A Qualitative Causal Approach to Determining Adequate Training Data Quantity for Machine Learning

F. Mignet[1], F. Slijkhuis[1], A. Abouhafc[1], G. Pavlin[1], K. B. Laskey[2]

[1]Thales Nederland, Delft, The Netherlands, [gregor.pavlin, franck.mignet, filip.slijkhuis]@nl.thalesgroup.com
[2]George Mason University, Fairfax, VA, USA, klaskey@gmu.edu

***Abstract*—This paper proposes an improved analysis of the Qualitative Models of Data Generating Processes (QM-DGP). The approach supports (i) determination of the complexity of a Machine Learning problem and (ii) a coarse determination of the quantities of training data that are needed to train good quality models. Compared to the previously published approach to the QM-DGP analysis, this paper introduces a more thorough and theoretically sound treatment of the learning complexity. Firstly, the approach provides more rigorous determination of the complexity of the data generating processes (DGP). Secondly, the determination of the learning complexity and the required training data volumes is based on sound statistical principles for the estimation of the distributions over categorical variables. The effectiveness of the proposed method was experimentally confirmed in controlled settings. Different ground truth models were used to sample test and training data. The approach correctly predicts the size of the training data sets for which machine learning yields models supporting classification close to Bayes Error. While the majority of the experiments were carried out on probabilistic graphical models (PGM), the experiments with Neural Networks confirmed that the QM-DGP approach is not limited to PGMs.**

I. Introduction

This paper builds on the ideas introduced in [1], where it was shown that the causal knowledge captured in a qualitative data generation process could be exploited for improved machine learning in multiple ways.

Firstly, the architecture of a model must correspond to the complexity of the data generating processes. The architecture must be sufficiently expressive to capture the relevant dependencies between the modelled variables. In this way the architecture has the capacity to absorb the relevant knowledge from the training data.

Secondly, a key to obtaining good models is to obtain training data collected under all relevant conditions, i.e. contexts, that are expected to be encountered during the real world operation of the learned models.

The number of different contexts and the complexity of the data generating process (DGP) determine the modelling complexity and the size of the training sets required for the training of adequate models. The Qualitative Models of the Data Generating Processes (QM-DGP), introduced in [1] provide a guidance for a systematic analysis of the complexity of the learning tasks in a relevant class of problems. QM-DGPs are causal models capturing the dependency structure among variables that influence data generation. As shown in [1], such models can be used to determine the model architectures and facilitate a coarse estimation of the size of the required training data sets.

This paper focuses on several important aspects that, while mentioned in [1], deserve a more thorough treatment. Firstly, an improved process for determining the model complexity is proposed. The emphasis is on using the dependency structure of the QM-DGP to identify groups of variables that are tightly interconnected within groups and not between groups. The resulting decoupling eases the combinatorics and allows for achieving good performance with smaller training sets. The process is formally described and a procedure is provided.

Secondly, the heuristic method of [1] is replaced by a theoretically justified method to determine the required sample size for estimating the distributions over categorical variables.

Thirdly, more rigorous metrics supporting empirical evaluation of the effectiveness of the proposed methods are introduced. Several convergence criteria are discussed and it is shown that the method for the estimation of the training data size yields reasonable results, in that the estimates of data size lead to models with near optimal classification performance.

Finally, the introduced process for the QM-DGP based analysis was evaluated in controlled experiments. The test and training data is sampled from synthetic ground truth probabilistic models of different complexity. For each version of the ground truth model we can reliably estimate the minimum possible classification error rates and the highest possible data likelihood for the underlying data distributions. These values are then used to investigate the performance of the trained models for different sized training data sets.

