On the Correlation between Individual Fairness and Predictive Accuracy in Probabilistic Models

Alessandro Antonucci[1,*], Eric Rossetto[1] and Ivan Duvnjak[2]

[1]*IDSIA (USI-SUPSI), Lugano, Switzerland*
[2]*SUPSI, Lugano, Switzerland*

Abstract

We investigate individual fairness in generative probabilistic classifiers by analysing the robustness of posterior inferences to perturbations in private features. Building on established results in robustness analysis, we hypothesise a correlation between robustness and predictive accuracy—specifically, instances exhibiting greater robustness are more likely to be classified accurately. We empirically assess this hypothesis using a benchmark of fourteen datasets with fairness concerns, employing Bayesian networks as the underlying generative models. To address the computational complexity associated with robustness analysis over multiple private features with Bayesian networks, we reformulate the problem as a *most probable explanation* task in an auxiliary Markov random field. Our experiments confirm the hypothesis about the correlation, suggesting novel directions to mitigate the traditional trade-off between fairness and accuracy.

Keywords

Algorithmic Fairness, Bayesian Networks, Markov Random Fields, Robustness Analysis.

1. Introduction

The widespread adoption of machine learning and knowledge-based systems for decision support and recommendation has made the development of tools and metrics for algorithmic fairness increasingly urgent [1]. Algorithmic fairness generally refers to the identification and mitigation of biases in model predictions with respect to sensitive, or *private*, features. However, simply omitting these private features from the model is often insufficient: it can significantly degrade predictive accuracy and fails to guarantee fairness, as other features may remain highly correlated with the excluded ones [2].

In this paper, we focus on generative probabilistic models that produce predictions in the form of posterior distributions over the target variable, conditioned on a complete assignment of both private and public features. For such models, biases can be naturally characterized through the robustness of predictions to perturbations in private features. We demonstrate how this notion of robustness aligns with established formulations of inference under non-ignorable missingness [3]. Prior work has further shown that robustness can be positively correlated with predictive accuracy [4]. This connection is particularly compelling, as it suggests a mitigation strategy for the classical trade-off between fairness and accuracy [5]: at the *individual* level [6], fair classifications may also be accurate, allowing standard predictive models to be employed without modification for the robust instances.

To validate such a conjecture we consider an extensive benchmark including fourteen datasets related to fairness. In our study, we adopt Bayesian networks as probabilistic generative models, compute the robustness, based on a fairness deviation measure [7] for each test instance, and evaluate the correlation with the predictive performance by using the Brier score. The empirical results we observe are in line with those reported in the earlier literature on robustness (e.g., [8]): the most robust, and hence, in our setup, fair, instances are also "easier" to classify. To the best of our knowledge, our work is the first attempt to apply concepts from robust probabilistic methods to fairness analysis.

We leverage this novel connection to offer a more interpretable perspective on fairness. Future work will investigate how this link might be harnessed to actively control fairness during training, as well as evaluate its effectiveness in comparison to existing methods (e.g., [9]).

The robustness measure we adopt could, in principle, require a brute-force computation over all possible joint assignments of the private features. However, we reformulate the inference as a *most probable explanation* (MPE) task within an auxiliary Markov random field derived from the original Bayesian network. This reformulation yields computational benefits, which we both formally establish and validate empirically.

The code used for the experiments is available for reproducibility in a dedicated Github repository.[1] The repository includes references to the original datasets, preprocessing scripts, discretised versions of the datasets, and a custom implementation of some inference algorithms.

The paper is organised as follows. In Section 2, we define the necessary background on probabilistic graphical models. Our approach to fairness analysis is discussed in Section 3. The mapping to Markov random fields is derived in Section 4. The benchmark we adopt is detailed in Section 5, while the results of our experiments are in Section 6. A critical discussion on the limitations of the present work and possible strategies to bypass them is in Section 7, while our conclusions are in Section 8. Technical results and additional experimental analyses are gathered in the appendix.

2. Basics

Let us first review the necessary background information on probabilistic graphical models. We remind the reader to [10] for deeper insights.

Variables, Potentials, and Tables. We denote variables by uppercase letters and their states with lowercase letters. A subscripted variable in Ω is used instead for the sample space. Thus $v \in \Omega_V$ is a generic state of the variable V. We focus on discrete variables only. A non-negative function $\phi : \Omega_V \to \mathbb{R}$ is called a *potential* over V. A potential over V that is also normalised to one is called *probability mass function* (PMF) and denoted as $P(V)$. PMF P gives the probability $P(v)$ to $V = v$ for each $v \in \Omega_V$. A *conditional probability table* (CPT) for V given W is a collection of PMFs over V indexed by the values of W and it is denoted as $P(V|W)$. We also define potentials over joint variables. As an example, the CPT $P(V|W)$ is a special example of a multi-variate potential $\phi(V, W)$.

Bayesian Networks (BNs). Consider a set of variables $\boldsymbol{V}$. A BN B over $\boldsymbol{V}$ is a collection of CPTs, one for each $V \in \boldsymbol{V}$, say $\{P(V|\mathrm{Pa}_V)\}_{V \in \boldsymbol{V}}$ where $\mathrm{Pa}_V \subseteq \boldsymbol{V} \setminus \{V\}$ for each $V \in \boldsymbol{V}$. From the BN specification B, we induce a directed graph $\mathcal{G}_B$ whose nodes are in one-to-one correspondence with the variables in $\boldsymbol{V}$ and such that there are directed arcs from W to V for each $W \in \mathrm{Pa}_V$ and $V \in \boldsymbol{V}$. Here we only consider *acyclic* BNs, i.e., BNs such that $\mathcal{G}_B$ does not contain directed loops. An acyclic BN B compactly defines a joint PMF $P(\boldsymbol{V})$ as follows:

$$P(\boldsymbol{v}) = \prod_{V \in \boldsymbol{V}} P(v|\mathrm{pa}_V), \tag{1}$$

for each $\boldsymbol{v} \in \Omega_{\boldsymbol{V}}$, where the values v and pa_V are those consistent with $\boldsymbol{v}$. Given a dataset of i.i.d. joint observations of $\boldsymbol{V}$, there are various *structural learning* techniques to obtain a directed acyclic graph over the variables. Typical approaches are maximising likelihood-based scores. Once the graph is detected, the BN CPTs can be quantified from the data by means of a count function n, i.e.,

$$P(v|\mathrm{pa}_V) = \frac{n(v, \mathrm{pa}_V) + \frac{s}{|\Omega_V|}}{\sum_{v \in \Omega_V} n(v, \mathrm{pa}_V) + s}, \tag{2}$$

where a Laplace smoothing with equivalent sample size $s > 0$ is preferred to a frequentist estimate.

---

[1] github.com/IDSIA-papers/fairness.

