Probabilistic Multi-Dimensional Classification with Incomplete Data at Prediction Time

Thu-Ha Do **Vu-Linh Nguyen** **Yves Grandvalet**
Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France

Abstract

Multi-dimensional classification (MDC) extends multi-class and multi-label learning by predicting several class variables per instance. We revisit probabilistic MDC methods with mixed features (discrete and continuous), focusing on their strengths and limits for handling incomplete data at prediction time. We present theoretical results leading to a new probabilistic approach with efficient learning and prediction algorithms that address scalability and robustness issues. Experiments demonstrate its benefits in different missingness scenarios.

1 Introduction

Multi-label classification predicts multiple binary outcomes per instance, while multi-dimensional classification (MDC) generalizes this to categorical outcomes with more than two values. MDC applies to various domains (Gil-Begue et al., 2021; Jia and Zhang, 2024; Ma and Chen, 2018; Nguyen et al., 2023), such as predicting multiple related health indices. The categorical indices can be nominal, mutually exclusive disease subtypes (Din et al., 2019; Ferreira et al., 2020; Garcia-Aymerich et al., 2011; Hoshida et al., 2007), or ordinal, such as graded disease severity (Dadu et al., 2022; Lynch et al., 2015). Another example is predicting antimicrobial resistance phenotypes (susceptible, intermediate, resistant) for multiple drugs based on the genomic sequences of bacterial strains (Do et al., 2024; Moradigaravand et al., 2018).

Learning and inference in probabilistic graphical models are generally NP-hard (Cooper, 1990; Chickering, 1996). Moreover, MDC faces amplified challenges restricting its applicability, as the number of class configurations grows exponentially with the number and cardinality of class variables. This amplifies several issues related to data sparsity, scalability, and interpretability (Gil-Begue et al., 2021).

In this context, discriminative MDCs (DMDCs) (Nguyen et al., 2023) balance scalability and expressiveness. They handle nicely mixed features (discrete and continuous) by constraining discrete–continuous dependencies. Though still NP-hard in the total number of discrete variables, DMDCs partition inputs into discrete and continuous features with conditional distributions. These models predict accurately when all features are observed, but performance may drop when discrete features are missing at prediction time.

We extend DMDCs by relaxing the structural constraint on discrete–continuous links, which allows for the processing of incomplete test data using three prediction principles: optimistic, pessimistic, and averaging. Our algorithms maintain the complexity level of DMDC, yet improve robustness under missingness in discrete variables. We analyze Bayesian optimal predictions under each principle, for the commonly used Hamming and 0/1 losses, showing equivalence to MAP inference under the optimistic and averaging principles in several cases.

Section 2 introduces probabilistic MDCs and our Hybrid MDCs, Section 3 covers learning and prediction, Section 4 reports experiments, and Section 5 concludes. Proofs of propositions are deferred to the appendix.

2 Probabilistic MDCs

Probabilistic MDCs (PMDCs), which are essentially Bayesian networks (BNs), are implemented in three stages: representation, learning, and inference. This section discusses how representation affects learning and inference. We first present the two standard approaches, *i.e.*, generative and discriminative, before

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

introducing our hybrid approach.

2.1 Generalities

Let $\mathcal{X} := \mathcal{X}^1 \times \ldots \times \mathcal{X}^Q$ denote the feature space, and let $\mathbf{X} = \{X^1, \ldots, X^Q\}$ be a finite set of features. In the mixed-feature setting, the feature variables consist of two disjoint sets: discrete features $\mathbf{X}^{\mathcal{D}}$, and continuous features $\mathbf{X}^{\mathcal{C}}$, such that $\mathbf{X} = \mathbf{X}^{\mathcal{C}} \cup \mathbf{X}^{\mathcal{D}}$, where $\mathbf{X}^{\mathcal{D}} = \{X^1, \ldots, X^P\}$ and $\mathbf{X}^{\mathcal{C}} = \{X^{P+1}, \ldots, X^Q\}$.

Next, let $\mathcal{Y} := \mathcal{Y}^1 \times \ldots \times \mathcal{Y}^K$ be the set of possible outcomes, and let $\mathbf{Y} = \{Y^1, \ldots Y^K\}$ be the set of the K class variables. For any $k \in [K] := \{1, \ldots, K\}$, let $\mathcal{Y}^k := \{y_1^k, \ldots, y_{M_k}^k\}$ be the set of possible outcomes that Y^k can take. The cardinality of the class variable Y^k is $M_k = |\mathcal{Y}^k|$, for $k \in [K]$. Finally, we also define the shorthand notations $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, $\mathbf{Z} = \mathbf{Y} \cup \mathbf{X}$ and $\mathbf{Z}^{\mathcal{D}} = \mathbf{Y} \cup \mathbf{X}^{\mathcal{D}}$.

The distribution of a PMDC is represented by a directed acyclic graph (DAG) G, where nodes are variables and directed edges model conditional dependencies between them, and are parameterized by $\boldsymbol{\theta}$. For all $\boldsymbol{z} \in \mathcal{Z}$, the distribution factorizes as:

$$\begin{aligned} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{z}) &= \prod_{Z^j \in \mathbf{Z}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(Z^j = z^j \,|\, \mathrm{pa}(Z^j) = \mathrm{pa}(z^j)) \\ &=: \prod_{z^j \in \mathcal{Z}^j} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j)) \ , \end{aligned}$$

where $\mathrm{pa}(Z^j)$ denotes the parent set of Z^j in the DAG G, while $\mathrm{pa}(z^j)$ is the specific configuration of $\mathrm{pa}(Z^j)$ specified by $\boldsymbol{z}$. The set of distributions that can be represented can be restricted by constraining the type or number of dependencies between variables, leading to different PMDC types as detailed below.

2.2 Generative and Discriminative PMDCs

Generative PMDCs (GMDCs) (Van Der Gaag and De Waal, 2006) extend classical BNs. Their DAGs are restricted by excluding arcs from features to class variables, as illustrated in Figure 1. They are adjusted by maximizing the joint likelihood over data

$$\prod_{\boldsymbol{z}_n \in \mathcal{S}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{z}_n) \ ,$$

where $\mathcal{S} := \{\boldsymbol{z}_n\}_{n=1}^N = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ is the training set. While the model is expressive, its inference is NP-hard, and scalability requires additional structural constraints to mitigate the computational complexity of both learning and inference (Benjumeda et al., 2018; Bielza et al., 2011; Van Der Gaag and De Waal, 2006).

Discriminative PMDCs (DMDC) (Nguyen et al., 2023) extend BN classifiers. Their DAGs are restricted by excluding arcs from class to feature variables, as illustrated in Figure 1. They are adjusted by maximization of the conditional likelihood over data

$$\prod_{\boldsymbol{z}_n \in \mathcal{S}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n) \ .$$

The decomposability of the conditional log-likelihood, combined with efficient structure learning algorithms (Bartlett and Cussens, 2017; Cussens, 2020), allows DMDC to handle mixed data. For mixed data, DMDCs further forbid arcs between continuous and discrete features. Inference remains NP-hard in the number of class variables but is practical for typical MDC tasks.

2.3 Hybrid PMDCs

GMDCs model the joint distribution $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathbf{Z})$ and are the most general family, but their application to mixed data is tricky. Extensions to continuous variables usually require strong assumptions, such as Gaussianity of variables and linearity of relationships. Discretization is a common but crude workaround (Sucar, 2021). Moreover, GMDCs optimize the joint likelihood, which is not discriminant and may thus perform poorly for classification purposes, particularly in M-open problems (Roos et al., 2005).

DMDCs are discriminant in the sense that they model the conditional distribution $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathbf{Y} \mid \mathbf{X}^{\mathcal{D}}, \mathbf{X}^{\mathcal{C}})$, and they are computationally efficient. Yet they cannot handle missing discrete features at prediction time, which requires $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathbf{Y}, \mathbf{X}^{\mathcal{D}} \mid \mathbf{X}^{\mathcal{C}})$.

2.3.1 Structural Constraints

To address these limitations, we introduce Hybrid PMDCs (HMDCs), a minimal extension of DMDCs that enables inference with missing discrete features or class variables. HMDCs relax DMDCs' structural constraints by allowing arcs into discrete features from class or continuous variables, while still prohibiting arcs into continuous features from discrete variables. The model is trained by maximizing the conditional likelihood

$$\prod_{\boldsymbol{z}_n \in \mathcal{S}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}_n, \boldsymbol{x}_n^{\mathcal{D}} \,|\, \boldsymbol{x}_n^{\mathcal{C}}) \ , \tag{1}$$

where $\boldsymbol{x}_n^{\mathcal{D}}$ and $\boldsymbol{x}_n^{\mathcal{C}}$ are respectively the discrete and continuous features of $\boldsymbol{x}_n$. An example of HMDC is shown in Figure 1.

This design, inspired by causal sufficiency (Pearl and Mackenzie, 2018), ensures that each potentially missing variable has at least one observed parent, allowing the model to predict missing features directly rather

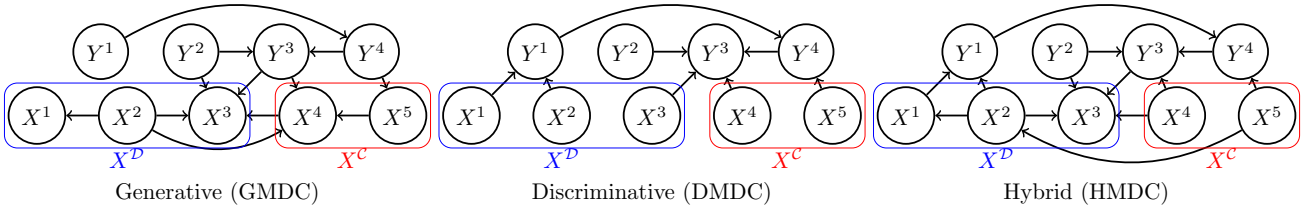


Figure 1: Examples of DAGs illustrating the PMDC types: GMDC, DMDC, and HMDC

than imputing them, thereby improving robustness in MAP-based inference. As an intermediate between GMDCs and DMDCs, HMDCs trade off representational flexibility and computational complexity: exact inference remains NP-hard in the number of class variables and discrete features. To mitigate this, grouping methods (Parviainen and Kaski, 2017) can be applied; here, we adopt simple random grouping, leaving optimal grouping strategies for future work.

2.3.2 Learning Objective

HMDCs are adjusted by maximizing the conditional likelihood (1), which is equivalent to maximizing the conditional log-likelihood

$$L(G, \boldsymbol{\theta} \,|\, \mathcal{S}) = \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log \left(\boldsymbol{p}_{(G,\boldsymbol{\theta})} \left(\boldsymbol{y}_n, \boldsymbol{x}_n^{\mathcal{D}} \,|\, \boldsymbol{x}_n^{\mathcal{C}} \right) \right) \ . \quad (2)$$

This design, positioned between GMDCs and DMDCs, targets deployment with mixed features and possibly missing discrete variables. Like DMDCs, and unlike GMDCs, HMDCs do not model the distribution of continuous features. Although not purely discriminative, HMDCs can be viewed as a mixture of discriminant models via the decomposition

$$\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}_n, \boldsymbol{x}_n^{\mathcal{D}} \,|\, \boldsymbol{x}_n^{\mathcal{C}}) = \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n^{\mathcal{D}}, \boldsymbol{x}_n^{\mathcal{C}}) \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{x}_n^{\mathcal{D}} \,|\, \boldsymbol{x}_n^{\mathcal{C}}) \ ,$$

where the discriminant components $\boldsymbol{p}(\boldsymbol{y}_n \mid \boldsymbol{x}_n^{\mathcal{D}}, \boldsymbol{x}_n^{\mathcal{C}})$ are weighted by $\boldsymbol{p}(\boldsymbol{x}_n^{\mathcal{D}} \mid \boldsymbol{x}_n^{\mathcal{C}})$. This mixture structure provides the flexibility to predict outcomes with missing discrete variables directly, without requiring their imputation. Compared to GMDCs, HMDCs avoid modeling continuous inputs, reducing both training complexity and data requirements.

3 Fitting and Using Hybrid MDCs

This section elaborates on the learning and prediction phases of HMDCs.

3.1 Learning

The goal of the learning phase is to maximize the conditional likelihood (2). To further mitigate overfitting in this data-hungry setting, we optimize a regularized conditional log-likelihood

$$L_\Omega(G, \boldsymbol{\theta} \,|\, \mathcal{S}) = L(G, \boldsymbol{\theta} \,|\, \mathcal{S}) - \Omega(G, \mathcal{S}) \ , \quad (3)$$

where $\Omega(G, \mathcal{S})$ is a decomposable regularization term

$$\Omega(G, \mathcal{S}) = \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S}) \ ,$$

which quantifies the complexity of the model. The regularized conditional log-likelihood (3) generalizes (2), with common choices for $\Omega(G, \mathcal{S})$ including the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which differ in penalty strength (Murphy, 2023, Sec. 3.8.7). They correspond to

$$\begin{aligned} \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S}) &= -Dim(G) \ \text{(AIC)} \ , \\ \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S}) &= -0.5 Dim(G) \log(N) \ \text{(BIC)} \ , \\ \text{with } \ Dim(G) &= \sum_{j \in \mathcal{D}} (|Z^j| - 1)|\mathrm{pa}(Z^j)| \ , \end{aligned}$$

where $|Z^j|$ is the cardinality of Z^j and $|\mathrm{pa}(Z^j)|$ is the number of possible configurations of $\mathrm{pa}(Z^j)$. In this paper, we adopt BIC and leave an extensive study of alternatives as a follow-up work.

In the following, we show that (3) can be optimized exactly under standard assumptions (Nguyen et al., 2023), and that existing learning strategies extend to our more general setting with minimal changes.

Proposition 1. *For any $\boldsymbol{z} \in \mathcal{Z}$, the structure constraints of HMDCs ensure that we have*

$$\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}, \boldsymbol{x}^{\mathcal{D}} \,|\, \boldsymbol{x}^{\mathcal{C}}) = \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j)) \ .$$

Proposition 1 implies that the estimation of $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}, \boldsymbol{x}^{\mathcal{D}} \mid \boldsymbol{x}^{\mathcal{C}})$ is not affected by arcs among continuous features $\mathbf{X}^{\mathcal{C}}$, allowing to exclude them in the DAG search, without loss of predictive power.

Proposition 2. *Let $\mathcal{S}$ be any finite training set. For any $Z^j \in \mathbf{Z}^{\mathcal{D}}$, let*

$$L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) := \ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) - \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S}) \tag{4}$$

$$\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) := \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\left(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\right) . \tag{5}$$

which only depends on the local structure and parameters $(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j)$ defining the local model at each node. The structure constraints of HMDCs ensure that, for any HMDC $(G, \boldsymbol{\theta})$, we have the decomposition:

$$L_\Omega(G, \boldsymbol{\theta} \,|\, \mathcal{S}) = \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \ .$$

Proposition 3. *For any $Z^j \in \mathbf{Z}^{\mathcal{D}}$, let $2^{\mathbf{Z}\setminus\{Z^j\}}$ be the set of all the subsets of $\mathbf{Z}\setminus\{Z^j\}$. Let $\mathcal{P}a(Z^j) \subset 2^{\mathbf{Z}\setminus\{Z^j\}}$ be any set of possible parent sets of Z^j. Given a training set $\mathcal{S}$, the task of learning an optimal HMDC, specified by the pair $(G^*, \boldsymbol{\theta}^*)$ that maximizes the regularized conditional log-likelihood in* (3)*, can be decomposed into a set of local learning problems and a BN structure learning problem.*

Concretely, training the model reduces to training a collection of local probabilistic multi-class classifiers. Each local classifier

$$\mathbf{h}_{\mathrm{pa}(Z^j)} : \bigtimes_{Z^k \in \mathrm{pa}(Z^j)} \mathcal{Z}^k \longrightarrow \Delta^{|\mathcal{Z}^j|} \ , \tag{6}$$

maps every configuration of the parent set $\mathrm{pa}(Z^j) \in \mathcal{P}a(Z^j)$ to a probability distribution over the levels of $Z^j \in \mathbf{Z}^{\mathcal{D}}$, with $\Delta^{|\mathcal{Z}^j|}$ being the probability simplex of dimension $|\mathcal{Z}^j|$. Each classifier is parameterized by parameters

$$\boldsymbol{\theta}_j^* \in \operatorname*{argmax}_{\boldsymbol{\theta}_j \in \Theta_j} \ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \ , \tag{7}$$

with $\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S})$ defined in (5)*. Each possible parent set $\mathrm{pa}(Z^j)$ is then associated with its optimal score $L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^* \,|\, \mathcal{S})$* (4)*. Finally, any scoring-based BN structure learning technique can be employed to find an optimal DAG G^* specified by $|\mathbf{Z}^{\mathcal{D}}|$ parent sets, one per $Z^j \in \mathbf{Z}^{\mathcal{D}}$, denoted by $\mathrm{pa}^*(Z^j)$. Together with their optimal parameter sets $\{\boldsymbol{\theta}_j^* \,|\, Z^j \in \mathbf{Z}^{\mathcal{D}}\} := \boldsymbol{\theta}^*$, we have an optimal HMDC $(G^*, \boldsymbol{\theta}^*)$.*

The assumption that $\mathcal{P}a(Z^j) \subset 2^{\mathbf{Z}\setminus\{Z^j\}}$ can be any given set of possible parent sets of Z^j allows us to be more flexible in defining the hypothesis space $\boldsymbol{p}_{(G,\boldsymbol{\theta})}$. For example, if one prefers to only seek the most promising $\boldsymbol{p}_{(G,\boldsymbol{\theta})}$ that does not exceed a certain level of complexity, one may try to encode the complexity level via the possible parent sets of Z^j, for any $Z^j \in \mathbf{Z}^{\mathcal{D}}$. As a consequence of Proposition 3, once the optimal local criteria (5) are computed, any scoring-based structure learning algorithm for discrete BNs can be employed to find an optimal DAG over $\mathbf{Z}^{\mathcal{D}}$.

Following (Nguyen et al., 2023), we employ GOBNILP (Bartlett and Cussens, 2017; Cussens, 2020), which is a globally optimal BN learning using integer linear programming, to solve the structure learning problem, and we also adapt the pruning rule (de Campos et al., 2018) to prune away unnecessary large candidate parent sets $\mathrm{pa}(Z^j) \in \mathcal{P}a(Z^j)$ that cannot be the parent set of Z^j in any optimal DAG G^*. This pruning does not require the fitting of the local classifier $\mathbf{h}_{\mathrm{pa}(Z^j)}$.

Proposition 4. *Let $Z^j \in \mathbf{Z}^{\mathcal{D}}$. For any $\mathrm{pa}'(Z^j) \supset \mathrm{pa}(Z^j)$, if*

$$L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \geq -\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S}) \ ,$$

then $\mathrm{pa}'(Z^j)$ cannot be a parent set of Z^j in any optimal DAG G^*.

This property can help reduce redundant computations, leveraging GOBNILP's flexibility in handling additional constraints on the set of candidate DAGs.

So far, we have assumed that for any possible pair $(Z^j, \mathrm{pa}(Z^j))$, the conditional distribution $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(Z^j \,|\, \mathrm{pa}(Z^j))$ using can be modeled by a probabilistic classifier $\mathbf{h}_{\mathrm{pa}(Z^j)}$ (6). Since $\mathrm{pa}(Z^j)$ may include both discrete and continuous variables, a theoretically sound approach is required to define and train these local probabilistic multi-class classifiers $\mathbf{h}_{\mathrm{pa}(Z^j)}$ in the presence of mixed features $\mathrm{pa}(Z^j)$. We adopt the input partitioning strategy of Nguyen et al. (2023) to construct $\mathbf{h}_{\mathrm{pa}(Z^j)}$to reduce the complexity of the learning phase while still enabling feature selection locally via either the construction of the local regularization terms or the choice of probabilistic multi-class classifiers like random forests, which typically select subset of features during training. This may help to discard irrelevant continuous features from $\mathrm{pa}(Z^j)$, $Z^j \in \mathbf{Z}^{\mathcal{D}}$.

For each possible candidate pair $(Z^j, \mathrm{pa}(Z^j))$, let $\mathrm{pa}_{\mathcal{D}}(Z^j)$ denote the set of the discrete variables among the parents of Z, $\mathrm{pa}_{\mathcal{D}}(Z^j) := \mathrm{pa}(Z^j) \cap \mathbf{Z}^{\mathcal{D}}$, and let

$$\mathcal{P}a_{\mathcal{D}}(Z^j) := \bigtimes_{Z^k \in \mathrm{pa}_{\mathcal{D}}(Z^j)} \mathcal{Z}^k \ ,$$

be the set of all possible configurations of $\mathrm{pa}_{\mathcal{D}}(Z^j)$. We partition $\mathcal{X}^{\mathcal{C}}$ into $|\mathcal{P}a_{\mathcal{D}}(Z^j)|$ regions, one per possible configuration $\mathrm{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)$, where a local probabilistic multi-class classifier

$$\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)} : \mathcal{X}^{\mathcal{C}} \longrightarrow \Delta^{|\mathcal{Z}^j|} \tag{8}$$

is used to model $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(Z^j \,|\, \mathrm{pa}_{\mathcal{D}}(Z^j) = \mathrm{pa}_{\mathcal{D}}(z^j), \mathbf{X}^{\mathcal{C}})$. For any $\mathrm{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)$, let

$\ell(\mathrm{pa}_{\mathcal{D}}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S})$

$$= \sum_{\substack{\boldsymbol{z}_n \in \mathcal{S} \\ \mathrm{pa}_{\mathcal{D}}(z_n^d) = \mathrm{pa}_{\mathcal{D}}(Z^j)}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^d \mid \mathrm{pa}_{\mathcal{D}}(Z^j), \boldsymbol{x}_n^{\mathcal{C}})\Big) \ . \quad (9)$$

The $|\mathcal{P}a_{\mathcal{D}}(Z^j)|$ classifiers (8) can be estimated independently because

$$\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) = \sum_{\mathrm{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)} \ell(\mathrm{pa}_{\mathcal{D}}(z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \ .$$

Altogether, the entire training phase is summarized in Algorithm 1. As a reminder, in this algorithm, we use GOBNILP (Bartlett and Cussens, 2017; Cussens, 2020) to find an optimal DAG G^* using $L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^* \,|\, \mathcal{S})$ (4), but any scoring-based BN structure learning algorithm could be employed in place of GOBNILP.

