

Probabilistic Multi-dimensional Classification with Incomplete Data

Thu Ha Do

Université de Technologie de Compiègne

Kim-Dung Tran

Université de Technologie de Compiègne

Vu-Linh Nguyen

Université de Technologie de Compiègne

Yves Grandvalet

`yves.grandvalet@utc.fr`

Université de Technologie de Compiègne

Sébastien Destercke

Université de Technologie de Compiègne

Research Article

Keywords: Probabilistic multi-dimensional classification, Probabilistic graphical models, Mixed data, Incomplete data, Supervised learning

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10976302/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Probabilistic Multi-dimensional Classification with Incomplete Data

Thu Ha Do¹, Kim-Dung Tran¹, Vu-Linh Nguyen¹,
Yves Grandvalet^{1*}, Sébastien Destercke¹

^{1*}Université de technologie de Compiègne, CNRS, Heudiasyc,
Compiègne, France.

*Corresponding author(s). E-mail(s): yves.grandvalet@utc.fr;
Contributing authors: thu-ha.do@utc.fr; kim-dung.tran@utc.fr;
vu-linh.nguyen@utc.fr; sebastien.destercke@utc.fr;

Abstract

Multi-dimensional classification (MDC) generalizes multi-class and multi-label classification by considering the simultaneous prediction of multiple class variables for each instance. In this paper, we revisit probabilistic MDC approaches for mixed data, where both discrete and continuous attributes are present, and examine their potential strengths and limitations when extended to handle incomplete data, a setting that remains largely underexplored. We then propose a novel approach to provide more scalable and robust probabilistic classifiers. We detail both the learning and prediction phases of our approach, deriving theoretical results that underpin the design of efficient algorithms for each phase. Empirical evidence is then presented to illustrate the robustness of this solution under different missingness scenarios.

Keywords: Probabilistic multi-dimensional classification, Probabilistic graphical models, Mixed data, Incomplete data, Supervised learning

1 Introduction

Multi-label classification predicts several binary outcomes simultaneously for each instance; multi-dimensional classification extends this framework to handle multiple outcomes that can each take more than two categorical values. MDC applies to various domains ([Gil-Begue et al, 2021](#); [Jia and Zhang, 2024](#); [Ma and Chen, 2018](#)), such

as predicting multiple related health indices. The categorical indices can be nominal, such as 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).

Graphical models are a go-to choice to perform MDC, called Probabilistic MDC (PMDC), yet 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.

In this context, discriminative MDC (DMDC) (Nguyen et al, 2023) balances scalability and expressiveness. This model handles properly mixed features (discrete and continuous) by constraining discrete-continuous dependencies. It optimizes a joint conditional distribution; however, inference remains NP-hard in the number of discrete variables. This model can achieve strong predictive performance when all features are observed, but it is not designed to natively handle incomplete data.

Yet, data may be missing either during training or at prediction time in many practical applications. Such a setting has already been investigated, mostly under assumptions that exclude mixed feature types and multiple class variables (García-Laencina et al, 2010). This work tackles the underexplored challenge of missing discrete variables in mixed data scenarios, which is particularly challenging in settings where both scalability and model expressivity are critical.

We extend DMDC by relaxing the structural constraint on discrete-continuous links, which allows for the processing of incomplete data. We provide optimality results that extend the ones already known for DMDC (Nguyen et al, 2023), with an implementation that can simply rely on any learning algorithm for probabilistic multi-class classification, provided it supports instance weights. Our prediction algorithms retain the computational complexity of DMDC while improving robustness to missing discrete variables. We also derive the Bayesian-optimal predictors for the Hamming and 0/1 losses, showing that they reduce to maximum a posteriori (MAP) inference in several cases.

This work extends our previous contribution (Do et al, 2026), which introduced Hybrid MDCs (HMDCs) and developed learning and inference under the assumption that missingness occurs only at prediction time. The present paper substantially expands that work in several respects. First, it introduces a new learning framework for incomplete training data. Second, it provides a more comprehensive experimental evaluation covering a wider range of missing-data scenarios during both training and prediction. For readability, all formal results (propositions) are stated without proofs, which are deferred to the Appendix B.

The paper is organized as follows. Section 2 reviews probabilistic MDCs for mixed data, covering the general framework, existing classifier subfamilies, and the hypothesis space of the proposed model, Hybrid MDCs (HMDCs). Section 3 then presents the

missingness scenarios and the principles governing the different options for learning and inference with incomplete data. Sections 4 and 5 respectively discuss the learning and prediction tasks for HMDCs under these various scenarios. Section 6 presents the experimental setup and results. Finally, Section 7 concludes the paper.

2 Probabilistic MDCs

PMDCs are essentially Bayesian networks (BNs). As such, their three key components are *representation*, *learning*, and *inference*. This section discusses how representation affects learning and inference, first reviewing the standard generative and discriminative approaches, then introducing our hybrid method.

2.1 Generalities

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

Next, let $\mathcal{Y} := \mathcal{Y}^1 \times \dots \times \mathcal{Y}^K$ be the set of possible outcomes, and let $\mathbf{Y} = \{Y^1, \dots, Y^K\}$ be the set of the K class variables. For any $k \in [K] := \{1, \dots, K\}$, let $\mathcal{Y}^k := \{y_1^k, \dots, y_{M_k}^k\}$ 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 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathbf{Z} = \mathbf{Y} \cup \mathbf{X}$ and $\mathbf{Z}^D = \mathbf{Y} \cup \mathbf{X}^D$.

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 θ . For all $\mathbf{z} \in \mathcal{Z}$, the distribution factorizes as:

$$\begin{aligned} p_{(G, \theta)}(\mathbf{z}) &= \prod_{Z^j \in \mathbf{Z}} p_{(G, \theta)}(Z^j = z^j \mid \text{pa}(Z^j) = \text{pa}(z^j)) \\ &=: \prod_{z^j \in \mathcal{Z}^j} p_{(G, \theta)}(z^j \mid \text{pa}(z^j)) , \end{aligned}$$

where $\text{pa}(Z^j)$ denotes the parent set of Z^j in the DAG G , while $\text{pa}(z^j)$ is the specific configuration of $\text{pa}(Z^j)$ specified by $\mathbf{z}$. 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) are BNs whose DAGs are restricted by excluding arcs originating from features and directed to class variables, as illustrated in Figure 1. They are adjusted by maximizing the joint likelihood over data

$$\prod_{n=1}^N p_{(G, \theta)}(\mathbf{z}_n) ,$$

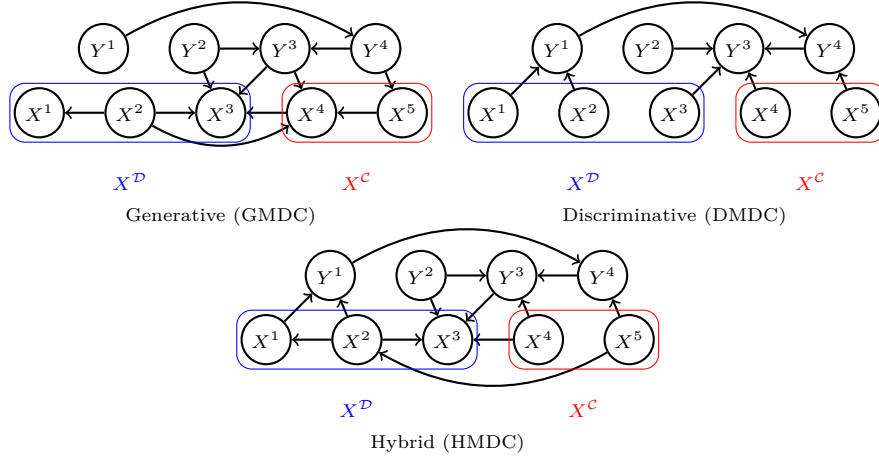


Fig. 1: Examples of DAGs illustrating the PMDC types: GMDC, DMDC, and HMDC

where $\mathcal{S} := \{\mathbf{z}_n\}_{n=1}^N = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ is the training set. While the model is expressive, exact MAP inference over $\mathcal{Y}$ 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) are BN classifiers whose DAGs are restricted by excluding class-to-feature arcs and feature-to-feature arcs from class to feature variables, as illustrated in Figure 1. They are adjusted by maximizing the conditional likelihood over the training data

$$\prod_{n=1}^N p_{(G, \theta)}(\mathbf{y}_n | \mathbf{x}_n) .$$

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. Exact MAP inference remains NP-hard in the number of class variables, although it remains computationally tractable for the MDC problems considered in this work.

2.3 Hybrid PMDCs

GMDCs model the joint distribution $p_{(G, \theta)}(\mathbf{Z})$ 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 often crude a workaround (Sucar, 2021).

Moreover, GMDCs optimize the joint likelihood, which is not discriminant and may thus perform poorly for classification purposes, where the data distribution cannot be represented with the BN (Roos et al, 2005).

DMDCs are discriminant in the sense that they model the conditional distribution $p_{(G,\theta)}(\mathbf{Y} | \mathbf{X}^{\mathcal{D}}, \mathbf{X}^{\mathcal{C}})$, and they are computationally efficient. Yet they cannot naturally handle missing discrete features at prediction time, as this requires an estimate of $p_{(G,\theta)}(\mathbf{Y}, \mathbf{X}^{\mathcal{D}} | \mathbf{X}^{\mathcal{C}})$.

2.3.1 Structural Constraints

To address these limitations, we introduce Hybrid PMDCs (HMDCs), a minimal extension of DMDCs that can naturally handle 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

$$\prod_{n=1}^N p_{(G,\theta)}(\mathbf{y}_n, \mathbf{x}_n^{\mathcal{D}} | \mathbf{x}_n^{\mathcal{C}}) , \quad (1)$$

where $\mathbf{x}_n^{\mathcal{D}}$ and $\mathbf{x}_n^{\mathcal{C}}$ are respectively the discrete and continuous features of $\mathbf{x}_n$. 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 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 discrete variables, namely 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

$$L_{\Omega}(G, \theta | \mathcal{S}) = \sum_{n=1}^N \log (p_{(G,\theta)}(\mathbf{y}_n, \mathbf{x}_n^{\mathcal{D}} | \mathbf{x}_n^{\mathcal{C}})) \quad (2)$$

with Ω the parameter space.

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

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

where the discriminant components $\mathbf{p}(\mathbf{y}_n | \mathbf{x}_n^{\mathcal{D}}, \mathbf{x}_n^{\mathcal{C}})$ are weighted by $\mathbf{p}(\mathbf{x}_n^{\mathcal{D}} | \mathbf{x}_n^{\mathcal{C}})$. 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 Missingness

HMDC enables us to extend existing approaches for handling incomplete discrete data to the MDC setting. Our framework covers several practical situations:

- (i) missing observations of some classes: for some instances, some class variables are observed, and the others are missing;
- (ii) incomplete discrete features: for some instances, some discrete features are missing;
- (iii) combined scenarios.

We consider missingness during both the training and inference phases. Throughout this paper, we assume that data are Missing At Random (MAR). Although the validity of the MAR assumption depends on the application, it is widely adopted in the missing-data literature. It avoids specifying a particular missingness mechanism (which would itself be application-dependent), and provides a principled basis for theoretical analysis and experimental evaluation. Moreover, addressing MAR missingness is a necessary first step before tackling more complex missingness mechanisms.

Before presenting the proposed learning and inference procedures, we introduce the notation used throughout the paper for MDC with mixed features and incomplete data. We emphasize that all subsequent notions are defined on an instance-by-instance basis, *i.e.*, they are specific to the instance $\mathbf{z}$.

For an instance $\mathbf{z}$, let $\mathbf{Y}^{\mathcal{M}}$ and $\mathbf{Y}^{\mathcal{O}}$ denote the sets of missing and observed class variables, respectively. Similarly, let $\mathbf{X}^{\mathcal{M}}$ and $\mathbf{X}^{\mathcal{O}}$ denote the sets of missing and observed discrete features, with $\mathbf{X}^{\mathcal{D}} = \mathbf{X}^{\mathcal{M}} \cup \mathbf{X}^{\mathcal{O}}$. We define $\mathbf{Z}^{\mathcal{M}} = \mathbf{Y}^{\mathcal{M}} \cup \mathbf{X}^{\mathcal{M}}$ and $\mathbf{Z}^{\mathcal{O}} = \mathbf{Y}^{\mathcal{O}} \cup \mathbf{X}^{\mathcal{O}}$ as the sets of missing and observed discrete variables, respectively, so that $\mathbf{Z}^{\mathcal{D}} = \mathbf{Z}^{\mathcal{M}} \cup \mathbf{Z}^{\mathcal{O}}$. We also define the shorthand notations $\mathbf{z}^{\mathcal{M}} = \mathbf{y}^{\mathcal{M}} \oplus \mathbf{x}^{\mathcal{M}}$ and $\mathbf{z}^{\mathcal{O}} = \mathbf{y}^{\mathcal{O}} \oplus \mathbf{x}^{\mathcal{O}}$, where $\oplus$ stands for vector concatenation.

Under missing data, the conditional log-likelihood in (2) is no longer directly applicable. Nevertheless, several decision principles can be considered for resolving the uncertainty induced by the missing data (Guillaume et al, 2017; Hüllermeier and Beringer, 2006; Hüllermeier et al, 2019; Koller and Friedman, 2009; Nguyen et al, 2018, 2021; Utkin and Coolen, 2011; Cooper and Herskovits, 1992): the *optimistic*, *pessimistic*, and *averaging* principles. Let

$$S = \prod_{\mathbf{z} \in \mathcal{S}} \prod_{j \in \mathcal{M}} |\mathcal{Z}^j| \quad (3)$$

be the number of completed data sets (*i.e.*, imputed data sets) $\{\hat{\mathcal{S}}^{(s)}\}_{s=1}^S$ that can be constructed from the incomplete training set $\mathcal{S}$. The optimistic, pessimistic, and averaging principles can be implemented as follows, respectively:

$$(G^{\text{opt}}, \theta^{\text{opt}}) \in \operatorname{argmax}_{(G, \theta)} \max_{s \in \{1, \dots, S\}} L_{\Omega}(G, \theta | \hat{\mathcal{S}}^{(s)}) \quad (4)$$

$$(G^{\text{pes}}, \theta^{\text{pes}}) \in \operatorname{argmax}_{(G, \theta)} \min_{s \in \{1, \dots, S\}} L_{\Omega}(G, \theta | \hat{\mathcal{S}}^{(s)}) \quad (5)$$

$$(G^{\text{ave}}, \theta^{\text{ave}}) \in \operatorname{argmax}_{(G, \theta)} \sum_{s=1}^S L_{\Omega}(G, \theta | \hat{\mathcal{S}}^{(s)}) . \quad (6)$$

Note that the optimistic and pessimistic principles are akin to imputation-based approaches, as they return both an optimal classifier and a completed dataset, either maximally or minimally consistent with the selected model. The averaging principle only returns an optimal classifier and can be linked to a Laplacian assumption where all completions are uniformly distributed, which usually do not correspond to any specific completed dataset. In this work, we will mainly focus on the optimistic and averaging principles, for two main reasons: they are computationally friendlier than the max/min formulation of the pessimistic one, and they correspond better to situations where missingness is non-adversarial, and where one seeks to obtain the best predictive power from the model rather than a model robust to possibly adversarial imputations.