Given a node W, the *Markov blanket* V_W of W is a subset of V including the variable W, its parents Pa_W, the children Ch_W of W, and the parents of those children. We denote as $\mathcal{G}_W$ the subgraph of $\mathcal{G}$ including only the variables in V_W. An example is in Figure 1. A BN over V_W and based on $\mathcal{G}_W$ can be obtained by taking the CPTs of W and its children from B and arbitrarily setting the marginal PMFs for the parents of W and the parents of its children. In the following we will use uniform PMFs for these marginals and denote such a BN as B_W.
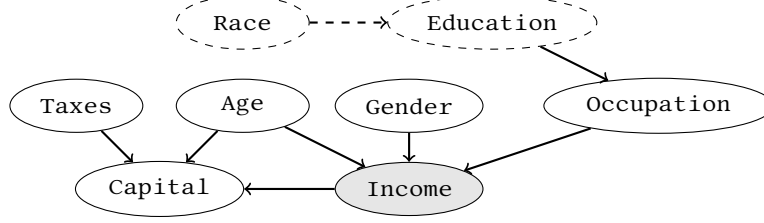

Figure 1: The directed graph of a BN. The nodes outside the Markov blanket of node `Income` are dashed.

Markov Random Fields (MRFs).   A MRF M over V is a collection of k potentials $\{\phi_i(V_i)\}_{i=1}^{k}$, with $V_i \subseteq V$ for each $i = 1, \ldots, k$ and such that for each $V \in V$ there is at least an index i for which $V \in V_i$. Given the MRF M, we induce an undirected graph $\mathcal{G}_M$ whose k nodes are in correspondence with the joint variables $\{V_i\}_{i=1}^{k}$ and such that there is an edge between two nodes if and only if the corresponding joint variables share at least one variable. An example is in Figure 2. A MRF defines a joint PMF $P(V)$ as follows:

$$P(v) \propto \prod_{i=1}^{n} \phi_i(v_i), \tag{3}$$

for each possible value v of V, where the values of v_i are those consistent with v.
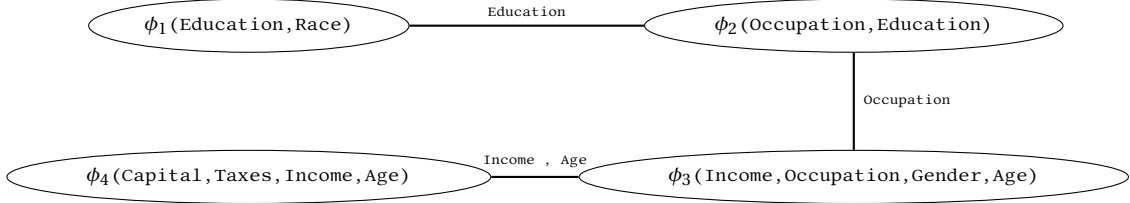

Figure 2: The undirected graph of a MRF. Labels over edges highlight the variables shared among potentials.

A BN can be intended as an MRF whose potentials are the CPTs. Because of Eq. (1), for those models, there is no normalisation constant to compute in Eq. (3). As an example the MRF induced by the BN in Figure 1 is based on the undirected graph in Figure 2.

Inference.   A typical inference tasks in MRFs and BNs is *updating*. This consists in the computation of the posterior PMF $P(V_Q|v_E)$ for a queried variable $V_Q \in V$, given an evidence about some other variables $V_E = v_E$ with $V_E \subseteq V \setminus \{V_Q\}$. *Marginal MAP* (MMAP) consists instead of finding the most probable configuration of a joint queried variable $V_Q \subseteq V$ given the evidence, i.e.,

$$v_Q^* = \arg\max_{v_Q} P(v_Q|v_E). \tag{4}$$

If, in particular, $V = (V_Q, V_E)$, i.e., all the variables are either queried either observed, the task is called *most probable explanation* (MPE). Moreover, notice that the maximisation in Eq. (4) is with respect to v_Q, the probability of the evidence $P(v_E)$ is a positive constant and the task in Eq. (4) can be equivalently solved by replacing the conditional probability in the r.h.s. with a joint probability.

Both updating and MPE tasks can be solved exactly by *variable elimination* schemes. Given an elimination order over V, this consists in processing the potential obtained by multiplying all the potentials, including the variable we want to eliminate. The processing consists in the marginalisation of the variables not involved in the query, the instantiation of the observed variables and the queries of the updating task, and the maximisation of the queried variables of the MPE. The maximum *arity* of the potentials created in this way is called *treewidth*. Inference with this scheme is exponential with respect to this parameter.

3. (Probabilistic) Fairness Analysis

We are interested in evaluating the presence of biases in a training dataset $\mathscr{D}$ over the discrete variables $V := (Y, X, Z)$. The dataset refers to a classification task where the target variable is Y and the features are distinguished between private, those in X, and public, those in Z. A test dataset $\mathscr{D}'$ over the same variables is also assumed available. We learn from $\mathscr{D}$ a generative probabilistic model in the form of a joint PMF $P(Y, X, Z)$. In our classification task, we want to identify the most probable value of Y given a test instance $(\hat{x}, \hat{z})$ of the features taken from $\mathscr{D}'$, i.e.,

$$y^* := \arg\max_{y \in \Omega_Y} P(y | \hat{x}, \hat{z}), \tag{5}$$

where a zero-one loss function is considered just for the sake of light notation. Besides achieving a good predictive level on the test instances, we might be interested in bounding the sensitivity of inferences to changes in the values of the private features and, ideally, keep these changes as small as possible. This could be done at the predictive level, i.e., checking whether the same class as in Eq. (5) is obtained for perturbed values of the private features $\hat{x}$. However, a deeper analysis can be achieved by a dissimilarity measure δ for the PMFs over the target Y. We thus consider the following optimisation:

$$x^* := \arg\max_{x \in \Omega_X} \delta \left[P(Y | x, \hat{z}), P(Y | \hat{x}, \hat{z}) \right], \tag{6}$$

and call *fairness robustness level* (FRL) of the instance $(\hat{x}, \hat{z})$ the corresponding maximum, i.e.,

$$\rho(\hat{x}, \hat{z}) := \delta \left[P(Y | x^*, \hat{z}), P(Y | \hat{x}, \hat{z}) \right]. \tag{7}$$

It is a simple remark that $\rho = 0$ implies a context-specific irrelevance of the private features to the target given a particular instance of the public features.

Let us consider, for the sake of simplicity, the case of a Boolean target Y and take as distance δ the Manhattan distance, i.e.,

$$\delta \left[P(Y), P'(Y) \right] := \frac{|P(y_0) - P'(y_0)| + |P(y_1) - P'(y_1)|}{2} = |P(y_0) - P'(y_0)|, \tag{8}$$

where y_0 is a short form for $Y = 0$ and similarly for y_1. Because of Eq. (8), Eq. (6) rewrites as:

$$x^* := \arg\max_{x \in \Omega_X} \left| P(y_0 | x, \hat{z}) - P(y_0 | \hat{x}, \hat{z}) \right|. \tag{9}$$

It is a trivial exercise to notice that considering y_1 instead of y_0 in Eq. (9) would not affect x^*. This implies that the FRL corresponds to the *fairness deviation measure* often used in the literature to rank unfairness [11]. Rather than directly addressing the optimisation in Eq. (9), we can focus on the two following optimisations:

$$\overline{x} = \arg\max_{x \in \Omega_X} P(y_0 | x, \hat{z}), \tag{10}$$

$$\underline{x} = \arg\min_{x \in \Omega_X} P(y_0 | x, \hat{z}). \tag{11}$$

This is possible because of the following result whose proof is in the appendix.