Proposition 5. *Given any finite training set $\mathcal{S}$, Algorithm 1 returns an optimal HMDC $\boldsymbol{p}_{(G^*,\boldsymbol{\theta}^*)}$ that optimizes the regularized conditional log-likelihood* (3) *if, for any possible pair $(Z^j, \mathrm{pa}(Z^j))$, and any $\mathrm{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$* (8) *that optimizes* (9) *under the feature selection option can be found.*

Algorithm 1 Learning an optimal HMDC $(G^*, \boldsymbol{\theta}^*)$

input: Training data $\mathcal{S} := \{\boldsymbol{z}_n\}_{n=1}^N = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$, possible parent sets $\mathcal{P}a(Z^j)$, $Z^j \in \mathbf{Z}^{\mathcal{D}}$, and hypothesis spaces for local classifiers $\mathbf{h}_{\mathrm{pa}(Z^j)}$ (8).
for $Z^j \in \mathbf{Z}^{\mathcal{D}}$ **do**
  Sort $\mathrm{pa}(Z^j) \in \mathcal{P}a(Z^j)$ s.t. $|\mathrm{pa}(Z^j)|$ increases
  for $\mathrm{pa}(Z^j) \in \mathcal{P}a(Z^j)$ **do**
    for $\mathrm{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)$ **do**
      Train a local classifier $\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$ (8)
    end for
    Store optimal set $\boldsymbol{\theta}_j^*$ (7) of $\mathbf{h}_{\mathrm{pa}(Z^j)}$ (6)
    Compute optimal score $L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^* \,|\, \mathcal{S})$ (4)
    Prune all $\mathrm{pa}'(Z^j) \supset \mathrm{pa}(Z^j)$ satisfying Proposition 4
  end for
end for
Find optimal DAG G^* using $L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^* \,|\, \mathcal{S})$ (4).
for $Z^j \in \mathbf{Z}^{\mathcal{D}}$ **do**
  Extract $\mathrm{pa}^*(Z^j)$ from G^* and its optimal parameter set $\boldsymbol{\theta}_j^*$.
end for
return: $(G^*, \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}_j^* \,|\, Z^j \in \mathbf{Z}^{\mathcal{D}}\}$

While doing both structure and parameter learning can be costly, it is needed to detect and incorporate conditional (in)dependencies $Z^i \to \mathbf{X}^C \leftarrow Z^j$ and $Z^i \leftarrow \mathbf{X}^C \to Z^j$, which might be difficult to detect even for domain experts, into the DAG structure, thanks to the relation between unconstrained DAGs and their I-maps (Koller and Friedman, 2009) under structural constraints imposed for Hybrid PMDCs. Moreover, as will be shown later, when the base classifiers $\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$, such as logistic regression, can provide good estimates of the local conditional probability distribution, the DAGs given by our framework seem to reflect the dependencies in the dataset reasonably. This is reflected via the sparseness of the DAGs as well as the robustness of the predictive performance.

3.2 Prediction

We now address the prediction of class variables for query instances $\boldsymbol{z}$ where some discrete features and class variables in $\mathbf{Z}^{\mathcal{D}}$ are missing. This framework encompasses several practical scenarios:

(i) Partial observation of classes: some class variables are observed and can be leveraged to enhance the prediction of the missing ones;

(ii) Incomplete data at prediction time: high-quality and complete training data may be available (for instance, from well-monitored cohorts in clinical studies), while incomplete instances are encountered during the prediction phase;

(iii) Combined scenarios: where both partial observations of class variables and incomplete feature data occur simultaneously.

Our framework is designed to handle all these cases, providing robust predictions despite missing discrete features or class variables.

To simplify the notations, we will not explicitly mention the query $\boldsymbol{z}$ in either subscripts or superscripts, but we emphasize that all subsequent notions are defined in an instance-wise manner, *i.e.*, they are tailored specifically to the query $\boldsymbol{z}$. We denote $\mathbf{Y}^{\mathcal{M}}$ as the set of missing class variables, $\mathbf{Y}^{\mathcal{O}}$ as the set of observed class variables, $\mathbf{X}^{\mathcal{M}}$ as the set of missing discrete features, and $\mathbf{X}^{\mathcal{O}}$ as the set of observed discrete feature variables ($\mathbf{X}^{\mathcal{D}} = \mathbf{X}^{\mathcal{M}} \cup \mathbf{X}^{\mathcal{O}}$). We denote $\mathbf{Z}^{\mathcal{M}} = \mathbf{Y}^{\mathcal{M}} \cup \mathbf{X}^{\mathcal{M}}$ as the set of all missing variables, and $\mathbf{Z}^{\mathcal{O}} = \mathbf{Y}^{\mathcal{O}} \cup \mathbf{X}^{\mathcal{O}}$ as the set of all observed discrete variables ($\mathbf{Z}^{\mathcal{M}} \cup \mathbf{Z}^{\mathcal{O}} = \mathbf{Y} \cup \mathbf{X}^{\mathcal{D}}$).

To address this challenging setting, we assume that data are missing at random (MAR). While the realism of MAR depends on the application, it is widely accepted and general. It eliminates the need to define a specific missingness mechanism (whose realism would have the same limitation) and enables experimental evaluation. In any case, mastering this type of missingness is necessary before considering more complex scenarios. Under this assumption, Bayesian optimal predictions can be formulated using different decision-making principles (Cooper and Herskovits, 1992; Guillaume et al., 2017; Hüllermeier et al., 2019; Koller and Friedman, 2009): the optimistic, pessimistic, and averaging principles.

Let $u : \mathcal{Y}^{\mathcal{M}} \times \mathcal{Y}^{\mathcal{M}} \longrightarrow \mathbb{R}_+$ be any utility quantifying the reward for predicting $\hat{\boldsymbol{y}}^{\mathcal{M}}$ when the ground-truth is $\boldsymbol{y}^{\mathcal{M}}$. We also define the shorthand notations $\boldsymbol{z}^{\mathcal{M}} = \boldsymbol{y}^{\mathcal{M}} \cup \boldsymbol{x}^{\mathcal{M}}$ and $\boldsymbol{z}^{\mathcal{O}} = \boldsymbol{y}^{\mathcal{O}} \cup \boldsymbol{x}^{\mathcal{O}}$. Let

$$\mathbf{E}_{\boldsymbol{p}_{(G,\boldsymbol{\theta})}}\left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}}\right] := \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}} \mid \boldsymbol{x}^{\mathcal{C}})$$

be the expected utility. The optimistic, pessimistic, and averaging principles can be implemented as follows, respectively:

$$\hat{\boldsymbol{y}}_u^{\text{opt}} \in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G,\boldsymbol{\theta})}}\left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}}\right] \ , \quad (10)$$

$$\hat{\boldsymbol{y}}_u^{\text{pes}} \in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \min_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G,\boldsymbol{\theta})}}\left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}}\right] \ , \quad (11)$$

$$\hat{\boldsymbol{y}}_u^{\text{ave}} \in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G,\boldsymbol{\theta})}}\left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}}\right] \ . \quad (12)$$

In this work, we do not focus on the pessimistic principle, as finding its optimal solution is notably challenging. Instead, we present algorithmic solutions for deriving optimal predictions under the optimistic and averaging principles. Our discussion specifically addresses their application in conjunction with commonly used evaluation metrics: subset 0/1 accuracy (13) and Hamming accuracy (14):

$$u_{0/1}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) := [\![\hat{\boldsymbol{y}}^{\mathcal{M}} = \boldsymbol{y}^{\mathcal{M}}]\!] \ , \quad (13)$$

$$u_H(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) := \frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{Y \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y} = y]\!] \ . \quad (14)$$

3.3 Algorithmic Solutions

Let $G^{\mathcal{D}}$ be the sub-DAG of G^* over $\mathbf{Z}^{\mathcal{D}}$, which is the same for any query $\boldsymbol{z}$. For each $Z^j \in \mathbf{Z}^{\mathcal{D}}$, let $\text{pa}^*_{\mathcal{D}}(Z^j)$ be the parent set of Z^j within G^* restricted to $\mathbf{Z}^{\mathcal{D}}$. Let $\mathcal{P}a^*_{\mathcal{D}}(Z^j)$ be the set of possible configurations of $\text{pa}^*_{\mathcal{D}}(Z^j)$. For any query $\boldsymbol{z}$, we can construct a discrete BN $\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}$ over $\mathbf{Z}^{\mathcal{D}}$, where for each $Z^j \in \mathbf{Z}^{\mathcal{D}}$, the conditional probability table (CPT) $\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(Z^j \,|\, \text{pa}^*_{\mathcal{D}}(Z^j))$ consisting of $|\mathcal{P}a^*_{\mathcal{D}}(Z^j)|$ rows, one per $\text{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a^*_{\mathcal{D}}(Z^j)$:

$$\begin{aligned}\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(Z^j \,|\, \text{pa}_{\mathcal{D}}(z^j)) &:= \boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}(Z^j \,|\, \text{pa}_{\mathcal{D}}(z^j), \boldsymbol{x}^{\mathcal{C}}) \\ &= \mathbf{h}_{\text{pa}_{\mathcal{D}}(z^j)}(\boldsymbol{x}^{\mathcal{C}}) \,,\end{aligned}$$

where $\mathbf{h}_{\text{pa}_{\mathcal{D}}(z^j)}$ is the optimal local classifier associated to the configuration $\text{pa}_{\mathcal{D}}(z^j)$ defined in (8).

The following propositions ensure that optimal predictions under the averaging principle, when coupled with the commonly used 0/1 accuracy (13) and Hamming accuracy (14), can be found by performing MAP queries on $\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}$. Similarly, optimal predictions under the optimistic principle, when coupled with the commonly used 0/1 accuracy (13) and Hamming accuracy (14), can be respectively estimated and computed by performing MAP queries on $\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}$ learned during the training phase (Gil-Begue et al., 2021; Koller and Friedman, 2009; Nguyen et al., 2023).

Proposition 6. *Let u be the 0/1 accuracy $u_{0/1}$ (13). For any $\boldsymbol{z} \in \mathcal{Z}$, and for any HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, we have*

$$\hat{\boldsymbol{y}}_u^{\text{opt}} \cup \hat{\boldsymbol{x}}^{\mathcal{M}} \in \operatorname*{argmax}_{\boldsymbol{z}^{\mathcal{M}} \in \mathcal{Z}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{M}} \,|\, \boldsymbol{z}^{\mathcal{O}}) \ , \quad (15)$$

$$\hat{\boldsymbol{y}}_u^{\text{ave}} \in \operatorname*{argmax}_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{y}^{\mathcal{M}} \,|\, \boldsymbol{z}^{\mathcal{O}}) \ . \quad (16)$$

Proposition 7. *Let u be the Hamming accuracy u_H (14). For any $\boldsymbol{z} \in \mathcal{Z}$, for any HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, $\hat{\boldsymbol{y}}_u^{\text{opt}}$ (10) can be estimated by finding, $\forall\, Y^k \in \mathbf{Y}^{\mathcal{M}}$,*

$$\hat{y}_u^{\text{opt},k} \cup \hat{\boldsymbol{x}}^{\mathcal{M}} \in \operatorname*{argmax}_{y^k \cup \boldsymbol{x}^{\mathcal{M}} \in \mathcal{Y}^k \times \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(y^k, \boldsymbol{x}^{\mathcal{M}} \,|\, \boldsymbol{z}^{\mathcal{O}}) \ ,$$

and $\hat{\boldsymbol{y}}_u^{\text{ave}}$ (12) can be obtained by finding, $\forall\, Y^k \in \mathbf{Y}^{\mathcal{M}}$,

$$\hat{y}_u^{\text{ave},k} \in \operatorname*{argmax}_{y^k \in \mathcal{Y}^k} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(y^k \,|\, \boldsymbol{z}^{\mathcal{O}}) \ . \quad (17)$$

From now on, we refer to the optimistic and averaging approaches as HMDC-OP and HMDC-AV, respectively. Given that HMDC-OP can be costly when the total number of missing discrete variables and their cardinalities increase, we also propose HMDC-MI, which is an approximation of HMDC-OP. HMDC-MI first imputes the missing discrete features using the most probable levels from their observed distributions in the training data, and then uses these imputed values, together with the observed discrete features and class variables, as evidence in MAP queries to predict the missing class variables.

4 Experiments

This section presents the experimental setting and results.

4.1 Experimental Settings

We experiment on three mixed-feature MDC tabular datasets from the literature (Jia and Zhang, 2024): Adult, Default, and Thyroid (see Table 1).

On the Thyroid dataset, we group the 22 discrete features into 8 (meta) discrete features, each with a cardinality of 8, except for the last one that has a cardinality of 6. We also exclude the optimistic approach from the comparison on this dataset, since performing MAP queries becomes computationally infeasible (given our computational resources), due to the multiple discrete variables with relatively large cardinalities.

Table 1: Summary of dataset characteristics: number of instances N, number of class variables K, cardinalities of class variables $\{|\mathcal{Y}^k|\}_{k=1}^K$, number of discrete features $|\mathbf{X}^{\mathcal{D}}|$, number of continuous features $|\mathbf{X}^{\mathcal{C}}|$

Dataset	N	K	$\{\lvert\mathcal{Y}^k\rvert\}_{k=1}^K$	$\lvert\mathbf{X}^{\mathcal{D}}\rvert$	$\lvert\mathbf{X}^{\mathcal{C}}\rvert$
Adult	18,419	4	[7, 7, 5, 2]	5	5
Default	28,779	4	[2, 7, 4, 2]	6	14
Thyroid	9,172	7	[5, 5, 3, 2, 4, 4, 3]	22	7

We follow a 10-fold cross-validation procedure. For each train-test split, the training data are used to train an HMDC using Algorithm 1, while the discrete features and class variables of the test data are made incomplete by introducing missing values under the MAR assumption. We evaluate performance across varying levels of missingness: 30% and 80% for discrete features, and 30%, 70%, 80%, and 90% for class variables, resulting in 8 distinct missingness scenarios. To assess whether our modeling of conditional dependencies alleviates class imbalance in MDC, we also adapt the balanced accuracy measure introduced for multiclass classification (Gösgens et al., 2021):

$$\mathrm{Rec}(y_m^k) := \frac{1}{|\mathcal{T}(y_m^k)|} \sum_{\mathcal{T}(y_m^k)} [\![\hat{y}^k = y_m^k]\!] \ ,$$
$$\mathrm{BA}(Y^k) := \frac{1}{|\mathcal{Y}^k|} \sum_{y_m^k \in \mathcal{Y}^k} \mathrm{Rec}(y_m^k) \ , \tag{18}$$

where, for each possible outcome $y_m^k \in \mathcal{Y}^k$, the set of test instances where Y^k is missing and the correct value y_m^k is denoted by $\mathcal{T}(y_m^k)$. For each train-test split, we compute the balanced accuracies, the average 0/1 accuracy (13) and the average Hamming accuracy (14) on the test set. We then calculate the average and standard deviation of these metrics across the 10 folds.

We conduct two experiment sets where the number of possible discrete parents is limited to at most two and three, $\forall Z^j \in \mathbf{Z}^{\mathcal{D}}$. The source code used in our experiments has been made public at `https://github.com/telagment/HMDC_tabular`.

4.2 Base Classifiers and Baselines

We implement HMDCs with three base classifiers, used to estimate the local conditional distributions $\boldsymbol{p}_{(G,\boldsymbol{\theta})}(Z^j \,|\, \mathrm{pa}_{\mathcal{D}}(Z^j) = \mathrm{pa}_{\mathcal{D}}(z^j), \mathbf{X}^{\mathcal{C}})$: *Logistic Regression* (LR), *Naive Bayes* (NB), and *Random Forest* (RF). We compare the performance of the HMDC-OP, HMDC-MI, and HMDC-AV approaches, which are defined in Section 3.3, against three baselines: The first baseline is a Random classifier (RandC) that randomly predicts an outcome vector drawn uniformly from the set of possible outcomes. The second baseline is Mode imputation (MI) that predicts the most frequently observed outcome in $\mathcal{S}$ for each missing class variable. We also include a Reference Classifier (RefC), similar to HMDC-MI but with missing discrete features replaced by their true values. RefC, having access to ground truth, is expected to provide an upper bound on performance for approaches dealing with missing features.

As shown in Appendix D, the correlations between discrete variables of the 3 datasets are often high. However, as discussed in Section 3.1, it may be the case that the conditional dependencies may be low and, in such cases, the learning algorithm should return reasonably sparse DAGs. To have an idea of the strength of the conditional dependencies, we include another baseline, namely Binary Relevance (BR), which forces the empty DAG structure. BR solves classification tasks on the class variables independently, given the continuous features. It is known that, despite its simplicity, BR can be competitive (optimal) on datasets with weak (no) conditional dependencies.

4.3 Results

For conciseness and readability, we summarize in Figure 2 the results with at most two discrete parents, focusing on LR classifiers, which perform best due to their probabilistic nature. Comprehensive results, including optimal DAGs $G^{\mathcal{D}}$, standard deviations, and results with RF and NB, are provided in Appendix E. Overall, standard deviations are small relative to the performance gaps between methods. RF and NB yield lower scores but follow similar trends across missingness rates and datasets. Results for experiments allowing up to three discrete parents show comparable patterns and are deferred to Appendix E.

As expected, RefC, which has access to the missing discrete features, achieves the best performance across most settings (but for the Default dataset with the RF base classifier), and its results are invariant to rates of missing features. A similar invariance holds for RandC. Since its results are notably worse than the MI approach, they are deferred to the appendix, and MI serves as a more informative baseline for the expected lower bound of predictive performance.

We observe consistent trends in the predictive performance of HMDC-based methods compared to baselines. Our proposed robust inference methods, HMDC-OP and HMDC-AV, consistently outperform both the MI and HMDC-MI. The predictive performances of HMDC-OP and HMDC-AV are much closer to RefC (which has access to ground-truth values of missing features) than to MI, occasionally surpassing

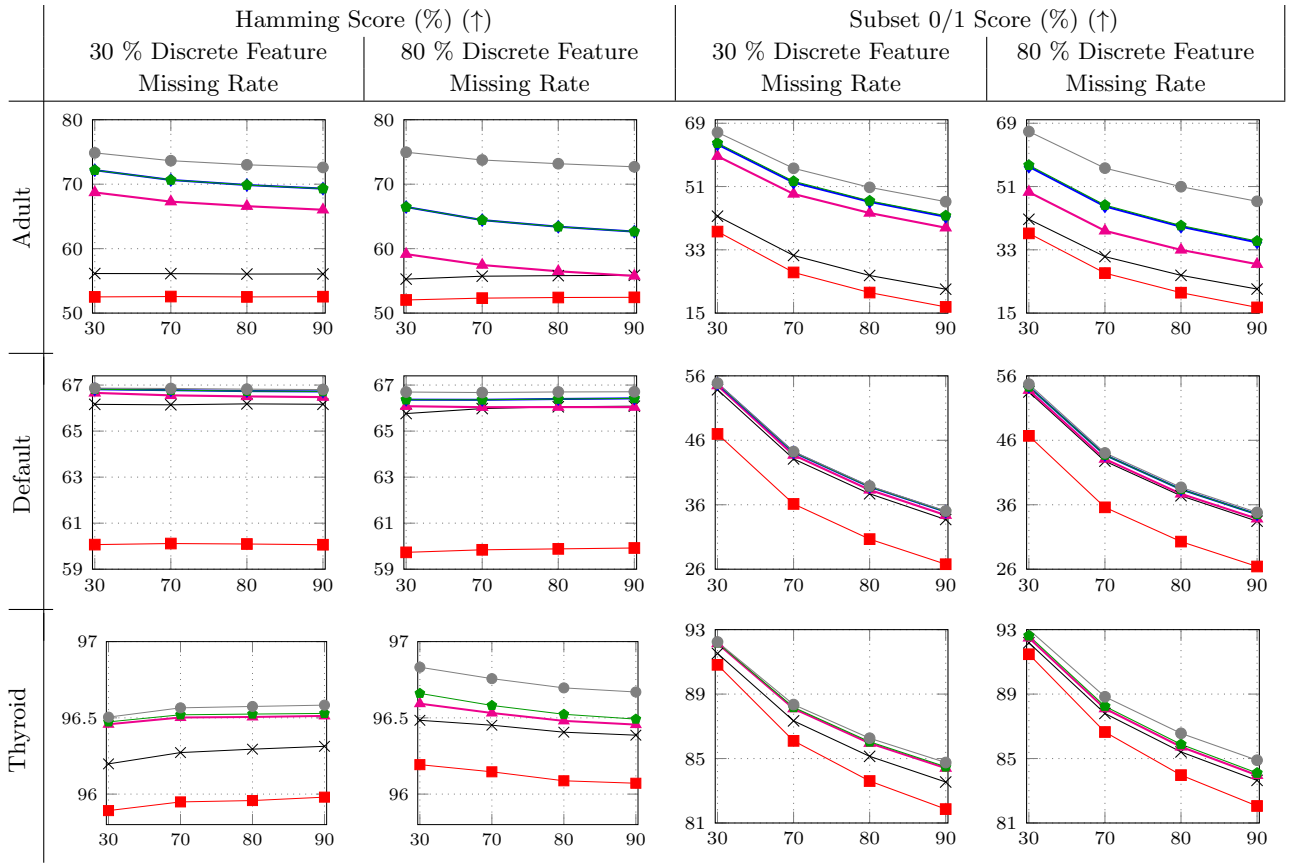


Figure 2: Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds, plotted against the percentage of missing class variables (in %, x-axis) for the three datasets. The missing rate for discrete features is either 30 % or 80 %; logistic regression is used as the base classifier. Probabilistic models are represented as follows: ■ MI, × BR, ▲ HMDC-MI, ♦ HMDC-OP, ⬟ HMDC-AV and ● RefC.

RefC. Overall, HMDC-OP and HMDC-AV have indistinguishably strong predictive scores, but HMDC-AV is computationally cheaper to use.

The results indicate that the choice of base classifier has a substantial impact on predictive performance. Overall, LR yields the most promising results, followed by RF, while NB performs the worst. This disparity may be attributed to the use of explicit and implicit regularization in LR and RF, respectively, which promotes stable training of local discriminative classifiers (Nguyen et al., 2018), favoring the identification of an optimal HMDC, as discussed in Proposition 5. In contrast, the poor performance of NB may stem from a mismatch between its training objective and the evaluation criteria used in our framework.

