. Ivanov et al. (2019) propose a variational autoencoder (Kingma and Welling, 2014) with arbitrary conditioning. The authors use stochastic variational Bayes during the training together with the reparameterisation trick as well as a masking distribution. Belghazi et al. (2019) propose a generative adversarial network (Goodfellow et al., 2014) to learn every single conditional distribution: they use gradient based regularisation as well as a masking distribution during training. It is possible to apply the masking schemes proposed in section 3 to these other types of universal marginalisers. Masking the input layer is a special case of dropout (Srivastava et al., 2014; Baldi and Sadowski, 2014), which is used to regularize over-parameterised neural networks by reducing the importance of individual neurons in the network during training.

3. Masking distributions

We consider a variety of masking distributions $M(\mathbf{b})$ that, together with the generative model $P(\mathbf{x})$, produce the data that we use to train the UM as $\mathbf{x} * \mathbf{b} \sim M(\mathbf{b})P(\mathbf{x})$, where $*$ denotes element-wise product.

Uniform power setwise (uniform): Masks are sampled uniformly from the power set of possible masks $\mathbf{b} \in \{0, 1\}^n$.

Uniform sizewise (sizewise): The distribution of the number of ones in $\mathbf{b}$ sampled using the uniform method above is Binomial$(n, 0.5)$, and hence the samples contain on average $n/2$ ones; see Figure 2 in the supplementary material for an illustration of this. If more samples from the tails of the Binomial are desired, a distribution that selects $\mathbf{b}$ uniformly in terms of the number of ones k therein can be achieved by first sampling a mask size $k \sim \text{Uniform}(\{0, \ldots, n\})$, and then sampling a permutation of the array $(1, \ldots, 1, 0, \ldots, 0)$ with k ones and $n - k$ zeros.

Independent nodewise (nodewise): Nodes are masked individually and independently rather than sampling an entire mask. Specifically, for each batch of samples, we sample $p \sim \text{Uniform}[0, p_{\max}]$ and set $M(b_i = 1) = p$ for that batch, where $p_{\max}$ is a hyperparameter of the model. As with sizewise masking above, this procedure controls the number of evidence sets of different sizes seen during training: each sampled value of p effectively specifies a maximum evidence set size for the masks in that batch. In this work we use $p_{\max} = 1$ in order to sample from all possible evidence sets during training, but with a different distribution to the two methods described above.

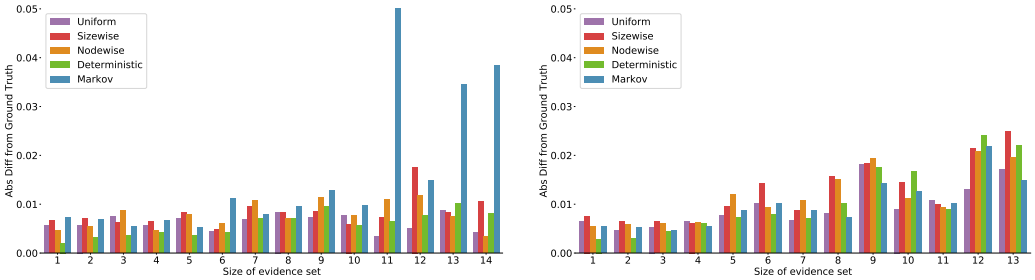
Deterministic cycle (deterministic): This is a deterministic version of the nodewise method above: instead of sampling the probability p of observing the nodes individually from a uniform distribution each batch, a list of values in $[0, 1]$ is preselected, and we cycle through these values proceeding to the next one each batch. The rationale is to avoid clusters of p values that can arise from sampling from the uniform distribution. Note that this method together with the above three are all **structure-agnostic** masking methods.

Markov blanket (markov): For each mask, all nodes are initially masked, then one disease node is chosen at random, and the nodes in its Markov blanket are unmasked using the independent nodewise scheme above. The Markov blanket of a node X is given by $\text{MB}(X) = \text{Pa}(X) \cup \text{Ch}(X) \cup \text{Pa}(\text{Ch}(X))$, where $\text{Pa}(X)$ and $\text{Ch}(X)$ correspond to the parents and children of X respectively. This is an example of a **structure-dependent** masking method.