Proposition 1. *The solutions of the optimisations in Eqs.* (9), (10) *and* (11) *are linked by the following relation:*

$$x^* := \begin{cases} \overline{x} & \text{if } P(y_0|\hat{x}, \hat{z}) < \frac{1}{2}\left[P(y_0|\overline{x}, \hat{z}) + P(y_0|\underline{x}, \hat{z})\right], \\ \underline{x} & \text{otherwise}. \end{cases} \tag{12}$$

It is important to notice that the task in Eq. (10) is substantially different from an MPE task as in Eq. (4) because the queried variables in Eq. (10) are after the conditioning bar (see discussion in Section 2). Thus, while in MPE we can equivalently address the task in the joint model instead of the conditional one becase the evidence is always the same, the same cannot be done with Eq. (10), and similarly for Eq. (11). These tasks can be regarded as a *conservative* inference based on the law proposed in [3]. Yet, no standard BN algorithm can solve the task, while a brute-force approach requires a number of BN inferences exponential with respect to $|X|$. In [12], the MMAP-like optimisation has been mapped to a *credal network* inference, whose complexity might be still demanding [13]. Here we leverage on the fact that we cope with complete observations to map the task to a proper MPE task in an auxiliary MRF. This is detailed in the next section.

4. Robustness of BN Classifiers as a MRF MPE

Thanks to Proposition 1, we can solve the optimisation in Eq. (9) and hence compute the FRL in Eq. (7) by solving the optimisations in Eqs. (10) and (11). In this section we show how these BN tasks can be equivalently addressed in an auxiliary MRF M whose specification is discussed here below.

First, it is a trivial exercise to notice that, for the classification task in Eq. (5), the BN can be equivalently restricted to the BN over the Markov blanket of Y. In the following, we therefore assume that the BN coincides with the Markov blanket of Y.

Let X' and Z' denote, respectively, the private and public parents of Y and by X'' and Z'' the private and public children of Y. We define a potential $\phi_Y(X', Z')$ as:

$$\phi_Y(x', z') := \frac{P(y_1|x', z')}{P(y_0|x', z')} \tag{13}$$

for each $x' \in \Omega_{X'}$ and $z' \in \Omega_{Z'}$. Similarly, we define a potential for each $X \in X''$ and $Z \in Z''$ as follows:

$$\phi_X(x, \mathrm{pa}'_X) := \frac{P(x|\mathrm{pa}'_X, y_1)}{P(x|\mathrm{pa}'_X, y_0)}, \tag{14}$$

$$\phi_Z(z, \mathrm{pa}'_Z) := \frac{P(z|\mathrm{pa}'_Z, y_1)}{P(z|\mathrm{pa}'_Z, y_0)}, \tag{15}$$

for each $x \in \Omega_X$, $\mathrm{pa}'_X \in \Omega_{\mathrm{Pa}'_X}$, $z \in \Omega_Z$, and $\mathrm{pa}'_Z \in \Omega_{\mathrm{Pa}'_Z}$, where $\mathrm{Pa}'_X := \mathrm{Pa}_X \setminus \{Y\}$ and the same notation is also used for Z. Overall, Eq. (13) together with Eqs. (14) and (15) define a MRF $\phi(X, Z)$ based on $1 + |X''| + |Z''|$ potentials. The following result, whose proof is in the appendix, holds.

Proposition 2. *Consider the MRF induced by the potentials in Eq.* (13) *and Eqs.* (14) *and* (15). *Let* $\phi(X, Z)$ *denote the joint PMF induced by such a MRF. The BN optimisation in Eq.* (11) *is equivalent to a MPE task in this MRF, namely:*

$$\underline{x} = \arg \max_{x \in \Omega_X} \phi(x, \hat{z}). \tag{16}$$

For the task in Eq. (10) *we have instead:*

$$\overline{x} = \arg \min_{x \in \Omega_X} \phi(x, \hat{z}). \tag{17}$$

Solving the task in the MRF has the same worst-case complexity, exponential with respect to $|X|$, as trying a brute-force in the BN according to Eqs. (10) and (11). Yet, generally speaking the treewidth of the MRF can be smaller than $|X|$. In particular, any $X'' \in X''$ that is not a parent of one of the children of Y can be eliminated by processing ϕ_Y alone, this potentially inducing a significant reduction of the computational complexity required to solve the task. This claim will be empirically validated in Section 6.

5. A Benchmark of Datasets for Fairness Analysis

The main goal of this paper is to investigate whether the FRL defined in Eq. (7) has a relation with the predictive performance of a probabilistic classifier. For a first empirical analysis, we consider fourteen datasets providing a suitable benchmark for fairness-aware machine learning analyses. These datasets have been also considered in [14], where the choice of their private features is discussed and detailed. Table 1 contains some relevant information about the datasets. Apart from `Communities & Crime`, all datasets have a binary target variable Y. The target of that dataset is instead an integer corresponding to the number of violent crimes in a selected area. A binary Y is obtained by a median-based discretisation (sample median of 374 crimes). Continuous features are similarly processed by quantile-based discretisation leading to quaternary variables. Rows with missing values are finally ignored.

For the experiments, we consider a ten-fold cross validation scheme. The folds are obtained by stratification with respect to the target values. For the sake of reproducibility, the code used to process the original datasets as well as the corresponding discretised (and anonymised) datasets are available in the Github repository of the paper (see Footnote 1). The results of our experiments on these data are discussed in the next section.

Table 1
The datasets considered for the experiments. Besides the name of the dataset and the target variable, we report the number of public and private features, the total number of records, the imbalance level Δ describing the relative occurrence of the least frequent class, and the number k of folds (out of ten) on which fairness is achieved by *design* (see Section 6).

Dataset	Y	$|\boldsymbol{X}|$	$|\boldsymbol{Z}|$	$|\mathscr{D}|$	Δ	k
`Adult`	Income (50k)	3	9	48'842	.24	0
`Bank marketing`	Deposit Subscription	2	13	45'211	.12	5
`Communities & Crime`	Violent Crimes	1	20	1'994	.50	10
`COMPAS recid.`	Two-year Recidivism	2	8	7'214	.45	1
`COMPAS violent recid.`	Two-year Violent Recidivism	2	8	4'738	.16	3
`Credit card clients`	Default Payment	3	20	30'000	.0005	10
`Diabetes`	Re-admission	1	16	101'766	.46	10
`Dutch census`	Occupation	1	10	60'420	.48	10
`German credit`	Creditability	2	19	1'000	.30	10
`KDD Census-Income`	Income (50k)	2	21	200'000	.07	0
`Law school`	Pass the bar exam	2	7	20'512	.05	1
`OULAD`	Final Result	3	9	31'482	.46	10
`Student-Mathematics`	Final Grade	2	31	395	.33	10
`Student-Portuguese`	Final Grade	2	31	649	.15	10

6. Experiments

For a deeper characterisation of the relation between the FRL in Eq. (7) and the predictive performance of the classifier, together with the zero-one accuracy of the single prediction in Eq. (5), we also consider the more informative *Brier score*, that is defined here as the square of one minus the probability assigned by the classifier to the actual class, i.e.,

$$\beta(\hat{y}, \hat{x}, \hat{z}) := [1 - P(\hat{y}|\hat{x}, \hat{z})]^2 . \tag{18}$$

A zero Brier score indicates perfect confidence in the correct class. It is easy to notice that, with binary classes, the prediction is correct if and only if $\beta(\hat{y}, \hat{x}, \hat{z}) < 0.25$.