On the Thyroid dataset, MI is a strong baseline, achieving a Hamming score of approximately 96%. Nevertheless, HMDCs still perform better, close to RefC, which achieves a Hamming score above 98%. The DAGs learned using RFs (shown in the appendix) are extremely sparse, which likely explains why the Hamming scores (computed only on the class variables that are missing) remain roughly constant across varying levels of missingness in both the class variables (x-axis) and discrete features (30% and 80% missingness settings). Since the discrete features were grouped into meta-features for this dataset, these results suggest that coupling HMDCs with the grouping of variables in DAGs, as proposed in (Parviainen and Kaski, 2017), could be a promising direction to manage computational complexity in both training and inference.

Regarding the effect of varying the proportion of missing data, the results show that the Hamming score of HMDC decreases moderately as the proportion of missing values in discrete features increases, but changes only slightly when the proportion of missing class values changes. In contrast, the Subset 0/1 decreases significantly as the proportion of missing class variable values increases, while exhibiting only minor changes when the missing rate of discrete features increases.

The graphs reported in Appendix E.4.1 show that LR

provides reasonably sparse DAGs. Moreover, the results given in Appendix E.4.2, where up to three parents are allowed among discrete variables, further highlight the stability of the learning phase. This could benefit the trade-off between model complexity and predictive performance. Unfortunately, we are unable to extend the analysis on this desirable property of our proposal due to the shortage of benchmark datasets with diverse levels of (conditional) dependencies.

Regarding balanced accuracy (18), results in Figure 3 and Appendix E.1 show that our HMDC-based line models consistently outperform BR and other baselines on Adult and Thyroid, and remain competitive on Default (with very sparse DAGs). This gain likely stems from the ability to take into account the dependencies among the discrete variables when modeling the joint conditional probability distribution. By capturing these relationships, the model can better identify plausible configurations of the class variables, thereby improving the prediction of minority outcomes.

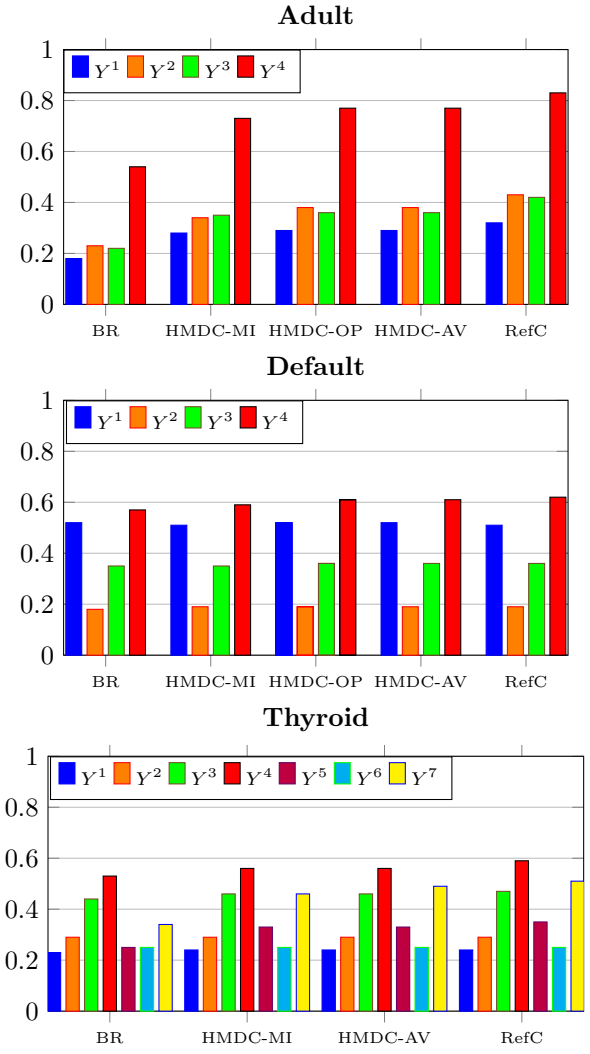


Figure 3: Balanced accuracies (↑) with LR as base learner; Adult, Default, Thyroid datasets with 90% missing class variables and 30% missing features.

Overall, class variables with lower cardinality tend to yield higher balanced accuracy due to their easier predictability. This pattern can be observed for Y^4 in Adult; Y^1, Y^4 in Default; and Y^3, Y^4, Y^7 in Thyroid, in both Figure 3 and Appendix E.1. Binary Relevance is not affected by changes in missing values due to the independence assumption. The per-class balanced accuracy of the HMDC methods decreases as the proportion of missing discrete feature values increases, while remaining relatively stable when the proportion of missing class values changes. In particular, HMDC-MI exhibits the greatest sensitivity to variations in the missingness rate of discrete features, likely because the method infers outcomes from unreliable or misleading evidence.

5 Conclusion

We presented Hybrid MDCs, a model family designed for the challenging MDC setting with mixed features, where both discrete features and class variables may be missing at prediction time. We detailed their learning and prediction procedures, supported by theoretical results that guide the design of efficient algorithms. Our experiments demonstrate the robustness of Hybrid MDCs in different scenarios, across varying levels of missingness in discrete features and class variables. As a follow-up work, we plan to extend the approach to also tackle incomplete data during the training phase.

Acknowledgements

This work was partially supported by the Junior Professor Chair in Trustworthy AI (Ref.ANR-R311CHD) and the French National Research Agency (ANR) under the France 2030 program, grant reference ANR-23-IACL-0007.

References

Mark Bartlett and James Cussens. Integer linear programming for the Bayesian network structure learning problem. *Artificial Intelligence*, 244:258–271, 2017.

Marco Benjumeda, Concha Bielza, and Pedro Larrañaga. Tractability of most probable explanations in multidimensional Bayesian network classifiers. *International Journal of Approximate Reasoning*, 93:74–87, 2018.

Concha Bielza, Guangdi Li, and Pedro Larranaga. Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011.

David Maxwell Chickering. Learning Bayesian networks is NP-complete. *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130, 1996.

Gregory F Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405, 1990.

Gregory F Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9:309–347, 1992.

James Cussens. GOBNILP: Learning Bayesian network structure with integer programming. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models (PGM)*, volume 138 of *Proceedings of Machine Learning Research*, pages 605–608. PMLR, 2020.

Anant Dadu, Vipul Satone, Rachneet Kaur, Sayed Hadi Hashemi, Hampton Leonard, Hirotaka Iwaki, Mary B Makarious, Kimberley J Billingsley, Sara Bandres-Ciga, Lana J Sargent, Alastair J Noyce, Ali Daneshmand, Cornelis Blauwendraat, Ken Marek, Sonja W Scholz, Andrew B Singleton, Mike A Nalls, Roy H Campbell, and Faraz Faghri. Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinson's Disease*, 8(1):172, 2022.

Cassio P de Campos, Mauro Scanagatta, Giorgio Corani, and Marco Zaffalon. Entropy-based pruning for learning Bayesian networks using BIC. *Artificial Intelligence*, 260:42–50, 2018.

Lennox Din, Mohammad Sheikh, Nikitha Kosaraju, Karin Ekstrom Smedby, Sasha Bernatsky, Sonja I Berndt, Christine F Skibola, Alexandra Nieters, Sophia Wang, James D McKay, et al. Genetic overlap between autoimmune diseases and non-Hodgkin lymphoma subtypes. *Genetic epidemiology*, 43(7): 844–863, 2019.

Van Hoan Do, Son Hoang Nguyen, Duc Quang Le, Tam Thi Nguyen, Canh Hao Nguyen, Tho Huu Ho, Nam S Vo, Trang Nguyen, Hoang Anh Nguyen, Minh Duc Cao, et al. Panka: Leveraging population pangenome to predict antibiotic resistance. *iScience*, 2024.

Daniel Ferreira, Agneta Nordberg, and Eric Westman. Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology*, 94(10): 436–448, 2020.

Judith Garcia-Aymerich, Federico P Gómez, Marta Benet, Eva Farrero, Xavier Basagana, Angel Gayete, Carles Paré, Xavier Freixa, Jaume Ferrer, Antoni Ferrer, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax*, 66(5): 430–437, 2011.

Santiago Gil-Begue, Concha Bielza, and Pedro Larrañaga. Multi-dimensional Bayesian network classifiers: A survey. *Artificial Intelligence Review*, 54(1):519–559, 2021.

Martijn Gösgens, Anton Zhiyanov, Alexey Tikhonov, and Liudmila Prokhorenkova. Good classification measures and how to find them. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, pages 17136–17147, 2021.

Romain Guillaume, Inés Couso, and Didier Dubois. Maximum likelihood with coarse data based on robust optimisation. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, volume 62, pages 169–180. PMLR, 2017.

Yujin Hoshida, Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Subclass mapping: Identifying common subtypes in independent disease data sets. *PloS One*, 2(11):e1195, 2007.

Eyke Hüllermeier, Sébastien Destercke, and Ines Couso. Learning from imprecise data: Adjustments of optimistic and pessimistic variants. In *Proceedings of the 13th International Conference on Scalable Uncertainty Management (SUM)*, volume 11940, pages 266–279. Springer, 2019.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification: Paradigm, algorithms and beyond. *Vicinagearth*, 1(1):3, 2024.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

David A Lynch, John HM Austin, James C Hogg, Philippe A Grenier, Hans-Ulrich Kauczor, Alexander A Bankier, R Graham Barr, Thomas V Colby, Jeffrey R Galvin, Pierre Alain Gevenois, et al. CT-definable subtypes of chronic obstructive pulmonary

disease: A statement of the Fleischner society. *Radiology*, 277(1):192–205, 2015.

Zhongchen Ma and Songcan Chen. Multi-dimensional classification via a metric approach. *Neurocomputing*, 275:1121–1131, 2018.

Danesh Moradigaravand, Martin Palm, Anne Farewell, Ville Mustonen, Jonas Warringer, and Leopold Parts. Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data. *PLoS computational biology*, 14(12):e1006258, 2018.

Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press, 2023.

Vu-Linh Nguyen, Sébastien Destercke, Marie-Hélène Masson, and Eyke Hüllermeier. Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5089–5095, 2018.

Vu-Linh Nguyen, Yang Yang, and Cassio P. de Campos. Probabilistic multi-dimensional classification. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *Proceedings of Machine Learning Research*, pages 1522–1533. PMLR, 2023.

Pekka Parviainen and Samuel Kaski. Learning structures of Bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88: 110–127, 2017.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.

Teemu Roos, Hannes Wettig, Peter Grünwald, Petri Myllymäki, and Henry Tirri. On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, 59(3):267–296, 2005.

Luis Enrique Sucar. *Probabilistic graphical models*. Advances in Computer Vision and Pattern Recognition. Springer, 2021.

Linda C Van Der Gaag and Peter R De Waal. Multi-dimensional Bayesian network classifiers. In *Proceedings of the third European workshop on Probabilistic Graphical Models (PGM)*, pages 107–114. Prague, 2006.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Notations and Acronyms

Symbol	Description
X	Feature (random) variable
Y	Class (random) variable
$\mathbf{X}$	Feature vector $\mathbf{X} = \{X^1, \ldots, X^Q\}$
$\mathbf{Y}$	Class vector $\mathbf{Y} = \{Y^1, \ldots, Y^K\}$
$\mathbf{X}^{\mathcal{D}}$	Set of discrete feature variables
$\mathbf{X}^{\mathcal{C}}$	Set of continuous feature variables
$\mathbf{Z}$	Union of disjoint sets $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$
$\mathbf{Z}^{\mathcal{D}}$	Union of the disjoint sets $\mathbf{X}^{\mathcal{D}}$ and $\mathbf{Y}$, $\mathbf{Z}^{\mathcal{D}} = \mathbf{X}^{\mathcal{D}} \cup \mathbf{Y}$
Q	Number of feature variables
K	Number of class variables
z	Particular value of variable $Z = z$
$\boldsymbol{x}$	Particular set value of feature variables, $\mathbf{X} = \boldsymbol{x}$
$\boldsymbol{y}$	Particular set value of class variables, $\mathbf{Y} = \boldsymbol{y}$
$\boldsymbol{z}$	Particular set value of variables $\mathbf{Z} = \boldsymbol{z}$
$\boldsymbol{z}_n$	Values of the joint variables $\mathbf{Z}$ corresponding with training instance n
$\boldsymbol{x}_n^{\mathcal{D}}$	Discrete features of $\boldsymbol{x}_n$
$\boldsymbol{x}_n^{\mathcal{C}}$	Continuous features of $\boldsymbol{x}_n$
$\mathcal{M}$	Superscript indicating missing term, *e.g.*, $\mathbf{X}^{\mathcal{M}}$, $\mathbf{Y}^{\mathcal{M}}$
$\mathcal{O}$	Superscript indicating observed term, *e.g.*, $\mathbf{X}^{\mathcal{O}}$, $\mathbf{Y}^{\mathcal{O}}$
n	Training data index
N	Number of training data instances
$\mathcal{S}$	Training set $\{\boldsymbol{z}_n \,\vert\, n = 1 \ldots, N\} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) \,\vert\, n = 1 \ldots, N\}$
$\mathcal{T}$	Test set
G	Directed Acyclic Graph (DAG)
$\boldsymbol{\theta}$	A parameter set
$\mathcal{H}$	Hypothesis space consists of possible pair $(G, \boldsymbol{\theta})$
G^*	Optimal DAG structure
$\boldsymbol{\theta}^*$	Optimal parameter set
$\mathrm{pa}(Z)$	Parent set of Z in the DAG G
$\mathrm{pa}(z)$	Specific configuration of $\mathrm{pa}(Z)$ specified by $\boldsymbol{z}$
$\boldsymbol{p}_{(G,\boldsymbol{\theta})}$	Admissible distribution
$\Omega(G, \mathcal{S})$	Decomposable regularization term
$\mathbf{h}_{\mathrm{pa}(Z^j)}$	Probabilistic classifier (6)
$\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$	Probabilistic classifier (8)
$\mathrm{pa}_{\mathcal{D}}(Z^j)$	Set of the discrete variables among the parents of Z
$\mathcal{P}a(Z^j) \subset 2^{\mathbf{Z} \setminus \{Z^j\}}$	Set of possible parent sets of Z^j

Acronym	Description
MDC	Multi-Dimensional Classification
PMDC	Probabilistic Multi-Dimensional Classification
BN	Bayesian Network
DAG	Directed Acyclic Graph
BOPs	Bayes-Optimal Predictions
MAP	Maximum A Posteriori
MAR	Missing At Random
GMDC	Generative Multi-Dimensional Classification
DMDC	Discriminative Multi-dimensional Classification
HMDC	Hybrid Multi-Dimensional Classification
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
LR	Logistic Regression
NB	Naive Bayes
RF	Random Forest

B Proofs for Propositions of Section 3.1 (Learning Phase)

Proposition 1. *For any $\boldsymbol{z} \in \mathcal{Z}$, the structure constraints of HMDCs ensure that, for any HMDC $\boldsymbol{p}_{(G,\boldsymbol{\theta})}$, we have*

$$\begin{aligned}\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}, \boldsymbol{x}^{\mathcal{D}} \,|\, \boldsymbol{x}^{\mathcal{C}}) &= \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j)) \\ &:= \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(Z^j = z^j \,|\, \mathrm{pa}(Z^j) = \mathrm{pa}(z^j)) \ , \ .\end{aligned}$$

Proof. The structure constraints of HMDCs ensure that, for any $X^i \in \mathbf{X}^{\mathcal{C}}$, $\mathrm{pa}(X^i) \subset \mathbf{X}^{\mathcal{C}}$. Therefore, we have

$$\begin{aligned}\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}, \boldsymbol{x}^{\mathcal{D}} \,|\, \boldsymbol{x}^{\mathcal{C}}) &= \frac{\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}, \boldsymbol{x}^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}{\sum_{\bar{\boldsymbol{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}^{\mathcal{D}}}, \boldsymbol{x}^{\mathcal{C}})} \\ &= \frac{\prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j)) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(x^i \,|\, \mathrm{pa}(x^i))}{\sum_{\bar{\boldsymbol{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \left(\prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\bar{z}^j \,|\, \mathrm{pa}(\bar{z}^j)) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(x^i \,|\, \mathrm{pa}(x^i)) \right)} \\ &= \frac{\left(\prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j)) \right) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(x^i \,|\, \mathrm{pa}(x^i))}{\left(\sum_{\bar{\boldsymbol{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\bar{z}^j \,|\, \mathrm{pa}(\bar{z}^j)) \right) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(x^i \,|\, \mathrm{pa}(x^i))} \\ &= \frac{\prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j))}{\sum_{\bar{\boldsymbol{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\bar{z}^j \,|\, \mathrm{pa}(\bar{z}^j))} \\ &= \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j))\end{aligned} \tag{19}$$

because by the definition of Bayesian Networks, we have

$$\sum_{\bar{\boldsymbol{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\bar{z}^j \,|\, \mathrm{pa}(\bar{z}^j)) = \sum_{\bar{\boldsymbol{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(\bar{\boldsymbol{z}}^{\mathcal{D}} \,|\, \boldsymbol{x}^{\mathcal{C}}) = 1 \,. \tag{20}$$

□

Proposition 2. *Given any finite training set $\mathcal{S}$, the structure constraints of HMDCs ensure that, for any HMDC $(G, \boldsymbol{\theta})$, we have the following decomposition:*

$$L_\Omega(G, \boldsymbol{\theta} \,|\, \mathcal{S}) = \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \ ,$$

where

$$L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \coloneqq \ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) - \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S})$$
$$\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \coloneqq \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\Big) \ ,$$

which only depends on the local structure and parameters $(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j)$ *defining the local model at each node.*

Proof. As a consequence of Proposition 1, we have

$$\begin{aligned}
L_\Omega((G,\boldsymbol{\theta}) \,|\, \mathcal{S}) &= \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}_n, \boldsymbol{x}_n^{\mathcal{D}} \,|\, \boldsymbol{x}_n^{\mathcal{C}})\Big) - \Omega(G, \mathcal{S}) \\
&= \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\left(\prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\right) - \Omega(G, \mathcal{S}) \\
&= \sum_{\boldsymbol{z}_n \in \mathcal{S}} \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\Big) - \Omega(G, \mathcal{S}) \\
&= \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\Big) - \Omega(G, \mathcal{S}) \\
&= \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\Big) - \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S}) \\
&= \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \left(\sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\Big) - \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S})\right) \\
&= \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \left(\underbrace{\sum_{\boldsymbol{z}_n \in \mathcal{S}} \log\Big(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(z_n^j \,|\, \mathrm{pa}(z_n^j))\Big)}_{=:\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S})} - \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S})\right) \\
&= \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \left(\underbrace{\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S}) - \omega(Z^j, \mathrm{pa}(Z^j), \mathcal{S})}_{=:L_\Omega(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S})}\right) .
\end{aligned} \tag{21}$$

□

Proposition 3. *For any* $Z^j \in \mathbf{Z}^{\mathcal{D}}$*, let* $2^{\mathbf{Z}\setminus\{Z^j\}}$ *be the set of all the subsets of* $\mathbf{Z} \setminus \{Z^j\}$*. Let* $\mathcal{P}a(Z^j) \subset 2^{\mathbf{Z}\setminus\{Z^j\}}$ *be any set of possible parent sets of* Z^j*. Given a training set* $\mathcal{S}$*, the task of learning an optimal HMDC, specified by the pair* $(G^*, \boldsymbol{\theta}^*)$ *that maximizes the regularized conditional log-likelihood in* (3)*, can be decomposed into a set of local learning problems and a Bayesian network structure learning problem.*

Specifically, it reduces to training a collection of local probabilistic multi-class classifiers. Let $\Delta^{|\mathcal{Z}^j|}$ *be the probability simplex over* $|\mathcal{Z}^j|$ *outcomes,each classifier*

$$\mathbf{h}_{\mathrm{pa}(Z^j)} : \bigtimes_{Z^k \in \mathrm{pa}(Z^j)} \mathcal{Z}^k \longrightarrow \Delta^{|\mathcal{Z}^j|} \ ,$$

estimates the conditional distribution of a variable $Z^j \in \mathbf{Z}^{\mathcal{D}}$ *given each valid parent set* $\mathrm{pa}(Z^j) \in \mathcal{P}a(Z^j)$ *and is specified by an optimal parameter set*

$$\boldsymbol{\theta}_j^* \in \operatorname*{argmax}_{\boldsymbol{\theta}_j \in \Theta_j} \ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \ ,$$

with $\ell(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S})$ *defined in* (5)*. Each possible parent set* $\mathrm{pa}(Z^j)$ *is then associated with its optimal score* $L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^* \,|\, \mathcal{S})$ (4)*. Finally, any scoring-based Bayesian network structure learning technique can be employed*

to find an optimal DAG G^* over $\mathbf{Z}^{\mathcal{D}}$ specified by $|\mathbf{Z}^{\mathcal{D}}|$ parent sets, one per $Z^j \in \mathbf{Z}^{\mathcal{D}}$, denoted by $\mathrm{pa}^*(Z^j)$. Together with their optimal parameter sets $\{\boldsymbol{\theta}_j^* \,|\, Z^j \in \mathbf{Z}^{\mathcal{D}}\} := \boldsymbol{\theta}^*$, we have an optimal HMDC $(G^*, \boldsymbol{\theta}^*)$.

Proof. For any $Z^j \in \mathbf{Z}^{\mathcal{D}}$, let $\mathcal{P}a(Z^j) \subset 2^{\mathbf{Z}\setminus\{Z^j\}}$ be any set of possible parent sets of Z^j. Each possible DAG G should be specified by a set of parent sets

$$\left\{\mathrm{pa}(Z^j) | Z^j \in \mathbf{Z}^{\mathcal{D}}\right\} \in \bigtimes_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \mathcal{P}a(Z^j)\,, \tag{22}$$

which forms a DAG. For each possible DAG G, there are (possibly infinite) parameter sets $\boldsymbol{\theta} := \left\{\boldsymbol{\theta}_j | Z^j \in \mathbf{Z}^{\mathcal{D}}\right\}$, which can be employed to estimate

$$\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\boldsymbol{y}, \boldsymbol{x}^{\mathcal{D}} \,|\, \boldsymbol{x}^{\mathcal{C}}) = \prod_{Z^j \in \mathbf{Z}^{\mathcal{D}}} \boldsymbol{p}_{(G,\boldsymbol{\theta})}(z^j \,|\, \mathrm{pa}(z^j)) \ . \tag{23}$$

As a consequence of proposition 2, for each fixed possible DAG G, an optimal parameter set $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}_j^* \,|\, Z^j \in \mathbf{Z}^{\mathcal{D}}\}$, which optimizes the local CL (3), can be found by independently finding, $\forall Z^j \in \mathbf{Z}^{\mathcal{D}}$, optimal parameter set $\boldsymbol{\theta}_j^*$ (7), which specifies the probabilistic multi-class classifier $\mathbf{h}_{\mathrm{pa}(Z^j)}$ (6).