4 Learning Phase

4.1 Learning from Complete Data

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

$$L_{\Omega}(G, \theta | \mathcal{S}) = L(G, \theta | \mathcal{S}) - \Omega(G, \mathcal{S}) , \quad (7)$$

where $\Omega(G, \mathcal{S})$ is a decomposable regularization term that quantifies the model complexity:

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

Common choices for $\Omega(G, \mathcal{S})$ are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which differ in penalty strength (Murphy, 2023, Sec. 3.8.7). We adopt definitions that account only for the complexity of the sub-DAG induced by the discrete variables when computing these penalties, since they are used exclusively to constrain this part of the Bayesian network:

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

where $\text{pa}^{\mathcal{D}}(Z^j)$ is the set of the discrete variables among the parents of Z^j , $|Z^j|$ is the cardinality of Z^j , and $|\text{pa}^{\mathcal{D}}(Z^j)|$ is the number of possible configurations of $\text{pa}^{\mathcal{D}}(Z^j)$. Overall, these choices result in $\Omega(G, \mathcal{S}) = -\text{Dim}(G)$ for AIC and $\Omega(G, \mathcal{S}) = -0.5 \text{Dim}(G) \log(N)$ for BIC, where $\text{Dim}(G) = \sum_{j \in \mathcal{D}} (|Z^j| - 1) |\text{pa}^{\mathcal{D}}(Z^j)|$. In this paper, we adopt BIC and additionally report experimental results obtained with AIC.

In the following, we show that (7) can be optimized exactly under standard assumptions, and that existing learning strategies (Nguyen et al, 2023) extend to our more general setting with minimal changes.

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

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

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

Proposition 2. *For any $j \in \mathcal{D}$ and any finite training set $\mathcal{S}$, let*

$$\ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) := \sum_{n=1}^N \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j))) \quad (8)$$

$$L_{\Omega}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) := \ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) - \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) , \quad (9)$$

the structure constraints of HMDCs ensure that, for any HMDC $(G, \boldsymbol{\theta})$, we have the decomposition:

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

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

Proposition 3. *Let $2^{\mathbf{Z} \setminus Z^j}$ denote the power set of $\mathbf{Z} \setminus Z^j$, and let $\mathcal{P}_a(Z^j) \subset 2^{\mathbf{Z} \setminus \{Z^j\}}$ denote the set of possible parent sets of Z^j , for each $j \in \mathcal{D}$. Given a training set $\mathcal{S}$, maximizing the regularized conditional log-likelihood (7) over all HMDCs decomposes into independent local learning problems, one for each admissible parent set, together with a BN structure learning problem. Concretely, training the model reduces to training a collection of local probabilistic multi-class classifiers. Each local classifier*

$$\mathbf{h}_{\text{pa}(Z^j)} : \bigtimes_{\{k: Z^k \in \text{pa}(Z^j)\}} \mathcal{Z}^k \longrightarrow \Delta^{|\mathcal{Z}^j|} \quad (10)$$

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

$$\boldsymbol{\theta}_j^* \in \underset{\boldsymbol{\theta}_j \in \Theta_j}{\operatorname{argmax}} \ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) , \quad (11)$$

with $\ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S})$ defined in (8). Each admissible parent set $\text{pa}(Z^j)$ is then associated with its optimal score $L_\Omega(\text{pa}(Z^j), \boldsymbol{\theta}_j^* | \mathcal{S})$ (9). Finally, any score-based BN structure learning algorithm can be used to identify an optimal DAG G^* , defined by the optimal parent sets $\text{pa}^*(Z^j)$ for each $j \in \mathcal{D}$. Combined with the corresponding optimal parameters $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}_j^* | j \in \mathcal{D}\}$, this yields the optimal HMDC $(G^*, \boldsymbol{\theta}^*)$.

Allowing arbitrary parent sets $\mathcal{P}_a(Z^j) \subset 2^{\mathbf{Z} \setminus \{Z^j\}}$ results in a very large hypothesis space, which may be computationally prohibitive and yield models that are difficult to interpret. Model complexity constraints can be incorporated by restricting the admissible parent sets for each $Z^j \in \mathbf{Z}^{\mathcal{D}}$, enabling the search for the best model within a prescribed complexity class.

We employ GOBNILP (Bartlett and Cussens, 2017; Cussens, 2020), which is a globally optimal BN learning framework using integer linear programming, to solve the structure learning problem. We also adapt the pruning rule of de Campos et al (2018) to discard unnecessarily large candidate parent sets $\text{pa}(Z^j) \in \mathcal{P}_a(Z^j)$ that cannot serve as the parent set of Z^j in any optimal DAG G^* . The following proposition implies that this pruning does not require the fitting of the local classifier $\mathbf{h}_{\text{pa}(Z^j)}$.

Proposition 4. For any $j \in \mathcal{D}$ and any $\text{pa}'(Z^j) \supset \text{pa}(Z^j)$, if

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

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

This property reduces 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 $(Z^j, \text{pa}(Z^j))$, the conditional distribution $\mathbf{p}_{(G, \boldsymbol{\theta})}(Z^j | \text{pa}(Z^j))$ can be modeled by a probabilistic classifier $\mathbf{h}_{\text{pa}(Z^j)}$ (10). A theoretically sound approach is required to define and train these local probabilistic multi-class classifiers $\mathbf{h}_{\text{pa}(Z^j)}$ when $\text{pa}(Z^j)$ includes both discrete and continuous variables. We adopt an input partitioning strategy (Nguyen et al, 2023) to learn $\mathbf{h}_{\text{pa}(Z^j)}$. It enables local feature selection, either through the construction of local regularization terms or by employing probabilistic multiclass classifiers, such as random forests, which typically select a subset of features during training. As a result, irrelevant continuous features can be effectively discarded from $\text{pa}(Z^j)$.

For each possible candidate parent set $\text{pa}(Z^j)$, let $\text{pa}^{\mathcal{D}}(Z^j)$ denote the set of the discrete variables among the parents of Z^j , and let

$$\mathcal{P}_a^{\mathcal{D}}(Z^j) := \bigtimes_{\{k: Z^k \in \text{pa}^{\mathcal{D}}(Z^j)\}} \mathcal{Z}^k ,$$

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

$$\mathbf{h}_{\text{pa}^{\mathcal{D}}(z^j)} : \mathcal{X}^{\mathcal{C}} \rightarrow \Delta^{|\mathcal{Z}^j|} \quad (12)$$

is used to model $\mathbf{p}_{(G, \theta)}(Z^j | \text{pa}^{\mathcal{D}}(Z^j) = \text{pa}^{\mathcal{D}}(z^j), \mathbf{X}^{\mathcal{C}})$.

For any $\text{pa}^{\mathcal{D}}(z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)$, let

$$\ell(\text{pa}^{\mathcal{D}}(Z^j), \theta_j | \mathcal{S}) = \sum_{\{z_n \in \mathcal{S} : \text{pa}^{\mathcal{D}}(z_n^d) = \text{pa}^{\mathcal{D}}(Z^j)\}} \log \left(\mathbf{p}_{(G, \theta)}(z_n^d | \text{pa}^{\mathcal{D}}(Z^j), \mathbf{x}_n^{\mathcal{C}}) \right) . \quad (13)$$

The $|\mathcal{Pa}^{\mathcal{D}}(Z^j)|$ classifiers (12) can be estimated independently because

$$\ell(\text{pa}(Z^j), \theta_j | \mathcal{S}) = \sum_{\text{pa}^{\mathcal{D}}(z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)} \ell(\text{pa}^{\mathcal{D}}(z^j), \theta_j | \mathcal{S}) .$$

Altogether, the entire training phase is summarized in Algorithm 1. In this algorithm, any score-based BN structure learning algorithm can be used to identify an optimal DAG G^* using $L_{\Omega}(\text{pa}(Z^j), \theta_j^* | \mathcal{S})$ (9). In this paper, we use GOBNILP.

Proposition 5. *Given any finite training set $\mathcal{S}$, Algorithm 1 returns an optimal HMDC (G^*, θ^*) maximizing the regularized conditional log-likelihood (9), provided that, for every possible pair $(Z^j, \text{pa}(Z^j))$, and every $\text{pa}^{\mathcal{D}}(z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)$, the corresponding local classifier $\mathbf{h}_{\text{pa}^{\mathcal{D}}(z^j)}$ (12) can be learned optimally by maximizing (13) under the feature-selection option.*

While doing both structure and parameter learning can be costly, it is needed to detect and incorporate conditional (in)dependencies $Z^i \rightarrow \mathbf{X}^{\mathcal{C}} \leftarrow Z^j$ and $Z^i \leftarrow \mathbf{X}^{\mathcal{C}} \rightarrow Z^j$, 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 $\mathbf{h}_{\text{pa}^{\mathcal{D}}(z^j)}$, such as logistic regression, can provide good estimates of the local conditional probability distribution, the DAGs provided by our framework seem to capture reasonable dependencies in the datasets.

4.2 Learning from Incomplete Data

4.2.1 EM Approach

A standard approach to handling missing data is the EM algorithm (Dempster et al, 1977), which relies on expected sufficient statistics. Lauritzen (1995) first applied EM to parameter estimation in Bayesian networks with incomplete data, assuming a known structure. Structure learning from incomplete data has since followed two main directions (Scutari, 2020). The first iteratively imputes missing values and applies conventional structure learning methods for complete data (Friedman, 1997, 1998; Adel and de Campos, 2017). The second develops scoring functions that can be evaluated directly on incomplete data, such as the variational approximation of BIC and $\mathbf{p}(\mathcal{S} | G)$ proposed by Beal and Ghahramani (2003). For learning both structure and

Algorithm 1 Learning an optimal HMDC (G^*, θ^*)

```

1: input: Training data  $\mathcal{S} = \{z_n\}_{n=1}^N$ , possible parent sets  $\mathcal{P}_a(Z^j)$ ,  $j \in \mathcal{D}$ , and hypothesis
   spaces for local classifiers  $\mathbf{h}_{\text{pa}(Z^j)}$  (12).
2: for  $j \in \mathcal{D}$  do
3:   Sort  $\text{pa}(Z^j) \in \mathcal{P}_a(Z^j)$  s.t.  $|\text{pa}(Z^j)|$  increases
4:   for  $\text{pa}(Z^j) \in \mathcal{P}_a(Z^j)$  do
5:     for  $\text{pa}^{\mathcal{D}}(z^j) \in \mathcal{P}_a^{\mathcal{D}}(Z^j)$  do
6:       Train a local classifier  $\mathbf{h}_{\text{pa}^{\mathcal{D}}(z^j)}$  (12)
7:     end for
8:     Store optimal set  $\theta_j^*$  (11) of  $\mathbf{h}_{\text{pa}(Z^j)}$  (10)
9:     Compute optimal score  $L_{\Omega}(\text{pa}(Z^j), \theta_j^* | \mathcal{S})$  (9)
10:    Prune all  $\text{pa}'(Z^j) \supset \text{pa}(Z^j)$  satisfying Proposition 4
11:  end for
12: end for
13: Find optimal DAG  $G^*$  using  $L_{\Omega}(\text{pa}(Z^j), \theta_j^* | \mathcal{S})$  (9).
14: for  $j \in \mathcal{D}$  do
15:   Extract  $\text{pa}^*(Z^j)$  from  $G^*$  and its optimal parameter set  $\theta_j^*$ .
16: end for
17: return:  $(G^*, \theta^*)$ , where  $\theta^* := \{\theta_j^* | j \in \mathcal{D}\}$ 

```

parameters, two EM variants are commonly considered. Soft EM computes expected sufficient statistics by averaging over all possible completions, but its computational cost grows exponentially with the number of missing values. Hard EM instead imputes the most likely completion at each iteration (Sriperumbudur et al, 2010), making it substantially more scalable. Although EM-based methods for Bayesian network learning have been studied extensively, scalability remains a major challenge, particularly for Soft EM (Bodewes and Scutari, 2021). These methods can nevertheless be extended naturally to our setting. In particular, Hard EM becomes

$$(\mathbf{G}^{\text{EM}}, \theta^{\text{EM}}) \in \underset{(\mathbf{G}, \theta)}{\text{argmax}} \max_{s \in \{1, \dots, S\}} L_{\Omega}(\mathbf{G}, \theta | \hat{\mathcal{S}}^{(s)}) , \quad (14)$$

which is exactly the optimistic approach.

To the best of our knowledge, no EM-like algorithm guarantees convergence to the global maximizer of the regularized log-likelihood for Bayesian networks with mixed discrete and continuous variables. The main difficulty is ensuring that each local model is optimized at every iteration, a prerequisite for convergence of the global model.

4.2.2 Averaging Approach

We now show that an HMDC optimal with respect to the averaging principle (6) can be obtained under a set of commonly adopted assumptions, analogous to those typically made for complete training data.

Proposition 6. Let $\bar{\mathcal{S}}$ be the extended data set that consists of the unique elements of the completed data sets $\hat{\mathcal{S}}^{(s)}$

$$\bar{\mathcal{S}} = \bigcup_{s=1}^S \hat{\mathcal{S}}^{(s)} ,$$

and for any $\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}$, we define

$$w(\bar{\mathbf{z}}_m) = \frac{1}{S} \sum_{s=1}^S \mathbb{I}[\bar{\mathbf{z}}_m \in \hat{\mathcal{S}}^{(s)}] .$$

Under the averaging principle, the learning criterion

$$L_{\Omega}^{\text{ave}}(G, \boldsymbol{\theta} | \mathcal{S}) := \frac{1}{S} \sum_{s=1}^S L_{\Omega}(G, \boldsymbol{\theta} | \hat{\mathcal{S}}^{(s)})$$

decomposes as

$$L_{\Omega}^{\text{ave}}(G, \boldsymbol{\theta} | \mathcal{S}) = \sum_{j \in \mathcal{D}} L_{\Omega}^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) , \quad (15)$$

where

$$L_{\Omega}^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) := \ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) - \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) , \quad (16)$$

$$\ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) := \sum_{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}} w(\bar{\mathbf{z}}_m) \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j | \text{pa}(\bar{\mathbf{z}}_m^j))) . \quad (17)$$

Moreover, for any $\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}$, we have

$$w(\bar{\mathbf{z}}_m) = \sum_{\mathbf{z} \in \mathcal{S}} \frac{\mathbb{I}[\bar{\mathbf{z}}_m^{\mathcal{O}} = \mathbf{z}^{\mathcal{O}}]}{\prod_{j \in \mathcal{M}} |\mathcal{Z}^j|} . \quad (18)$$

Proposition 7. For any $j \in \mathcal{D}$, 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 for Z^j . Given a training set $\mathcal{S}$, the task of learning an optimal HMDC $(G^{\text{ave}}, \boldsymbol{\theta}^{\text{ave}})$ maximizing the regularized conditional log-likelihood (15) can be decomposed into independent local learning problems, followed by a Bayesian network structure learning problem.

Specifically, it reduces to training a collection of local probabilistic multi-class classifiers: each classifier

$$\mathbf{h}_{\text{pa}(Z^j)} : \bigtimes_{\{k: Z^k \in \text{pa}(Z^j)\}} \mathcal{Z}^k \longrightarrow \Delta^{|\mathcal{Z}^j|} \quad (19)$$

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

$$\boldsymbol{\theta}_j^{\text{ave}} \in \operatorname{argmax}_{\boldsymbol{\theta}_j \in \Theta_j} \ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) , \quad (20)$$

with $\ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S})$ defined in (17). Each possible parent set $\text{pa}(Z^j)$ is then associated with its optimal score $L_\Omega(\text{pa}(Z^j), \boldsymbol{\theta}_j^{\text{ave}} | \mathcal{S})$ (16). Finally, any score-based Bayesian network structure learning technique can be employed to identify an optimal DAG G^{ave} specified by the optimal parent sets $\text{pa}^{\text{ave}}(Z^j)$. This DAG G^{ave} , together with the optimal parameters $\boldsymbol{\theta} := \{\boldsymbol{\theta}_j^{\text{ave}} | j \in \mathcal{D}\}$, define an optimal HMDC $(G^{\text{ave}}, \boldsymbol{\theta}^{\text{ave}})$.