4. Experiments

In this section, we compare the performance of the UM trained with loss function (1) using the different masking distributions described in Section 3. The target (synthetic) Noisy-OR BN consists of $n = 24$ nodes arranged in three layers of eight nodes each, representing risk factors, symptoms and diseases respectively. See Appendix B for details about the architecture and training of the UM, and Appendix C for the structure of the target BN.

We tested the UM's ability to reproduce conditioned marginals $P(X|\mathbf{X_b} = \mathbf{x_b})$ for evidence sets $\{\mathbf{X_b} = \mathbf{x_b}\}$ which were randomly chosen under two different observation models. In the first case, we assume that evidence is observed under the uniform power setwise



(a) Uniform evidence (b) Markov evidence

Figure 1: Absolute difference $|Q_\theta(X = 0|\mathbf{x_b}) - P(X = 0|\mathbf{x_b})|$ between the UM's prediction and the ground truth as a function of evidence set size k over $\{\mathbf{x_b}\}$, where the states $\mathbf{x}$ are picked arbitrarily and $\mathbf{b} \sim M(\mathbf{b})$, M is either the uniform power setwise masking distribution or the Markov blanket one. The queried disease node X is fixed across bars and plots.

masking scheme. In the second case, the test set construction uses the Markov masking assumptions—evidence is picked only from a randomly chosen disease node's Markov blanket. The assignments of the unmasked nodes in both cases were chosen arbitrarily. We constructed these test sets in order to capture the behaviour of each masking scheme when used for *training* a UM under different *prediction*-time observation patterns, mimicking the clinical diagnosis scenario. Our comparison looks at the absolute difference $|Q_\theta(X = 0|\mathbf{x_b}) - P(X = 0|\mathbf{x_b})|$ between the UM's prediction and the ground truth for a given disease node X as a function of evidence set size k. The ground truth $P(X = 0|\mathbf{x_b})$ can be obtained in small Bayesian networks using exact inference methods. In Figure 1(a), we see that in the case of uniform evidence all masking methods are competitive for small k, with deterministic masking being better on average in that regime. Markov masking progressively gets worse as k increases—the reason for this behaviour is that the Markov masking never chooses some of the masking patterns during training. When we mask the evidence with respect to the Markov distribution (Figure 1(b)), the Markov masking becomes competitive again across different k.

The UM in Figure 1 was trained with a large number of samples for each masking scheme, which makes the effect of the structure-agnostic masking distributions less apparent in the context of our small BN; see Appendix D for examples of the UM trained for fewer epochs where differences are more pronounced.

We conjecture that using different masking schemes for training and corruption has an effect due to the inherent structure in the dataset, for example when using data from a Bayesian network or image data. This effect could potentially be less pronounced if the dataset does not contain any structure. In Appendix D, we conducted an additional experiment to verify this claim empirically. We benchmarked some of our proposed masking schemes with a different universal marginaliser that uses a variational autoencoder architecture from (Ivanov et al., 2019). The yeast dataset (Dua and Graff, 2017) was used, which

is an example of independent and iid data and different masking and corruption schemes were picked.

5. Conclusions

A number of structure-agnostic masking schemes are presented and its performance has been evaluated using a structured synthetic dataset from a Bayesian network. The main conclusion of this work is that the choice of masking scheme employed during training impacts both predictive performance and training efficiency, this choice should be informed by the quantities of interest at prediction time for maximal benefit. Model generalisation is also affected by the choice of masking method when the corruption process differs from the masking used during training or is unknown, as in real world scenarios.

A natural extension to this work is to benchmark the structure-dependent and structure-agnostic masking schemes on real-world clinical case data and larger Bayesian networks.