Once the local optimal parameter sets θ_j^* are learned for all the possible pair $(Z^j, \mathrm{pa}(Z^j))$, for any $Z^j \in \mathbf{Z}^{\mathcal{D}}$, each possible parent set $\mathrm{pa}(Z^j)$ is then associated with its optimal score $L_\Omega((\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^*) \,|\, \mathcal{S})$ (4).

Therefore, finding an optimal HMDC reduces to finding an optimal DAG G^* over $\mathbf{Z}^{\mathcal{D}}$ such that

$$G^* \in \operatorname*{argmax}_G \sum_{Z^j \in \mathbf{Z}^{\mathcal{D}}} L_\Omega(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j^*) \,|\, \mathcal{S})\,, \tag{24}$$

which can be done by using any scoring-based Bayesian network structure learning technique.

Once G^* is found, for each $Z^j \in \mathbf{Z}^{\mathcal{D}}$, let $\mathrm{pa}^*(Z^j)$ its parent set in G^*. We can retrieve the corresponding parameter set $\boldsymbol{\theta}_j^*$. Finally, the optimal parameter set $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}_j^* \,|\, Z^j \in \mathbf{Z}^{\mathcal{D}}\}$ of the optimal HMDC $\boldsymbol{p}_{(G^*,\boldsymbol{\theta}^*)}$ can be formed. □

Proposition 4. *Let $Z^j \in \mathbf{Z}^{\mathcal{D}}$. For any $\mathrm{pa}'(Z^j) \supset \mathrm{pa}(Z^j)$, if*

$$L_\Omega(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j \,|\, \mathcal{S}) \geq -\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S}) \ ,$$

then $\mathrm{pa}'(Z^j)$ cannot be a parent set of Z^j in any optimal DAG G^*.

Proof. For any $\mathrm{pa}'(Z^j) \supset \mathrm{pa}(Z^j)$, if

$$L_\Omega(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S}) \geq -\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S}) \ , \tag{25}$$

we have

$$\begin{aligned}
\underbrace{L_\Omega(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathrm{pa}(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S})}_{\geq -\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S})} - L_\Omega(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathrm{pa}'(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S}) &\geq -\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S}) - \underbrace{L_\Omega(\boldsymbol{p}_{(G,\boldsymbol{\theta})}(\mathrm{pa}'(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S})}_{= \ell(\mathrm{pa}'(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S}) - \omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S})} \\
&= -\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S}) - \ell(\mathrm{pa}'(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S}) + \omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S}) \\
&= -\ell(\mathrm{pa}'(Z^j), \boldsymbol{\theta}_j) \,|\, \mathcal{S}) \geq 0\,.
\end{aligned}$$

Then $\mathrm{pa}'(Z^j)$ cannot be a parent set of Z_d in any optimal DAG G^*. This is because if assuming contrary to the claim, there must be an optimal DAG G^* which takes $\mathrm{pa}'(Z^j)$ as the parent set of Z^j. If we replace $\mathrm{pa}'(Z^j)$ by $\mathrm{pa}(Z^j)$, we will have another DAG with a smaller complexity whose optimal regularized conditional log-likelihood in (3) is at least the optimal regularized conditional log-likelihood in (3) attained when using DAG G^*. This is a contradiction. □

Proposition 5. *Given any finite training set $\mathcal{S}$, Algorithm 1 returns an optimal HMDC $\boldsymbol{p}_{(G^*,\boldsymbol{\theta}^*)}$ that optimizes the regularized conditional log-likelihood (3) if, for any possible pair $(Z^j, \mathrm{pa}(Z^j))$, and any $\mathrm{Pa}_{\mathcal{D}}(Z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$ (8) that optimizes (9) can be found.*

Proof. For any possible pair $(Z^j, \mathrm{pa}(Z^j))$, the optimal parameter set $\boldsymbol{\theta}^*_j$ (7) of $\mathbf{h}_{\mathrm{pa}(Z^j)=\mathrm{pa}(z^j)}$ (6) is found if, for any possible configuration $\mathrm{pa}_{\mathcal{D}}(z^j) \in \mathcal{P}a_{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$ (8) that optimizes (9) can be found. If the local optimal classifier $\mathbf{h}_{\mathrm{pa}_{\mathcal{D}}(z^j)}$ (8) is learned under the feature selection option, the irrelevant continuous features are discarded from $\mathrm{pa}(Z^j)$. This is equivalent to exhaustively assessing all the possible parent sets $\mathrm{pa}'(Z^j) \subset \mathrm{pa}(Z^j)$, where $\mathrm{pa}'(Z^j) := \mathrm{pa}_{\mathcal{D}}(Z^j) \cup \mathbf{Z}^{\mathcal{C}}$, $\forall\ \mathbf{Z}^{\mathcal{C}} \subset \mathbf{X}^{\mathcal{C}}$ and select the best one. The non-optimal parent sets $\mathrm{pa}'(Z^j)$ can be discarded without discarding any possible optimal DAG since all the $\mathrm{pa}'(Z^j) = \mathrm{pa}_{\mathcal{D}}(Z^j) \cup \mathbf{Z}^{\mathcal{C}}$, $\mathbf{Z}^{\mathcal{C}} \subset \mathbf{X}^{\mathcal{C}}$, have the same regularized score $\omega(Z^j, \mathrm{pa}'(Z^j), \mathcal{S})$ defined independently from $|\mathbf{Z}^{\mathcal{C}}|$.

Applying the pruning rule (detailed in Proposition 4) in the line 10 of Algorithm 1 does not discard any optimal DAG. Therefore, the for-loop in lines 2–12 returns, for each $Z^j \in \mathbf{Z}^{\mathcal{D}}$, the optimal score $L_\Omega((\mathrm{pa}(Z^j), \boldsymbol{\theta}^*_j) \,|\, \mathcal{S})$ (4) of all the possible parent sets $\mathrm{pa}(Z^j) \in \mathcal{P}a(Z^j)$, except those parent sets that can not belong to any optimal DAG, which are pruned in the line 10 of Algorithm 1.

GOBNILP (and any other exact Bayesian Network structure learning algorithms), if it converges, should return an optimal DAG G^* by using $L_\Omega((\mathrm{pa}(Z^j), \boldsymbol{\theta}^*_j) \,|\, \mathcal{S})$ (4) as the input.

Altogether, the HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, where $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}^*_j \,|\, Z^j \in \mathbf{Z}^{\mathcal{D}}\}$ returned by Algorithm 1 should optimize the regularized conditional log-likelihood (3). □

C Proofs for Propositions of Section 3.2 (Prediction Phase)

Proposition 6. *Let u be the 0/1 accuracy $u_{0/1}$ (13). To simplify the notations, for any $\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}$, we shall denote by $\hat{\boldsymbol{z}}^{\mathcal{M}} = \hat{\boldsymbol{y}}^{\mathcal{M}} \cup \boldsymbol{x}^{\mathcal{M}}$. For any $\boldsymbol{z} \in \mathcal{Z}$, and for any HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, we have*

$$\hat{\boldsymbol{y}}^{\mathrm{opt}}_u \cup \hat{\boldsymbol{x}}^{\mathcal{M}} \in \operatorname*{argmax}_{\hat{\boldsymbol{z}}^{\mathcal{M}} \in \mathcal{Z}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{z}}^{\mathcal{M}} \,|\, \boldsymbol{z}^{\mathcal{O}}) \ ,$$
$$\hat{\boldsymbol{y}}^{\mathrm{ave}}_u \in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}} \,|\, \boldsymbol{z}^{\mathcal{O}}) \ .$$

Proof. To simplify the notations, for any $\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}$, we shall denote by $\hat{\boldsymbol{z}}^{\mathcal{M}} = \hat{\boldsymbol{y}}^{\mathcal{M}} \cup \boldsymbol{x}^{\mathcal{M}}$.

We will first proceed with the optimistic principle. For any $\boldsymbol{z} \in \mathcal{Z}$, and for HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, we have

$$\begin{aligned}
\hat{\boldsymbol{y}}^{\mathrm{opt}}_u &\in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}} \left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}} \right] \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} [\![\hat{\boldsymbol{y}}^{\mathcal{M}} = \hat{\boldsymbol{y}}^{\mathcal{M}}]\!] \boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}} \mid \boldsymbol{x}^{\mathcal{C}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} [\![\hat{\boldsymbol{y}}^{\mathcal{M}} = \boldsymbol{y}^{\mathcal{M}}]\!] \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \overbrace{\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \underbrace{\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{O}})}_{\text{A constant}}}^{\text{Using Bayes' Theorem}} \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \ .
\end{aligned} \tag{26}$$

Therefore, $\hat{\boldsymbol{y}}^{\mathrm{opt}}_u$ can be found by finding

$$\hat{\boldsymbol{y}}^{\mathrm{opt}}_u \cup \hat{\boldsymbol{x}}^{\mathcal{M}} \in \operatorname*{argmax}_{\hat{\boldsymbol{z}}^{\mathcal{M}} \in \mathcal{Z}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{z}}^{\mathcal{M}} \,|\, \boldsymbol{x}^{\mathcal{C}}) \ .$$

We now proceed with the averaging principle. For any $\boldsymbol{z} \in \mathcal{Z}$, and for HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, we have

$$\hat{\boldsymbol{y}}^{\mathrm{ave}}_u \in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}} \left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}} \right] \tag{27}$$

$$= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \ .$$
$$= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \ .$$

□

Proposition 7. *Let u be the Hamming accuracy u_H (14). For any $\boldsymbol{z} \in \mathcal{Z}$, for any HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, $\hat{\boldsymbol{y}}_u^{\text{ave}}$ (12) can be found by finding*

$$\hat{y}_u^{\text{ave},k} \in \operatorname*{argmax}_{\hat{y}^k \in \mathcal{Y}^k} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k \mid \boldsymbol{z}^{\mathcal{O}}) \ , \forall Y^k \in \mathbf{Y}^{\mathcal{M}} \ ,$$

and $\hat{\boldsymbol{y}}_u^{\text{opt}}$ (10) can be estimated by finding, $\forall Y^k \in \mathbf{Y}^{\mathcal{M}}$,

$$\hat{y}_u^{\text{opt},k} \cup \hat{\boldsymbol{x}}^{\mathcal{M}} \in \operatorname*{argmax}_{\hat{y}^k \cup \boldsymbol{x}^{\mathcal{M}} \in \mathcal{Y}^k \times \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \ .$$

Proof. To simplify the notations, for any $\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}$, we shall denote by $\hat{\boldsymbol{z}}^{\mathcal{M}} = \hat{\boldsymbol{y}}^{\mathcal{M}} \cup \boldsymbol{x}^{\mathcal{M}}$.

We will first proceed with the optimistic principle. For any $\boldsymbol{z} \in \mathcal{Z}$, and for HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, we have

$$\begin{aligned}
\hat{\boldsymbol{y}}_u^{\text{opt}} &\in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}} \left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}} \right] \qquad (28) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \right) \boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}} \mid \boldsymbol{x}^{\mathcal{C}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \right) \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \right) \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \right) \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \right) \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \right) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \left(\sum_{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} [\![\hat{y}^k = y^k]\!] \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \right) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \left(\sum_{\substack{\boldsymbol{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}} \\ \hat{y}^k = y^k}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{x}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \right) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}}, \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \underbrace{\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{O}})}_{\text{Using Bayes' Theorem}}
\end{aligned}$$

$$
\begin{aligned}
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \left(\sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \right) \underbrace{\boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\boldsymbol{z}^{\mathcal{O}})}_{\text{A constant}} \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \quad .
\end{aligned}
$$

which can be estimated (since for different $Y^k \in \mathbf{Y}^{\mathcal{D}}$, $\hat{y}_u^{\text{opt},k}$ (29) may be attained with different $\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}$) by finding

$$
\hat{y}_u^{\text{opt},k} \in \operatorname*{argmax}_{\hat{y}^k \in \mathcal{Y}^k} \max_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}), Y^k \in \mathbf{Y}^{\mathcal{M}} . \tag{29}
$$

For any $Y^k \in \mathbf{Y}^{\mathcal{M}}$, $\hat{y}_u^{\text{opt},k}$ (29) can be found by finding

$$
\hat{y}_u^{\text{opt},k} \cup \hat{\boldsymbol{x}}^{\mathcal{M}} \in \operatorname*{argmax}_{\hat{y}^k \cup \boldsymbol{x}^{\mathcal{M}} \in \mathcal{Y}^k \times \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \quad . \tag{30}
$$

We now proceed with the averaging principle. For any $\boldsymbol{z} \in \mathcal{Z}$, and for HMDC $\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}$, we have

$$
\begin{aligned}
\hat{\boldsymbol{y}}_u^{\text{ave}} &\in \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\boldsymbol{p}_{(G^*, \boldsymbol{\theta}^*)}} \left[u(\hat{\boldsymbol{y}}^{\mathcal{M}}, \boldsymbol{y}^{\mathcal{M}}) \,|\, \boldsymbol{x}^{\mathcal{C}} \right] \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \sum_{\boldsymbol{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k, \boldsymbol{x}^{\mathcal{M}} \mid \boldsymbol{z}^{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{Y^k \in \mathbf{Y}^{\mathcal{M}}} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k \mid \boldsymbol{z}^{\mathcal{O}}) \quad ,
\end{aligned} \tag{31}
$$

which can be found by finding, for any $Y^k \in \mathbf{Y}^{\mathcal{M}}$,

$$
\hat{y}_u^{\text{ave},k} \in \operatorname*{argmax}_{\hat{y}^k \in \mathcal{Y}^k} \boldsymbol{p}_{(G^{\mathcal{D}}, \boldsymbol{x}^{\mathcal{C}})}(\hat{y}^k \mid \boldsymbol{z}^{\mathcal{O}}) .
$$

□

D Characteristics of the Datasets

This section reports summaries of the distributions and pairwise association of the discrete features and class variables in the three datasets: Adult, Default, and Thyroid.

As mentioned in Section 3.1, the pairwise associations on the discrete variables do not necessarily indicate the sparseness of the optimal DAGs learned when conditioning on the continuous features. Nevertheless, it might be reasonable to expect sparse DAGs on these 3 datasets because the satisfactory performance of BR suggests weak to moderate conditional dependencies.

D.1 Adult DataSet

The Adult dataset has 4 class variables and 5 discrete features. For each class variable and discrete feature, we report the proportions of the outcomes in Table 2 and 3, respectively. The discrete features have a high number of modalities (*e.g.*, $|\mathcal{X}_4| = 40$). The association matrix shown in Figure 4 indicates that class variables and discrete features exhibit strong dependencies.

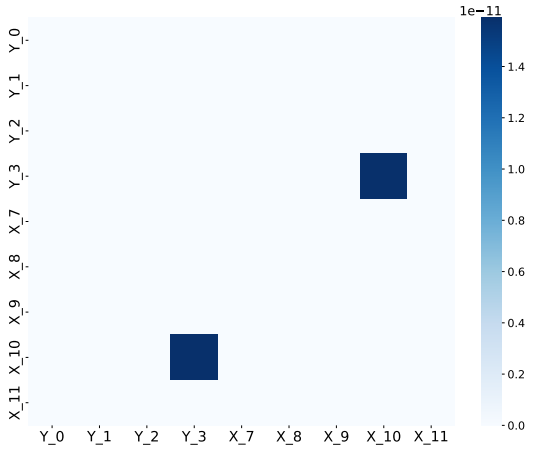


Figure 4: p-values for all pairwise chi-square tests across discrete variables of the Adult dataset

Table 2: Proportion of each outcome for the 4 class variables of the Adult dataset

Class	**Proportion of each outcome (%)**
Y^1	[10.57, 20.61, 35.31, 7.63, 16.83, 8.94, 0.11]
Y^2	[25.01, 45.07, 12.16, 3.00, 7.66, 0.17, 6.93]
Y^3	[65.69, 22.95, 7.07, 2.36, 1.92]
Y^4	[64.16, 35.84]

Table 3: Proportion of each outcome for the 5 discrete features of the Adult dataset

Feature	**Proportion of each outcome (%)**
X^1	[17.01, 31.14, 2.87, 7.42, 1.45, 20.87, 3.43, 2.10, 1.80, 4.02, 2.63, 0.98, 2.55, 0.18, 1.09, 0.48]
X^2	[13.02, 13.95, 2.88, 17.36, 11.09, 9.50, 4.32, 5.07, 4.53, 2.6, 10.85, 4.13, 0.08, 0.55]
X^3	[26.16, 39.69, 4.80, 10.77, 15.27, 3.32]
X^4	[88.57, 0.31, 0.54, 0.74, 1.38, 0.50, 0.06, 0.24, 0.30, 0.41, 0.17 1.49, 0.15, 0.18, 0.14, 0.14, 0.11, 0.11, 0.28, 0.37, 0.11, 0.30, 0.20, 0.08, 0.15, 0.20, 0.61, 0.55, 0.36, 0.04, 0.08, 0.08, 0.03, 0.14, 0.12, 0.09, 0.44, 0.13, 0.04, 0.05]
X^5	[74.31, 25.69]

D.2 Default Dataset

The Default dataset has 4 class variables and 6 discrete features. For each class variable and discrete feature, we report the proportions of the outcomes in Table 4 and Table 5, respectively. The association matrix shown in Figure 5 indicates that class variables and discrete features exhibit strong dependencies.

Table 4: Proportion of each outcome for the 4 class variables of the Default dataset

Class	Proportion of each outcome (%)
Y^1	[39.16, 60.84]
Y^2	[0.05, 35.88, 46.36, 16.14, 0.43, 0.96, 0.18]
Y^3	[0.18, 45.57, 53.19, 1.06]
Y^4	[79.56, 20.44]

Table 5: Proportion of the outcomes for the 6 discrete features of the Default dataset

Feature	Proportion of each outcome (%)
X^1	[9.58, 19.46, 50.80, 11.79, 8.37]
X^2	[13.12, 20.91, 54.20, 11.77]
X^3	[14.14, 20.53, 54.30, 11.04]
X^4	[15.02, 19.61, 56.30, 9.06]
X^5	[15.72, 19.02, 57.67, 7.58]
X^6	[16.88, 19.69, 55.29, 8.13]

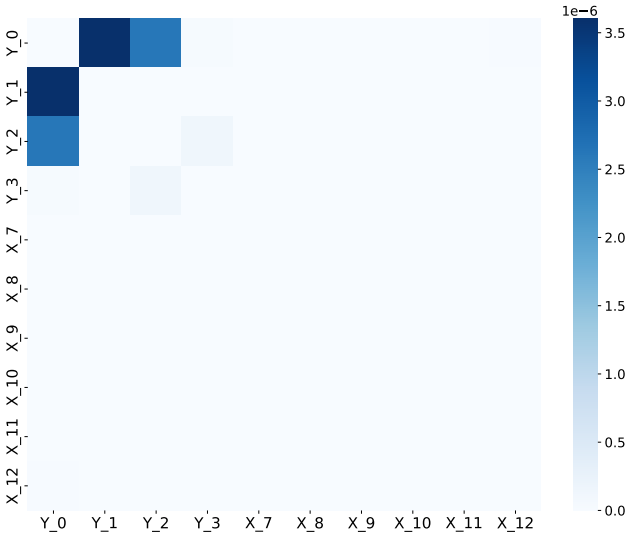


Figure 5: p-values for all pairwise chi-square tests across discrete variables of the Default dataset

D.3 Thyroid Dataset

The Thyroid dataset has 7 class variables and 22 discrete features. For each class variable and discrete feature, we report the proportions of the outcomes in Table 6 and Table 7, respectively. The data show extreme imbalance. The association matrix shown in Figure 6 indicates a mix of strong and weak dependencies across class variables and discrete features.

Table 6: Proportion of each outcome for the 7 class variables of the Thyroid dataset

Class	Proportion of each outcome (%)
Y^1	[97.37, 2.10, 0.23, 0.20, 0.10]
Y^2	[92.73, 0.01, 2.61, 4.57, 0.09]
Y^3	[95.49, 4.04, 0.47]
Y^4	[93.75, 6.25]
Y^5	[96.13, 1.26, 1.41, 1.20]
Y^6	[99.63, 0.16, 0.05, 0.15]
Y^7	[96.93, 2.15, 0.93]

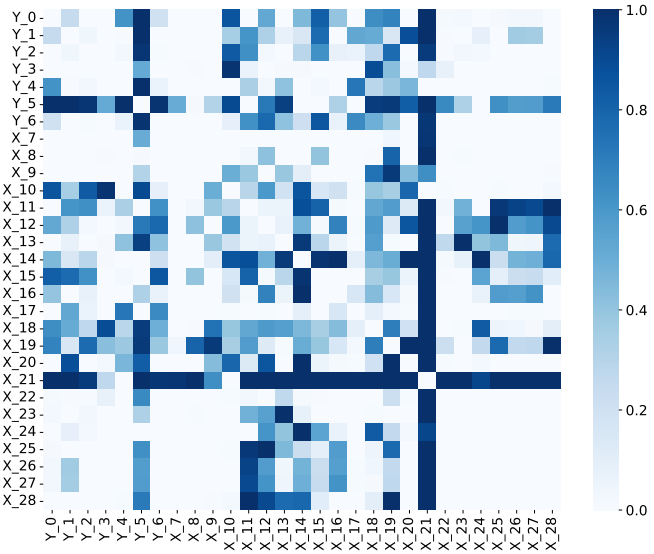


Figure 6: p-values for all pairwise chi-square tests across discrete variables of the Thyroid dataset

Table 7: Proportion of each outcome for the 22 discrete features of the Thyroid dataset

Feature	Proportion of each outcome (%)
X^1	[0.03, 2.78, 10.42, 26.10, 0.77, 59.89]
X^2	[30.44, 69.56]
X^3	[86.48, 13.52]
X^4	[98.33, 1.67]
X^5	[98.74, 1.26]
X^6	[96.25, 3.75]
X^7	[98.83, 1.17]
X^8	[98.54, 1.46]
X^9	[98.16, 1.84]
X^{10}	[93.13, 6.87]
X^{11}	[92.90, 7.10]
X^{12}	[98.99, 1.01]
X^{13}	[99.08, 0.92]
X^{14}	[97.37, 2.63]
X^{15}	[99.98, 0.02]
X^{16}	[95.44, 4.56]
X^{17}	[9.18, 90.82]
X^{18}	[28.39, 71.61]
X^{19}	[4.82, 95.18]
X^{20}	[8.82, 91.18]
X^{21}	[8.74, 91.26]
X^{22}	[96.19, 3.81]

E Additional Experimental Results

In this section, we report additional results in terms of per-class balanced accuracy, Hamming score, and subset 0/1 score. A random selection of the learned optimal structures is also displayed.