Proposition 8. For any $j \in \mathcal{D}$ and $\text{pa}'(Z^j) \supset \text{pa}(Z^j)$, if

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

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

We then adopt the input partitioning strategy previously described for complete data to learn $\mathbf{h}_{\text{pa}(Z^j)}$, for all $j \in \mathcal{D}$. Altogether, the entire training phase is summarized in Algorithm 2.

Proposition 9. Given any finite training set $\mathcal{S}$, Algorithm 2 returns an optimal HMDC $(G^{\text{ave}}, \boldsymbol{\theta}^{\text{ave}})$ maximizing the regularized conditional log-likelihood (15), provided that, for every pair $(Z^j, \text{pa}(Z^j))$, and every parent configuration $\text{pa}^{\mathcal{D}}(Z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\text{pa}^{\mathcal{D}}(Z^j)}$ satisfying (12) can be learned by maximizing

$$\ell^{\text{ave}}(\mathcal{Pa}^{\mathcal{D}}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) = \sum_{\{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}: \text{pa}^{\mathcal{D}}(\bar{\mathbf{z}}_m^j) = \mathcal{Pa}^{\mathcal{D}}(Z^j)\}} w(\bar{\mathbf{z}}_m) \log \left(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j | \mathcal{Pa}^{\mathcal{D}}(Z^j), \bar{\mathbf{x}}_m^c) \right) , \quad (21)$$

under the feature-selection option.

5 Inference Phase

5.1 Inference from Incomplete Data

We now address the prediction of class variables for query instances $\mathbf{z}$ where some discrete features and class variables are missing.

Let $u : \mathcal{Y}^{\mathcal{M}} \times \mathcal{Y}^{\mathcal{M}} \rightarrow \mathbb{R}_+$ be any utility quantifying the reward for predicting $\hat{\mathbf{y}}^{\mathcal{M}}$ when the ground-truth is $\mathbf{y}^{\mathcal{M}}$. The expected utility is defined as

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

For reasons mentioned in Section 3, we will only consider the optimistic and averaging principles, that can respectively be implemented as follows:

$$\hat{\mathbf{y}}^{\text{opt}} \in \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\mathbf{p}_{(G, \boldsymbol{\theta})}} [u(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) | \mathbf{x}^c] , \quad (22)$$

Algorithm 2 Learning an optimal HMDC $(G^{\text{ave}}, \theta^{\text{ave}})$ from incomplete training data

```

1: input: Training data  $\mathcal{S} := \{z_n\}_{n=1}^N = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ , possible parent sets  $\mathcal{P}_a(Z^j)$ ,
    $j \in \mathcal{D}$ , and hypothesis spaces for local classifiers  $\mathbf{h}_{\text{pa}(Z^j)}$  (12).
2: Initialize  $\bar{\mathcal{S}} = \emptyset$ ,  $\text{We}\mathcal{S} = \emptyset$ 
3: for  $n \in [N]$  do
4:   Create  $\prod_{j \in \mathcal{M}} |\mathcal{Z}^j|$  completed  $\bar{z}_m$  from  $z$  with the weight  $\text{We} = 1/(\prod_{j \in \mathcal{M}} |\mathcal{Z}^j|)$ 
5:   Increase the weight of seen instances  $\bar{z}$  by  $\text{We}$ 
6:   Enlarge  $\bar{\mathcal{S}}$  and  $\text{We}\mathcal{S}$  with the emerging instances  $\bar{z}$  and their weight  $\text{We}$ 
7: end for
8: for  $j \in \mathcal{D}$  do
9:   Sort  $\text{pa}(Z^j) \in \mathcal{P}_a(Z^j)$  in the increasing order of  $|\text{pa}(Z^j)|$ 
10:  for  $\text{pa}(Z^j) \in \mathcal{P}_a(Z^j)$  do
11:    for  $\text{pa}^{\mathcal{D}}(Z^j) \in \mathcal{P}_a^{\mathcal{D}}(Z^j)$  do
12:      Train a local classifier  $\mathbf{h}_{\text{pa}^{\mathcal{D}}(Z^j)}$  (12) to optimize (21)
13:    end for
14:    Store the optimal parameter set  $\theta_j^{\text{ave}}$  (20) of  $\mathbf{h}_{\text{pa}(Z^j)}$  (19)
15:    Compute the optimal score  $L_{\Omega}^{\text{ave}}(\text{pa}(Z^j), \theta_j^{\text{ave}} | \mathcal{S})$  (16)
16:    Prune all  $\text{pa}'(Z^j) \supset \text{pa}(Z^j)$  satisfying Proposition 8
17:  end for
18: end for
19: Find an optimal DAG  $G^{\text{ave}}$  using  $L_{\Omega}^{\text{ave}}(\text{pa}(Z^j), \theta_j^{\text{ave}} | \mathcal{S})$  (16).
20: for  $j \in \mathcal{D}$  do
21:   Extract  $\text{pa}^{\text{ave}}(Z^j)$  from  $G^{\text{ave}}$  and its optimal parameter set  $\theta_j^{\text{ave}}$ .
22: end for
23: return:  $(G^{\text{ave}}, \theta^{\text{ave}})$ , where  $\theta^{\text{ave}} := \{\theta_j^{\text{ave}} | j \in \mathcal{D}\}$ 

```

$$\hat{\mathbf{y}}^{\text{ave}} \in \underset{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}}{\text{argmax}} \sum_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{p_{(G, \theta)}} [u(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) | \mathbf{x}^{\mathcal{C}}] . \quad (23)$$

We study optimal prediction under the optimistic and averaging principles for the two evaluation metrics most commonly used in MDC:

$$\text{subset 0/1 accuracy} \quad u_{0/1}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) := \mathbb{I}[\hat{\mathbf{y}}^{\mathcal{M}} = \mathbf{y}^{\mathcal{M}}] , \quad (24)$$

$$\text{Hamming accuracy} \quad u_H(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) := \frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{m \in \mathcal{M}} \mathbb{I}[\hat{y}^m = y^m] . \quad (25)$$

5.2 Algorithmic Solutions

Let $G^{\mathcal{D}}$ denote the sub-DAG of G^* induced by the discrete variables $\mathbf{Z}^{\mathcal{D}}$, which is the same for any query $\mathbf{z}$. For each $j \in \mathcal{D}$, let $\text{pa}^{\mathcal{D}^*}(Z^j)$ denote the parents of Z^j in G^* . Let $\mathcal{P}_a^{\mathcal{D}^*}(Z^j)$ denote the set of all possible configurations of $\text{pa}^{\mathcal{D}^*}(Z^j)$. Given a query $\mathbf{z}$, we construct the discrete BN $(G^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})$ over $\mathbf{Z}^{\mathcal{D}}$. For each $j \in \mathcal{D}$, the conditional probability table (CPT) $p_{(G^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})}(Z^j | \text{pa}^{\mathcal{D}^*}(Z^j))$ contains $|\mathcal{P}_a^{\mathcal{D}^*}(Z^j)|$ rows, one for

each parent configuration $\text{pa}^{\mathcal{D}}(z^j) \in \mathcal{Pa}^{\mathcal{D}^*}(Z^j)$:

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

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

The following propositions ensure that optimal predictions under the averaging or the optimistic principles, when coupled with the commonly used 0/1 and Hamming accuracies, can be computed by MAP queries on $(G^{\mathcal{D}}, \mathbf{x}^c)$.

Proposition 10. *Let u be the 0/1 accuracy $u_{0/1}$ (24). For any $\mathbf{z} \in \mathcal{Z}$, and for any HMDC $(G^*, \boldsymbol{\theta}^*)$, we have*

$$\hat{\mathbf{y}}^{\text{opt}} \oplus \hat{\mathbf{x}}^{\mathcal{M}} \in \underset{\mathbf{z}^{\mathcal{M}} \in \mathcal{Z}^{\mathcal{M}}}{\text{argmax}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) , \quad (26)$$

$$\hat{\mathbf{y}}^{\text{ave}} \in \underset{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}}{\text{argmax}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{y}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) . \quad (27)$$

Proposition 11. *Let u be the Hamming accuracy u_H (25). For any $\mathbf{z} \in \mathcal{Z}$, and for any HMDC $(G^*, \boldsymbol{\theta}^*)$, $\hat{\mathbf{y}}^{\text{opt}}$ (22) is estimated by finding, $\forall k \in \mathcal{M}$,*

$$\hat{\mathbf{y}}^{\text{opt}, k} \oplus \hat{\mathbf{x}}^{\mathcal{M}} \in \underset{(y^k, \mathbf{x}^{\mathcal{M}}) \in \mathcal{Y}^k \times \mathcal{X}^{\mathcal{M}}}{\text{argmax}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(y^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) ,$$

and $\hat{\mathbf{y}}^{\text{ave}}$ (23) is obtained by finding, $\forall k \in \mathcal{M}$,

$$\hat{\mathbf{y}}^{\text{ave}, k} \in \underset{y^k \in \mathcal{Y}^k}{\text{argmax}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(y^k | \mathbf{z}^{\mathcal{O}}) . \quad (28)$$

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.

Inference with complete data can be viewed as a special case of inference with incomplete data, where $\mathbf{x}^{\mathcal{M}} = \emptyset$ and $\mathbf{y}^{\mathcal{M}} = \mathbf{y}$.

6 Experiments

6.1 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. On the Adult dataset,

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 $|\mathbf{X}^D|$, number of continuous features $|\mathbf{X}^C|$

Dataset	N	K	$\{ \mathcal{Y}^k \}_{k=1}^K$	$ \mathbf{X}^D $	$ \mathbf{X}^C $
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 partition the 40 categories of the discrete variable into two groups: the majority category and all remaining categories, which account for (88.57%) and (11.43%) of the observations, respectively. 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.

We limit each discrete variable to have at most two parents, as increasing this number to three substantially increases the computational cost without yielding a significant improvement in predictive performance (Do et al, 2026).

In addition to the Hamming (25) and 0/1 (24) accuracy measures, to assess whether modeling conditional dependencies alleviates class imbalance in MDC, we adapt the balanced accuracy measure proposed for multiclass classification by Gösgens et al (2021):

$$\begin{aligned} \text{Rec}(y_m^k) &:= \frac{1}{|\mathcal{T}(y_m^k)|} \sum_{\mathcal{T}(y_m^k)} \mathbb{I}[\hat{y}^k = y_m^k] , \\ \text{BA}(Y^k) &:= \frac{1}{|\mathcal{Y}^k|} \sum_{y_m^k \in \mathcal{Y}^k} \text{Rec}(y_m^k) , \end{aligned} \quad (29)$$

where $\mathcal{T}(y_m^k)$ denotes the set of test instances for which Y^k is missing and whose true value is y_m^k . All experiments use 10-fold cross-validation. For each fold and missingness scenario, we report the average of the three accuracy metrics across the 10 folds.

We conduct two sets of experiments. First, we consider the setting in which missingness occurs only during inference (Section 5.1), while learning is performed on complete data (Section 4.1). For each train-test split, an HMDC is learned from the complete training data using Algorithm 1. Missing values are then introduced into discrete features and class variables of the test set under the MAR assumption. We consider missingness rates of 30% and 80% for discrete features, and 30%, 70%, 80%, 90% and 100% for class variables, resulting in 10 distinct missingness scenarios.

Second, we consider missingness during both learning (Section 4.2) and inference (Section 5.1). For each train-test split, missing values are introduced into both the training and test sets according to a MAR mechanism. Missing values in the training data are handled using Algorithm 2. During learning, we consider two missingness settings, corresponding to 5% and 10% missing rates for both discrete features and class variables.

Overall, to facilitate a clearer and more concise presentation and comparison of the results, both experimental settings are expressed within the more general framework in which missingness may occur during both learning and inference. The learning-phase missingness settings are therefore (0%, 0%), corresponding to complete training data, (5%, 5%) and (10%, 10%). The results are presented in Section 6.4 and the Supplementary Results provided online, together with the source code used in our experiments, at https://github.com/telagment/HMDC_Missing.

6.2 Base Classifiers

The local conditional distributions $p_{(G,\theta)}(Z^j | \text{pa}_{\mathcal{D}}(Z^j) = \text{Pa}(Z^j), \mathbf{X}^c)$ can be estimated using standard base classifiers, such as *Logistic Regression* (LR), Naive Bayes, and Random Forest. The choice of base classifier substantially impacts predictive performance. We use LR throughout our experiments because LR consistently outperforms Naive Bayes and Random Forest while yielding reasonably sparse DAGs (Do et al, 2026). Ensemble learning methods generally achieve better predictive performance than individual models. We therefore investigate whether bagging can improve performance when used with logistic regression as the base learner. Stratified bagging trains multiple base models on class-stratified bootstrap samples and aggregates their predictions to reduce variance and improve robustness. We implement a bagging version of logistic regression and refer to it as *Stratified Bagging Logistic Regression* (SBLR).

We compare the performance of the optimistic (HMDC-OP), approximate optimistic (HMDC-MI), and averaging (HMDC-AV) approaches against two 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.

6.3 Baselines and Oracle Classifiers

As shown in Appendix C, the pairwise correlations between discrete variables are often strong for the three datasets. However, as discussed in Section 4, the conditional dependencies may still be sparse. In such cases, a learning algorithm should return sparse DAGs. To assess the role of the conditional dependencies, we include another baseline: Binary Relevance (BR), which enforces an empty DAG structure. BR decomposes the MDC problem into independent classification tasks with all discrete variables as targets for prediction, one for each discrete variable based on the continuous variables. Despite its simplicity, BR can be competitive (optimal) on datasets with weak (no) conditional dependencies.

We also include a few oracle classifiers, which are expected to provide upper bounds on possible performances:

- First, we include a Reference Classifier (RefC), which is similar to HMDC-MI, except that missing discrete features are replaced by their true values. Since RefC has access to the true values of missing features, it is expected to provide

an optimistic estimate of achievable performance for approaches dealing with missing features.

- Additionally, we implement four reference HMDC-based classifiers, denoted by R-HMDC-MI, R-HMDC-OP, R-HMDC-AV, and [R-RefC](#). These classifiers use DAGs selected by optimizing the conditional log-likelihood score on the test set. They can therefore be viewed as oracle classifiers that have access to test-set information when selecting the dependency structure. They can therefore serve as reference models quantifying the performance gap by selecting models with BIC. Note that we use the log-likelihood score for this oracle selection because Hamming and subset 0/1 scores are not decomposable over a Bayesian network and thus cannot be directly optimized during the learning phase.

In the following sections, we report the average hamming and subset 0/1 scores. To reduce visual clutter and improve readability, each figure shows only the best-performing result among R-HMDC-MI, R-HMDC-OP, and R-HMDC-AV, denoted by [R-HMDC](#).

6.4 Experimental Results

In this section, we present results for general scenarios in which missingness occurs during both the learning and inference phases, restricting DAGs to at most two discrete parent variables. We examine how missingness introduced during learning interacts with inference-time missingness and assess its impact on predictive performance.

Figures 2–4 summarize the results for each of the three datasets for models trained on either complete or incomplete data and evaluated under different missingness settings at inference time. Comprehensive results, including the remaining experiments with both LR and SBLR classifiers and the corresponding optimal DAGs G^D are provided as [Supplementary Results](#).