Finally, we have used our masking schemes for other universal marginalisers such as the variational autoencoder with arbitrary conditioning (Ivanov et al., 2019) to see the effect of masking schemes for iid data. Further experiments can be conducted with different neural network architectures such as the generative adversarial network from Belghazi et al. (2019) and more complex datasets to study the performance of universal marginalisers for structured data. We believe this is an interesting avenue of future work.

References

Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.

E. Bauer, D. Koller, and Y. Singer. Update rules for parameter estimation in Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 3–13. Morgan Kaufmann Publishers Inc., 1997.

M. I. Belghazi, Y. Oquad, M. LeCun, and D. Lopez-Paz. Learning about an exponential amount of conditional distributions. In *Neural Information Processing Systems*, 2019.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems 26*, pages 899–907. 2013.

L. Douglas, I. Zarov, K. Gourgoulias, C. Lucas, C. Hart, A. Baker, M. Sahani, Y. Perov, and S. Johri. A universal marginaliser for amortized inference in generative models. In *Approximate inference NIPS workshop 2017*, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

M. Germain, K. Gregor, I. Murray, and H. Larochelle. MADE: masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015.

S. Gershman and N. Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the Cognitive Science Society*, 2014.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv preprint arXiv:1502.03167*, 2015.

Oleg Ivanov, Michael Figurnov, and Dmitry P. Vetrov. Variational autoencoder with arbitrary conditioning. In *ICLR*, 2019.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ArXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

H. E. Pople. Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnosis. *Artificial Intelligence in Medicine*, 1980.

M. Shwe and G. Cooper. An empirical analysis of likelihood-weighting simulation on a large multiply connected medical belief network. *Computers and Biomedical Research*, 1991.

M. A. Shwe, B. Middleton, D. E. Heckerman, Henrion M., E. J. Horvitz, H. P. Lehmann, and G. F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. *Method Inf. Med*, 1991a.

M. A. Shwe, B. Middleton, D. E. Heckerman, Henrion M., E. J. Horvitz, H. P. Lehmann, and G. F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. II.Evaluation of diagnostic performance. *Method Inf. Med*, 1991b.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

S. Tong and D. Koller. Active learning for parameter estimation in Bayesian networks. In *Neural Information Processing Systems*, 2000.

P. Vincent, H. Larochelle, I. Lajoie, and Manzagol P. A. Bengio, J. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 2010.