We adopt BNs as probabilistic classifiers and the *Pyagrum*[2] Python library to learn BNs from the

---

data and compute the queries of interest. In particular, we employ a *tabu* local search with list size equal to 10 and a maximum number of consequent iterative changes equal to 100.

We focus on *individual* fairness by evaluating complete feature instances. In this context (see Section 4), features that lie outside the Markov blanket of the target variable are irrelevant for prediction. When private features are entirely irrelevant to the target, the classifier achieves fairness *by design*, without any trade-off against the accuracy, and the FRL is trivially zero on all test instances. In our cross-validation scheme, with a same dataset, we might obtain different BN topologies for the different folds. The last column of Table 1 reports the number of folds for which the BN is fair by design.

We compute the Brier score of each test instance. The low number of private features ($|\boldsymbol{X}| \leq 3$) allows to obtain the corresponding FRL by a brute-force approach on the BN according to Eq. (6). Yet, for a comparison, we obtain the same results by also solving the MPE task in the auxiliary MRF as discussed in Section 4. In the latter case, we use a custom implementation of the variable elimination algorithm, that also allows to solve the minimisation task in Eq. (17). To solve also the latter task, we replace each potential with its opposite.
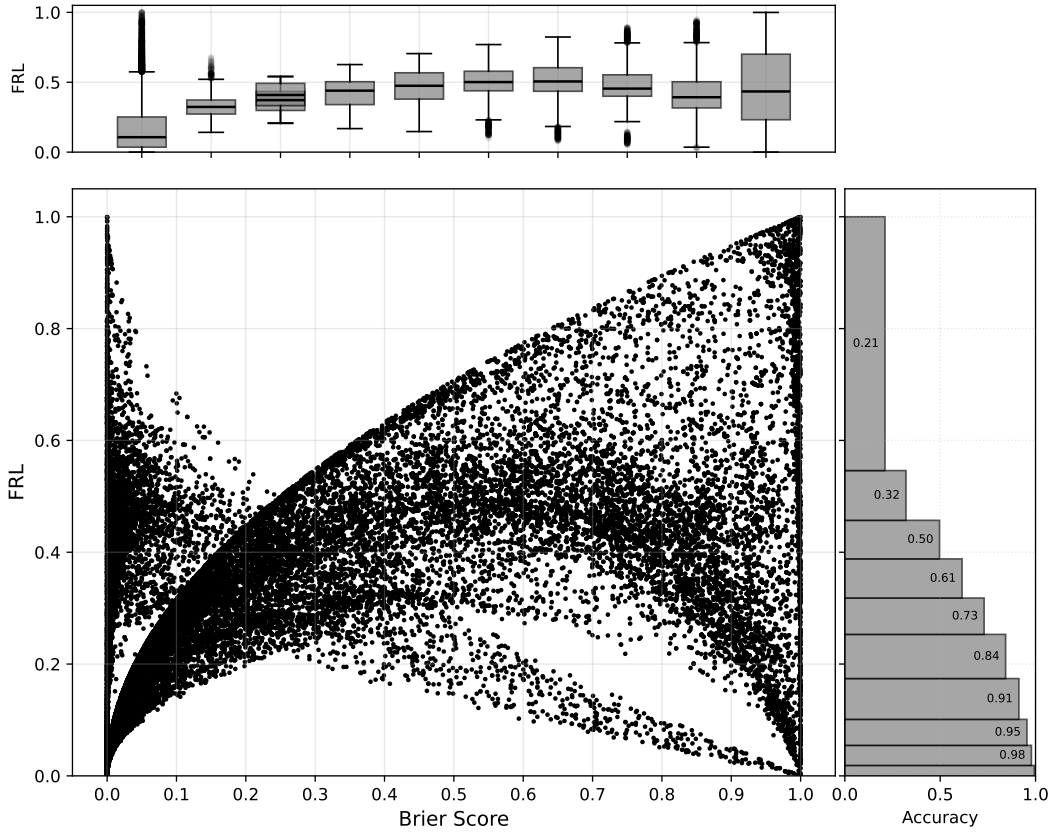

Figure 3: Results of the experiments with the `Adult` dataset.

A first example of our empirical analysis is provided by Figure 3, where the results for the `Adult` dataset are summarised. The scatter plot displays the Brier score versus the FRL value for each test instance. Instances satisfying $\beta(\hat{y}, \hat{x}, \hat{z}) < 0.25$—positioned in the left region of the plot—correspond to cases that are accurately classified (approximately 84% of the total sample). For these points, because of Eq. (9), the FRL cannot exceed the value of $P(\hat{y}|\hat{x}, \hat{z}) = 1 - \sqrt{\beta(\hat{y}, \hat{x}, \hat{z})}$. This explain why all these points are located under a parabolic curve. Similar considerations can be done for the incorrectly

classified instances corresponding to $\beta(\hat{y}, \hat{x}, \hat{z}) > 0.25$. This also explains the smaller FRL ranges we observe for less extreme values of the Brier score, as they can be observed in the boxplots obtained by equal-frequency binning with respect to the Brier score and depicted on the upper part of the figure.

The most notable aspect of Figure 3 is the set of histograms on the right. These depict predictive accuracy across ten equally sized subgroups of instances, defined by the quantiles of the FRL distribution. The results reveal a clear decline in prediction accuracy as FRL increase. This trend is consistent with the conjectured relationship between accuracy and robustness/fairness introduced in Section 1.