Overall, HMDC-based methods using the BIC pruning score achieve better predictive performance than those using the AIC pruning score. The only exception is the Thyroid dataset, for which the two pruning scores yield comparable performance. Among methods using the BIC pruning score, HMDC-based methods, i.e., [HMDC-MI](#), [HMDC-OP](#), [HMDC-AV](#), and [RefC](#), generally outperform their corresponding oracle HMDC-based variants, i.e., R-HMDC-MI, R-HMDC-OP, R-HMDC-AV, and [R-RefC](#). In addition, with AIC, each HMDC-based method and its corresponding reference HMDC-based method yield almost identical predictive performance. The observation that an oracle variant may perform similarly to, or even slightly worse than, its corresponding non-oracle variant can be explained. Although the oracle classifier is allowed to select a dependency structure that yields a higher test-set log-likelihood, this criterion does not directly optimize either the Hamming or subset 0/1 score. Therefore, an improvement in test-set log-likelihood does not necessarily translate into improved predictive performance under these evaluation metrics.

The Supplementary Results reveal a consistent pattern in the effect of the local base classifier and its bagging variant on predictive performance. Overall, LR achieves the most promising results, performing comparably to SBLR while requiring less computation time. This behavior is consistent with the general observation that bagging tends to be more beneficial for high-variance learners. Since LR is a relatively

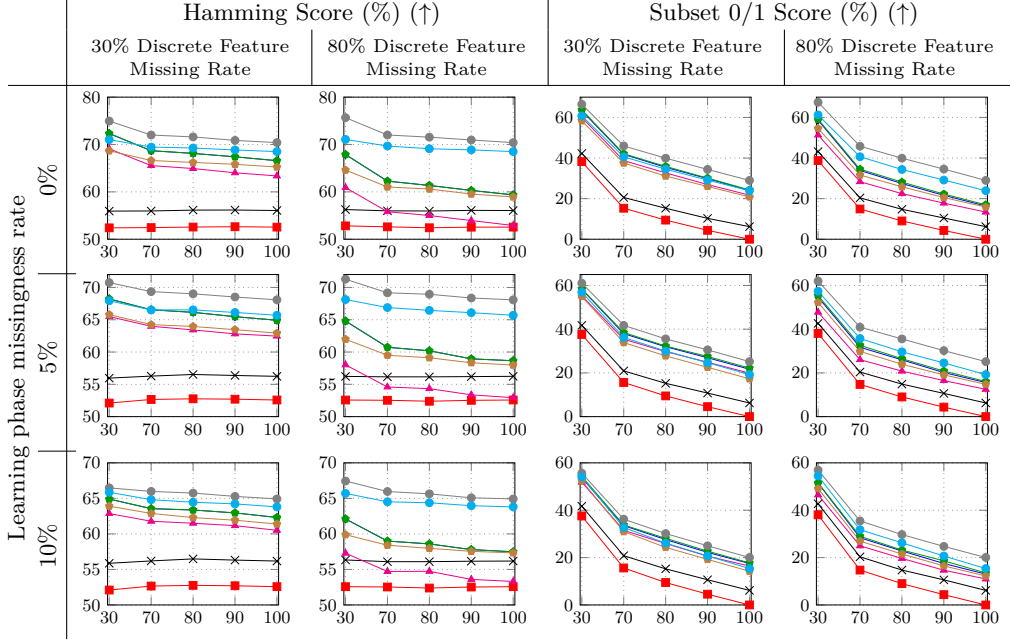


Fig. 2: Adult dataset, with LR base classifier and BIC regularization score. Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds, plotted against the percentage of missing class variables (in %, x-axis) in the inference phase. The learning phase missingness rates are 0% (complete training data), 5% and 10% for both discrete features and class variables. In the inference phase, the missing rate for discrete features is either 30% or 80%. The probabilistic models are displayed as follows: ■ MI, × BR, ▲ HMDC-MI, ◆ HMDC-OP, ◆ HMDC-AV, and ● RefC; ◆ R-HMDC, ● R-RefC.

stable, low-variance classifier, SBLR is therefore expected to provide only marginal improvements over LR.

As expected, RefC, which has access to the missing discrete features, achieves the best performance across most settings, and its results are invariant to rates of missing features at inference time. A similar invariance holds for RandC. Since its results are notably worse than the MI approach, they are deferred to the Supplementary Results, 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 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 at inference time) than to MI, occasionally surpassing RefC. Overall, HMDC-OP and HMDC-AV have indistinguishably strong predictive scores, but HMDC-AV is computationally cheaper to use.

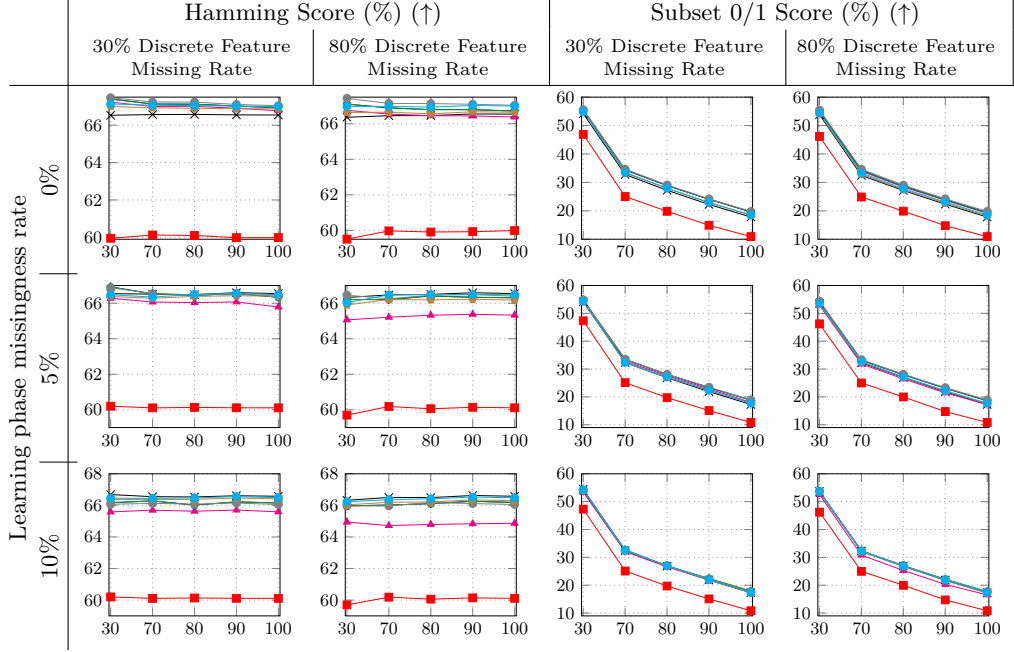


Fig. 3: Default dataset, with LR base classifier and BIC regularization score. Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds, plotted against the percentage of missing class variables (in %, x-axis) in the inference phase. The learning phase missingness rates are 0% (complete training data), 5% and 10% for both discrete features and class variables. In the inference phase, the missing rate for discrete features is either 30% or 80%. The probabilistic models are displayed as follows: ■ MI, × BR, ▲ HMDC-MI, ◆ HMDC-OP, ◆ HMDC-AV, and ● RefC; ◆ R-HMDC, ● R-RefC.

On the Thyroid dataset, MI provides a strong baseline, with a Hamming score of approximately 96%, largely due to the high class imbalance, with a single major outcome, for each class variable. Nevertheless, HMDCs still perform better, close to RefC, which achieves a Hamming score above 98%. 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 missingness at inference time, the Hamming scores of HMDC decrease moderately as the proportion of missing discrete features increases, but change only slightly with the proportion of missing class values. In contrast, the Subset 0/1 scores decrease substantially as the proportion of missing class values increases, while changing only slightly with the proportion of missing discrete features.

Regarding balanced accuracy (29), the results in Figures 5–7 and the Appendix C.4 show that our HMDC-based models consistently outperform BR and other baselines

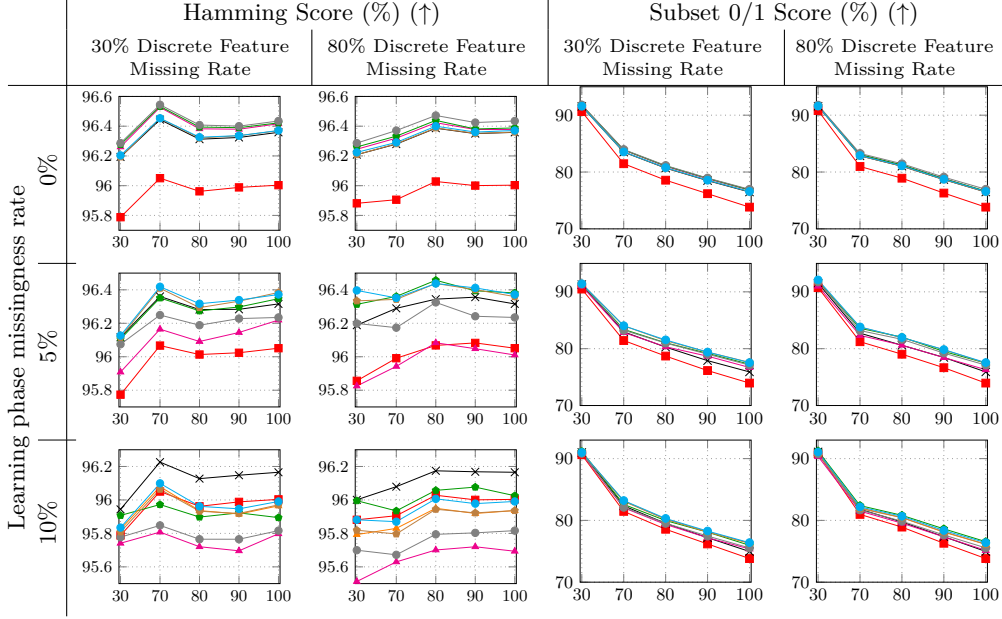


Fig. 4: Thyroid dataset, with LR base classifier and BIC regularization score. Average Hamming and subset 0/1 accuracy scores (in %, y-axis) over 10 cross-validation folds, plotted against the percentage of missing class variables (in %, x-axis) in the inference phase. The learning phase missingness rates are 0% (complete training data), 5% and 10% for both discrete features and class variables. In the inference phase, the missing rate for discrete features is either 30% or 80%. The probabilistic models are displayed as follows: ■ MI, × BR, ▲ HMDC-MI, ◆ HMDC-OP, ◆ HMDC-AV, and ● RefC; ◆ R-HMDC, ● R-RefC.

on Adult and Thyroid, and remain competitive on Default (with very sparse DAGs). This gain likely stems from the ability to account for dependencies among the discrete variables when modeling the joint conditional probability distribution. By capturing these dependencies, the model can better identify plausible configurations of the class variables, thereby alleviating the bias toward majority outcomes and improving predictions for minority outcomes.

Overall, class variables with lower cardinality tend to yield higher balanced accuracy, as they are generally easier to predict. 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. Within the tested range, the missingness rates during the learning phase have little effect the performance of BR. 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, as shown in the Supplementary Results. In particular, **HMDC-MI** shows high sensitivity to changes in the missingness rate of discrete features, likely because the method infers outcomes from unreliable or misleading evidence.

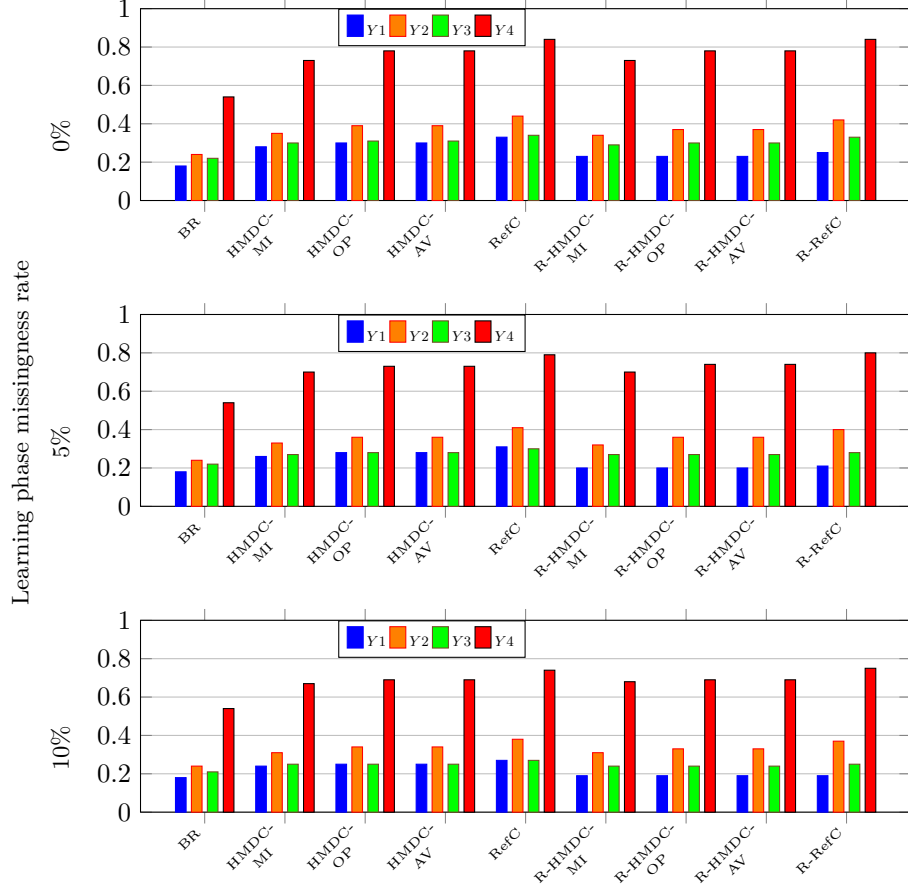


Fig. 5: Adult dataset, with LR base classifier and BIC regularization score. Balanced accuracies ($\uparrow$), with missingness rates in the inference phase of 90% for class variables and 30% for discrete features. The learning phase missingness rates are 0% (complete training data), 5% and 10% for both discrete features and class variables.