Appendix A. Distribution of possible masks of a given size

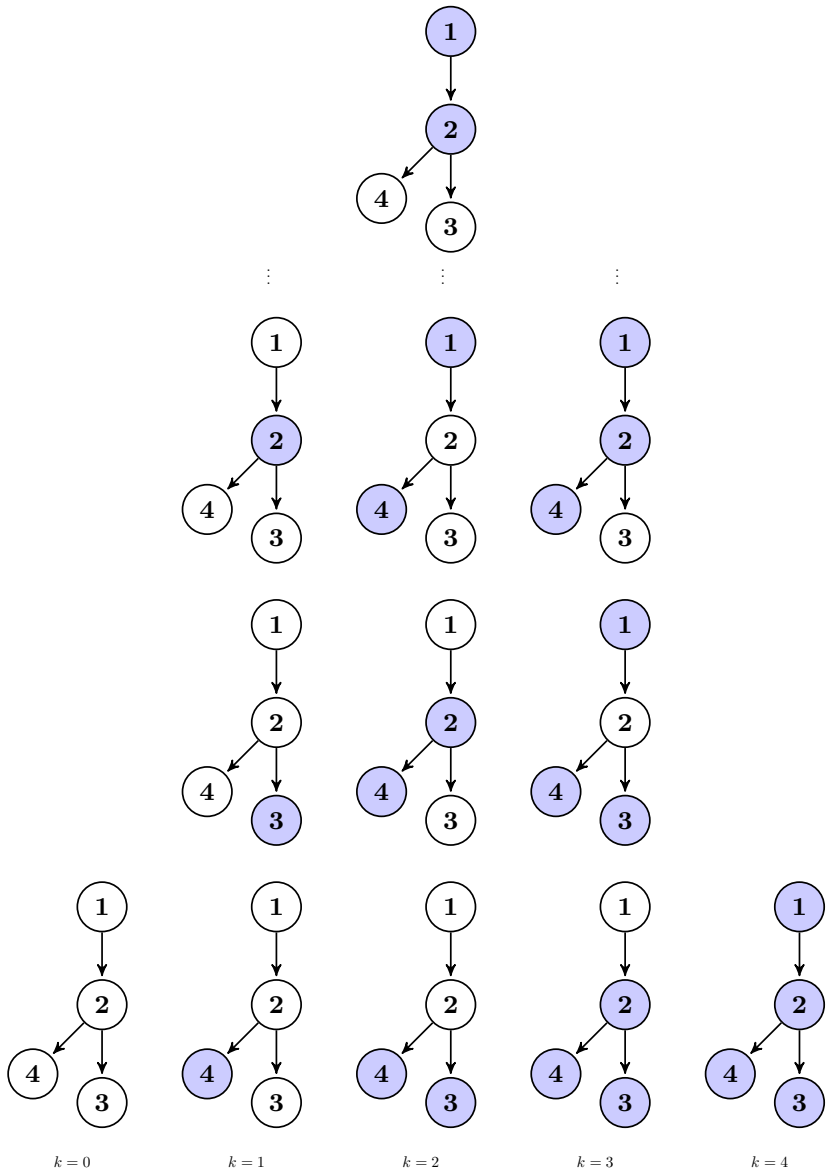


Figure 2: Illustration of the non-uniform distribution of the possible masks of size k for $k \in \{0, \ldots, 4\}$, for a BN of $n = 4$ nodes.

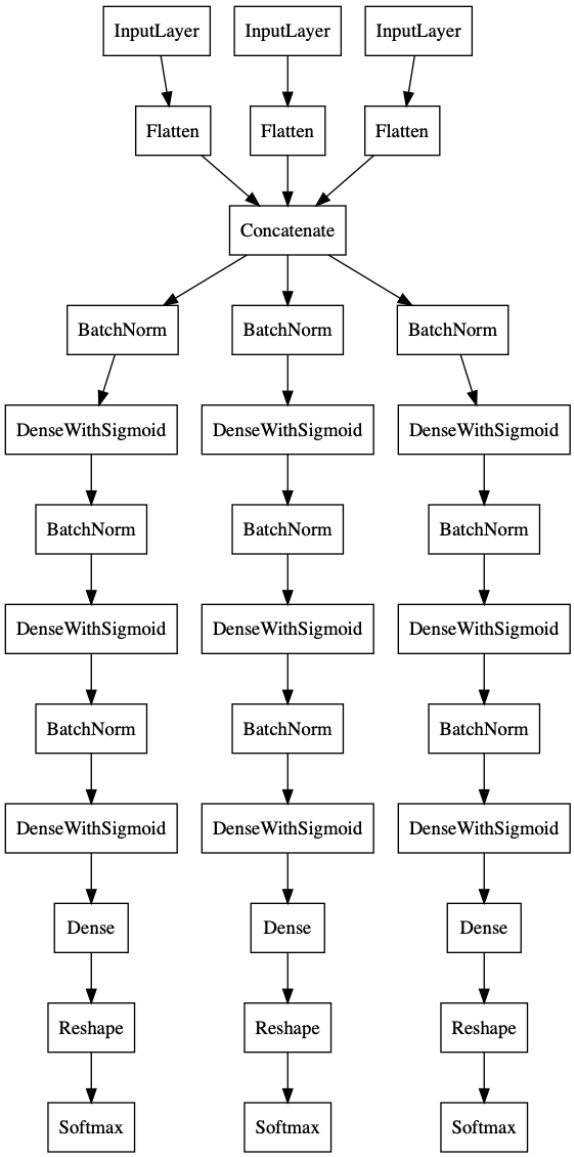


Figure 3: Neural network architecture. The three branches correspond to the R, D and S nodes of a three-layer RDS BN.