The optimal DAGs obtained when allowing at most two or three parents among discrete variables, are respectively given in Section E.4.1 and E.4.2. They indicate that with Logistic regression (LR) as the base classifier, the optimal DAGs always have at most 2 discrete parents. Therefore, the slightly different results between the two settings might merely stem from the non-optimality of the base learner. This is because if we can ensure the optimality of the base learner, Proposition 5 should ensure that any DAG that contains at least one parent set of cardinality of at least 3 should be sub-optimal.

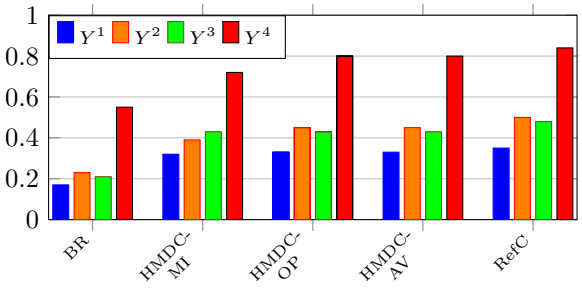
E.1 Per-Class Balanced Accuracy

We report the per-class balanced accuracy for five methods (BR, HMDC-MI, HMDC-OP, HMDC-AV, and RefC) assuming at most two possible parents. The results for up to three possible parents are similar and are omitted for brevity.

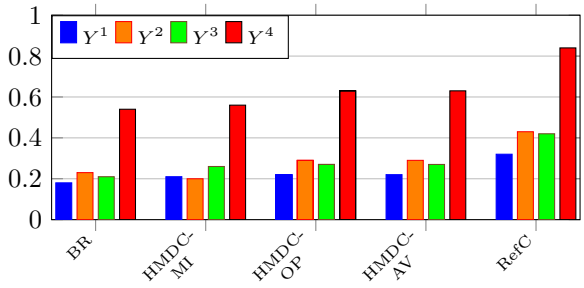
On the Adult dataset, Figure 7 shows that HMDC-OP and HMDC-AV consistently outperform BR.

On the Default dataset, Figure 8 shows that HMDC-OP and HMDC-AV are competitive with BR, for LR and RF employed as base learners. However, they are slightly worse than BR with NB as the base classifier. This might be explained by the fact that the optimal DAGs provided by NB, given in Section E.4.1, are denser than those that are provided by LR and RF, and may be overly complicated.

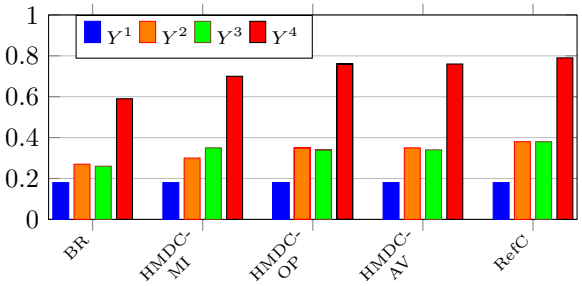
Similar trends are observed on the Thyroid dataset. HMDC-OP and HMDC-AV are either competitive with or outperform BR with LR and RF as base learners. However, they are either competitive or worse than BR with NB as base learner. Again, the optimal DAGs provided by NB, given in Section E.4.1, suggest that NB may produce overly complicated DAGs.



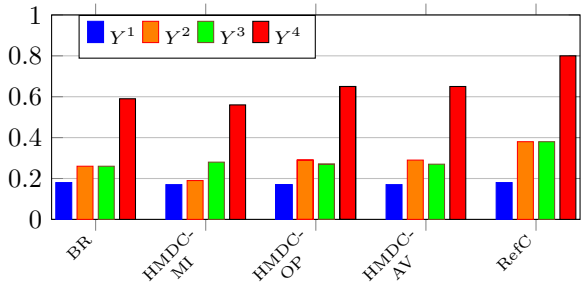
(a) LR - 30% class and 30% discrete feat. missing



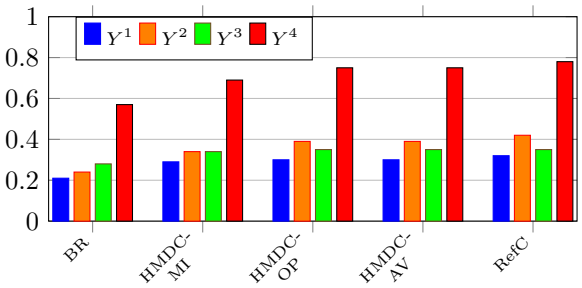
(d) LR - 90% class and 80% discrete feat. missing



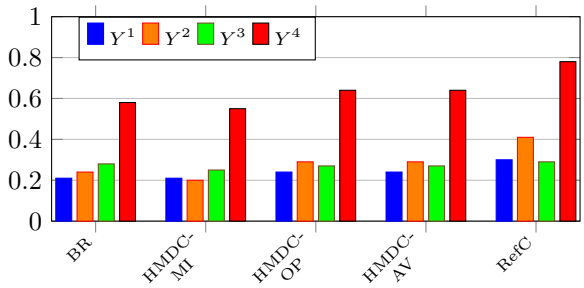
(b) NB - 30% class and 30% discrete feat. missing



(e) NB - 90% class and 80% discrete feat. missing

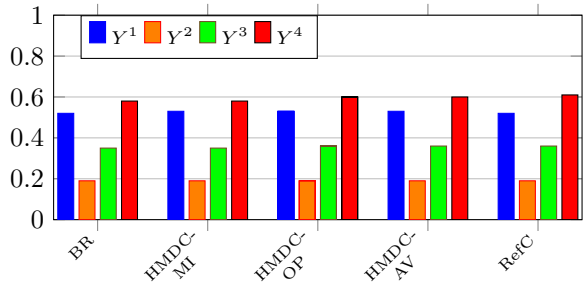


(c) RF - 30% class and 30% discrete feat. missing

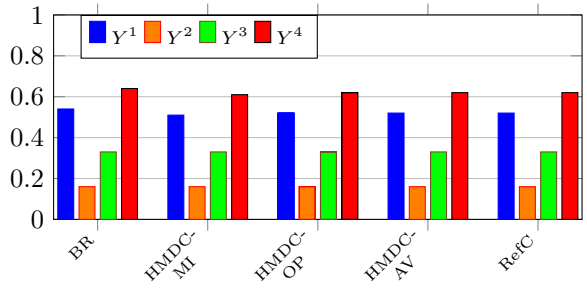


(f) RF - 90% class and 80% discrete feat. missing

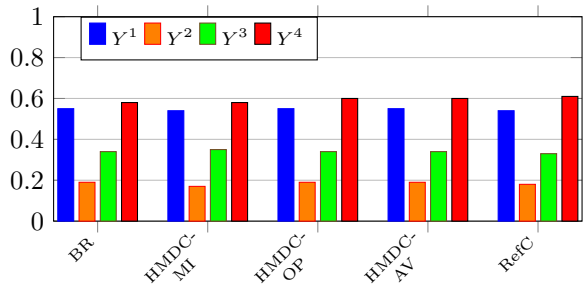
Figure 7: Per-Class balanced accuracies (↑) using LR (a, d); NB (b, e); RF (c, f); as the base learner on the Adult dataset, when allowing at most two parents among discrete variables



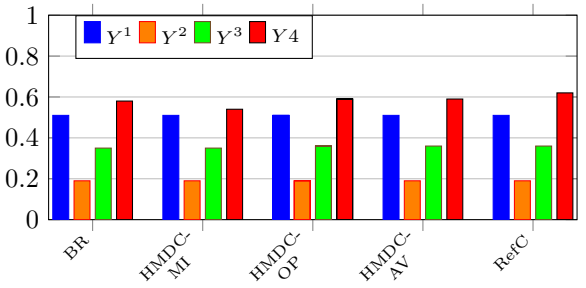
(a) LR - 30% class and 30% discrete feat. missing



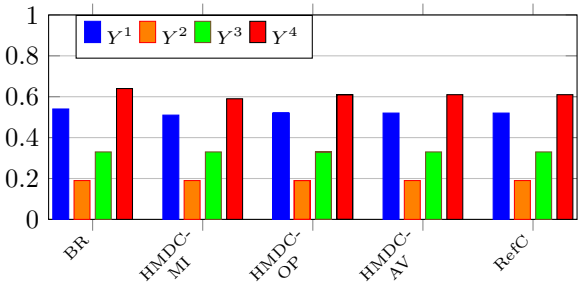
(b) NB - 30% class and 30% discrete feat. missing



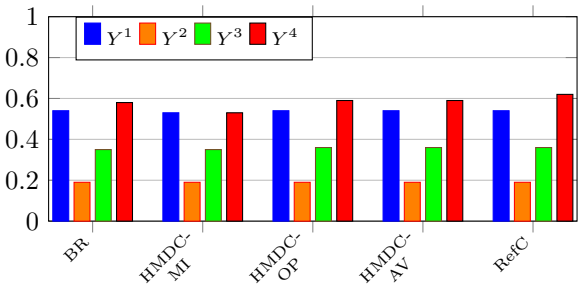
(c) RF - 30% class and 30% discrete feat. missing



(d) LR - 90% class and 80% discrete feat. missing

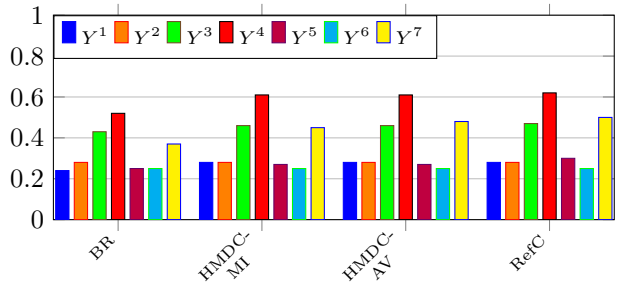


(e) NB - 90% class and 80% discrete feat. missing

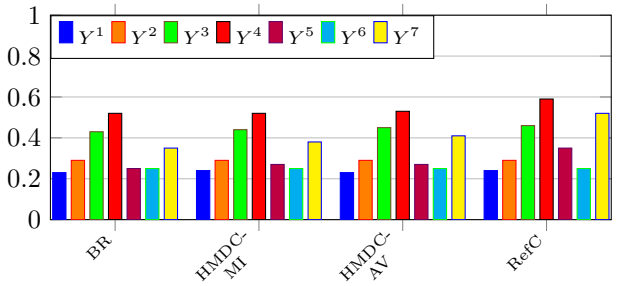


(f) RF - 90% class and 80% discrete feat. missing

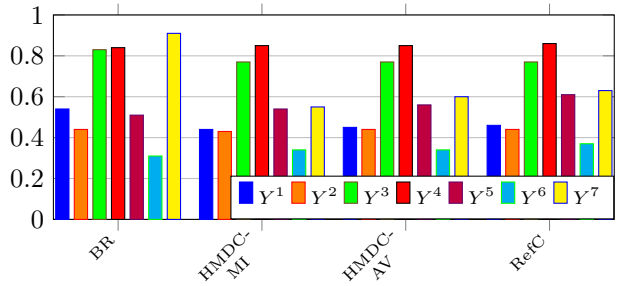
Figure 8: Per-Class balanced accuracies (↑) using LR (a, d); NB (b, e); RF (c, f); as the base learner on the Default dataset, when allowing at most two parents among discrete variables



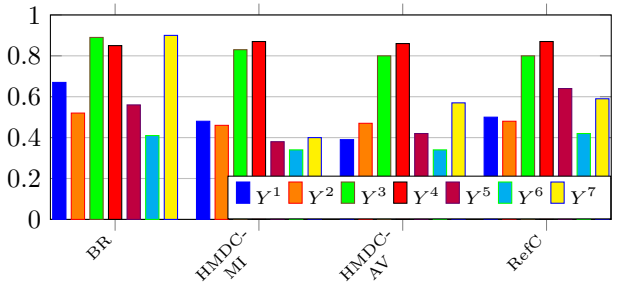
(a) LR - 30% class and 30% discrete feat. missing



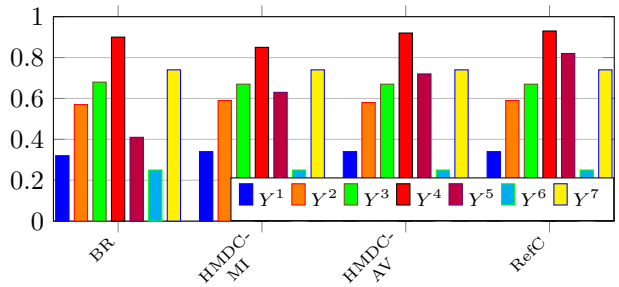
(d) LR - 90% class and 80% discrete feat. missing



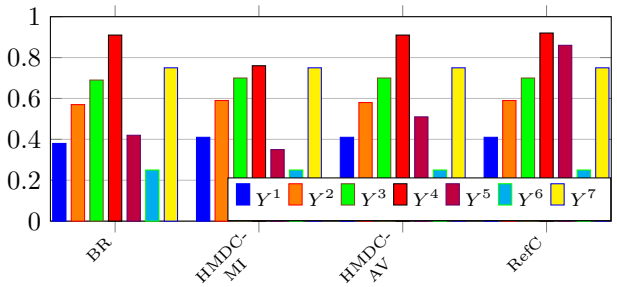
(b) NB - 30% class and 30% discrete feat. missing



(e) NB - 90% class and 80% discrete feat. missing



(c) RF - 30% class and 30% discrete feat. missing



(f) RF - 90% class and 80% discrete feat. missing

Figure 9: Per-Class balanced accuracies (↑) using LR (a, d); NB (b, e); RF (c, f); as the base learner on the Thyroid dataset, when allowing at most two parents among discrete variables

E.2 Additional Results for Hamming and Subset 0/1 Accuracies

E.2.1 Results with at Most 2 Discrete Parents

Figure 10 presents the average scores with Naive Bayes and Random Forest employed as base learners, on three datasets, Adult, Default, and Thyroid, when allowing at most 2 discrete parents. Together with the Figure 2 in Section 4.3, they provide the entire results on the Hamming and Subset 0/1 accuracies provided by MI, BR, HMDC-MI, HMDC-OP, HMDC-AV, and RefC.

E.2.2 Results with at Most 3 Discrete Parents

Figure 11–12 presents the average scores on three datasets, Adult, Default, and Thyroid, when allowing at most three parents among discrete variables (palim=3). The results are quite similar to those obtained when allowing at most 2 parents (palim=2).

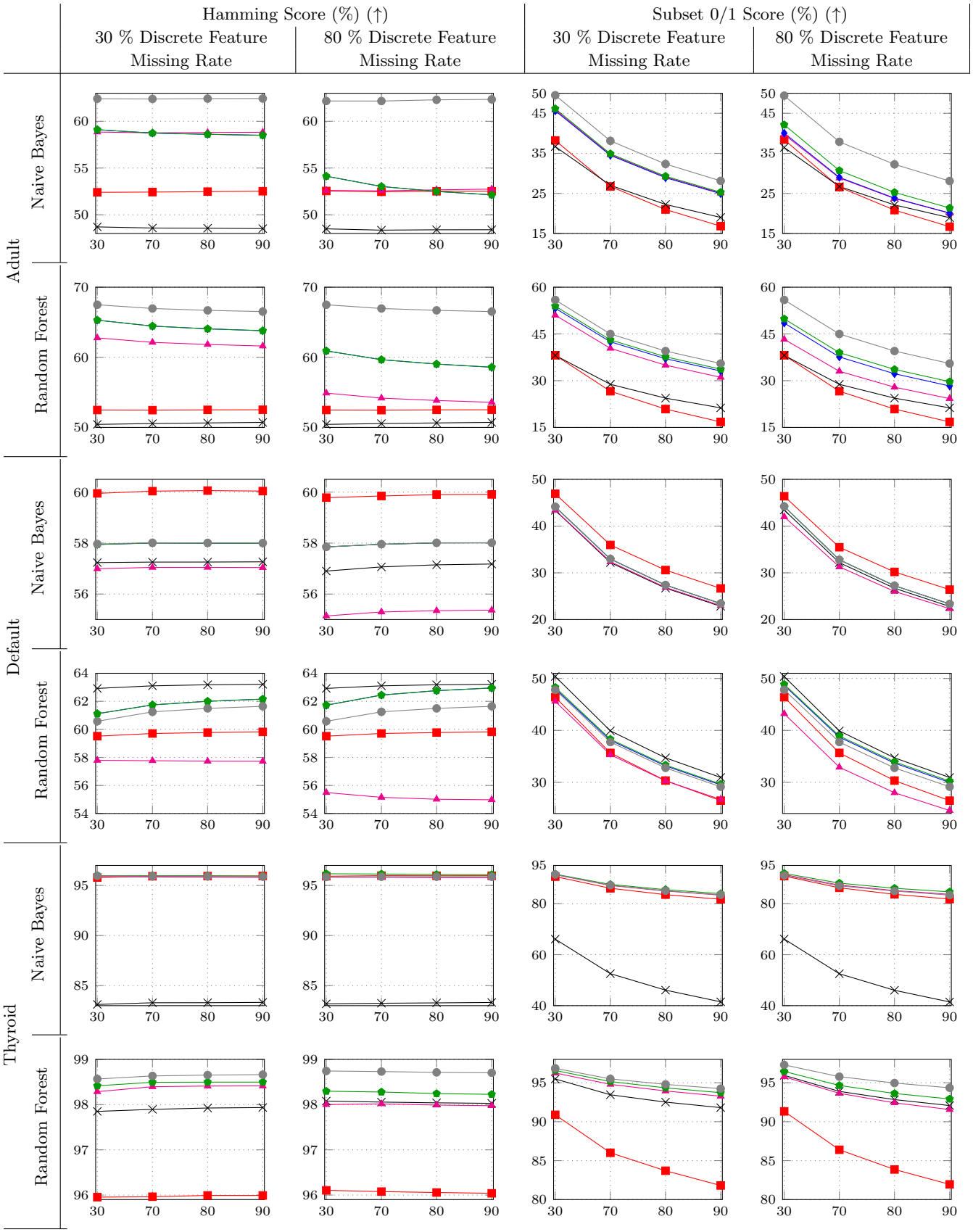


Figure 10: Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds plotted against the percentage of missing class variables (in %, x-axis) for the three datasets, when allowing at most two parents among discrete variables. The missing rate for discrete features is either 30 % or 80 %; Naive Bayes and Random Forests are used as base classifiers. Probabilistic models are represented as follows: ■ MI, × BR, ▲ HMDC-MI, ♦ HMDC-OP, ⬟ HMDC-AV and ● RefC.

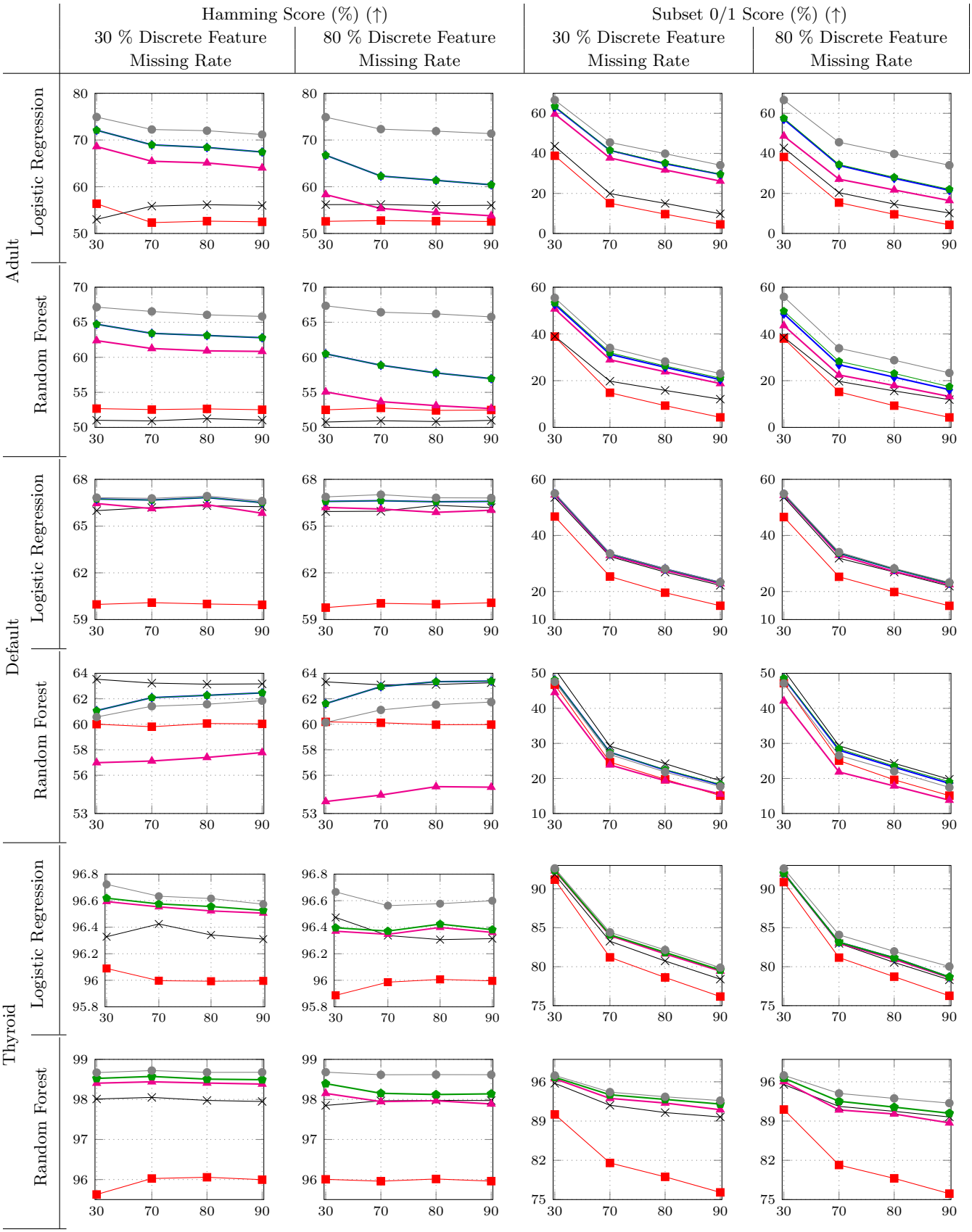


Figure 11: Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds plotted against the percentage of missing class variables (in %, x-axis) for the three datasets, when allowing at most three parents among discrete variables. The missing rate for discrete features is either 30 % or 80 %; Logistic Regression and Random Forests are used as base classifiers. Probabilistic models are represented as follows: ■ MI, × BR, ▲ HMDC-MI, ♦ HMDC-OP, ⬟ HMDC-AV and ● RefC.