II. Qualitative Models of Data Generation Processes

We assume a learning task to obtain a classifier using model $\mathcal{M}$ defined over variables $\mathcal{V}_{\mathcal{M}}$. The model relates classification variable $X_c \in \mathcal{V}_{\mathcal{M}}$ and the observed variables $X_i \in \mathcal{O} \subset \mathcal{V}_{\mathcal{M}}$. $\mathcal{O}$ corresponds to the set of all observed features that are influenced by X_c. The classifier uses $\mathcal{M}$ to predict the states of the classification variable X_c given observations of the features represented by variables $X_i \in \mathcal{O}$. The training of $\mathcal{M}$ is based

on a set of training data $\mathcal{D}_{train}$, a set of records $d_k \in \mathcal{D}_{train}$, each d_k consisting of observations of the states of variables X_c and $\mathcal{O}$ obtained for model training. This paper assumes the observations collected in a record d_k are a result of a causal data generation process (DGP) that can be seen as an ancestral sampling process [2] from some distribution $P(\mathcal{V})$. $\mathcal{V}$ is the set of variables corresponding to phenomena that occur in the underlying DGP and it is assumed that $\mathcal{V}_{\mathcal{M}} \subseteq \mathcal{V}$. Each sample d_k corresponds to the observations of X_c and $\mathcal{O}$ in a specific situation and $\mathcal{D}_{train}$ is obtained by repeated sampling from $P(\mathcal{V})$.

To obtain a good quality model $\mathcal{M}$, its complexity must reflect the complexity of the correlations in the underlying DGP. While $\mathcal{M}$ should be as simple as possible, it must be rich enough, i.e. support the representation of all relevant combinations of values of the variables of the DGP to capture the relevant knowledge about the relations between modelled variables $V_{\mathcal{M}}$ from $\mathcal{D}_{train}$ during an ML process. **The complexity of the DGP correlations is in this paper defined as the minimum number of parameters required to specify a model that faithfully represents these correlations.** Moreover, the complexity of a model $\mathcal{M}$ representing a DGP influences the size of training data set $\mathcal{D}_{train}$ required for the training of good quality models.

A. Correlations in a Data Generation Process

The qualitative dependencies in a DGP governed by $P(\mathcal{V})$ can be expressed with a Qualitative Model of a Data Generation Process (QM-DGP). This is a directed acyclic graph that captures the relations in the data generating process in a qualitative manner. That is, the model contains nodes representing the phenomena of interest and directed links representing direct dependencies between these phenomena; but no strength of these dependencies is considered. Figure 1.a shows an example of such a model.

A QM-DGP graph typically consists of a combination of three types of gates: chain, fork and collider [3]. This paper pays a special attention to colliders as they typically can greatly increase the modelling complexity.

The set of variables directly influencing a random variable X_i is denoted by $\pi(X_i)$. In the graphical representation, variable X_i and $\pi(X_i)$ correspond to the nodes. $\pi(X_i)$ are called parent nodes of X_i, while X_i is a child node of any variable in $\pi(X_i)$. If $\pi(X_i) = \emptyset$, then X_i corresponds to a root node.

A QM-DGP accurately represents a DGP over a set of variables $\mathcal{V}$ if it is an Independence Map (I-MAP) of the underlying probability distribution $P(\mathcal{V})$ [4]. In that case we say that QM-DGP is *faithful* with respect to the underlying distribution $P(\mathcal{V})$. Given that a QM-DGP is faithful as a representation of the data generating process, its topology reveals useful information about the process complexity. Namely, the QM-DGP topology provides information on the number of different conditions under which the distribution over the states of any random variable $X_i \in \mathcal{V}$ has to be estimated.

To facilitate further analysis of the process complexity, we first review a few concepts introduced in [1]. One of them is the the notion of *DGP context variables*. In common parlance, the term context refers to "the interrelated conditions in which something exists or occurs".[1] The DGP context variables $\mathcal{C} \in \mathcal{V}$ in a QM-DGP are intended to represent the background conditions under which the process operates. Generally, a DGP context variable $Y_i \in \mathcal{C}$ is either a root node in the DGP graph or, if not a root node, has no non-context nodes as parents. In addition, a DGP context variable will not be d-separated [4] from X_c or $X_i \in \mathcal{O}$ by other context variables. This means it has an influence on the class and/or feature variables that is not accounted for by the other context variables. The set of DGP context variables satisfying this condition for the set of observable variables $\mathcal{O}$ is called the DGP relevant context $\mathcal{C}_{\mathcal{O}} \subseteq \mathcal{C}$, i.e. for any $Y_j \in \mathcal{C}_{\mathcal{O}}$ there exists at least one path between Y_j and X_c and/or at least one $X_i \in \mathcal{O}$. Such context variables influence the occurrence of the states of the variables that are represented in the trained model $\mathcal{M}$. Fig. 1.a shows an example where $\mathcal{O} = \{X_1, X_2\}$ and the corresponding relevant DGP context is $\mathcal{C}_{\mathcal{O}} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$.