Appendix B. Neural network architecture and training

Figure 3 shows the neural network (NN) architecture, which consists of an input layer followed by a number of distinct multi-layer perceptron (MLP) branches, one for each layer of descendants in the BN, e.g. three branches for the R, D and S layers of a three-layer RDS BN network.

Each MLP branch consists of 3 fully-connected layers of 512 hidden units each, to which batch normalisation (Ioffe and Szegedy, 2015) before—and sigmoid activation after—is applied; followed by a final fully-connected layer of $N \times 2$ hidden units, where $N \in \{N_R, N_D, N_S\}$, and $\{N_R, N_D, N_S\}$ correspond to the numbers of R, D and S nodes in the BN respectively. Lastly softmax activation

is applied on a per node basis. This is done because of the 2 × Boolean parameterization of the node assignments, where (F, F) corresponds to unobserved/masked, (T, F) to observed and false, and (F, T) to observed and true. Hence the NN ends up learning two marginals for each node: one corresponding to the probability that the node is false, and another corresponding to the probability that it is true. The softmax activation ensures these two probabilities sum to 1.

The NN is trained using the Adam optimizer (Kingma and Ba, 2014), with the default parameter values and a learning rate of 10^{-4}.

Appendix C. Noisy-OR model

The number of parameters in a BN can be potentially large if the network is dense. Namely, each node i has $O(2^{|\mathrm{Pa}_{\mathcal{G}}(i)|})$ parameters where $|\mathrm{Pa}_{\mathcal{G}}(i)|$ is the number of parents of node i. In order to have a more compact representation for its parameters, only storing $O(|\mathrm{Pa}_{\mathcal{G}}(i)|)$ parameters per node, the Noisy OR model was proposed (Shwe et al., 1991a).

The top layer of nodes corresponds to risk factors, the second layer, to disease nodes and the third layer, to symptom nodes. The probability of a symptom node being false, given that only one disease is true and the rest are false, is $P(S_j = 0|\text{only } D_i = 1) = \gamma_{i0} \times \gamma_{ij}$, where γ_{i0} is the leak probability which corresponds to the case where all diseases are false. Analogously, the probability of a disease being false, given that only one risk factor is true, is $P(D_i = 0|\text{only } RF_r = 1) = \lambda_{r0} \times \lambda_{ir}$, where λ_{r0} is the leak probability for the case where all risk factors are false. The prior for risk factor r being false is denoted by $P(RF_r = 0) = \pi_j$. The general form for the probability of a risk factor, disease or symptom node being false given the values of its parents can be written in the following way:

$$P(RF_1 = r_1, \ldots, RF_n = r_n) = \prod_{j=1}^{n} \pi_j^{1-r_j} (1 - \pi_j)^{r_j}$$

$$P(D_i = 0|RF_1 = r_1, \ldots, RF_n = r_n) = \lambda_{i0} \prod_{j=1}^{n} \lambda_{ij}^{r_j}$$

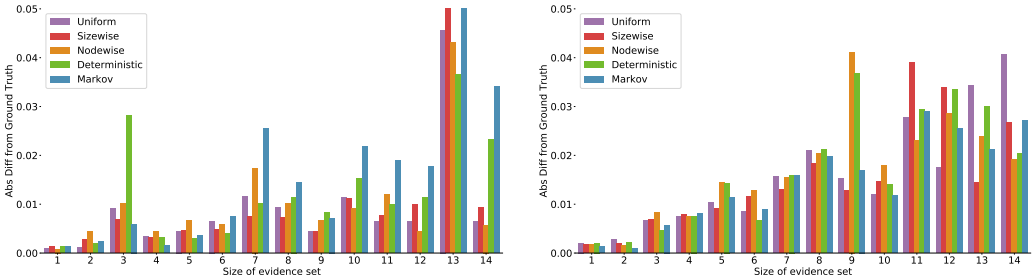
$$P(S_j = 0|D_1 = d_1, \ldots, D_m = d_m) = \gamma_{j0} \prod_{k=1}^{m} \gamma_{jk}^{d_k}.$$