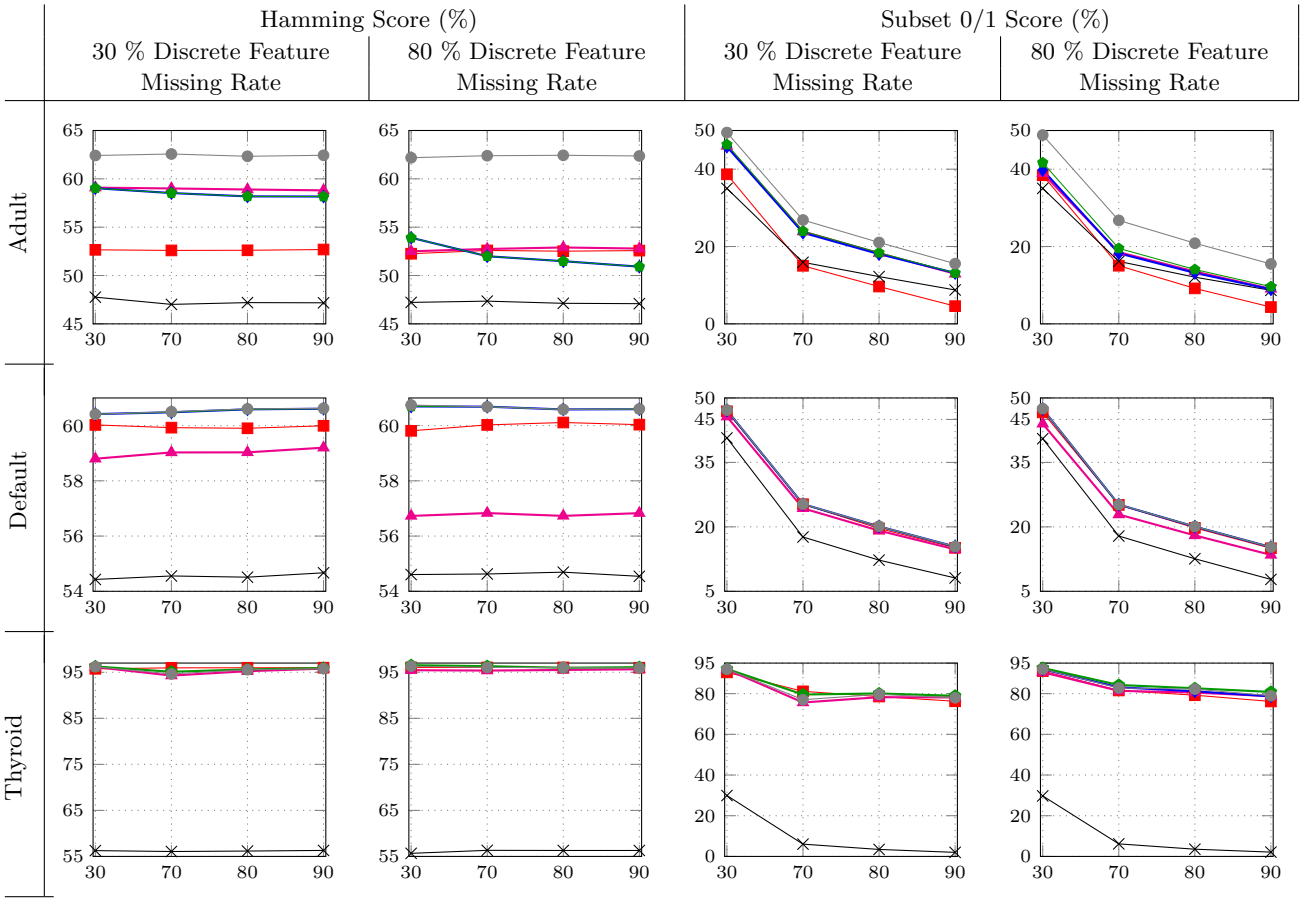


Figure 12: Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds plotted against the percentage of missing class variables (in %, x-axis) for the three datasets, with the Naive Bayes classifier, when allowing at most three parents among discrete variables. The missing rate for discrete features is either 30 % or 80 %. Probabilistic models are represented as follows: ■ MI, × BR, ▲ HMDC-MI, ♦ HMDC-OP, ⬟ HMDC-AV and ● RefC.

E.3 Detailed Experimental Results

The average scores and standard deviations achieved by Logistic Regression, Naive Bayes, and Random Forest on three datasets—Adult, Default, and Thyroid when allowing at most 2 discrete parents are given in Table 8–16. As mentioned in the main text, RandC is consistently worse than the others for both Hamming and 0/1 accuracies. In these tables, bold numbers indicate the best average results (for a given missingness configuration and performance score).

Similar results when allowing at most 3 discrete parents are given in Table 17–25.

Table 8: Detailed results for Adult with Logistic Regression when allowing at most two parents among discrete variables

Model	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	25.08±1.02	24.96±0.86	24.94±0.77	24.93±0.73	24.68±0.98	24.78±0.84	24.78±0.77	24.79±0.70
	Subset 0/1	15.68±1.10	9.73±6.00	7.16±6.11	5.56±5.97	15.34±0.82	9.55±5.82	7.02±5.95	5.46±5.82
MI	Hamming	52.04±1.28	52.32±1.00	52.42±0.84	52.45±0.74	52.51±0.50	52.57±0.50	52.51±0.50	52.55±0.48
	Subset 0/1	37.67±1.39	26.37±11.36	20.80±12.18	16.62±12.79	38.18±1.06	26.57±11.66	20.84±12.51	16.76±12.94
BR	Hamming	55.27±0.91	55.72±0.93	55.81±0.82	55.86±0.77	56.14±1.09	56.12±0.86	56.06±0.76	56.07±0.70
	Subset 0/1	41.69±1.08	31.06±10.68	25.74±11.56	21.84±12.09	42.57±1.20	31.39±11.24	25.70±12.22	21.80±12.57
HMDC-MI	Hamming	59.15±1.37	57.46±2.06	56.49±2.20	55.77±2.30	68.72±0.73	67.29±1.59	66.59±1.65	66.03±1.75
	Subset 0/1	49.41±1.61	38.41±11.12	32.98±11.91	28.91±12.50	59.63±1.03	48.89±10.79	43.47±11.68	39.28±12.46
HMDC-OP	Hamming	**66.46±1.17**	**64.43±2.24**	**63.40±2.34**	**62.64±2.42**	**72.18±0.69**	**70.66±1.63**	**69.88±1.74**	**69.31±1.84**
	Subset 0/1	56.66±1.27	45.36±11.35	39.58±12.37	35.12±13.22	62.98±1.13	52.14±10.88	46.65±11.80	42.41±12.60
HMDC-AV	Hamming	**66.46±1.17**	**64.43±2.24**	**63.40±2.34**	**62.64±2.42**	**72.18±0.69**	**70.66±1.63**	**69.88±1.74**	**69.31±1.84**
	Subset 0/1	**57.06±1.33**	**45.73±11.39**	**39.90±12.44**	**35.49±13.21**	**63.34±1.07**	**52.50±10.89**	**46.93±11.89**	**42.70±12.65**
RefC	Hamming	74.96±1.04	73.75±1.45	73.18±1.47	72.70±1.53	74.86±0.70	73.65±1.33	73.01±1.43	72.60±1.46
	Subset 0/1	66.65±1.09	56.23±10.46	50.95±11.36	46.79±12.20	66.41±1.10	56.19±10.26	50.77±11.37	46.70±12.13

Table 9: Detailed results for Adult with Random Forest when allowing at most two parents among discrete variables

Model	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	24.35±0.68	24.61±0.68	24.69±0.65	24.72±0.63	24.35±0.68	24.61±0.68	24.69±0.65	24.72±0.63
	Subset 0/1	14.76±0.67	9.15±5.63	6.70±5.76	5.21±5.61	14.76±0.67	9.15±5.63	6.70±5.76	5.21±5.61
MI	Hamming	52.48±0.76	52.46±0.62	52.50±0.56	52.52±0.52	52.48±0.76	52.46±0.62	52.50±0.56	52.52±0.52
	Subset 0/1	38.16±1.13	26.61±11.59	20.90±12.45	16.76±12.95	38.16±1.13	26.61±11.59	20.90±12.45	16.76±12.95
BR	Hamming	50.43±1.07	50.55±0.92	50.64±0.89	50.72±0.84	50.43±1.07	50.55±0.92	50.64±0.89	50.72±0.84
	Subset 0/1	38.11±1.09	28.83±9.34	24.40±9.89	21.28±10.14	38.11±1.09	28.83±9.34	24.40±9.89	21.28±10.14
HMDC-MI	Hamming	54.90±1.10	54.18±1.15	53.84±1.10	53.58±1.08	62.77±0.84	62.13±1.02	61.83±1.01	61.61±0.99
	Subset 0/1	43.27±1.35	33.01±10.31	27.92±11.09	24.26±11.52	51.02±1.27	40.38±10.69	34.99±11.61	31.08±12.12
HMDC-OP	Hamming	**60.92±1.08**	**59.66±1.59**	**59.04±1.63**	**58.58±1.63**	**65.30±1.18**	**64.45±1.25**	**64.06±1.21**	**63.79±1.17**
	Subset 0/1	48.51±1.19	37.62±10.95	32.23±11.75	28.30±12.25	53.24±1.27	42.47±10.82	37.06±11.70	33.11±12.24
HMDC-AV	Hamming	**60.92±1.08**	**59.66±1.59**	**59.04±1.63**	**58.58±1.63**	**65.30±1.18**	**64.45±1.25**	**64.06±1.21**	**63.79±1.17**
	Subset 0/1	**49.84±1.33**	**39.00±10.90**	**33.62±11.72**	**29.68±12.24**	**53.92±1.41**	**43.10±10.88**	**37.66±11.76**	**33.69±12.31**
RefC	Hamming	67.50±0.98	66.96±0.96	66.70±0.95	66.52±0.91	67.50±0.98	66.96±0.96	66.70±0.95	66.52±0.91
	Subset 0/1	55.91±1.30	44.99±10.98	39.51±11.86	35.52±12.40	55.91±1.30	44.99±10.98	39.51±11.86	35.52±12.40

Table 10: Detailed results for Adult with Naive Bayes when allowing at most two parents among discrete variables

Model	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	24.82±0.92	24.84±0.74	24.78±0.69	24.80±0.63	24.91±0.93	24.85±0.76	24.85±0.66	24.84±0.64
	Subset 0/1	15.72±0.49	9.78±5.96	7.19±6.09	5.61±5.95	15.71±1.08	9.69±6.07	7.08±6.18	5.49±6.02
MI	Hamming	52.55±0.76	52.46±0.69	52.51±0.59	52.52±0.53	52.41±0.68	52.43±0.64	52.48±0.56	52.52±0.51
	Subset 0/1	38.43±0.94	26.58±11.88	20.84±12.66	16.69±13.11	38.23±1.53	26.75±11.55	20.98±12.48	16.83±12.98
BR	Hamming	48.50±1.72	48.36±1.62	48.39±1.57	48.40±1.53	48.70±1.72	48.58±1.64	48.56±1.54	48.52±1.51
	Subset 0/1	36.46±1.73	26.73±9.84	22.13±10.35	18.95±10.54	36.70±1.89	27.02±9.83	22.25±10.50	19.01±10.70
HMDC-MI	Hamming	52.60±2.69	52.56±2.80	52.67±2.78	52.75±2.73	58.85±3.18	58.76±3.08	58.80±3.04	58.83±2.96
	Subset 0/1	39.70±2.46	28.86±11.07	23.73±11.62	20.07±11.91	45.87±3.73	34.69±11.64	29.10±12.42	25.09±12.83
HMDC-OP	Hamming	**54.13±2.10**	**53.03±2.16**	**52.50±2.12**	**52.13±2.09**	**59.10±2.59**	**58.74±2.53**	**58.61±2.42**	**58.50±2.35**
	Subset 0/1	40.05±2.55	29.01±11.22	23.82±11.76	20.11±12.06	45.61±3.14	34.50±11.41	28.92±12.25	24.94±12.65
HMDC-AV	Hamming	**54.13±2.10**	**53.03±2.16**	**52.50±2.12**	**52.13±2.09**	**59.10±2.59**	**58.74±2.53**	**58.61±2.42**	**58.50±2.35**
	Subset 0/1	**42.14±2.26**	**30.71±11.55**	**25.26±12.20**	**21.36±12.56**	**46.15±3.19**	**34.90±11.57**	**29.29±12.36**	**25.29±12.76**
RefC	Hamming	62.17±2.97	62.17±3.04	62.30±2.96	62.35±2.91	62.41±3.16	62.39±3.17	62.43±3.09	62.44±3.05
	Subset 0/1	49.41±3.10	37.89±11.91	32.26±12.63	28.10±13.13	49.52±3.87	38.12±11.91	32.34±12.78	28.13±13.28

Table 11: Detailed results for Default with Logistic Regression when allowing at most two parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	34.72±1.01	34.70±0.79	34.69±0.70	34.71±0.65	34.67±0.97	34.74±0.73	34.74±0.65	34.75±0.60
	Subset 0/1	22.79±0.67	15.00±7.81	11.38±8.18	9.08±8.12	23.02±0.74	15.03±8.02	11.46±8.27	9.14±8.21
MI	Hamming	59.73±0.46	59.84±0.46	59.88±0.41	59.92±0.40	60.07±0.57	60.12±0.50	60.09±0.43	60.06±0.42
	Subset 0/1	46.69±0.81	35.60±11.12	30.28±11.79	26.42±12.21	46.98±0.82	36.13±10.91	30.67±11.79	26.77±12.24
BR	Hamming	65.76±0.62	65.98±0.60	66.06±0.58	66.08±0.56	66.16±0.66	66.14±0.53	66.18±0.50	66.16±0.53
	Subset 0/1	53.50±0.87	42.76±10.79	37.39±11.65	33.45±12.19	53.84±0.99	43.11±10.77	37.70±11.66	33.71±12.25
HMDC-MI	Hamming	66.08±0.81	66.04±0.61	66.04±0.58	66.04±0.53	66.66±0.48	66.55±0.45	66.51±0.40	66.48±0.43
	Subset 0/1	53.78±1.03	43.09±10.74	37.69±11.65	33.83±12.13	54.50±0.93	43.71±10.83	38.31±11.70	34.38±12.22
HMDC-OP	Hamming	**66.36±0.76**	**66.36±0.58**	**66.39±0.52**	**66.42±0.47**	**66.82±0.49**	**66.78±0.43**	**66.74±0.37**	**66.72±0.37**
	Subset 0/1	**54.31±1.13**	**43.72±10.64**	**38.39±11.52**	34.52±12.03	54.76±0.84	44.13±10.66	**38.79±11.54**	**34.93±12.03**
HMDC-AV	Hamming	**66.36±0.76**	**66.36±0.58**	**66.39±0.52**	**66.42±0.47**	**66.82±0.49**	**66.78±0.43**	**66.74±0.37**	**66.72±0.37**
	Subset 0/1	54.30±1.12	**43.72±10.62**	**38.39±11.51**	34.53±12.02	**54.79±0.82**	**44.14±10.67**	**38.79±11.55**	**34.93±12.04**
RefC	Hamming	66.70±0.83	66.68±0.62	66.70±0.57	66.71±0.52	66.87±0.58	66.85±0.50	66.82±0.42	66.80±0.41
	Subset 0/1	54.72±1.10	44.04±10.73	38.71±11.58	34.82±12.09	54.88±0.91	44.25±10.66	38.91±11.53	35.03±12.04

Table 12: Detailed results for Default with Random Forest when allowing at most two parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	35.02±0.77	34.90±0.62	34.87±0.53	34.85±0.48	35.02±0.77	34.90±0.62	34.87±0.53	34.85±0.48
	Subset 0/1	23.53±0.96	15.46±8.11	11.77±8.43	9.39±8.38	23.53±0.96	15.46±8.11	11.77±8.43	9.39±8.38
MI	Hamming	59.52±1.11	59.71±0.86	59.77±0.74	59.82±0.67	59.52±1.11	59.71±0.86	59.77±0.74	59.82±0.67
	Subset 0/1	46.39±1.68	35.67±10.82	30.34±11.62	26.47±12.09	46.39±1.68	35.67±10.82	30.34±11.62	26.47±12.09
BR	Hamming	62.92±0.51	63.10±0.43	63.18±0.39	63.21±0.36	62.92±0.51	63.10±0.43	63.18±0.39	63.21±0.36
	Subset 0/1	50.38±0.74	39.91±10.48	34.72±11.28	30.94±11.77	50.38±0.74	39.91±10.48	34.72±11.28	30.94±11.77
HMDC-MI	Hamming	55.50±0.74	55.15±0.69	55.02±0.64	54.98±0.60	57.79±0.84	57.77±0.66	57.73±0.61	57.72±0.58
	Subset 0/1	43.22±0.90	32.91±10.34	28.00±10.94	24.58±11.17	45.60±1.15	35.32±10.33	30.29±11.04	26.70±11.41
HMDC-OP	Hamming	**61.72±0.89**	**62.45±0.99**	**62.76±0.95**	**62.95±0.90**	**61.12±0.96**	**61.75±0.96**	**62.00±0.90**	**62.16±0.85**
	Subset 0/1	48.68±0.95	38.70±10.03	33.70±10.82	29.99±11.37	**48.03±1.18**	38.10±9.98	33.16±10.74	29.51±11.26
HMDC-AV	Hamming	**61.72±0.89**	**62.45±0.99**	**62.76±0.95**	**62.95±0.90**	**61.12±0.96**	**61.75±0.96**	**62.00±0.90**	**62.16±0.85**
	Subset 0/1	**48.87±1.09**	**38.93±9.99**	**33.95±10.78**	**30.24±11.34**	**48.30±1.22**	**38.31±10.04**	**33.34±10.81**	**29.66±11.33**
RefC	Hamming	60.58±0.94	61.25±0.98	61.50±0.90	61.64±0.84	60.58±0.94	61.25±0.98	61.50±0.90	61.64±0.84
	Subset 0/1	47.81±1.17	37.75±10.10	32.78±10.84	29.15±11.31	47.81±1.17	37.75±10.10	32.78±10.84	29.15±11.31

Table 13: Detailed results for Default with Naive Bayes when allowing at most two parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	34.54±0.69	34.61±0.57	34.67±0.51	34.70±0.50	34.81±0.76	34.83±0.60	34.81±0.56	34.78±0.54
	Subset 0/1	22.96±0.86	15.14±7.86	11.58±8.15	9.23±8.15	22.84±0.85	15.01±7.86	11.47±8.15	9.15±8.12
MI	Hamming	**59.78±0.51**	**59.85±0.46**	**59.90±0.44**	**59.91±0.41**	**59.95±0.68**	**60.04±0.55**	**60.06±0.48**	**60.04±0.45**
	Subset 0/1	**46.39±0.75**	**35.47±10.93**	**30.19±11.64**	**26.41±12.03**	**46.91±0.97**	**35.97±10.99**	**30.59±11.77**	**26.67±12.25**
BR	Hamming	56.90±2.43	57.07±2.51	57.15±2.53	57.18±2.52	57.24±2.74	57.26±2.55	57.26±2.54	57.27±2.50
	Subset 0/1	43.44±2.76	32.18±11.68	26.71±12.43	22.80±12.81	43.37±3.18	32.19±11.60	26.73±12.40	22.84±12.77
HMDC-MI	Hamming	55.14±2.28	55.30±2.35	55.35±2.43	55.37±2.41	56.99±2.61	57.05±2.55	57.05±2.52	57.04±2.50
	Subset 0/1	42.00±2.64	31.28±11.11	26.02±11.89	22.34±12.22	43.41±3.07	32.38±11.48	26.88±12.35	22.99±12.74
HMDC-OP	Hamming	57.85±2.46	57.95±2.54	58.01±2.57	58.01±2.56	57.95±2.65	58.00±2.55	58.00±2.54	57.99±2.52
	Subset 0/1	44.20±2.81	32.79±11.83	27.25±12.58	23.33±12.93	44.11±3.13	32.95±11.60	27.38±12.48	23.42±12.89
HMDC-AV	Hamming	57.85±2.46	57.95±2.54	**58.01±2.57**	58.01±2.56	57.95±2.65	58.00±2.55	58.00±2.54	57.99±2.52
	Subset 0/1	44.21±2.81	32.79±11.84	27.25±12.58	23.33±12.94	44.14±3.14	32.96±11.61	27.39±12.49	23.43±12.90
RefC	Hamming	57.85±2.45	57.96±2.54	58.01±2.57	58.02±2.56	57.97±2.63	58.01±2.54	58.01±2.54	58.00±2.52
	Subset 0/1	44.21±2.81	32.79±11.83	27.26±12.58	23.34±12.93	44.15±3.15	32.98±11.61	27.40±12.49	23.44±12.90