Figure 4(a) presents the same accuracy histograms alongside the corresponding Brier score levels. Predictive performance deteriorates with increasing FRL values, as indicated by both the decline in accuracy and the rise in Brier losses (dashed lines). A similar pattern emerges in Figure 4(b), where the KDD Census-Income dataset is considered.
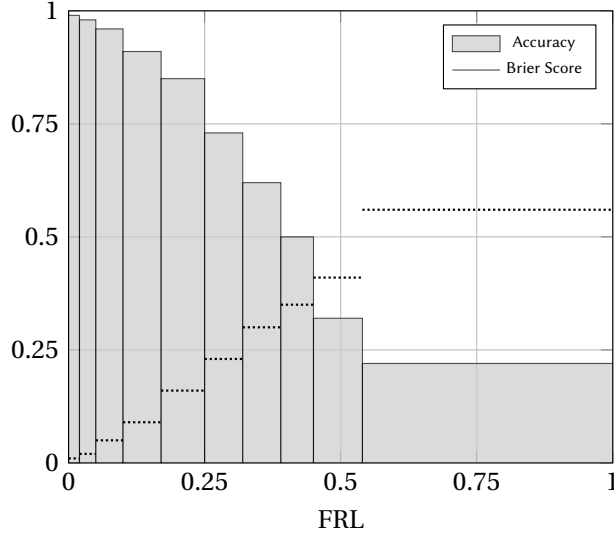
(a) Adult
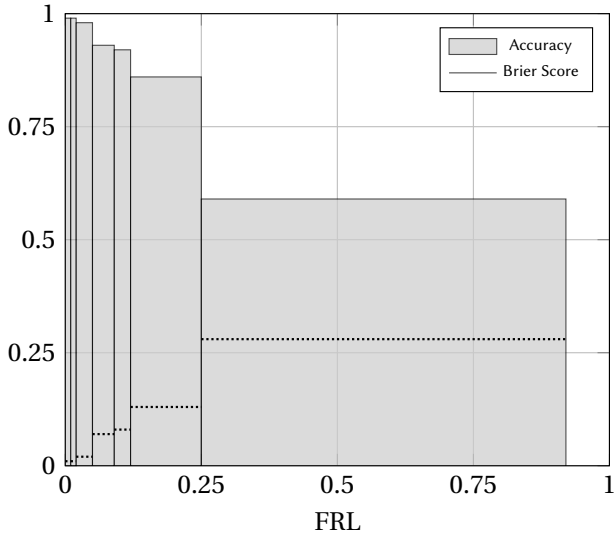
(b) KDD Census-Income

Figure 4: FRL versus accuracy and Brier score histograms for datasets where at least one private feature remains predictive of the target in each fold.

As shown in Table 1, Adult and KDD Census-Income are the only datasets of our benchmark not achieving fairness by design on any fold (i.e., such that $k = 0$), this meaning that no BN learned from these datasets makes all the private features irrelevant to the target.

Let us now consider the datasets for which fairness by design is achieved only on some folds, but not on all of them (i.e., $0 < k < 10$). On these folds we have FRL= 0 by construction. For an effective visualisation of the relation between FRL and the predictive accuracy (or Brier score) we use the same kind of histograms as in Figure 4, with an additional bar/line of fixed width on the left side, depicting the predictive performance on the instances with FRL= 0. The results are in Figure 5. Also for these models we observe a clear degradation of the predictive performance for increasing FRL levels.

(a) `Law school`

(b) `Bank marketing`

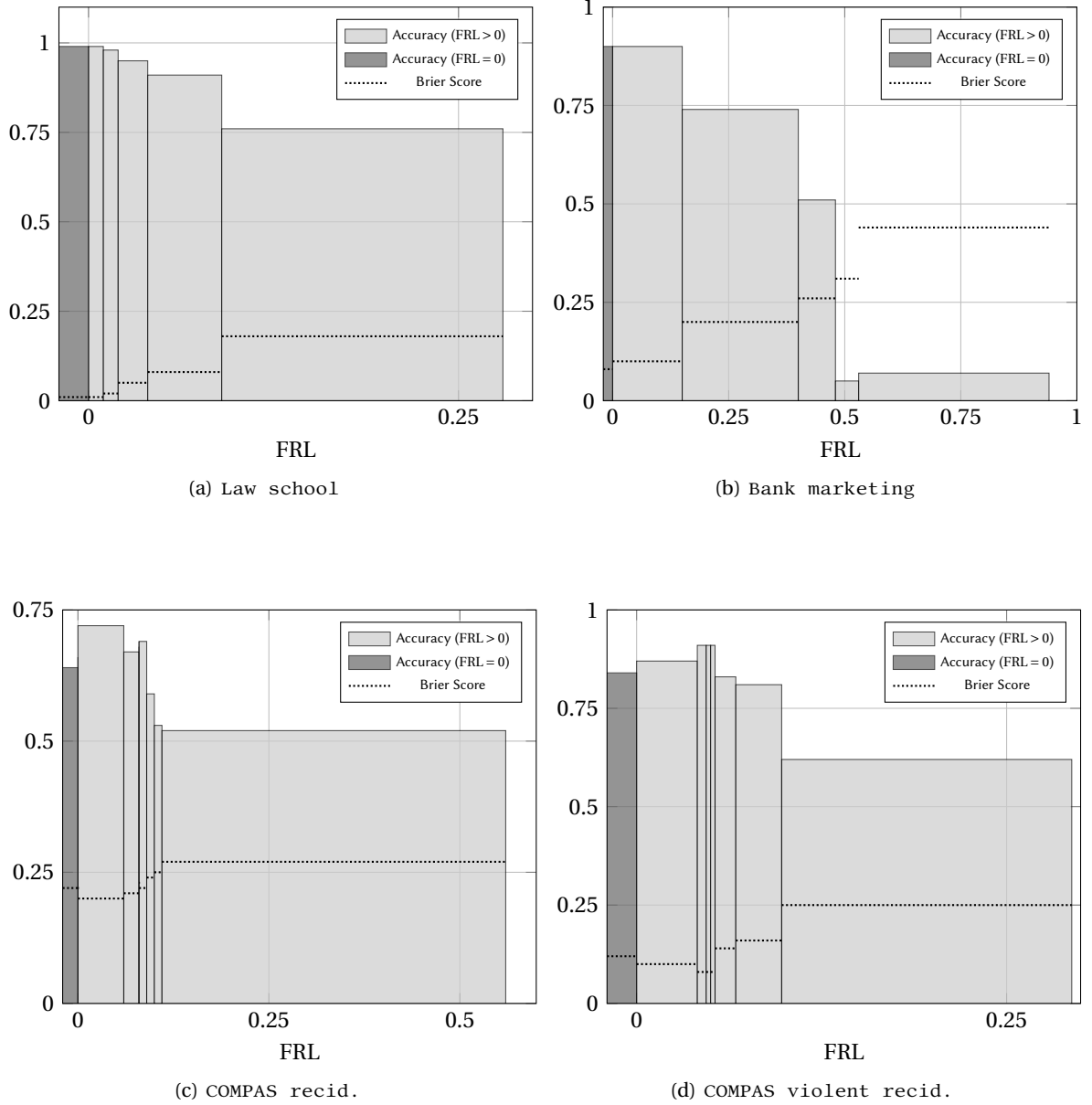(c) `COMPAS recid.`

(d) `COMPAS violent recid.`

Figure 5: FRL versus accuracy and Brier score histograms for datasets where at least one private feature remains predictive of the target in at least a fold.