Moreover, for each $\mathcal{M}$ we can obtain a **Reduced QM-DGP** relating $X_i \in \mathcal{O}$, X_c and $C_{\mathcal{O}}$. Fig. 1.b shows an example of such a reduced model that was obtained from the full QM-DGP in Fig. 1.a. In case the conditional probabilities for all relations in the QM-DGP were known, the model shown in Fig. 1.b could be obtained through marginalization of all latent variables along the paths connecting $\mathcal{C}_{\mathcal{O}}$, X_c and $X_i \in \mathcal{O}$ in Fig. 1.a. Moreover, if all context variables $\mathcal{C}_{\mathcal{O}}$ are removed from the reduced model, we obtain a *core model* $\mathcal{M}_{core}$ over a subset of variables $\mathcal{V}_{\mathcal{M}_{core}} \subset \mathcal{V}_{\mathcal{M}}$, corresponding to black nodes in Fig. 1.b. In this example $\mathcal{V}_{\mathcal{M}_{core}} = \{X_c, \mathcal{O}\}$. Note also that all core variables in $\mathcal{V}_{\mathcal{M}_{core}}$ are observed during the sampling of the training data. Furthermore, $\mathcal{M}_{core}$ is defined by specifying a joint probability distribution $P(\mathcal{V}_{\mathcal{M}_{core}})$. Since context variables influence the distribution over the core variables $\mathcal{V}_{\mathcal{M}_{core}}$, there are different versions of $P(\mathcal{V}_{\mathcal{M}_{core}})$, each corresponding to a different context. In Fig. 1.b joint distribution $P(X_c, X_1, X_2)$ over the core model depends on the DGP context, which is determined by the values of the context variables $\mathcal{C}_{\mathcal{O}} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$ (blue nodes).

B. Expressing the DGP complexity

In this paper the complexity of a DGP has been defined as the minimum number of parameters needed to specify a PGM model $\mathcal{M}$ that is faithful to the correlations in the DGP. The overall complexity depends on the topology of the PGM. Let's assume a simple DGP corresponding to a model consisting of a single collider variable X_i and a set of parent variables $\pi(X_i)$. A collider variable X_i is associated with a number of parameters that increases exponentially with the number of its parents $\pi(X_i)$, unless modeling assumptions are imposed to control the number of parameters. For simplicity of exposition it is assumed initially that all variables are discrete

[1] https://www.merriam-webster.com/dictionary/context

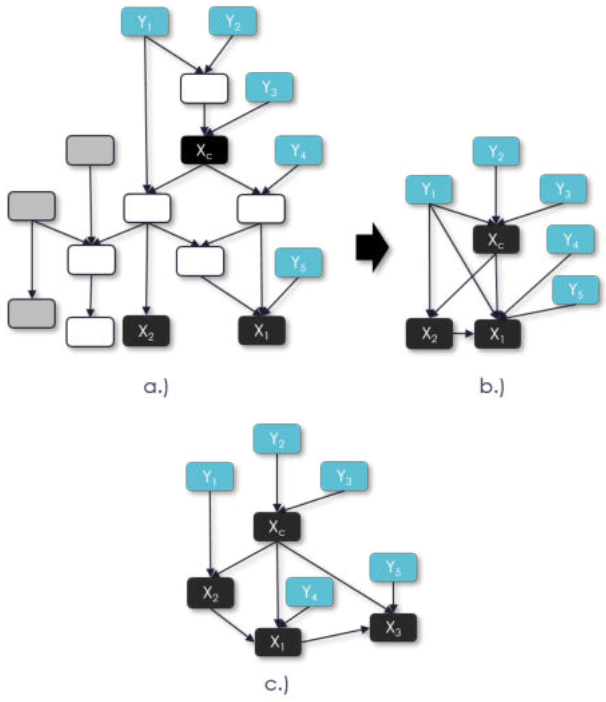

Fig. 1. a.) An example QM-DGP describing dependencies in $\mathcal{V}$, where the black nodes denote the class and the observed feature variables in the trained model $\mathcal{M}$, while the blue nodes represent the relevant context variables in the DGP. Gray nodes are observed context and feature variables that are irrelevant for the inference over X_c. b.) A reduced QM-DGP explicitly relating only $X_i \in \mathcal{O}$, X_c and $\mathcal{C}_\mathcal{O}$. c.) A reduced QM-DGP with multiple fragments.