Table 14: Detailed results for Thyroid with Logistic Regression when allowing at most two parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	33.20±1.96	33.29±1.78	33.36±1.69	33.42±1.65	33.65±1.52	33.54±1.50	33.48±1.50	33.50±1.51
	Subset 0/1	13.50±1.57	7.37±6.23	5.03±6.07	3.79±5.68	14.13±1.22	7.60±6.59	5.22±6.35	3.94±5.93
MI	Hamming	96.19±0.37	96.15±0.33	96.09±0.30	96.07±0.28	95.89±0.36	95.95±0.34	95.96±0.31	95.98±0.29
	Subset 0/1	91.47±0.83	86.65±4.98	83.97±5.59	82.06±5.92	90.82±0.82	86.10±4.86	83.60±5.36	81.86±5.60
BR	Hamming	96.48±0.36	96.45±0.32	96.41±0.30	96.39±0.29	96.20±0.39	96.27±0.35	96.29±0.31	96.31±0.30
	Subset 0/1	92.18±0.84	87.80±4.54	85.41±5.06	83.64±5.41	91.52±0.90	87.34±4.35	85.14±4.77	83.54±5.05
HMDC-MI	Hamming	96.59±0.35	96.53±0.31	96.48±0.29	96.46±0.28	96.46±0.42	96.50±0.37	96.51±0.33	96.51±0.31
	Subset 0/1	92.49±0.84	88.10±4.53	85.73±5.04	83.96±5.39	92.14±0.94	88.10±4.24	85.95±4.65	84.41±4.91
HMDC-AV	Hamming	**96.66±0.35**	**96.58±0.32**	**96.52±0.30**	**96.49±0.30**	**96.47±0.41**	**96.52±0.37**	**96.52±0.33**	**96.53±0.32**
	Subset 0/1	**92.62±0.81**	**88.24±4.51**	**85.88±5.03**	**84.10±5.40**	**92.19±0.91**	**88.18±4.22**	**86.04±4.63**	**84.49±4.91**
RefC	Hamming	96.83±0.40	96.76±0.36	96.70±0.34	96.67±0.33	96.50±0.40	96.57±0.38	96.57±0.33	96.58±0.32
	Subset 0/1	93.00±0.90	88.83±4.33	86.56±4.85	84.90±5.18	92.23±0.89	88.35±4.12	86.26±4.53	84.75±4.80

Table 15: Detailed results for Thyroid with Random Forest when allowing at most two parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	33.37±1.48	33.31±1.45	33.38±1.45	33.41±1.45	33.62±1.55	33.48±1.59	33.46±1.52	33.49±1.52
	Subset 0/1	13.57±1.75	7.42±6.27	5.09±6.09	3.85±5.70	13.93±1.19	7.53±6.46	5.12±6.28	3.88±5.85
MI	Hamming	96.10±0.46	96.08±0.36	96.06±0.34	96.04±0.32	95.95±0.45	95.96±0.36	95.99±0.32	95.99±0.30
	Subset 0/1	91.32±1.06	86.40±5.04	83.87±5.54	81.96±5.89	90.88±1.04	86.01±4.99	83.70±5.28	81.81±5.67
BR	Hamming	98.08±0.26	98.06±0.24	98.04±0.24	98.03±0.22	97.85±0.33	97.89±0.27	97.92±0.25	97.94±0.24
	Subset 0/1	95.95±0.44	93.89±2.15	92.84±2.37	92.07±2.47	95.48±0.64	93.46±2.12	92.52±2.22	91.81±2.33
HMDC-MI	Hamming	98.00±0.35	98.02±0.29	97.99±0.28	97.98±0.26	98.29±0.33	98.40±0.27	98.41±0.24	98.41±0.22
	Subset 0/1	95.76±0.61	93.64±2.25	92.42±2.63	91.56±2.78	96.26±0.71	94.79±1.60	93.94±1.82	93.28±2.00
HMDC-AV	Hamming	**98.30±0.31**	**98.28±0.25**	**98.24±0.24**	**98.23±0.23**	**98.42±0.35**	**98.49±0.28**	**98.50±0.24**	**98.50±0.22**
	Subset 0/1	**96.45±0.54**	**94.65±1.87**	**93.62±2.17**	**92.92±2.28**	**96.58±0.73**	**95.15±1.58**	**94.34±1.76**	**93.72±1.91**
RefC	Hamming	98.74±0.20	98.73±0.17	98.71±0.16	98.70±0.15	98.57±0.35	98.63±0.27	98.65±0.23	98.66±0.21
	Subset 0/1	97.29±0.41	95.78±1.59	94.96±1.79	94.35±1.91	96.85±0.73	95.50±1.49	94.79±1.63	94.24±1.74

Table 16: Detailed results for Thyroid with Naive Bayes when allowing at most two parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	33.63±1.37	33.49±1.54	33.48±1.59	33.49±1.56	33.23±1.75	33.37±1.65	33.44±1.59	33.46±1.54
	Subset 0/1	13.99±1.32	7.56±6.50	5.19±6.28	3.93±5.86	13.59±1.50	7.40±6.29	5.09±6.08	3.84±5.70
MI	Hamming	95.92±0.46	95.98±0.40	95.97±0.36	95.98±0.34	95.79±0.33	95.92±0.34	95.93±0.32	95.95±0.30
	Subset 0/1	90.84±1.15	86.23±4.82	83.67±5.42	81.81±5.77	90.63±0.71	86.04±4.73	83.55±5.32	81.71±5.66
BR	Hamming	83.18±1.55	83.24±1.39	83.28±1.31	83.32±1.26	83.12±0.97	83.29±1.12	83.29±1.13	83.33±1.10
	Subset 0/1	66.14±2.86	52.51±14.06	46.03±14.87	41.46±15.27	66.09±2.29	52.55±13.94	46.09±14.79	41.50±15.22
HMDC-MI	Hamming	95.82±0.56	95.81±0.48	95.78±0.46	95.78±0.45	95.90±0.39	95.84±0.39	95.82±0.38	95.80±0.36
	Subset 0/1	91.41±1.18	87.29±4.30	85.14±4.75	83.68±4.91	91.38±0.82	87.14±4.37	84.97±4.83	83.35±5.08
HMDC-AV	Hamming	**96.17±0.49**	**96.14±0.43**	**96.09±0.43**	**96.08±0.41**	**95.98±0.46**	**95.97±0.42**	**95.97±0.40**	**95.94±0.39**
	Subset 0/1	**91.89±1.12**	**88.07±4.02**	**86.04±4.48**	**84.64±4.64**	**91.58±0.92**	**87.56±4.18**	**85.51±4.63**	**83.92±4.94**
RefC	Hamming	95.85±0.47	95.88±0.41	95.88±0.41	95.87±0.41	95.96±0.48	95.92±0.43	95.91±0.41	95.89±0.39
	Subset 0/1	91.10±1.19	86.99±4.30	84.94±4.69	83.36±5.00	91.46±0.85	87.24±4.36	85.08±4.82	83.47±5.08

Table 17: Detailed results for Adult with Logistic Regression when allowing at most three parents among discrete variables

Model	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	25.06±0.65	24.84±0.40	24.86±0.65	24.87±0.57	25.22±0.80	24.75±0.63	24.78±0.60	24.80±0.61
	Subset 0/1	15.01±0.75	3.78±0.39	1.89±0.24	0.86±0.25	15.68±0.50	3.75±0.54	2.05±0.37	0.84±0.23
MI	Hamming	52.60±0.79	52.76±0.37	52.65±0.40	52.57±0.34	53.01±0.93	52.32±0.32	52.64±0.30	52.50±0.23
	Subset 0/1	38.20±1.07	15.40±0.66	9.57±0.78	4.31±0.69	38.82±0.97	15.13±0.81	9.65±0.71	4.55±0.42
BR	Hamming	56.17±1.01	56.20±0.41	55.96±0.90	56.03±0.37	56.36±1.35	55.85±0.70	56.16±0.49	55.99±0.47
	Subset 0/1	42.72±1.03	20.39±1.37	14.67±1.16	10.22±0.63	43.62±1.40	19.90±1.35	15.03±0.80	9.83±0.98
HMDC-MI	Hamming	58.32±0.92	55.34±1.02	54.50±0.78	53.78±0.53	68.60±1.28	65.45±0.53	65.09±0.83	64.02±0.35
	Subset 0/1	48.64±1.37	27.11±1.21	21.74±1.35	16.43±1.66	59.65±1.44	37.70±1.21	31.74±1.22	26.17±0.64
HMDC-OP	Hamming	**66.75±1.22**	**62.28±0.69**	**61.37±0.68**	**60.42±0.35**	**72.09±1.06**	**68.96±0.46**	**68.41±0.61**	**67.43±0.37**
	Subset 0/1	57.01±1.41	34.00±1.17	27.55±0.60	21.55±1.06	63.12±1.14	41.49±1.09	34.83±1.25	29.50±0.66
HMDC-AV	Hamming	**66.75±1.22**	**62.28±0.69**	**61.37±0.68**	**60.42±0.35**	**72.09±1.06**	**68.96±0.46**	**68.41±0.61**	**67.43±0.37**
	Subset 0/1	**57.54±1.39**	**34.46±1.18**	**28.03±0.75**	**22.06±0.73**	**63.48±1.10**	**41.57±0.80**	**35.19±1.03**	**29.56±0.61**
RefC	Hamming	74.88±0.97	72.33±0.63	71.90±0.59	71.37±0.39	74.91±0.95	72.25±0.52	71.99±0.47	71.17±0.30
	Subset 0/1	66.64±1.40	45.56±1.14	39.73±0.92	34.08±1.15	66.60±0.97	45.48±1.20	39.85±0.87	34.12±0.68

Table 18: Detailed results for Adult with Random Forest when allowing at most three parents among discrete variables

Model	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	24.95±0.78	24.96±0.64	24.87±0.63	24.82±0.54	24.88±0.84	24.94±0.63	24.72±0.51	24.87±0.47
	Subset 0/1	15.32±1.08	3.52±0.27	1.89±0.23	0.78±0.17	15.43±1.00	3.85±0.44	1.82±0.36	0.92±0.21
MI	Hamming	52.50±1.23	52.78±0.34	52.43±0.34	52.48±0.31	52.68±1.04	52.54±0.48	52.64±0.35	52.52±0.28
	Subset 0/1	38.03±1.38	15.19±0.60	9.30±0.59	4.28±0.55	38.86±1.44	14.84±0.61	9.33±0.66	4.31±0.29
BR	Hamming	50.77±0.98	50.95±0.90	50.84±0.41	51.00±0.48	51.02±0.62	50.93±0.49	51.25±0.42	51.04±0.51
	Subset 0/1	38.57±1.09	19.72±1.30	15.60±0.87	11.89±0.67	38.99±1.03	19.78±0.76	15.84±0.63	12.09±0.65
HMDC-MI	Hamming	55.06±1.07	53.68±0.79	53.11±0.50	52.69±0.46	62.38±0.99	61.25±0.61	60.92±0.85	60.84±0.50
	Subset 0/1	43.57±1.26	22.44±0.88	17.92±0.59	13.24±0.37	50.70±1.27	28.94±0.67	23.81±0.94	18.77±0.51
HMDC-OP	Hamming	**60.49±0.92**	**58.87±0.65**	**57.77±0.60**	**56.96±0.49**	**64.71±0.95**	**63.42±0.57**	**63.11±0.71**	**62.79±0.45**
	Subset 0/1	48.57±0.95	26.80±0.77	21.51±0.79	16.14±0.90	52.66±1.09	31.28±0.56	25.72±0.74	20.43±0.64
HMDC-AV	Hamming	**60.49±0.92**	**58.87±0.65**	**57.77±0.60**	**56.96±0.49**	**64.71±0.95**	**63.42±0.57**	**63.11±0.71**	**62.79±0.45**
	Subset 0/1	**49.74±1.01**	**28.28±0.66**	**23.11±0.73**	**17.44±0.94**	**53.26±1.06**	**31.96±0.44**	**26.30±0.76**	**21.07±0.61**
RefC	Hamming	67.35±0.92	66.43±0.71	66.20±0.59	65.76±0.41	67.15±0.92	66.54±0.76	66.06±0.63	65.83±0.74
	Subset 0/1	55.87±1.36	33.86±1.31	28.75±0.84	23.32±0.70	55.43±1.25	34.04±0.73	28.21±0.69	23.06±0.83

Table 19: Detailed results for Adult with Naive Bayes when allowing at most three parents among discrete variables

Model	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RandC	Hamming	25.22±0.76	24.70±0.58	24.77±0.57	24.89±0.54	24.72±0.84	24.88±0.52	24.86±0.53	24.95±0.46
	Subset 0/1	15.72±0.62	3.76±0.59	1.96±0.25	0.78±0.19	15.56±0.77	3.73±0.28	1.97±0.30	0.82±0.22
MI	Hamming	52.26±0.90	52.59±0.21	52.52±0.36	52.58±0.30	52.66±1.08	52.60±0.48	52.61±0.38	52.69±0.43
	Subset 0/1	38.43±1.01	15.04±0.59	9.17±0.51	4.34±0.57	38.67±1.26	15.02±0.95	9.65±0.68	4.58±0.65
BR	Hamming	47.22±3.71	47.35±3.53	47.12±3.62	47.09±3.58	47.76±3.48	47.01±3.59	47.20±3.83	47.18±3.62
	Subset 0/1	35.03±3.59	16.16±2.89	12.08±2.68	8.68±2.41	35.02±3.12	15.92±3.11	12.22±2.80	8.76±2.49
HMDC-MI	Hamming	52.50±3.04	**52.74±2.62**	**52.90±2.59**	**52.90±2.59**	**59.09±2.99**	**59.00±3.09**	**58.90±2.79**	**58.81±2.82**
	Subset 0/1	38.94±2.64	18.53±1.91	13.53±1.50	13.53±1.50	46.01±3.37	23.87±2.31	18.29±2.04	12.95±1.32
HMDC-OP	Hamming	**53.91±2.12**	51.99±1.64	51.48±1.42	51.48±1.42	59.03±2.47	58.53±2.37	58.18±2.18	58.17±2.08
	Subset 0/1	39.79±2.12	18.25±1.91	13.19±1.55	13.19±1.55	45.82±2.86	23.56±1.55	18.04±1.37	**13.16±0.90**
HMDC-AV	Hamming	**53.91±2.12**	51.99±1.64	51.48±1.42	51.48±1.42	59.03±2.47	58.53±2.37	58.18±2.18	58.17±2.08
	Subset 0/1	**41.67±2.02**	**19.52±1.69**	**14.06±1.30**	**14.06±1.30**	**46.35±2.68**	**23.97±1.51**	**18.40±1.28**	**13.16±0.92**
RefC	Hamming	62.18±3.23	62.39±2.83	62.43±2.72	62.43±2.72	62.41±3.09	62.57±3.11	62.33±2.98	62.44±2.84
	Subset 0/1	48.83±3.12	26.75±2.48	20.86±1.82	20.86±1.82	49.47±3.43	26.85±2.39	21.05±2.50	15.57±1.82

Table 20: Detailed results for Default with Logistic Regression when allowing at most three parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RanC	Hamming	34.75±0.79	34.76±0.50	34.60±0.34	34.68±0.31	35.02±0.82	34.82±0.46	34.75±0.49	34.70±0.43
	Subset 0/1	23.13±0.97	7.31±0.36	4.32±0.41	2.20±0.15	23.68±0.67	7.13±0.43	4.05±0.40	2.24±0.22
MI	Hamming	59.77±0.69	60.04±0.27	59.99±0.39	60.08±0.30	59.97±0.67	60.08±0.36	60.00±0.35	59.95±0.31
	Subset 0/1	46.56±0.88	25.25±0.50	19.83±0.38	14.91±0.48	46.73±1.20	25.35±0.38	19.58±0.62	14.94±0.54
BR	Hamming	65.94±0.59	65.95±0.30	66.34±0.50	66.19±0.38	65.99±0.64	66.19±0.54	66.31±0.42	66.24±0.56
	Subset 0/1	53.62±0.76	31.88±0.60	26.94±0.91	21.92±0.68	53.52±1.08	32.45±0.63	26.93±0.88	22.21±0.95
HMDC-MI	Hamming	66.20±1.12	66.09±0.57	65.88±0.49	66.02±0.80	66.45±0.81	66.12±0.98	66.38±0.87	65.82±1.32
	Subset 0/1	54.33±1.46	32.96±0.77	27.16±0.85	22.45±1.01	54.45±0.98	32.90±0.80	27.55±1.02	22.80±1.13
HMDC-OP	Hamming	**66.58±0.96**	**66.63±0.31**	**66.56±0.29**	**66.58±0.45**	**66.74±0.69**	**66.67±0.41**	**66.83±0.47**	**66.50±0.27**
	Subset 0/1	**54.71±1.22**	33.63±0.56	27.94±0.55	**22.96±0.80**	**54.85±0.85**	33.42±0.47	**28.05±0.79**	**23.21±0.94**
HMDC-AV	Hamming	**66.58±0.96**	**66.63±0.31**	**66.56±0.29**	**66.58±0.45**	**66.74±0.69**	**66.67±0.41**	**66.83±0.47**	**66.50±0.27**
	Subset 0/1	54.63±1.30	**33.64±0.59**	**27.96±0.58**	22.95±0.83	**54.85±0.89**	**33.43±0.47**	28.03±0.81	23.20±0.94
RefC	Hamming	66.87±0.95	67.02±0.33	66.82±0.30	66.81±0.57	66.83±0.65	66.78±0.38	66.93±0.49	66.61±0.28
	Subset 0/1	54.89±1.20	34.06±0.66	28.26±0.51	23.23±0.93	55.00±0.86	33.58±0.40	28.16±0.77	23.34±0.95

Table 21: Detailed results for Default with Random Forest when allowing at most three parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RanC	Hamming	35.22±0.64	34.95±0.34	34.85±0.39	34.72±0.33	34.74±0.86	34.79±0.37	34.75±0.35	34.73±0.32
	Subset 0/1	23.87±0.57	7.67±0.47	4.29±0.34	2.12±0.10	22.83±1.13	7.21±0.42	4.29±0.32	2.23±0.11
MI	Hamming	60.19±0.40	60.12±0.33	59.97±0.30	59.97±0.24	60.01±0.63	59.81±0.38	60.06±0.33	60.03±0.29
	Subset 0/1	47.20±0.65	25.08±0.64	19.61±0.70	15.10±0.41	46.74±1.14	24.65±0.62	19.75±0.53	15.10±0.53
BR	Hamming	63.33±0.61	63.09±0.33	63.12±0.25	63.26±0.30	63.53±0.60	63.23±0.49	63.14±0.38	63.17±0.27
	Subset 0/1	50.87±0.88	29.34±0.74	24.32±0.81	19.76±0.66	51.24±0.88	29.23±0.51	24.22±0.62	19.36±0.61
HMDC-MI	Hamming	53.94±0.73	54.45±0.64	55.11±0.61	55.06±0.57	56.98±0.86	57.12±0.63	57.40±0.46	57.78±0.61
	Subset 0/1	42.12±1.10	21.85±0.80	17.85±0.52	13.80±0.43	44.53±1.26	23.86±0.74	19.36±0.57	15.53±0.55
HMDC-OP	Hamming	**61.62±0.71**	**62.95±0.40**	**63.34±0.46**	**63.39±0.42**	**61.07±0.87**	**62.09±0.50**	**62.27±0.58**	**62.46±0.38**
	Subset 0/1	48.54±0.87	28.06±0.70	23.21±0.67	18.58±0.79	**48.11±0.93**	**27.48±0.83**	22.38±0.60	18.17±0.42
HMDC-AV	Hamming	**61.62±0.71**	**62.95±0.40**	**63.34±0.46**	**63.39±0.42**	**61.07±0.87**	**62.09±0.50**	**62.27±0.58**	**62.46±0.38**
	Subset 0/1	**48.62±0.88**	**28.48±0.76**	**23.57±0.93**	**19.04±0.74**	**48.11±0.91**	27.44±0.90	**22.52±0.58**	**18.32±0.48**
RefC	Hamming	60.14±0.80	61.12±0.50	61.54±0.47	61.75±0.33	60.57±0.78	61.42±0.64	61.57±0.52	61.85±0.47
	Subset 0/1	47.13±1.11	26.57±0.76	22.05±0.69	17.45±0.73	47.58±1.02	26.82±0.87	21.84±0.71	17.73±0.58

Table 22: Detailed results for Default with Naive Bayes when allowing at most three parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RanC	Hamming	34.78±0.42	34.65±0.47	34.88±0.44	34.84±0.30	34.85±0.78	34.91±0.40	34.64±0.43	34.70±0.45
	Subset 0/1	23.36±0.65	7.16±0.36	4.49±0.41	2.19±0.18	23.19±0.92	7.48±0.36	4.17±0.29	2.16±0.15
MI	Hamming	59.81±0.49	60.02±0.52	60.11±0.40	60.03±0.35	60.02±0.64	59.93±0.39	59.90±0.46	59.99±0.34
	Subset 0/1	46.65±0.78	25.11±0.96	19.76±0.66	15.01±0.56	46.89±0.93	25.27±0.64	19.63±0.49	15.11±0.46
BR	Hamming	54.61±2.93	54.63±2.93	54.69±3.09	54.54±3.09	54.43±2.68	54.56±3.16	54.51±3.04	54.67±3.05
	Subset 0/1	40.52±3.63	17.86±3.71	12.58±3.68	7.77±3.61	40.69±3.26	17.64±4.04	12.25±3.87	8.11±3.63
HMDC-MI	Hamming	56.73±0.87	56.83±0.46	56.73±0.53	56.83±0.41	58.80±0.73	59.03±0.47	59.03±0.56	59.20±0.43
	Subset 0/1	43.99±1.17	22.87±0.84	18.01±0.52	13.47±0.66	45.65±0.72	24.32±0.45	19.12±0.73	14.77±0.48
HMDC-OP	Hamming	**60.69±0.78**	**60.68±0.75**	**60.58±0.59**	**60.60±0.61**	**60.42±0.71**	**60.48±0.65**	**60.59±0.70**	**60.61±0.58**
	Subset 0/1	47.43±0.97	25.17±0.72	20.06±0.65	15.28±0.49	47.12±0.98	**25.29±0.57**	**20.11±0.92**	**15.41±0.47**
HMDC-AV	Hamming	**60.69±0.78**	**60.68±0.75**	**60.58±0.59**	**60.60±0.61**	**60.42±0.71**	**60.48±0.65**	**60.59±0.70**	**60.61±0.58**
	Subset 0/1	**47.44±0.96**	**25.20±0.71**	**20.09±0.67**	**15.29±0.50**	**47.17±0.95**	**25.29±0.57**	20.09±0.92	**15.41±0.47**
RefC	Hamming	60.74±0.76	60.68±0.75	60.59±0.59	60.61±0.62	60.41±0.73	60.50±0.65	60.61±0.69	60.62±0.58
	Subset 0/1	47.50±0.96	25.18±0.73	20.08±0.65	15.29±0.47	47.18±0.96	25.30±0.53	20.15±0.93	15.41±0.47