On the remaining eight datasets, fairness is achieved by design on each fold (i.e., $k = 10$, see Table 1) and we have zero FRL on all the instances in the dataset. Yet, for a further validation of our conjecture, we re-train the BN structures by forcing arcs between the target and all the features. This can be trivially done by constraining the local search meta-heuristics we adopt. The results, depicted in Figure 6, confirm the trend observed in the previous experiments. We notice that degradation of the performance for higher FRLs is typically less evident on datasets on which the predictive performance of the BNs is poorer (such as those in Figs. 6(e)-6(h)).
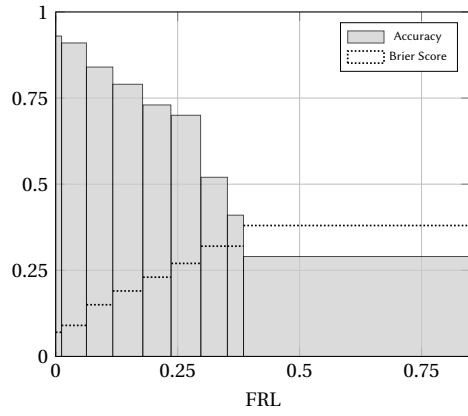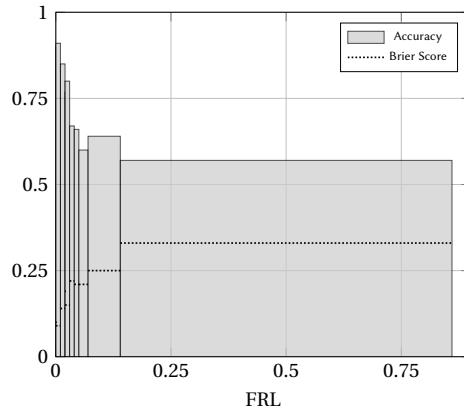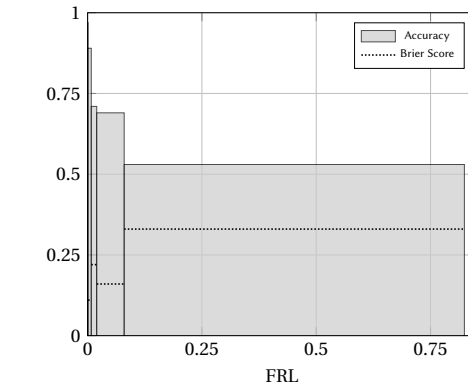
(a) Communities & Crime
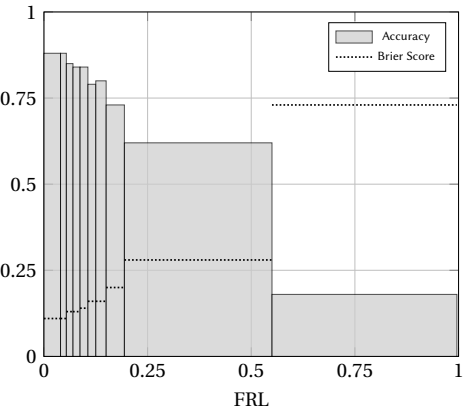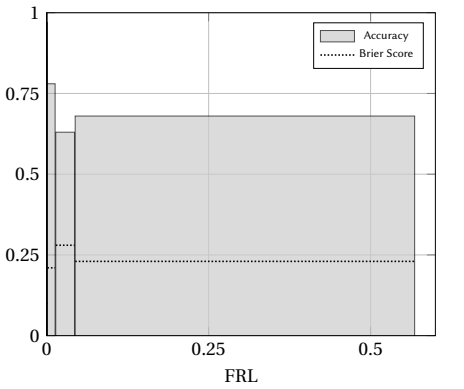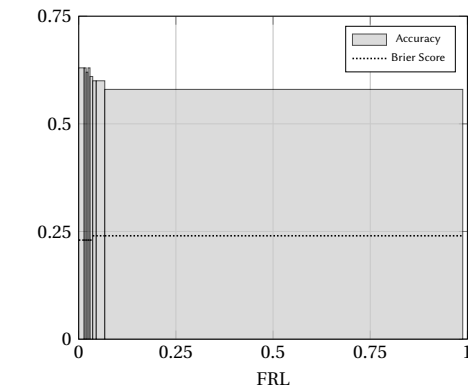
(b) Credit card clients
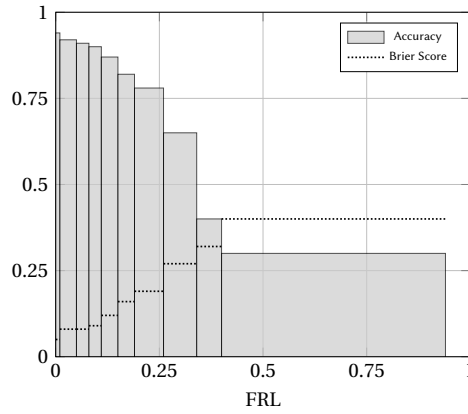
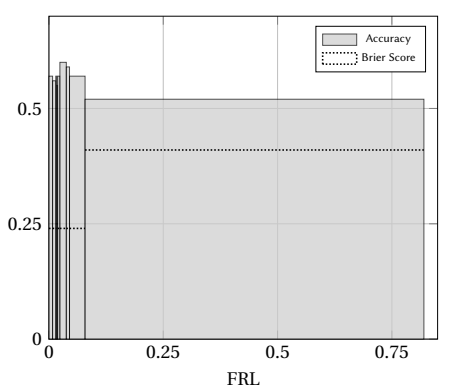(c) Dutch census

(d) German credit

(e) Student-Mathematics

(f) Student-Portuguese

(g) OULAD

(h) Diabetes

Figure 6: FRL vs. accuracy/Brier for datasets where the target is forced to be a parent of the features.

Finally, to evaluate the computational advantages of the MRF-based approach to the computation of the FRL with respect to the brute-force BN method, Figure 7 shows boxplots of the ratio between the execution times for each test instance across the two methods. The low number of private features (i.e., $|X| \leq 3$) makes the difference in performance between the two approaches small. Nevertheless, as expected, the datasets with three private features exhibits the largest discrepancy.

Figure 7: Relative execution times for FRL computation in the BN (Eq. (7)) with respect to the MRF approach (Section 4). The x-labels denote the dataset and the number of private features.

Finally, Figure 8 presents the boxplots of the same ratio while treating as private an increasing number of public features of the `Student-Mathematics` dataset. These results underscore the computational efficiency of the MRF-based method, particularly as the number of private features grows.