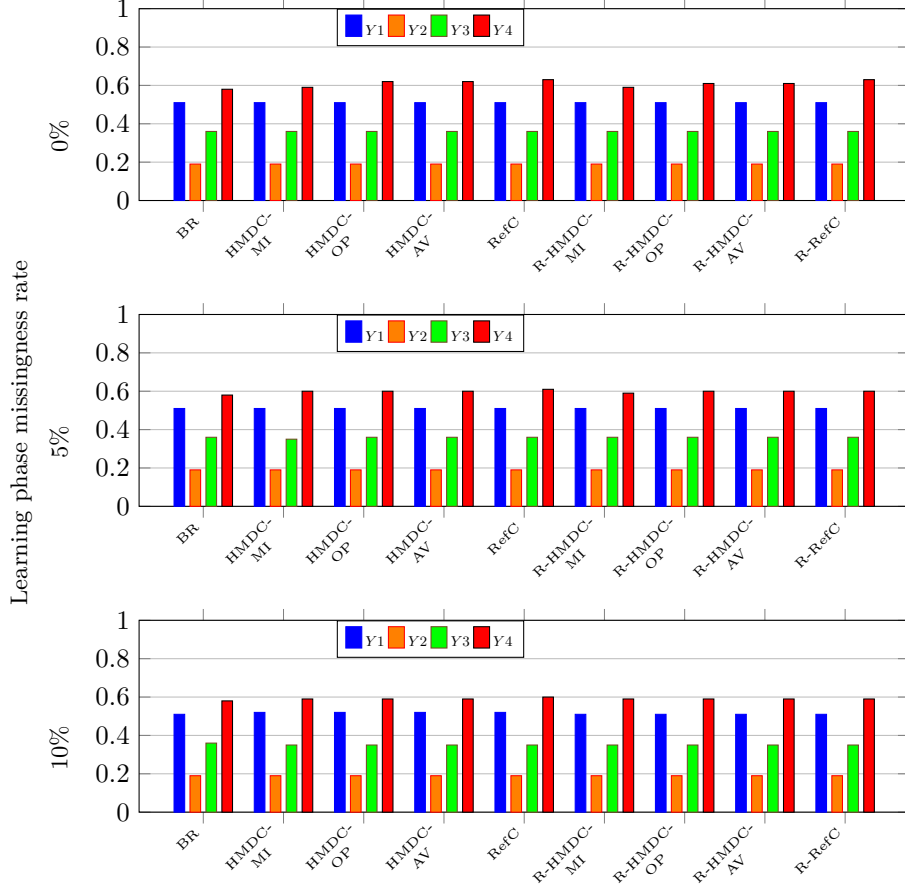


Fig. 6: Default dataset, with LR base classifier and BIC regularization score. Balanced accuracies ($\uparrow$), with missingness rates in the inference phase of 90% for class variables and 30% for discrete features. The learning phase missingness rates are 0% (complete training data), 5% and 10% for both discrete features and class variables.

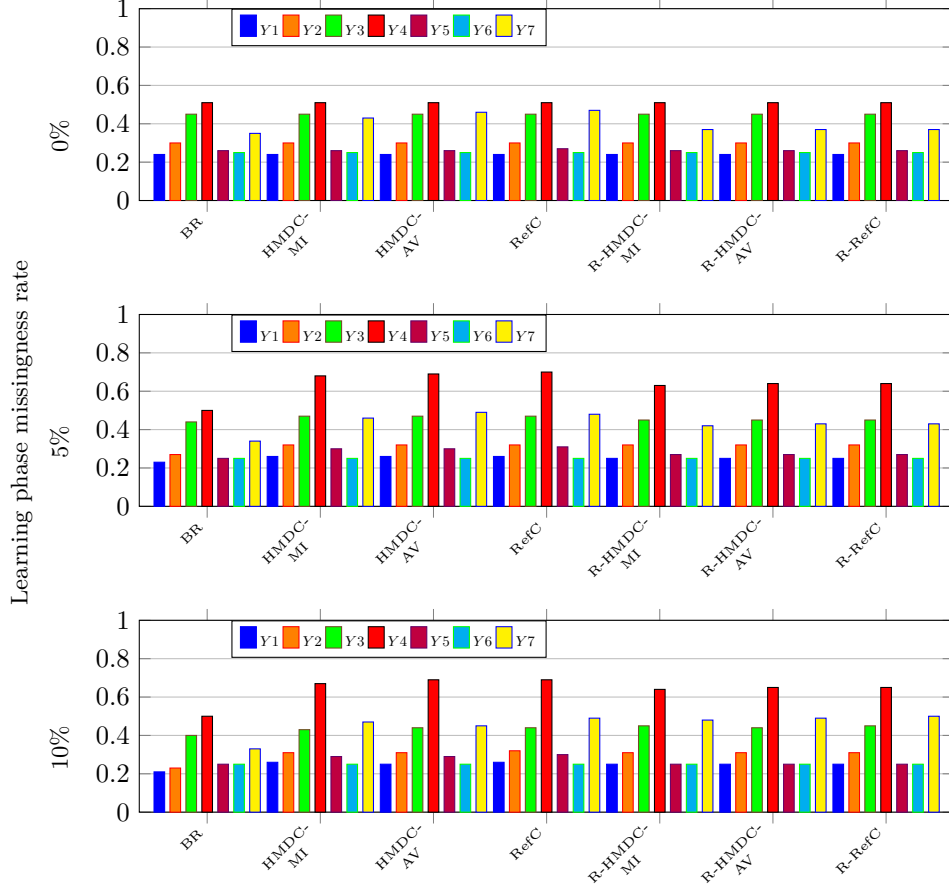


Fig. 7: Thyroid dataset, with LR base classifier and BIC regularization score. Balanced accuracies ($\uparrow$), with missingness rates in the inference phase of 90% for class variables and 30% for discrete features. The learning phase missingness rates are 0% (complete training data), 5% and 10% for both discrete features and class variables.

7 Conclusion

In this work, we presented Hybrid Probabilistic Multi-Dimensional Classifiers (HMDCs), a model tailored to address the challenging multi-dimensional classification setting with mixed data and incomplete discrete observations. The proposed model accommodates settings in which discrete features and class variables may be missing during both prediction and learning. We detailed the learning and prediction phases of this model and derived theoretical results that support the development of efficient algorithms for both phases. Experimental results demonstrated the robustness of the model in different scenarios, under varying levels of missingness in discrete features and class variables, and highlighted the potential of the proposed framework for practical MDC applications.

As a follow-up work, we plan to further investigate learning from incomplete training data, particularly through hard-EM-based strategies that offer another avenue for jointly estimating missing values and model parameters. Further evaluation on real-world applications would also provide valuable insight into the practical scope and limitations of the framework.

References

- Adel T, de Campos C (2017) Learning Bayesian networks with incomplete data by augmentation. In: AAAI Conference on Artificial Intelligence
- Bartlett M, Cussens J (2017) Integer linear programming for the Bayesian network structure learning problem. *Artificial Intelligence* 244:258–271
- Beal MJ, Ghahramani Z (2003) The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In: *Bayesian Statistics*, vol 7. Oxford University Press, p 453–464
- Benjumbeda M, Bielza C, Larrañaga P (2018) Tractability of most probable explanations in multidimensional Bayesian network classifiers. *International Journal of Approximate Reasoning* 93:74–87
- Bielza C, Li G, Larranaga P (2011) Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning* 52(6):705–727
- Bodewes T, Scutari M (2021) Learning Bayesian networks from incomplete data with the node-average likelihood. *International Journal of Approximate Reasoning* 138:145–160
- de Campos CP, Scanagatta M, Corani G, et al (2018) Entropy-based pruning for learning Bayesian networks using BIC. *Artificial Intelligence* 260:42–50
- Chickering DM (1996) Learning Bayesian networks is NP-complete. *Learning from Data: Artificial Intelligence and Statistics V* pp 121–130

- Cooper GF (1990) The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* 42(2-3):393–405
- Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine learning* 9(4):309–347
- Cussens J (2020) GOBNILP: Learning Bayesian network structure with integer programming. In: *International Conference on Probabilistic Graphical Models*, PMLR, pp 605–608
- Dadu A, Satone V, Kaur R, et al (2022) Identification and prediction of Parkinson’s disease subtypes and progression using machine learning in two cohorts. *npj Parkinson’s Disease* 8(1):172
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (methodological)* 39(1):1–22
- Din L, Sheikh M, Kosaraju N, et al (2019) Genetic overlap between autoimmune diseases and non-Hodgkin lymphoma subtypes. *Genetic epidemiology* 43(7):844–863
- Do TH, Nguyen VL, Grandvalet Y (2026) Probabilistic multi-dimensional classification with incomplete data at the prediction time. In: *The 29th International Conference on Artificial Intelligence and Statistics*
- Do VH, Nguyen SH, Le DQ, et al (2024) PanKA: Leveraging population pangenome to predict antibiotic resistance. *iScience* 27(9):110623
- Ferreira D, Nordberg A, Westman E (2020) Biological subtypes of Alzheimer disease: a systematic review and meta-analysis. *Neurology* 94(10):436–448
- Friedman N (1997) Learning belief networks in the presence of missing values and hidden variables. In: *International Conference on Machine Learning (ICML)*, pp 125–133
- Friedman N (1998) The Bayesian structural EM algorithm. In: *Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, UAI’98, pp 129–138
- Garcia-Aymerich J, Gómez FP, Benet M, et al (2011) Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 66(5):430–437
- García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR (2010) Pattern classification with missing data: a review. *Neural Computing and Applications* 19(2):263–282
- Gil-Begue S, Bielza C, Larrañaga P (2021) Multi-dimensional Bayesian network classifiers: A survey. *Artificial Intelligence Review* 54(1):519–559

- Gösgens M, Zhiyanov A, Tikhonov A, et al (2021) Good classification measures and how to find them. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), pp 17136–17147
- Guillaume R, Couso I, Dubois D (2017) Maximum likelihood with coarse data based on robust optimisation. In: Tenth international symposium on imprecise probability: theories and applications (ISIPTA), PMLR, pp 169–180
- Hoshida Y, Brunet JP, Tamayo P, et al (2007) Subclass mapping: identifying common subtypes in independent disease data sets. *PloS One* 2(11):e1195
- Hüllermeier E, Beringer J (2006) Learning from ambiguously labeled examples. *Intelligent Data Analysis* 10(5):419–439
- Hüllermeier E, Destercke S, Couso I (2019) Learning from imprecise data: Adjustments of optimistic and pessimistic variants. In: 13th International Conference on Scalable Uncertainty Management (SUM), pp 266–279
- Jia BB, Zhang ML (2024) Multi-dimensional classification: Paradigm, algorithms and beyond. *Vicinagearth* 1(1):3
- Koller D, Friedman N (2009) Probabilistic graphical models: Principles and Applications. MIT Press
- Lauritzen SL (1995) The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis* 19(2):191–201
- Lynch DA, Austin JH, Hogg JC, et al (2015) CT-definable subtypes of chronic obstructive pulmonary disease: A statement of the Fleischner Society. *Radiology* 277(1):192–205
- Ma Z, Chen S (2018) Multi-dimensional classification via a metric approach. *Neurocomputing* 275:1121–1131
- Moradigaravand D, Palm M, Farewell A, et al (2018) Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS computational biology* 14(12):e1006258
- Murphy KP (2023) Probabilistic machine learning: Advanced topics. MIT Press
- Nguyen VL, Destercke S, Masson MH (2018) Partial data querying through racing algorithms. *International Journal of Approximate Reasoning* 96:36–55
- Nguyen VL, Destercke S, Masson MH, et al (2021) Racing trees to query partial data. *Soft Computing* 25(14):9285–9305
- Nguyen VL, Yang Y, de Campos CP (2023) Probabilistic multi-dimensional classification. In: 39th Conference on Uncertainty in Artificial Intelligence (UAI), pp

- Parviainen P, Kaski S (2017) Learning structures of Bayesian networks for variable groups. *International Journal of Approximate Reasoning* 88:110–127
- Pearl J, Mackenzie D (2018) *The book of why: the new science of cause and effect*. Basic books
- Roos T, Wettig H, Grünwald P, et al (2005) On discriminative Bayesian network classifiers and logistic regression. *Machine Learning* 59(3):267–296
- Scutari M (2020) Bayesian network models for incomplete and dynamic data. *Statistica Neerlandica* 74(3):397–419
- Sriperumbudur BK, Gretton A, Fukumizu K, et al (2010) Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* 11:1517–1561
- Sucar LE (2021) *Probabilistic graphical models: Principles and Applications*. *Advances in Computer Vision and Pattern Recognition*, Springer
- Utkin LV, Coolen FP (2011) Interval-valued regression and classification models in the framework of machine learning. In: *7th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, PMLR, pp 371–380
- Van Der Gaag LC, De Waal PR (2006) Multi-dimensional Bayesian network classifiers. In: *Third European workshop on Probabilistic Graphical Models (PGM)*, Prague, pp 107–114

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 ANR23-IACL-0007.

Declarations

Author contributions statement

T.H.D contributed to Conceptualization, Methodology, Software, Validation, Visualization, and Writing – original draft. K.D.T. contributed to Methodology, Software, Validation, and Writing – review & editing. V.L.N contributed to Conceptualization, Methodology, Validation, Visualization, Supervision, and Writing – original draft. Y.G contributed to Methodology, Validation, Visualization, Supervision, and Writing – original draft. S.D. contributed to Methodology and Writing – review & editing.

Data availability

The three data sets used in the empirical study are publicly available at <https://palm.seu.edu.cn/zhangml/>. The source code used in our experiments is provided at https://github.com/telagment/HMDC_Missing.

Research funding

Thu-Ha DO received funding from French National Research Agency (ANR) under the France 2030 program, Grant ID ANR23-IACL-0007. Kim-Dung Tran received funding from Agence Nationale de la Recherche; Grant ID ANR-R311CHD. Vu-Linh Nguyen received funding from Agence Nationale de la Recherche; Grant ID ANR-R311CHD.

Conflict of interest

We declare that we have no conflict of interest.

Ethics approval

We declare that this research did not require Ethics approval.

Consent for publication

All the authors of this manuscript consent to its publication.

A Notations and Acronyms

Symbol	Description
X	Feature (random) variable
Y	Class (random) variable
$\mathbf{X}$	Feature vector $\mathbf{X} = \{X^1, \dots, X^Q\}$
$\mathbf{Y}$	Class vector $\mathbf{Y} = \{Y^1, \dots, Y^K\}$
$\mathbf{X}^{\mathcal{D}}$	Set of discrete feature variables
$\mathbf{X}^{\mathcal{C}}$	Set of continuous feature variables
$\mathbf{Z}$	Union of the 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 $Z = z$
$\mathbf{x}$	Particular set value of feature variables, $\mathbf{X} = \mathbf{x}$
$\mathbf{y}$	Particular set value of class variables, $\mathbf{Y} = \mathbf{y}$
$\mathbf{z}$	Particular set value of variables $\mathbf{Z} = \mathbf{z}$
$\mathbf{z}_n$	Values of the joint variables $\mathbf{Z}$ corresponding with training instance n
$\mathbf{x}_n^{\mathcal{D}}$	Discrete features of $\mathbf{x}_n$
$\mathbf{x}_n^{\mathcal{C}}$	Continuous features of $\mathbf{x}_n$
$\mathcal{M}$	Superscript indicating missing term, <i>e.g.</i> , $\mathbf{X}^{\mathcal{M}}$, $\mathbf{Y}^{\mathcal{M}}$
$\mathcal{O}$	Superscript indicating observed term, <i>e.g.</i> , $\mathbf{X}^{\mathcal{O}}$, $\mathbf{Y}^{\mathcal{O}}$
n	Training data index
N	Number of training data instances
$\mathcal{S}$	Training set $\{\mathbf{z}_n \mid n = 1 \dots, N\} = \{(\mathbf{x}_n, \mathbf{y}_n) \mid n = 1 \dots, N\}$
$\mathcal{T}$	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
$\text{pa}(Z)$	Parent set of Z in the DAG G
$\text{pa}(z)$	Specific configuration of $\text{pa}(Z)$ specified by $\mathbf{z}$
$\mathcal{P}(G, \boldsymbol{\theta})$	Admissible distribution
$\Omega(G, \mathcal{S})$	Decomposable regularization term
$\mathbf{h}_{\text{pa}(Z^j)}$	Probabilistic classifier (10)
$\mathbf{h}_{\text{pa}^{\mathcal{D}}(Z^j)}$	Probabilistic classifier (12)
$\text{pa}^{\mathcal{D}}(Z^j)$	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 Probabilistic Multi-Dimensional Classification
DMDC	Discriminative Probabilistic Multi-dimensional Classification
HMDC	Hybrid Probabilistic Multi-Dimensional Classification
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
LR	Logistic Regression
NB	Naive Bayes
RF	Random Forest

B Proofs for Formal Results

B.1 Proofs of Section 4 (Learning Phase)