The prior probabilities of diseases and risk factors can obtained from epidemiological data, where available. The conditional probabilities $P(S_j = 0|\text{only } D_i = 1)$ and $P(D_i = 0|\text{only } RF_r = 1)$ can be obtained through elicitation from multiple independent medical experts or from electronic medical records. The network structure expresses the corresponding clinical knowledge in graphical form, namely, whether or not there exists a direct relationship between a given pair of nodes.. See Shwe and Cooper (1991) for details about how to obtain the corresponding priors from QMR frequencies for a two layer BN and Shwe et al. (1991a), for another method to obtain the corresponding conditional probabilities. Bauer et al. (1997) and Tong and Koller (2000) propose an online parameter estimation and active learning methods respectively for parameter estimation in BNs. Pople (1980) discusses structure finding methods for BNs.

Appendix D. Additional Experiments

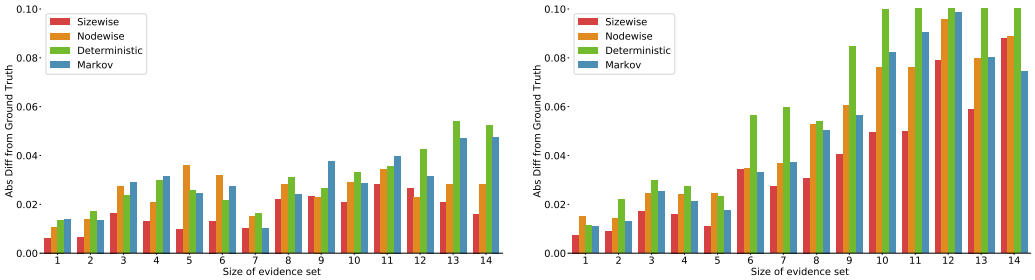
D.1. Masking schemes with data from a Bayesian network

In this appendix, we share a few more details from our experiments. Figure 4 is similar to the plot in the main paper, but for a different query disease node X. In this example it is somewhat more apparent that the deterministic-cycle masking is the best performing masking scheme for small



(a) Uniform evidence (b) Markov evidence

Figure 4: These plots correspond to a different query and are similar to Figure 1. Deterministic cycle masking is the best performing masking scheme for small evidence sets.



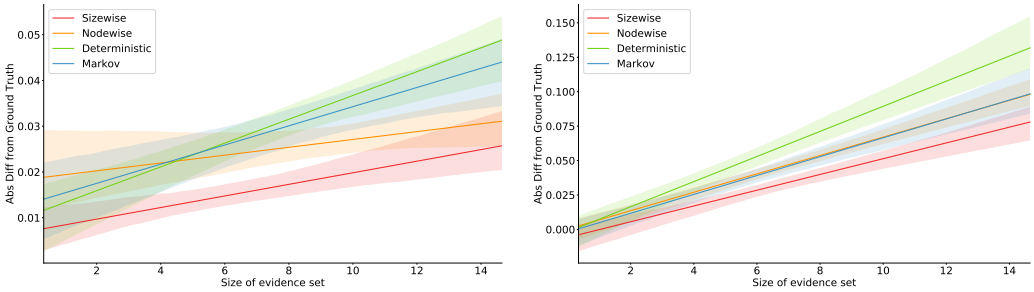
(a) Uniform evidence (b) Markov evidence

Figure 5: Models trained for 20 epochs exhibit poorer performance on average and the effect of different masking schemes is more pronounced. One epoch corresponds to the UM using 10 million samples. Note the different scale of the y-axis compared to Figures 1 and 4.

evidence sets. In Figure 4(a), the UM trained using Markov masking does not generalise well to uniformly-chosen evidence but is well-behaved over all the Markov-chosen evidence set sizes for which it was trained (Figure 4(b)).

During training with the different masking distributions, the UM obtained more than 10^{11} possible states of the 24-node BN during training, which is nearly four orders of magnitude larger than the size of the state space. In larger BNs with hundreds of nodes, the UM would not be able to get samples from substantial parts of the state space during training. In addition, for the masking distributions that satisfy $M(b_i) \in \{0, 1\}$ for all i, it is expected that a denoising-autoencoder will behave as a UM after obtaining a sufficiently large number of samples, making comparisons between masking schemes difficult. UMs trained with uniform masking exhibit good performance (Douglas et al., 2017) even for large BNs.