Table 23: Detailed results for Thyroid with Logistic Regression when allowing at most three parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RanC	Hamming	33.38±1.90	33.36±1.22	33.43±1.26	33.57±1.45	33.24±1.49	33.41±1.53	33.59±1.36	33.56±1.54
	Subset 0/1	14.04±1.91	1.16±0.22	0.48±0.12	0.12±0.10	13.85±1.17	1.22±0.41	0.38±0.16	0.13±0.08
MI	Hamming	**95.89±0.56**	**95.99±0.27**	96.01±0.23	96.00±0.22	96.09±0.38	96.00±0.31	95.99±0.19	96.00±0.23
	Subset 0/1	90.86±1.26	81.17±1.36	78.72±1.38	76.26±1.54	91.17±0.89	81.20±1.62	78.62±1.22	76.17±1.54
BR	Hamming	96.47±0.45	96.34±0.19	96.31±0.28	96.31±0.27	96.33±0.35	96.42±0.32	96.34±0.31	96.31±0.26
	Subset 0/1	92.11±1.04	82.98±1.04	80.55±1.54	78.31±1.81	91.83±0.79	83.27±1.48	80.74±1.77	78.40±1.64
HMDC-MI	Hamming	96.37±0.44	96.35±0.29	96.40±0.25	96.36±0.22	96.59±0.37	96.55±0.26	96.52±0.25	96.51±0.27
	Subset 0/1	91.95±1.06	83.02±1.42	80.94±1.35	78.59±1.57	92.31±0.84	83.98±1.34	81.64±1.55	79.44±1.70
HMDC-AV	Hamming	96.40±0.42	96.37±0.30	**96.42±0.25**	**96.38±0.23**	**96.62±0.37**	**96.58±0.25**	**96.56±0.25**	**96.53±0.28**
	Subset 0/1	91.99±1.02	**83.14±1.39**	**81.14±1.35**	**78.67±1.60**	**92.38±0.82**	**84.10±1.28**	**81.84±1.58**	**79.58±1.68**
RefC	Hamming	96.67±0.43	96.56±0.33	96.58±0.29	96.60±0.25	96.72±0.35	96.63±0.23	96.62±0.26	96.57±0.28
	Subset 0/1	92.61±0.99	84.09±1.65	81.97±1.63	80.03±1.62	92.62±0.78	84.41±1.21	82.15±1.64	79.87±1.72

Table 24: Detailed results for Thyroid with Random Forest when allowing at most three parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RanC	Hamming	33.70±1.74	33.59±1.39	33.52±1.59	33.70±1.51	34.02±1.37	33.71±1.32	33.48±1.47	33.53±1.43
	Subset 0/1	13.98±1.35	1.13±0.35	0.38±0.19	0.17±0.16	14.26±1.23	0.76±0.28	0.50±0.25	0.04±0.07
MI	Hamming	96.01±0.24	95.96±0.27	96.01±0.22	95.96±0.22	95.63±0.30	96.03±0.23	96.06±0.29	96.00±0.23
	Subset 0/1	91.06±0.51	81.17±1.43	78.78±1.45	76.06±1.57	90.18±0.68	81.53±1.25	79.06±1.78	76.29±1.45
BR	Hamming	97.85±0.34	97.97±0.17	97.98±0.17	97.97±0.20	98.01±0.36	98.05±0.26	97.98±0.20	97.95±0.17
	Subset 0/1	95.52±0.69	91.61±0.59	90.73±0.71	89.73±0.93	95.73±0.75	91.82±0.85	90.51±0.78	89.73±0.77
HMDC-MI	Hamming	98.15±0.20	97.95±0.33	97.97±0.23	97.89±0.22	98.41±0.37	98.44±0.20	98.41±0.11	98.38±0.15
	Subset 0/1	96.04±0.43	90.98±1.37	90.26±1.15	88.67±1.26	96.55±0.83	93.04±0.89	92.22±0.48	91.02±0.76
HMDC-AV	Hamming	**98.40±0.14**	**98.15±0.27**	**98.12±0.18**	**98.14±0.20**	**98.53±0.33**	**98.57±0.18**	**98.51±0.15**	**98.49±0.13**
	Subset 0/1	**96.66±0.35**	**92.53±1.04**	**91.49±0.72**	**90.39±1.09**	**96.80±0.74**	**93.68±0.87**	**92.86±0.69**	**92.03±0.64**
RefC	Hamming	98.68±0.13	98.62±0.21	98.62±0.13	98.62±0.14	98.67±0.31	98.72±0.15	98.68±0.12	98.68±0.11
	Subset 0/1	97.15±0.29	93.93±0.74	93.04±0.67	92.20±0.87	97.13±0.67	94.16±0.71	93.33±0.62	92.65±0.64

Table 25: Detailed results for Thyroid with Naive Bayes when allowing at most three parents among discrete variables

Category	Metric	Missing rate on discrete features = 80%				Missing rate on discrete features = 30%			
		30%	70%	80%	90%	30%	70%	80%	90%
RanC	Hamming	33.39±1.49	33.41±1.37	33.44±1.50	33.52±1.53	33.42±1.96	33.70±1.56	33.65±1.53	33.51±1.38
	Subset 0/1	13.49±1.04	1.26±0.29	0.29±0.22	0.13±0.12	14.11±2.21	1.31±0.39	0.48±0.25	0.10±0.12
MI	Hamming	96.06±0.27	96.04±0.30	**96.05±0.25**	95.98±0.26	95.76±0.34	**95.96±0.32**	**95.98±0.21**	**95.99±0.20**
	Subset 0/1	91.10±0.57	81.56±1.52	79.23±1.66	76.15±1.77	90.45±0.80	81.12±1.65	78.58±1.26	76.25±1.26
BR	Hamming	55.67±3.45	56.32±3.77	56.29±3.67	56.29±3.67	56.26±4.01	56.07±3.68	56.17±3.68	56.30±3.73
	Subset 0/1	29.74±2.89	6.17±1.12	3.55±0.70	2.13±0.56	29.92±3.38	6.04±1.05	3.43±0.49	2.00±0.44
HMDC-MI	Hamming	95.45±0.49	95.37±0.36	95.50±0.44	95.65±0.40	96.13±0.32	94.28±1.19	95.24±0.41	95.79±0.36
	Subset 0/1	90.52±0.97	81.37±1.53	80.53±1.48	78.74±1.70	91.89±0.56	75.64±5.56	78.25±1.73	78.12±1.86
HMDC-AV	Hamming	**96.54±0.37**	**96.32±0.34**	95.98±0.41	**96.11±0.35**	**96.27±0.36**	95.10±0.79	95.62±0.44	95.92±0.32
	Subset 0/1	**92.61±0.74**	**84.25±1.55**	**82.63±1.43**	**80.82±1.54**	**92.13±0.75**	**79.56±3.51**	**80.08±1.77**	**78.93±1.65**
RefC	Hamming	96.33±0.46	96.05±0.29	96.02±0.43	95.94±0.31	96.20±0.39	94.64±1.03	95.56±0.37	95.80±0.42
	Subset 0/1	92.05±1.02	82.78±1.26	82.22±1.60	78.96±1.74	91.98±0.84	77.01±4.85	79.64±1.56	77.92±2.14

E.4 Examples of Learned DAGs

The optimal DAGs learned from training data when allowing at most 2 and 3 discrete parents are given in Section E.4.1 and E.4.2, respectively. For each combination of base classifier, dataset, and missingness levels on the discrete features (30%, 80%) and the class variables (30%, 70%, 80%, and 90%), we plot the DAG $G^{\mathcal{D}}$ learned from the training data in the last train-test split. Overall, we observe that the DAGs produced by Random Forest (RF) and Naive Bayes (NB) are, respectively, the sparsest and densest.

As a consequence, using NB as the base classifier tends to increase the cost of performing MAP queries, while often providing poor classification performance. In contrast, using RF as the base classifier tends to lower the cost of performing MAP queries. However, this can degrade the classification performance of HMDCs, compared to the use of LR as the base classifier.

From a practical perspective, we suggest using LR as the base classifier when seeking a trade-off between the cost of performing MAP queries and the classification performance.

E.4.1 Results with at most 2 parents among discrete variables

Figures 13 to 15 present the DAG structures obtained when allowing at most 2 among discrete variables (palim = 2). Each DAG displayed is an example taken from the ten DAGs generated during the 10-fold cross-validation process in the training phase.

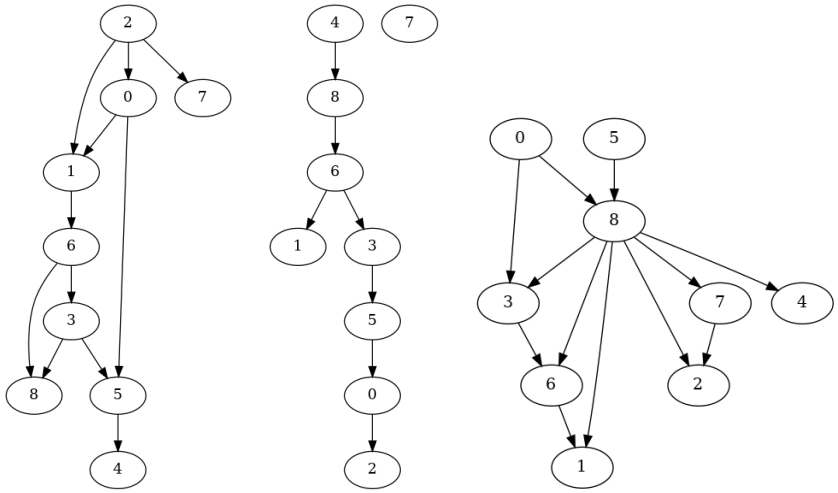


Figure 13: Adult dataset, class variables are labeled $\{0, 1, 2, 3\}$, at most two parents are allowed among discrete variables

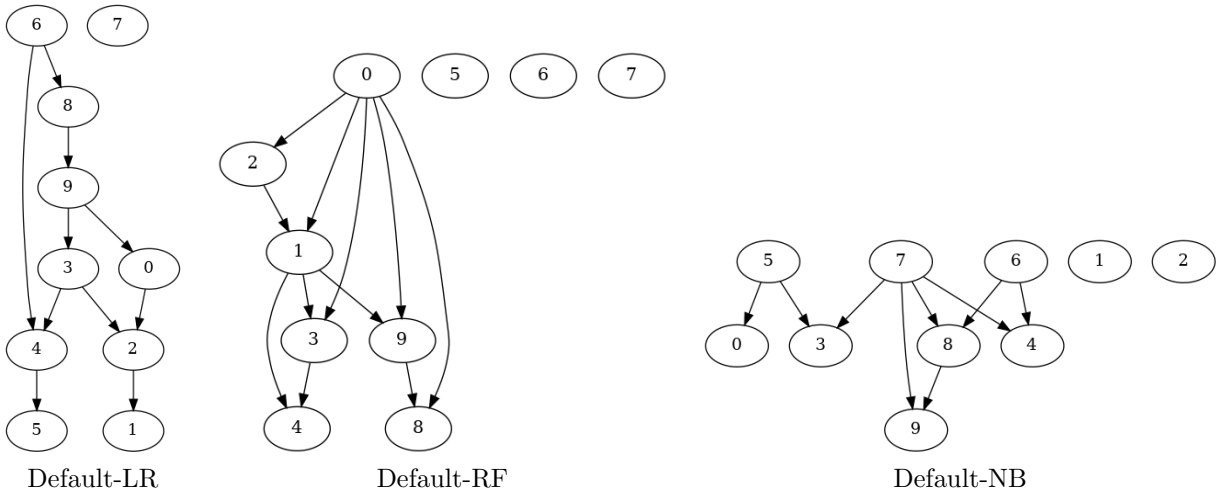
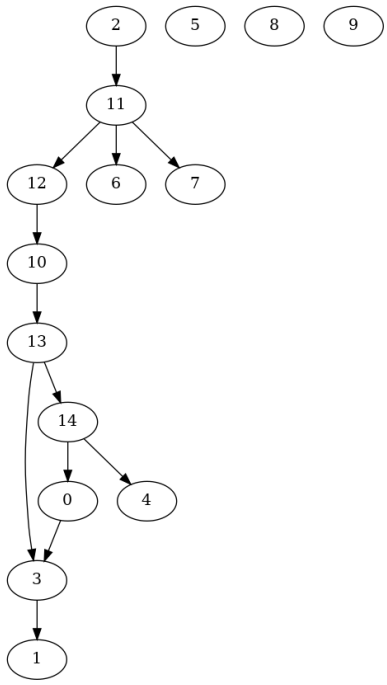
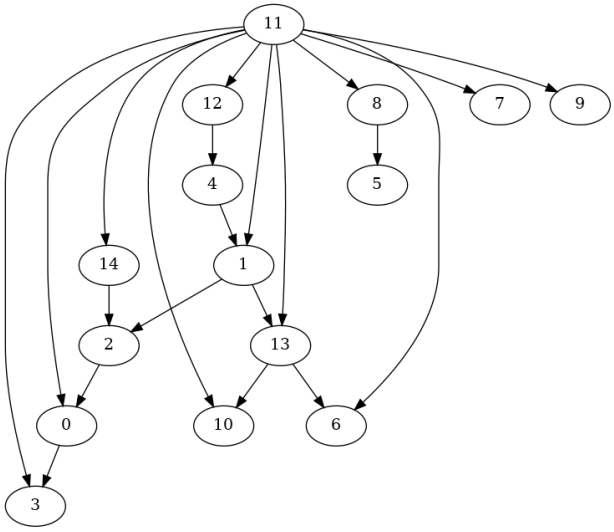


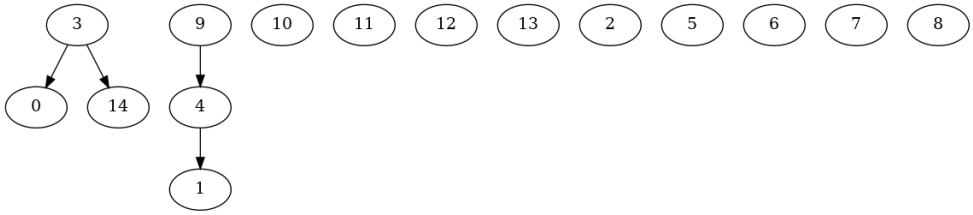
Figure 14: Default dataset, class variables are labeled $\{0, 1, 2, 3\}$, at most two parents are allowed among discrete variables



Thyroid–LR



Thyroid–NB

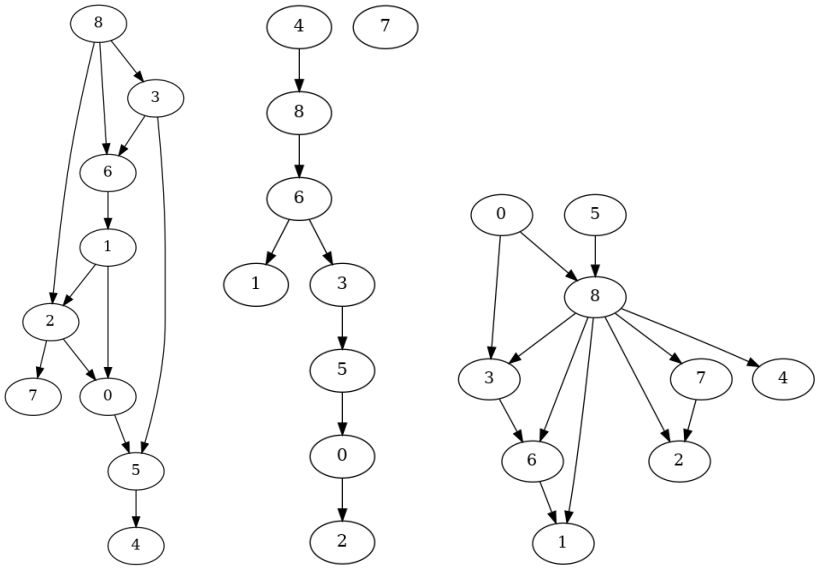


Thyroid–RF

Figure 15: Thyroid dataset, class variables are labeled $\{0, 1, 2, 3, 4, 5, 6\}$, at most two parents are allowed among discrete variables

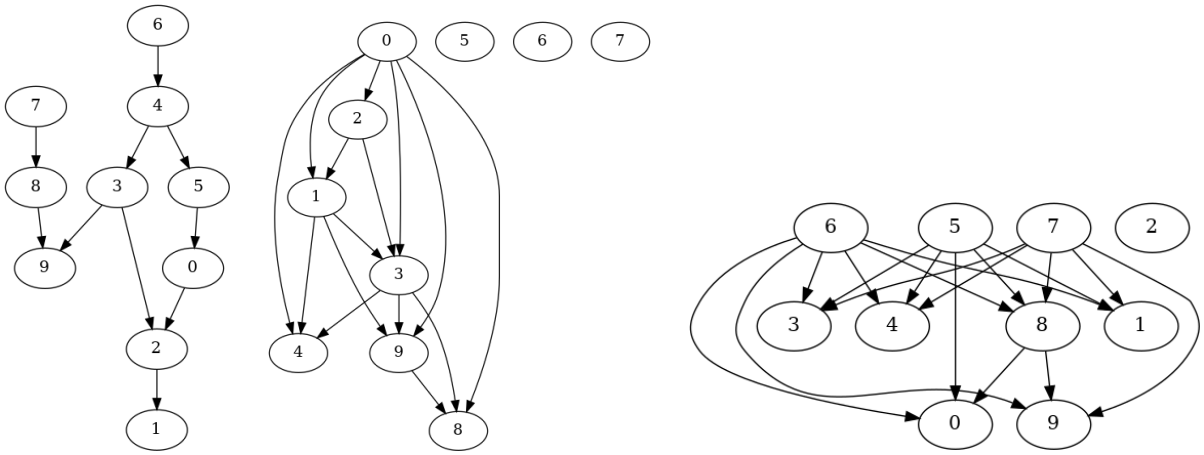
E.4.2 Results with at most 3 parents among discrete variables

Figures 16 to 18 present the DAG structures when allowing at most 3 parents among discrete variables (palim=3). Each DAG displayed is an example taken from the ten DAGs generated during the 10-fold cross-validation process in the training phase.



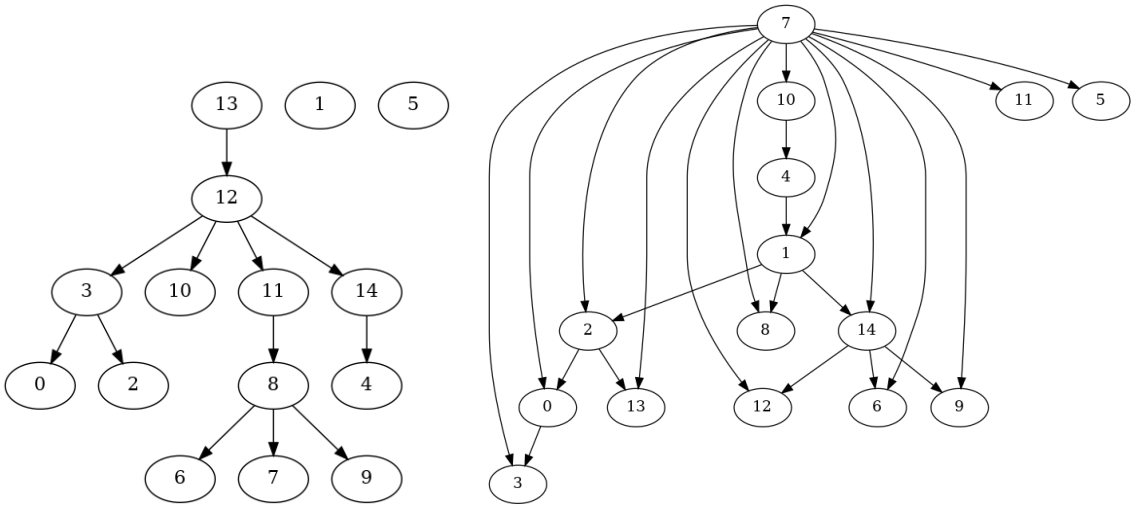
Adult-LR Adult-RF Adult-NB

Figure 16: Adult dataset, class variables are labeled $\{0, 1, 2, 3\}$, at most three parents are allowed among discrete variables

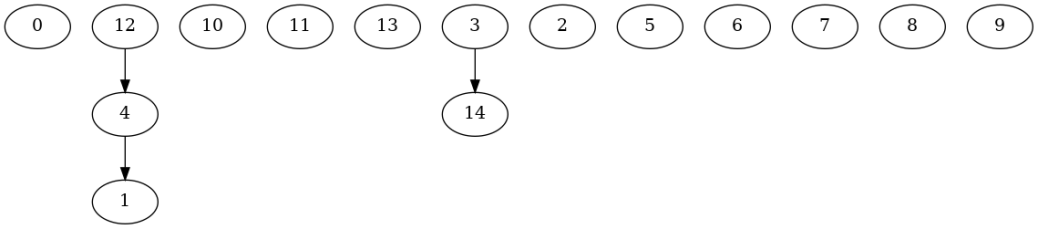


Default-LR Default-RF Default-NB

Figure 17: Default dataset, class variables are labeled $\{0, 1, 2, 3\}$, at most three parents are allowed among discrete variables



Thyroid–LR Thyroid–NB



Thyroid–RF

Figure 18: Thyroid dataset, class variables are labeled $\{0, 1, 2, 3, 4, 5, 6\}$, at most three parents are allowed among discrete variables