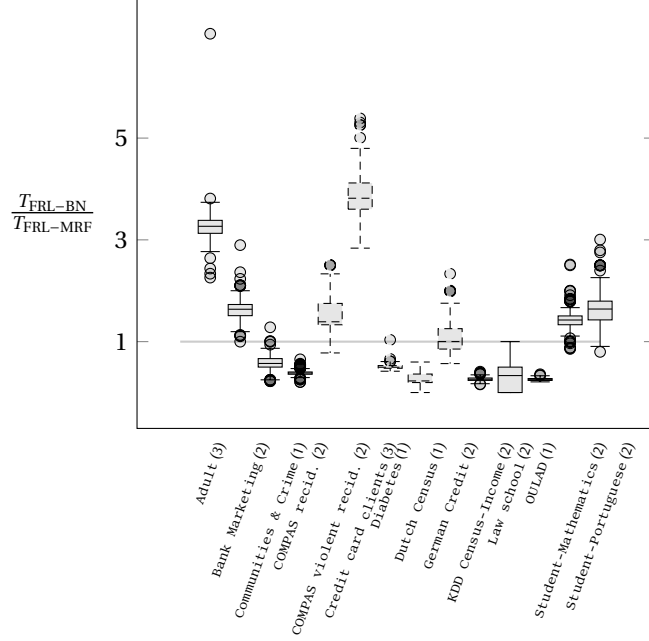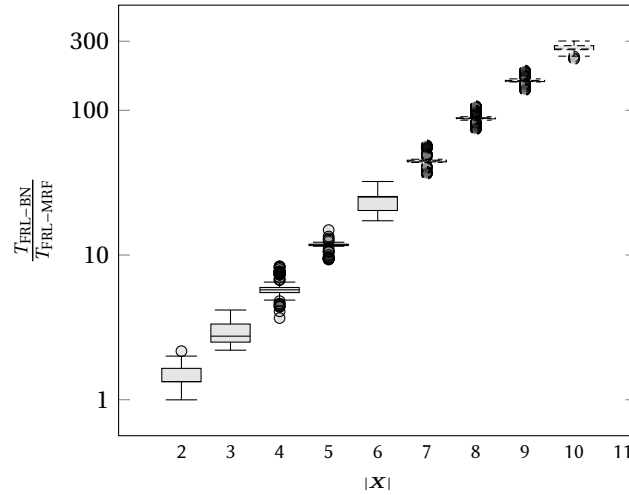**Figure 8:** Relative execution times for FRL computation in the BN (Eq. (7)) with respect to the MRF approach (Section 4) for in increasing number of features treated as private in the `Student-Mathematics` dataset.

7. Limitations and Outlooks

The analysis conducted in this study is purely *correlational*, and the directed graphical models employed here are not intended to capture causal relationships. Accordingly, the fairness deviation measure we derived via the FRL does not carry a causal interpretation, and the different instantiations of the private features should not be interpreted as interventions. The approach of Plečko and Bareinboim [15] on *counterfactual fairness* offers a promising foundation for integrating causal reasoning into the models discussed here.

Our work focuses on discrete features for the sake of a simple presentation. An extension to a hybrid setup mixing discrete and continuous features could be obtained within the framework of linear-Gaussian BNs [10], for which also extensions to the causal setup have been already discussed. The situation is different for the target variable: extending the framework to regression tasks with a continuous target would require a substantial reformulation of the concepts introduced in this paper. In contrast, our focus on binary classes can be naturally extended to non-binary targets by conducting classification and FRL evaluations through pairwise dominance tests applied to each pair of classes.

Regarding the measure considered to evaluate the robustness of the posterior inferences, the current focus on the Manhattan distance has clear computational advantages related to the linear structure of such measure. Coping with non-linear alternatives (such as the KL) could be much more challenging, while an extension to variance and similar measure of variation could be more manageable, and naturally allow for the extension to continuous targets.

Once our framework will be extended to hybrid models that incorporate both continuous and discrete variables, the experimental evaluation should be revisited. A more refined representation of the model variables is expected to enhance the overall predictive accuracy. Assessing whether the discriminative power of the FRL remains stable under these conditions constitutes a necessary future investigation. A validation under *controlled bias* scenarios[3] should be also considered.

Finally, extending the framework from individual to *group* fairness would be conceptually straightforward, although it would preclude the use of the MRF mapping. To maintain computational tractability and avoid costly inference procedures, more efficient models-such as probabilistic circuits [16]—could serve as practical alternatives to BNs.

8. Conclusions

We examined whether the correlation between robustness and predictive accuracy—commonly observed in the field of robust probabilistic models—also holds in a fairness-aware setting. The results of an extensive empirical evaluation are promising: predictive accuracy consistently declines for less robust instances. This observation opens new avenues for mitigating the classical fairness–accuracy trade-off. Specifically, instances with high fairness scores can be handled by standard models, while a fairness-constrained, potentially less accurate model could be selectively applied to non-robust instances. Exploring this targeted approach to fairness mitigation represents an exciting and necessary direction for future research in machine learning.

Acknowledgements

---

[3]See, e.g., github.com/rcrupiISP/BiasOnDemand.

Declaration on Generative AI

For the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: *Grammar and spelling check, Paraphrase and reword.* After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

A. Proofs

In this appendix, we gather the proofs of the two technical results of the paper.

Proof of Proposition 1. *The following straightforward result:*

$$\max_t |f(t)| = \max \left\{ \max_t [f(t)], \max_t [-f(t)] \right\}, \tag{19}$$

allows to rewrite the maximum of the optimisation in Eq. (9) as follows:

$$\max \left\{ \max_{\boldsymbol{x} \in \Omega_{\boldsymbol{X}}} \left[P(y_0|\boldsymbol{x}, \hat{\boldsymbol{z}}) - P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) \right], \max_{\boldsymbol{x} \in \Omega_{\boldsymbol{X}}} \left[P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\boldsymbol{x}, \hat{\boldsymbol{z}}) \right] \right\} \tag{20}$$

As $P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}})$ is a constant, we can move the maximisation with respect to $\boldsymbol{x}$ and obtain the maximum and the minimum in Eqs. (10) and (11), i.e.,

$$\max \left\{ \left[P(y_0|\overline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) \right], \left[P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\underline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) \right] \right\}. \tag{21}$$

The above maximum between two quantities can be clearly rewritten as:

$$\begin{cases} P(y_0|\overline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) & \text{if } P(y_0|\overline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) > P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\underline{\boldsymbol{x}}, \hat{\boldsymbol{z}}), \\ P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\underline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) & \text{otherwise}. \end{cases} \tag{22}$$

And hence:

$$\begin{cases} P(y_0|\overline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) & \text{if } P(y_0|\overline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) + P(y_0|\underline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) > 2P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}), \\ P(y_0|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) - P(y_0|\underline{\boldsymbol{x}}, \hat{\boldsymbol{z}}) & \text{otherwise} \end{cases}, \tag{23}$$

which can be seen as an equivalent formulation of Eq. (12), where the values of the maxima appear instead of the points where these maxima are achieved.

Proof of Proposition 2. *As we cope with a Boolean target variable, i.e., $|\Omega_Y| = 2$, the objective function of the optimisations in Eqs. (10) and (11) rewrites as follows:*

$$P(y_0|\boldsymbol{x}, \hat{\boldsymbol{z}}) = \frac{P(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}})}{P(\boldsymbol{x}, \hat{\boldsymbol{z}})} = \frac{P(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}})}{P(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}}) + P(y_1, \boldsymbol{x}, \hat{\boldsymbol{z}})} = \frac{1}{1 + \frac{P(y_1, \boldsymbol{x}, \hat{\boldsymbol{z}})}{P(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}})}} \tag{24}$$

As $f(t) = (1 + t)^{-1}$ is a monotone decreasing function of t, we can address Eq. (11) as follows:

$$\overline{\boldsymbol{x}} := \arg \max_{\boldsymbol{x} \in \Omega_{\boldsymbol{X}}} \frac{P(y_1, \boldsymbol{x}, \hat{\boldsymbol{z}})}{P(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}})}. \tag{25}$$