In order to simulate the scenario of training a large BN with the same small sythentic BN available, we trained another series of models for a much smaller number of epochs, in order to see if differences between the masking schemes are more pronounced. Figures 5 and 6 correspond to this experiment. Across the various queries considered during this experiment (most of which are



(a) Uniform evidence (b) Markov evidence

Figure 6: Linear model plots corresponding to the same experiment as Figure 5. For a small number of epochs, the sizewise masking distribution is the best scheme to use for both the uniform and the Markov evidence.

not shown), we noticed that the deterministic cycle scheme had the worst behaviour whereas the sizewise scheme was the most performant.

D.2. Masking schemes with iid data

Feature	MCAR-trained	Sizewise-trained
0	0.02748 (0.044)	0.02838 (0.046)
1	0.02160 (0.037)	0.02207 (0.038)
2	0.00936 (0.017)	0.00942 (0.017)
3	0.02511 (0.050)	0.02534 (0.051)
4	0.00386 (0.031)	0.00389 (0.031)
5	0.00856 (0.073)	0.00845 (0.073)
6	0.00404 (0.012)	0.00406 (0.012)
7	0.01597 (0.046)	0.01584 (0.046)

Table 1: MCAR corrupted data. Mean squared error of reconstruction per feature and standard deviation.

In this section, we included an experiment with the variational autoencoder with arbitrary conditioning (VAEAC) architecture (Ivanov et al., 2019), together with the yeast dataset (Dua and Graff, 2017). Ivanov et al. (2019) consider a single corruption distribution called "Missing completely at random" (MCAR) which corrupts the original data by randomly turning observed values into unobserved. The masking scheme MCAR with parameter equal to 0.5 corresponds to what our masking scheme called **uniform power setwise masking distribution**; see Section 3 for details. Their experimental setup goes as follows: a VAE is first trained to reconstruct the corrupted dataset with the use of another MCAR distribution that is different from the one used to corrupt the dataset. Since the VAE can utilize any reasonable masking distribution in order to recover the original values of the features, we extended this experiment. Specifically, the dataset was corrupted and reconstructed using different masking distributions for each step. Then, the mean-squared errors per feature were compared for each configuration of masking and corruption

Feature	MCAR-trained	Sizewise-trained
0	0.03181 (0.051)	0.03202 (0.051)
1	0.02286 (0.038)	0.02317 (0.038)
2	0.00980 (0.017)	0.01002 (0.017)
3	0.02619 (0.052)	0.02617 (0.051)
4	0.00370 (0.030)	0.00372 (0.030)
5	0.00878 (0.075)	0.00878 (0.075)
6	0.00410 (0.011)	0.00411 (0.011)
7	0.01539 (0.041)	0.01524 (0.042)

Table 2: Structured-corrupted data. Mean squared error of reconstruction per feature and standard deviation.

processes. In Table 1, the data were corrupted by an MCAR with probability of observing a single feature $p = 0.5$ per row of the dataset. The VAE that uses a sizewise distribution during training does not have a considerable advantage over the MCAR case.

We also picked the corruption distribution to be different from the possible masking schemes used for training. In Table 2, we consider a corruption distribution that induces some structure - we call this a structured-corrupted distrbution (SC). The corruption scheme SC always masks neighboring triplets of features, i.e., $(0, 1, 2)$, $(4, 5, 6)$, etc. as opposed to MCAR, which independently corrupts different features. The SC corruption distribution emulates situations where only a subset of relevant features is observed at any given time. As expected, the VAEAC using MCAR or uniform-sizewise during training exhibits larger error, on average, when the ground truth masking is SC. We conjecture that part of this error is because the VAE cannot capture couplings between the predicted distributions given evidence but further investigations are needed to verify such claim.