B.1.1 Proofs for Propositions of Section 4.1 (Learning from Complete Data)

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

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

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

$$\begin{aligned} \mathbf{p}_{(G,\theta)}(\mathbf{y}, \mathbf{x}^{\mathcal{D}} | \mathbf{x}^{\mathcal{C}}) &= \frac{\mathbf{p}_{(G,\theta)}(\mathbf{y}, \mathbf{x}^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})}{\sum_{\bar{\mathbf{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \mathbf{p}_{(G,\theta)}(\bar{\mathbf{y}}, \bar{\mathbf{x}}^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})} \quad (30) \\ &= \frac{\prod_{j \in \mathcal{D}} \mathbf{p}_{(G,\theta)}(z^j | \text{pa}(z^j)) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \mathbf{p}_{(G,\theta)}(x^i | \text{pa}(x^i))}{\sum_{\bar{\mathbf{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \left(\prod_{j \in \mathcal{D}} \mathbf{p}_{(G,\theta)}(\bar{z}^j | \text{pa}(\bar{z}^j)) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \mathbf{p}_{(G,\theta)}(x^i | \text{pa}(x^i)) \right)} \\ &= \frac{\left(\prod_{j \in \mathcal{D}} \mathbf{p}_{(G,\theta)}(z^j | \text{pa}(z^j)) \right) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \mathbf{p}_{(G,\theta)}(x^i | \text{pa}(x^i))}{\left(\sum_{\bar{\mathbf{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \prod_{j \in \mathcal{D}} \mathbf{p}_{(G,\theta)}(\bar{z}^j | \text{pa}(\bar{z}^j)) \right) \prod_{X^i \in \mathbf{X}^{\mathcal{C}}} \mathbf{p}_{(G,\theta)}(x^i | \text{pa}(x^i))} \\ &= \frac{\prod_{j \in \mathcal{D}} \mathbf{p}_{(G,\theta)}(z^j | \text{pa}(z^j))}{\sum_{\bar{\mathbf{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \prod_{j \in \mathcal{D}} \mathbf{p}_{(G,\theta)}(\bar{z}^j | \text{pa}(\bar{z}^j))} \end{aligned}$$

$$= \prod_{j \in \mathcal{D}} \mathbf{p}_{(G, \boldsymbol{\theta})}(z^j | \text{pa}(z^j))$$

because by the definition of Bayesian Networks, we have

$$\sum_{\bar{\mathbf{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \prod_{j \in \mathcal{D}} \mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{z}^j | \text{pa}(\bar{z}^j)) = \sum_{\bar{\mathbf{z}}^{\mathcal{D}} \in \mathcal{Z}^{\mathcal{D}}} \mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}^{\mathcal{D}} | \mathbf{x}^{\mathcal{C}}) = 1. \quad (31)$$

□

Proposition B.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:*

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

where

$$\begin{aligned} L_{\Omega}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) &:= \ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) - \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) \\ \ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) &:= \sum_{\mathbf{z}_n \in \mathcal{S}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j))) , \end{aligned}$$

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

Proof. As a consequence of Proposition 1, we have

$$\begin{aligned} L_{\Omega}((G, \boldsymbol{\theta}) | \mathcal{S}) &= \sum_{\mathbf{z}_n \in \mathcal{S}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\mathbf{y}_n, \mathbf{x}_n^{\mathcal{D}} | \mathbf{x}_n^{\mathcal{C}})) - \Omega(G, \mathcal{S}) \\ &= \sum_{\mathbf{z}_n \in \mathcal{S}} \log \left(\prod_{j \in \mathcal{D}} \mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j)) \right) - \Omega(G, \mathcal{S}) \\ &= \sum_{\mathbf{z}_n \in \mathcal{S}} \sum_{j \in \mathcal{D}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j))) - \Omega(G, \mathcal{S}) \\ &= \sum_{j \in \mathcal{D}} \sum_{\mathbf{z}_n \in \mathcal{S}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j))) - \Omega(G, \mathcal{S}) \\ &= \sum_{j \in \mathcal{D}} \sum_{\mathbf{z}_n \in \mathcal{S}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j))) - \sum_{j \in \mathcal{D}} \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) \\ &= \sum_{j \in \mathcal{D}} \left(\sum_{\mathbf{z}_n \in \mathcal{S}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(z_n^j | \text{pa}(z_n^j))) - \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) \right) \end{aligned} \quad (32)$$

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

□

Proposition B.3. Let $2^{\mathbf{Z} \setminus Z^j}$ denote the power set of $\mathbf{Z} \setminus Z^j$, and let $\mathcal{P}_a(Z^j) \subset 2^{\mathbf{Z} \setminus \{Z^j\}}$ denote the set of possible parent sets of Z^j , for each $j \in \mathcal{D}$. Given a training set $\mathcal{S}$, maximizing the regularized conditional log-likelihood (7) over all HMDCs decomposes into independent local learning problems, one for each admissible parent set, together with a BN structure learning problem. Concretely, training the model reduces to training a collection of local probabilistic multi-class classifiers. Each local classifier

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

maps every configuration of the parent set $\text{pa}(Z^j) \in \mathcal{P}_a(Z^j)$ 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 $|\mathcal{Z}^j|$. The optimal local classifier $\mathbf{h}_{\text{pa}(Z^j)}$ is parameterized by

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

with $\ell(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S})$ defined in (8). Each admissible parent set $\text{pa}(Z^j)$ is then associated with its optimal score $L_{\Omega}(\text{pa}(Z^j), \boldsymbol{\theta}_j^* | \mathcal{S})$ (9). Finally, any score-based BN structure learning algorithm can be used to identify an optimal DAG G^* , defined by the optimal parent sets $\text{pa}^*(Z^j)$ for each $j \in \mathcal{D}$. Combined with the corresponding optimal parameters $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}_j^* | j \in \mathcal{D}\}$, this yields the 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

$$\{\text{pa}(Z^j) | j \in \mathcal{D}\} \in \prod_{j \in \mathcal{D}} \mathcal{P}_a(Z^j) \quad (33)$$

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

$$\mathbf{p}_{(G, \boldsymbol{\theta})}(\mathbf{y}, \mathbf{x}^{\mathcal{D}} | \mathbf{x}^{\mathcal{C}}) = \prod_{j \in \mathcal{D}} \mathbf{p}_{(G, \boldsymbol{\theta})}(z^j | \text{pa}(z^j)) . \quad (34)$$

As a consequence of Proposition 2, for each fixed possible DAG G , an optimal parameter set $\boldsymbol{\theta}^* := \{\boldsymbol{\theta}_j^* | j \in \mathcal{D}\}$, which optimizes the local CL (7), can be found by independently finding, $\forall Z^j \in \mathbf{Z}^{\mathcal{D}}$, optimal parameter set $\boldsymbol{\theta}_j^*$ (11), which specifies the probabilistic multi-class classifier $\mathbf{h}_{\text{pa}(Z^j)}$ (10).

Once the local optimal parameter sets $\boldsymbol{\theta}_j^*$ are learned for all the possible pair $(Z^j, \text{pa}(Z^j))$, for any $Z^j \in \mathbf{Z}^{\mathcal{D}}$, each possible parent set $\text{pa}(Z^j)$ is then associated with its optimal score $L_{\Omega}(\text{pa}(Z^j), \boldsymbol{\theta}_j^* | \mathcal{S})$ (9).

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

$$G^* \in \underset{G}{\operatorname{argmax}} \sum_{j \in \mathcal{D}} L_{\Omega}(\mathbf{p}_{(G, \boldsymbol{\theta})}(\text{pa}(Z^j), \boldsymbol{\theta}_j^*) | \mathcal{S}), \quad (35)$$

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

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

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

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

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

$$L_{\Omega}(\mathbf{p}_{(G, \boldsymbol{\theta})}(\text{pa}(Z^j), \boldsymbol{\theta}_j) | \mathcal{S}) \geq -\omega(Z^j, \text{pa}'(Z^j), \mathcal{S}), \quad (36)$$

we have

$$\begin{aligned} & \underbrace{L_{\Omega}(\mathbf{p}_{(G, \boldsymbol{\theta})}(\text{pa}(Z^j), \boldsymbol{\theta}_j) | \mathcal{S}) - L_{\Omega}(\mathbf{p}_{(G, \boldsymbol{\theta})}(\text{pa}'(Z^j), \boldsymbol{\theta}_j) | \mathcal{S})}_{\geq -\omega(Z^j, \text{pa}'(Z^j), \mathcal{S})} \\ & \geq -\omega(Z^j, \text{pa}'(Z^j), \mathcal{S}) - \underbrace{L_{\Omega}(\mathbf{p}_{(G, \boldsymbol{\theta})}(\text{pa}'(Z^j), \boldsymbol{\theta}_j) | \mathcal{S})}_{=\ell(\text{pa}'(Z^j), \boldsymbol{\theta}_j) | \mathcal{S} - \omega(Z^j, \text{pa}'(Z^j), \mathcal{S})} \\ & = -\omega(Z^j, \text{pa}'(Z^j), \mathcal{S}) - \ell(\text{pa}'(Z^j), \boldsymbol{\theta}_j) | \mathcal{S} + \omega(Z^j, \text{pa}'(Z^j), \mathcal{S}) \\ & = -\ell(\text{pa}'(Z^j), \boldsymbol{\theta}_j) | \mathcal{S} \geq 0. \end{aligned} \quad (37)$$

Then $\text{pa}'(Z^j)$ 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 $\text{pa}'(Z^j)$ as the parent set of Z^j . If we replace $\text{pa}'(Z^j)$ by $\text{pa}(Z^j)$, we will have another DAG with a smaller complexity whose optimal regularized conditional log-likelihood in Equation (7) is at least the optimal regularized conditional log-likelihood in Equation (7) attained when using DAG G^* . This is a contradiction. $\square$

Proposition B.5. *Given any finite training set $\mathcal{S}$, Algorithm 1 returns an optimal HMDC $\mathbf{p}_{(G^*, \theta^*)}$ that optimizes the regularized conditional log-likelihood (7) if, for any possible pair $(Z^j, \text{pa}(Z^j))$, and any $\text{pa}^{\mathcal{D}}(Z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\text{pa}^{\mathcal{D}}(Z^j)}$ (12) that optimizes (13) can be found.*

Proof. For any possible pair $(Z^j, \text{pa}(Z^j))$, the optimal parameter set θ_j^* (11) of $\mathbf{h}_{\text{pa}(Z^j)=\text{pa}(z^j)}$ (10) is found if, for any possible configuration $\text{pa}^{\mathcal{D}}(z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\text{pa}^{\mathcal{D}}(z^j)}$ (12) that optimizes (13) can be found. If the local optimal classifier $\mathbf{h}_{\text{pa}^{\mathcal{D}}(z^j)}$ (12) is learned under the feature selection option, the irrelevant continuous features are discarded from $\text{pa}(Z^j)$. This is equivalent to exhaustively assessing all the possible parent sets $\text{pa}'(Z^j) \subset \text{pa}(Z^j)$, where $\text{pa}'(Z^j) := \text{pa}^{\mathcal{D}}(Z^j) \cup \mathbf{Z}^{\mathcal{C}}$, $\forall \mathbf{Z}^{\mathcal{C}} \subset \mathbf{X}^{\mathcal{C}}$ and select the best one. The non-optimal parent sets $\text{pa}'(Z^j)$ can be discarded without discarding any possible optimal DAG since all the $\text{pa}'(Z^j) = \text{pa}^{\mathcal{D}}(Z^j) \cup \mathbf{Z}^{\mathcal{C}}$, $\mathbf{Z}^{\mathcal{C}} \subset \mathbf{X}^{\mathcal{C}}$, have the same regularized score $\omega(Z^j, \text{pa}'(Z^j), \mathcal{S})$ 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 $L_{\Omega}((\text{pa}(Z^j), \theta_j^*) | \mathcal{S})$ (9) of all the possible parent sets $\text{pa}(Z^j) \in \mathcal{Pa}(Z^j)$, 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 $L_{\Omega}((\text{pa}(Z^j), \theta_j^*) | \mathcal{S})$ (9) as the input.

Altogether, the HMDC $\mathbf{p}_{(G^*, \theta^*)}$, where $\theta^* := \{\theta_j^* | j \in \mathcal{D}\}$ returned by Algorithm 1 should optimize the regularized conditional log-likelihood (7). $\square$

B.1.2 Proofs for Propositions of Section 4.2 (Learning from Incomplete Data)

Proposition B.6. *Let $\bar{\mathcal{S}}$ be the extended data set that consists of the unique elements of the completed data sets $\hat{\mathcal{S}}^{(s)}$*

$$\bar{\mathcal{S}} = \bigcup_{s=1}^S \hat{\mathcal{S}}^{(s)} ,$$

and for any $\bar{z}_m \in \bar{\mathcal{S}}$, we define

$$w(\bar{z}_m) = \frac{1}{S} \sum_{s=1}^S [\bar{z}_m \in \hat{\mathcal{S}}^{(s)}] .$$

Under the averaging principle, the learning criterion

$$L_{\Omega}^{\text{ave}}(G, \theta | \mathcal{S}) := \frac{1}{S} \sum_{s=1}^S L_{\Omega}(G, \theta | \hat{\mathcal{S}}^{(s)})$$

decomposes as

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

where

$$\begin{aligned} L_{\Omega}^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) &:= \ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) - \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) \\ \ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) &:= \sum_{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}} w(\bar{\mathbf{z}}_m) \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j | \text{pa}(\bar{\mathbf{z}}_m^j))) . \end{aligned}$$

Moreover, for any $\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}$, we have

$$w(\bar{\mathbf{z}}_m) = \sum_{\mathbf{z} \in \mathcal{S}} \frac{\mathbb{I}[\bar{\mathbf{z}}_m \in \mathbf{z}]}{\prod_{j \in \mathcal{M}} |\mathcal{Z}^j|} .$$

Proof. A simple proof is given for completeness. By definition, we have

$$\begin{aligned} L_{\Omega}^{\text{ave}}(G, \boldsymbol{\theta} | \mathcal{S}) &:= \frac{1}{S} \sum_{s=1}^S L_{\Omega}(G, \boldsymbol{\theta} | \hat{\mathcal{S}}^{(s)}) \\ &= \frac{1}{S} \sum_{s=1}^S \left(\sum_{\bar{\mathbf{z}}_m \in \hat{\mathcal{S}}^{(s)}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{y}}_m, \bar{\mathbf{x}}_m^{\mathcal{D}} | \bar{\mathbf{x}}_m^{\mathcal{C}})) - \underbrace{\Omega(G, \hat{\mathcal{S}}^{(s)})}_{=\Omega(G, \mathcal{S})} \right) \\ &= \frac{1}{S} \sum_{s=1}^S \left(\sum_{\bar{\mathbf{z}}_m \in \hat{\mathcal{S}}^{(s)}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{y}}_m, \bar{\mathbf{x}}_m^{\mathcal{D}} | \bar{\mathbf{x}}_m^{\mathcal{C}})) - \Omega(G, \mathcal{S}) \right) \\ &= \frac{1}{S} \sum_{s=1}^S \sum_{\bar{\mathbf{z}}_m \in \hat{\mathcal{S}}^{(s)}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{y}}_m, \bar{\mathbf{x}}_m^{\mathcal{D}} | \bar{\mathbf{x}}_m^{\mathcal{C}})) - \Omega(G, \mathcal{S}) \\ &= \sum_{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}} w(\bar{\mathbf{z}}_m) \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{y}}_m, \bar{\mathbf{x}}_m^{\mathcal{D}} | \bar{\mathbf{x}}_m^{\mathcal{C}})) - \Omega(G, \mathcal{S}) \\ &= \sum_{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}} w(\bar{\mathbf{z}}_m) \log \left(\prod_{j \in \mathcal{D}} \mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j | \text{pa}(\bar{\mathbf{z}}_m^j)) \right) - \Omega(G, \mathcal{S}) \\ &= \sum_{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}} w(\bar{\mathbf{z}}_m) \sum_{j \in \mathcal{D}} \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j | \text{pa}(\bar{\mathbf{z}}_m^j))) - \Omega(G, \mathcal{S}) \\ &= \sum_{j \in \mathcal{D}} \sum_{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}} \underbrace{w(\bar{\mathbf{z}}_m) \log(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j | \text{pa}(\bar{\mathbf{z}}_m^j)))}_{=:\ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S})} - \Omega(G, \mathcal{S}) \\ &= \sum_{j \in \mathcal{D}} \ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) - \sum_{j \in \mathcal{D}} \omega(Z^j, \text{pa}(Z^j), \mathcal{S}) \end{aligned}$$

$$= \sum_{j \in \mathcal{D}} \left(\underbrace{\ell^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}) - \omega(Z^j, \text{pa}(Z^j), \mathcal{S})}_{=: L_{\Omega}^{\text{ave}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S})} \right).$$