The BN factorisation in Eq. (1) allows to rewrite the objective function in Eq. (25) as follows:

$$\frac{P(y_1, \boldsymbol{x}, \hat{\boldsymbol{z}})}{P(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}})} = \frac{\prod_{V \in (Y, \boldsymbol{X}, \boldsymbol{Z})} P_1(v|\mathrm{pa}_v)}{\prod_{V \in (Y, \boldsymbol{X}, \boldsymbol{Z})} P_0(v|\mathrm{pa}_V)} = \prod_{V \in (Y, \boldsymbol{X}, \boldsymbol{Z})} \frac{P_1(v|\mathrm{pa}_v)}{P_0(v|\mathrm{pa}_V)}, \tag{26}$$

where P_1 denotes the elements of the BN CPTs when the states (v, pa_V) are those consistent with $(y_1, \boldsymbol{x}, \hat{\boldsymbol{z}})$ and P_0 when consistent with $(y_0, \boldsymbol{x}, \hat{\boldsymbol{z}})$. The product in the rightmost hand side of Eq. (26) can be restricted to ratios of CPT values including Y, i.e., the CPTs associated with Y and its children. This corresponds to the joint PMF of the MRF and proves Eq. (16). The result in Eq. (17) follows analogously.

B. FRL-vs-Brier-Score Curves

The scatter plot in Figure 9 details the output of an experiment with the `Bank marketing` dataset in terms of relation between FRL levels and Brier scores. Here, as well as in similar plots we obtained for the other datasets, the points seem to lie on different curves. In the figure, two groups of seven points, each belonging to a different curve, have been highlighted in black. The points in a same group have the same value of the private feature *marital status* and are obtained by changing the public features. Namely, the points with higher Brier scores are those associated with *single*, and the other ones with *married*. We also observe that each point in a group has a corresponding point in the other group with the same FRL value. These pairs correspond to instances with the same public states and different private state. The reason is that, as a straightforward consequence of the definition, two instances with the same public feature have the same FRL level.
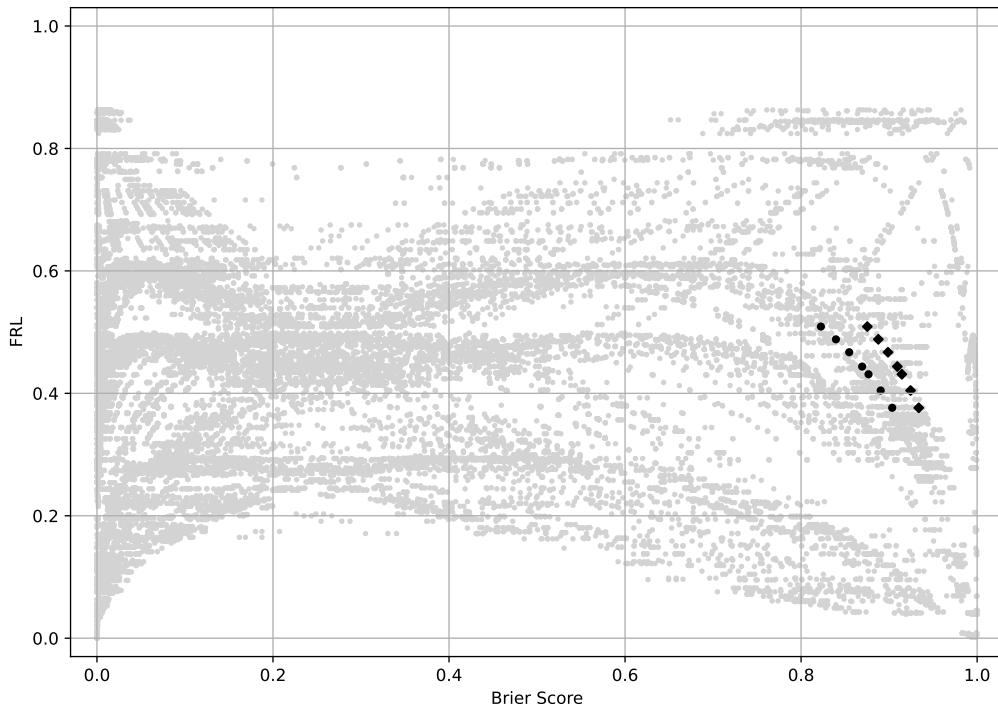

Figure 9: A FRL-vs-brier-score scatter plot for the the `Bank marketing` dataset.

References

[1] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annual review of statistics and its application 8 (2021) 141–163.

[2] R. Cummings, V. Gupta, D. Kimpara, J. Morgenstern, On the compatibility of privacy and fairness, in: Adjunct publication of the 27th conference on user modeling, adaptation and personalization, 2019, pp. 309–315.

[3] G. De Cooman, M. Zaffalon, Updating beliefs with incomplete observations, Artificial Intelligence 159 (2004) 75–125.

[4] J. V. Llerena, D. D. Mauá, A. Antonucci, Cautious classification with data missing not at random using generative random forests, in: J. Vejnarová, N. Wilson (Eds.), Symbolic and Quantitative

Approaches to Reasoning with Uncertainty, Springer International Publishing, Cham, 2021, pp. 284–298.

[5] A. F. Cooper, E. Abrams, N. Na, Emergent unfairness in algorithmic fairness-accuracy trade-off research, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 46–54.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[7] B. Woodworth, S. Gunasekar, M. I. Ohannessian, N. Srebro, Learning non-discriminatory predictors, in: S. Kale, O. Shamir (Eds.), Proceedings of the 2017 Conference on Learning Theory, volume 65 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1920–1953.

[8] L. Mattei, A. Antonucci, D. D. Mauá, A. Facchini, J. Villanueva Llerena, Tractable inference in credal sentential decision diagrams, International Journal of Approximate Reasoning 125 (2020) 26–48.

[9] H. Xu, X. Liu, Y. Li, A. Jain, J. Tang, To be robust or to be fair: Towards fairness in adversarial training, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 11492–11501.

[10] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT, 2009.

[11] N. Konstantinov, C. H. Lampert, Fairness through regularization for learning to rank, arXiv preprint arXiv:2102.05996 (2021).

[12] A. Antonucci, M. Zaffalon, Equivalence between Bayesian and credal nets on an updating problem, in: J. Lawry, E. Miranda, A. Bugarin, S. Li, M. A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Proceedings of third international conference on Soft Methods in Probability and Statistics (SMPS-2006), Springer, Bristol, United Kingdom, 2006, pp. 223–230.

[13] D. D. Mauá, C. P. de Campos, A. Benavoli, A. Antonucci, Probabilistic inference in credal networks: new complexity results, Journal of Artificial Intelligence Research 50 (2014) 603–637.

[14] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, WIREs Data Mining and Knowledge Discovery 12 (2022).

[15] D. Plečko, E. Bareinboim, Causal fairness analysis: A causal toolkit for fair machine learning, Foundations and Trends® in Machine Learning 17 (2024) 304–589.

[16] Y. Choi, A. Vergari, G. Van den Broeck, Probabilistic circuits: A unifying framework for tractable probabilistic models, UCLA Technical Reports (2020).