Moreover, for any $\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}$, we have

$$\begin{aligned} w(\bar{\mathbf{z}}_m) &= \frac{1}{S} \sum_{s=1}^S \mathbb{I}[\bar{\mathbf{z}}_m \in \hat{\mathcal{S}}^{(s)}] \\ &= \frac{1}{\prod_{\mathbf{z} \in \mathcal{S}} \prod_{j \in \mathcal{M}} |\mathcal{Z}^j|} \sum_{\substack{\mathbf{z} \in \mathcal{S} \\ \bar{\mathbf{z}}_m \in \mathbf{z}}} \prod_{\substack{\mathbf{z}' \in \mathcal{S} \\ \mathbf{z}' \neq \mathbf{z}}} \prod_{j \in \mathcal{M}'} |\mathcal{Z}^j| \\ &= \sum_{\substack{\mathbf{z} \in \mathcal{S} \\ \bar{\mathbf{z}}_m \in \mathbf{z}}} \frac{\prod_{\substack{\mathbf{z}' \in \mathcal{S} \\ \mathbf{z}' \neq \mathbf{z}}} \prod_{j \in \mathcal{M}'} |\mathcal{Z}^j|}{\prod_{\mathbf{z} \in \mathcal{S}} \prod_{j \in \mathcal{M}} |\mathcal{Z}^j|} \\ &= \sum_{\substack{\mathbf{z} \in \mathcal{S} \\ \bar{\mathbf{z}}_m \in \mathbf{z}}} \frac{1}{\prod_{j \in \mathcal{M}} |\mathcal{Z}^j|} \\ &= \sum_{\mathbf{z} \in \mathcal{S}} \frac{\mathbb{I}[\bar{\mathbf{z}}_m \in \mathbf{z}]}{\prod_{j \in \mathcal{M}} |\mathcal{Z}^j|}. \end{aligned}$$

□

Proposition B.7. For any $j \in \mathcal{D}$, 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 for Z^j . Given a training set $\mathcal{S}$, the task of learning an optimal HMDC $(G^{\text{ave}}, \boldsymbol{\theta}^{\text{ave}})$ maximizing the regularized conditional log-likelihood (15) can be decomposed into independent local learning problems, followed by a Bayesian network structure learning problem.

Specifically, it reduces to training a collection of local probabilistic multi-class classifiers: each classifier

$$\mathbf{h}_{\text{pa}(Z^j)} : \prod_{\{k: Z^k \in \text{pa}(Z^j)\}} \mathcal{Z}^k \longrightarrow \Delta^{|\mathcal{Z}^j|} \quad (38)$$

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

$$\boldsymbol{\theta}_j^{\text{ave}} \in \underset{\boldsymbol{\theta}_j \in \Theta_j}{\text{argmax}} \ell^{\text{wei}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S}), \quad (39)$$

with $\ell^{\text{wei}}(\text{pa}(Z^j), \boldsymbol{\theta}_j | \mathcal{S})$ defined in Equation (17). Each possible parent set $\text{pa}(Z^j)$ is then associated with its optimal score $L_{\Omega}(\text{pa}(Z^j), \boldsymbol{\theta}_j^{\text{ave}} | \mathcal{S})$ (16). Finally, any scoring-based Bayesian network structure learning technique can be employed to find an

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

Proof. Similar to the proof of Proposition 3. $\square$

Proposition B.8. For any $j \in \mathcal{D}$ and $\text{pa}'(Z^j) \supset \text{pa}(Z^j)$, if

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

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

Proof. Similar to the proof of Proposition 4. $\square$

Proposition B.9. Given any finite training set $\mathcal{S}$, Algorithm 2 returns an optimal HMDC $(G^{\text{ave}}, \boldsymbol{\theta}^{\text{ave}})$ maximizing the regularized conditional log-likelihood (15), provided that, for every pair $(Z^j, \text{pa}(Z^j))$, and every $\text{pa}^{\mathcal{D}}(Z^j) \in \mathcal{Pa}^{\mathcal{D}}(Z^j)$, a local optimal classifier $\mathbf{h}_{\text{pa}^{\mathcal{D}}(Z^j)}$ satisfying (12) can be learned by maximizing

$$\ell^{\text{ave}}(\mathcal{Pa}^{\mathcal{D}}(Z^j), \boldsymbol{\theta}_j \mid \mathcal{S}) = \sum_{\{\bar{\mathbf{z}}_m \in \bar{\mathcal{S}}: \text{pa}^{\mathcal{D}}(\bar{\mathbf{z}}_m^j) = \mathcal{Pa}^{\mathcal{D}}(Z^j)\}} w(\bar{\mathbf{z}}_m) \log \left(\mathbf{p}_{(G, \boldsymbol{\theta})}(\bar{\mathbf{z}}_m^j \mid \mathcal{Pa}^{\mathcal{D}}(Z^j), \bar{\mathbf{x}}_m^c) \right) , \quad (40)$$

under the feature-selection option.

Proof. Similar to the proof of Proposition 5. $\square$

B.2 Proofs of Section 5 (Inference and Prediction Phase)

B.2.1 Proofs for Propositions of Section 5.1 (Inference from Incomplete Data)

Proposition B.10. Let u be the 0/1 accuracy $u_{0/1}$ (24). For any $\mathbf{z} \in \mathcal{Z}$, and for any HMDC $(G^*, \boldsymbol{\theta}^*)$, we have

$$\begin{aligned} \hat{\mathbf{y}}^{\text{opt}} \oplus \hat{\mathbf{x}}^{\mathcal{M}} &\in \arg\max_{\hat{\mathbf{z}}^{\mathcal{M}} \in \mathcal{Z}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{z}}^{\mathcal{M}} \mid \mathbf{z}^{\mathcal{O}}) , \\ \hat{\mathbf{y}}^{\text{ave}} &\in \arg\max_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}} \mid \mathbf{z}^{\mathcal{O}}) . \end{aligned}$$

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

$$\begin{aligned} \hat{\mathbf{y}}^{\text{opt}} &\in \arg\max_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\mathbf{p}_{(G^*, \boldsymbol{\theta}^*)}} [u(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) \mid \mathbf{x}^c] \\ &= \arg\max_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \mathbb{I}[\hat{\mathbf{y}}^{\mathcal{M}} = \mathbf{y}^{\mathcal{M}}] \mathbf{p}_{(G^*, \boldsymbol{\theta}^*)}(\mathbf{z}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}} \mid \mathbf{x}^c) \\ &= \arg\max_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \mathbb{I}[\hat{\mathbf{y}}^{\mathcal{M}} = \mathbf{y}^{\mathcal{M}}] \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \end{aligned} \quad (41)$$

$$\begin{aligned}
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \overbrace{\mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}})}^{\text{Using Bayes' Theorem}} \underbrace{\mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{O}})}_{\text{A constant}} \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) .
\end{aligned}$$

Therefore, $\hat{\mathbf{y}}^{\text{opt}}$ can be found by solving

$$\hat{\mathbf{y}}^{\text{opt}} \oplus \hat{\mathbf{x}}^{\mathcal{M}} \in \operatorname{argmax}_{\hat{\mathbf{z}}^{\mathcal{M}} \in \mathcal{Z}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{z}}^{\mathcal{M}} | \mathbf{x}^c) .$$

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

$$\begin{aligned}
\hat{\mathbf{y}}^{\text{ave}} &\in \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\mathbf{p}_{(G^*, \boldsymbol{\theta}^*)}} [u(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) | \mathbf{x}^c] \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) . \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) .
\end{aligned} \tag{42}$$

□

Proposition B.11. *Let u be the Hamming accuracy u_H (25). For any $\mathbf{z} \in \mathcal{Z}$, for any HMDC $(G^*, \boldsymbol{\theta}^*)$, $\hat{\mathbf{y}}^{\text{opt}}$ (22) is estimated by finding, $\forall k \in \mathcal{M}$,*

$$\hat{\mathbf{y}}^{\text{opt}, k} \oplus \hat{\mathbf{x}}^{\mathcal{M}} \in \operatorname{argmax}_{(y^k, \mathbf{x}^{\mathcal{M}}) \in \mathcal{Y}^k \times \mathcal{X}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(y^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) ,$$

and $\hat{\mathbf{y}}^{\text{ave}}$ (23) is obtained by finding, $\forall k \in \mathcal{M}$,

$$\hat{\mathbf{y}}^{\text{ave}, k} \in \operatorname{argmax}_{y^k \in \mathcal{Y}^k} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^c)}(y^k | \mathbf{z}^{\mathcal{O}}) .$$

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

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

$$\begin{aligned}
\hat{\mathbf{y}}^{\text{opt}} &\in \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\mathbf{p}_{(G^*, \boldsymbol{\theta}^*)}} [u(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) | \mathbf{x}^c] \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\frac{1}{|\mathcal{Y}^{\mathcal{M}}|} \sum_{Y^k \in \mathcal{Y}^{\mathcal{M}}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] \right) \mathbf{p}_{(G^*, \boldsymbol{\theta}^*)}(\mathbf{z}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}} | \mathbf{x}^c)
\end{aligned} \tag{43}$$

$$\begin{aligned}
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{k \in \mathcal{M}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] \right) p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \frac{1}{|\mathbf{Y}^{\mathcal{M}}|} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{k \in \mathcal{M}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] \right) p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{k \in \mathcal{M}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] \right) p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{k \in \mathcal{M}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] \right) p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \left(\sum_{k \in \mathcal{M}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \right) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} \left(\sum_{\substack{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}} \\ \hat{\mathbf{y}}^k = y^k}} \mathbb{I}[\hat{\mathbf{y}}^k = y^k] p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \right) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} \left(\sum_{\substack{\mathbf{y}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}} \\ \hat{\mathbf{y}}^k = y^k}} p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{x}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \right) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^k, \mathbf{x}^{\mathcal{M}}, \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} \underbrace{p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{O}})}_{\text{Using Bayes' Theorem}} \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \left(\sum_{k \in \mathcal{M}} p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) \right) \underbrace{p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\mathbf{z}^{\mathcal{O}})}_{\text{A constant}} \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) .
\end{aligned}$$

which can be estimated (since for different $\mathbf{y}^k \in \mathbf{Y}^{\mathcal{D}}$, $\hat{\mathbf{y}}^{\text{opt},k}$ (44) may be attained with different $\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}$) by finding

$$\hat{\mathbf{y}}^{\text{opt},k} \in \operatorname{argmax}_{\hat{\mathbf{y}}^k \in \mathcal{Y}^k} \max_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}), k \in \mathcal{M}. \quad (44)$$

For any $k \in \mathcal{M}$, $\hat{\mathbf{y}}^{\text{opt},k}$ (44) can be found by finding

$$\hat{\mathbf{y}}^{\text{opt},k} \oplus \hat{\mathbf{x}}^{\mathcal{M}} \in \operatorname{argmax}_{\hat{\mathbf{y}}^k \cup \mathbf{x}^{\mathcal{M}} \in \mathcal{Y}^k \times \mathcal{X}^{\mathcal{M}}} p_{(G^{\mathcal{D}}, \mathbf{x}^c)}(\hat{\mathbf{y}}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) . \quad (45)$$

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

$$\begin{aligned}
\hat{\mathbf{y}}^{\text{ave}} &\in \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{E}_{\mathbf{p}_{(G^*, \theta^*)}} [u(\hat{\mathbf{y}}^{\mathcal{M}}, \mathbf{y}^{\mathcal{M}}) | \mathbf{x}^{\mathcal{C}}] \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})}(\hat{y}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} \sum_{\mathbf{x}^{\mathcal{M}} \in \mathcal{X}^{\mathcal{M}}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})}(\hat{y}^k, \mathbf{x}^{\mathcal{M}} | \mathbf{z}^{\mathcal{O}}) \\
&= \operatorname{argmax}_{\hat{\mathbf{y}}^{\mathcal{M}} \in \mathcal{Y}^{\mathcal{M}}} \sum_{k \in \mathcal{M}} \mathbf{p}_{(G^{\mathcal{D}}, \mathbf{x}^{\mathcal{C}})}(\hat{y}^k | \mathbf{z}^{\mathcal{O}}) ,
\end{aligned} \tag{46}$$

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

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

□

C Characteristics of the Datasets

This section reports summaries of the distributions and pairwise associations of the discrete features and class variables in the three datasets: Adult, Default, and Thyroid. We report normalized entropy to summarize the degree of imbalance of the discrete variables. For a discrete variable with K possible outcomes and class probabilities $p_1, \dots, p_K$, its normalized entropy is

$$H = -\frac{1}{\log K} \sum_{k=1}^K p_k \log p_k .$$

This quantity ranges from 0 to 1. A value close to 1 indicates that the categories are approximately equally represented, whereas a value close to 0 indicates that one category dominates the others and, hence, a highly imbalanced distribution.

As mentioned in Section 4, pairwise associations among the discrete variables do not necessarily reflect the sparsity of the optimal DAGs learned when conditioning on the continuous features. Nevertheless, sparse DAGs may reasonably be expected to perform well on these two datasets, as the satisfactory performance obtained for all discrete features by BR suggests that their conditional dependencies are weak to moderate.

C.1 Adult Dataset

The Adult dataset has 4 class variables and 5 discrete features. For each class variable and discrete feature, we report the Normalized entropy in Table 2. Most discrete

variables exhibit relatively high normalized entropy, indicating fairly balanced distributions, whereas X^4 has a much lower entropy, reflecting strong imbalance. The association matrix shown in Figure 8 indicates that class variables and discrete features exhibit strong dependencies.

Table 2: Normalized entropy of discrete variables of Adult dataset

Class	Normalized entropy (%)
Y^1	0.8480
Y^2	0.7502
Y^3	0.5999
Y^4	0.9413
X^1	0.7455
X^2	0.8881
X^3	0.8390
X^4	0.1965
X^5	0.8220

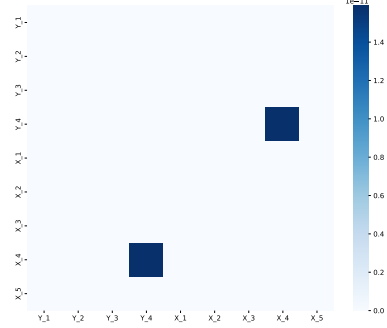


Fig. 8: p-values for all pairwise chi-square tests across discrete variables of the Adult dataset

C.2 Default Dataset

The Default dataset has 4 class variables and 6 discrete features. We report the Normalized entropy in Table 3. Most discrete features exhibit relatively high normalized entropy, indicating fairly balanced categorical distributions. Among the class variables, Y^1 has the highest entropy (0.9658), whereas Y^2 and Y^3 are notably lower (0.5662 and 0.5435). The association matrix shown in Figure 9 indicates that class variables and discrete features exhibit strong dependencies.

C.3 Thyroid Dataset

The Thyroid dataset has 7 class variables and 22 discrete features. The normalized entropy values reported in Table 4 and Table 5 are relatively low for all class variables, indicating highly imbalanced class distributions. In particular, Y^6 has the lowest entropy (0.0198), while Y^4 has the highest (0.3373), confirming the pronounced class imbalance in this dataset. The association matrix shown in Figure 10 indicates a mix of strong and weak dependencies across class variables and discrete features.

Table 3: Normalized entropy of discrete variables of Default dataset

Class	Normalized entropy (%)
Y^1	0.9658
Y^2	0.5662
Y^3	0.5435
Y^4	0.7306
X^1	0.8369
X^2	0.8494
X^3	0.8487
X^4	0.8261
X^5	0.8076
X^6	0.8310

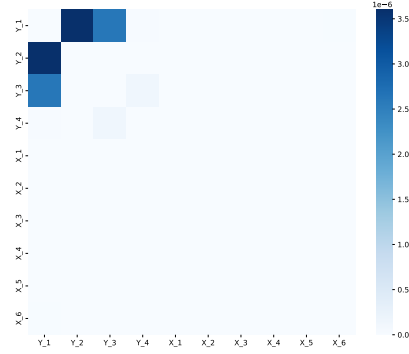


Fig. 9: p-values for all pairwise chi-square tests across discrete variables of the Default dataset

Table 4: Normalized entropy of the discrete variables of Thyroid dataset

Class	Normalized entropy
Y^1	0.0872
Y^2	0.1948
Y^3	0.1810
Y^4	0.3373
Y^5	0.1488
Y^6	0.0198
Y^7	0.1423

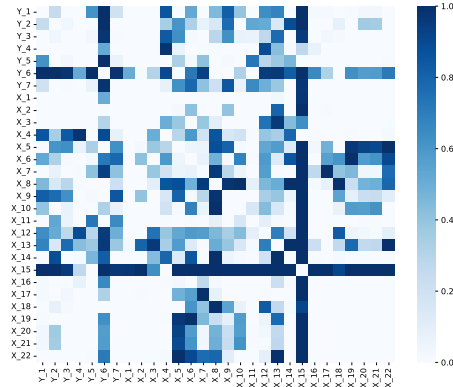


Fig. 10: p-values for all pairwise chi-square tests across discrete variables of the Thyroid dataset

Table 5: Normalized entropy of the discrete variables in the Thyroid dataset

Feature	Normalized entropy (%)	Feature	Normalized entropy (%)
X^1	0.5764	X^{12}	0.0815
X^2	0.8866	X^{13}	0.0754
X^3	0.5715	X^{14}	0.1755
X^4	0.1225	X^{15}	0.0027
X^5	0.0976	X^{16}	0.2674
X^6	0.2307	X^{17}	0.4424
X^7	0.0919	X^{18}	0.8607
X^8	0.1099	X^{19}	0.2787
X^9	0.1324	X^{20}	0.4304
X^{10}	0.3611	X^{21}	0.4277
X^{11}	0.3696	X^{22}	0.2335

C.4 Per-Class Balanced Accuracy

We report in Figures [11-13](#) the per-class balanced accuracy for five methods (BR, HMDC-MI, HMDC-OP, HMDC-AV, and RefC) with LR, using BIC as the regularization term, when restricting each node to at most two parents. Results obtained when allowing up to three parents per node are similar and are omitted for brevity. The other results are presented in the [Supplementary Results](#).

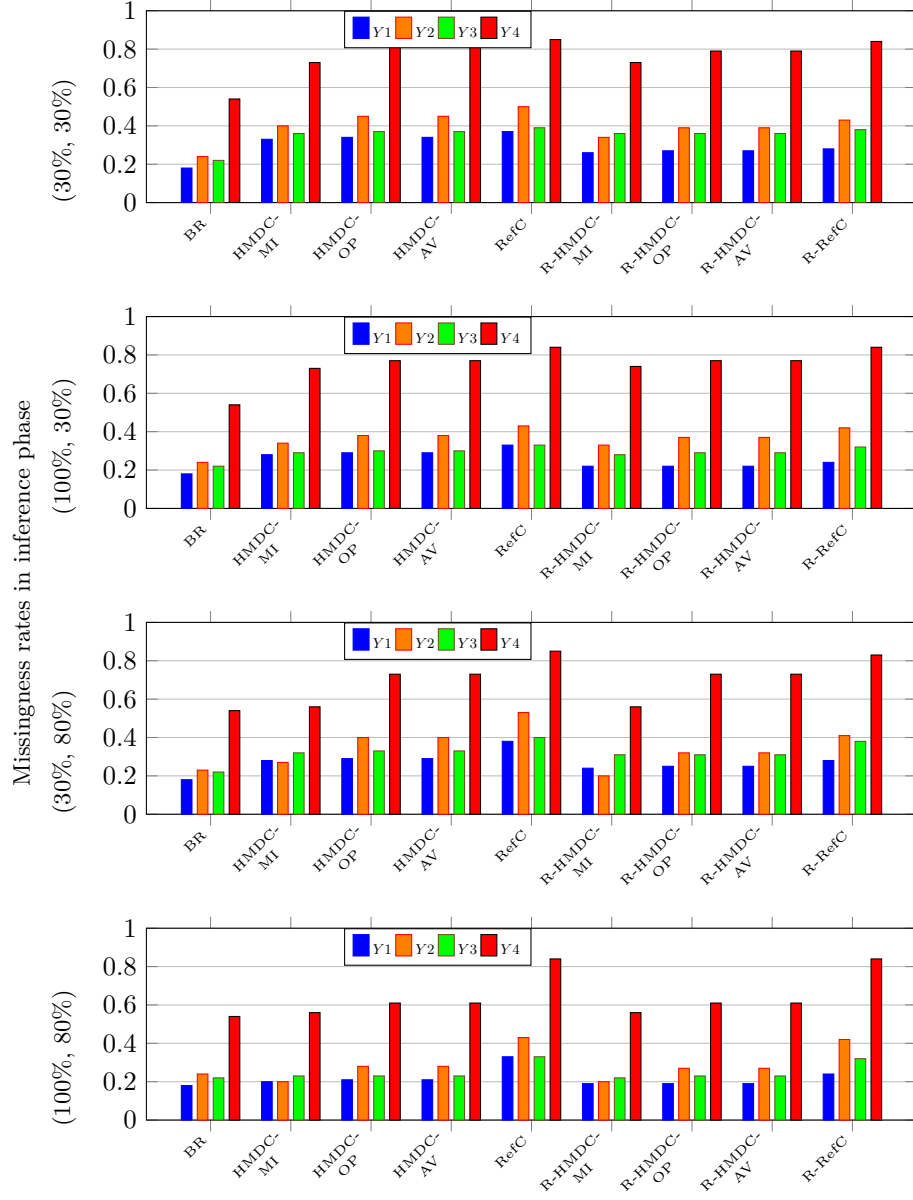


Fig. 11: Adult dataset, learning performed on complete data using LR as the base classifier and BIC regularization score. Balanced accuracies ($\uparrow$) under missingness rates of (30%, 30%), (100%, 30%), (30%, 80%), and (100%, 80%) in the inference phase, where the first and second values denote the missingness rates for class variables and discrete features, respectively.

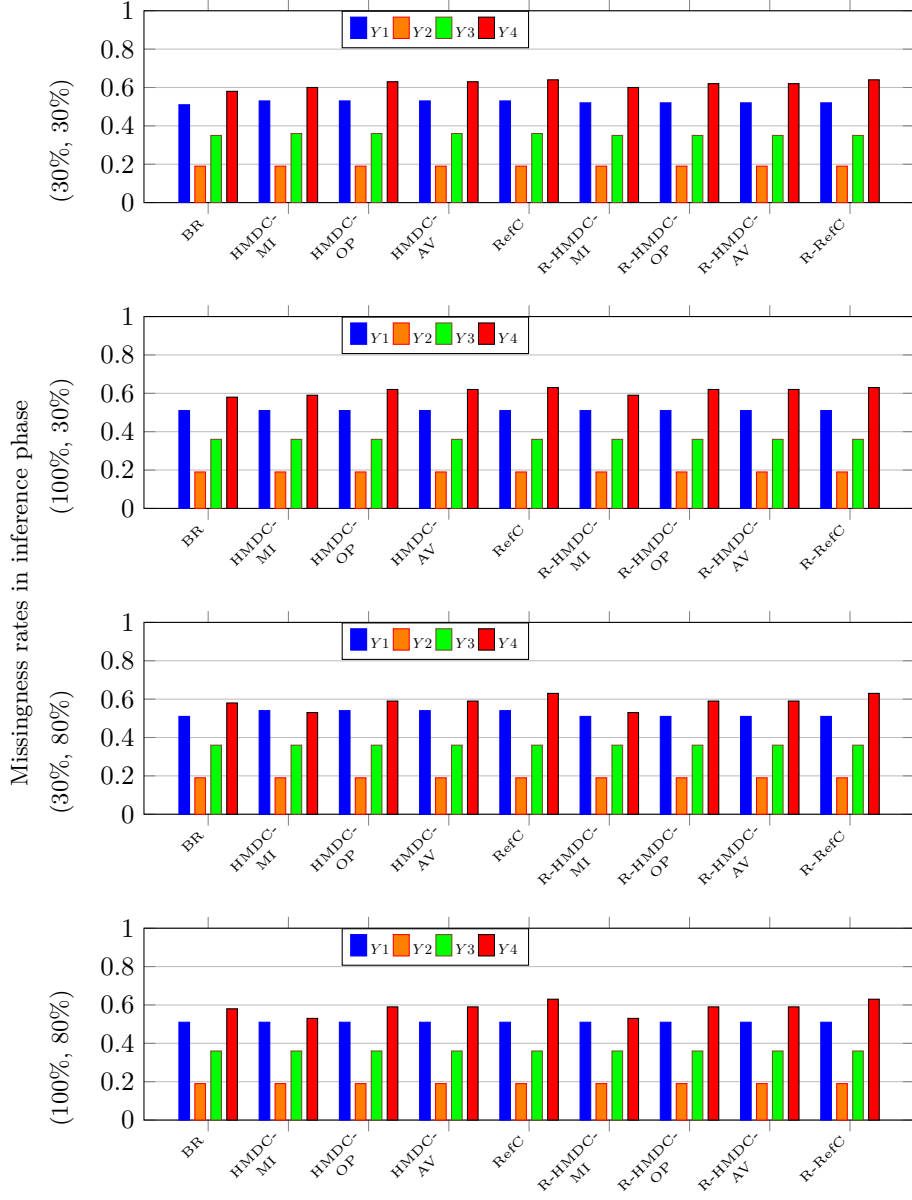


Fig. 12: Default dataset, learning performed on complete data using LR as the base classifier and BIC regularization score. Balanced accuracies ($\uparrow$) under missingness rates of (30%, 30%), (100%, 30%), (30%, 80%), and (100%, 80%) in the inference phase, where the first and second values denote the missingness rates for class variables and discrete features, respectively.

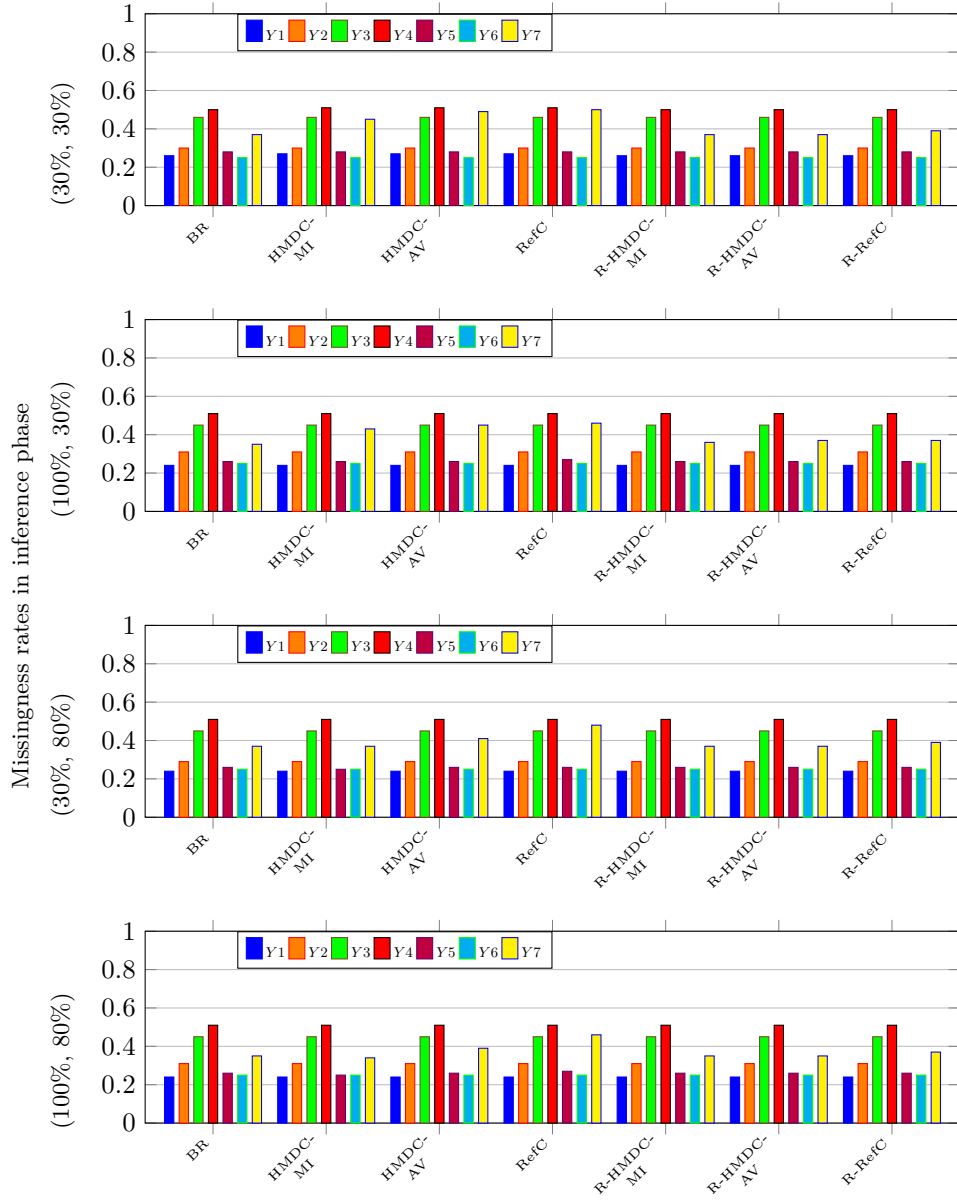


Fig. 13: Thyroid dataset, learning performed on complete data using LR as the base classifier and BIC regularization score. Balanced accuracies ($\uparrow$) under missingness rates of (30%, 30%), (100%, 30%), (30%, 80%), and (100%, 80%) in the inference phase, where the first and second values denote the missingness rates for class variables and discrete features, respectively.