and there are no assumptions constraining the distributions.[2] Thus, we assume that the DGP samples states of X_i from a discrete distribution $P(X_i|\pi(X_i))$. This distribution describes the chance that a certain combination of states of X_i and its parents $\pi(X_i)$ will materialize. The number of possible combinations of states n_{X_i} can be expressed as:

$$n_{X_i} = |X_i| \prod_{Y_j \in \pi(X_i)} |Y_j|, \tag{1}$$

where $|X_i|$ and $|Y_j|$ denote the number of states of variables X_i and Y_j, respectively. Note that this equation expresses an upper bound[3] on the complexity of the DGP governing the relations between X_i and its parents $\pi(X_i)$. Thus we can express the complexity of a DGP in terms of the number of possible state combinations, which facilitates further analysis.

As discussed in [1], this way of looking at the complexity can be extended to complex topologies of a QM-DGP. This approach is based on the identification of the relevant DGP context variables and the derivation of a core model $\mathcal{M}_{core}$ defined over the set of variables $\mathcal{V}_{\mathcal{M}_{core}}$.

Let n_{core} denote the number of possible combinations of values of X_c and $X_i \in \mathcal{O}$ over which the distribution $P(\mathcal{V}_{\mathcal{M}_{core}})$ is defined. Then the number of different combinations of states $n_{\mathcal{V}_k}$ of X_c, $\mathcal{O}$ and the relevant DGP context $\mathcal{C}_\mathcal{O}$ in the reduced model can be expressed as follows:

$$n_{\mathcal{V}_\mathcal{M}} = n_{core} \prod_{Y_j \in \mathcal{C}_\mathcal{O}} |Y_j|, \tag{2}$$

where $|Y_j|$ denote the cardinality of the discrete DGP context variables. Note that Equation (2) takes into account different contexts which may or may not be observed. Still, these variables must be considered as their states drive the DGP and thus influence the observable variables.

In the case of fully connected variables in $\mathcal{V}_{\mathcal{M}_{core}}$, n_{core} is simply the product of the numbers of states of all variables X_c and $\mathcal{O}$. However, the core model is likely to represent an underlying data structure that is not fully connected, especially if the QM-DGP has many variables. This can be exploited by some training techniques, such as the EM algorithm. In those cases, learning can be carried out on conditionally independent fragments $\mathcal{F}_i$, leading to good models using less training data compared to approaches that cannot exploit the structure.

A *fragment* $\mathcal{F}_i$ consists of a set of core variables $\mathcal{V}_{F_i} \subset \mathcal{V}_{\mathcal{M}_{core}}$, such that there exists one variable $X_i \in \mathcal{V}_{F_i}$ that is directly influenced by all other variables in $\mathcal{V}_{F_i}$. Thus $\mathcal{V}_{F_i} = X_i \cup \pi^{core}(X_i)$, where $\pi^{core}(X_i)$ denotes all X_i's parents that are included in $\mathcal{V}_{\mathcal{M}_{core}}$. Such a fragment is an irreducible modelling unit, as the distribution of X_i depends on all combinations of states of $\pi^{core}(X_i)$. For example, variable X_1 in Fig. 1.b has $\pi^{core}(X_1) = \{X_c, X_2\}$ while variable X_3 in Fig. 1.c has $\pi^{core}(X_3) = \{X_c, X_1\}$.

Moreover, a fragment $\mathcal{F}_i$ has a set $\mathcal{A}_i$ of *local fragment antecedents that* influences the probability distributions over the variables in $\mathcal{V}_{F_i}$. This set is defined as $\mathcal{A}_i = \pi(\mathcal{V}_{F_i})$, where $\pi(\mathcal{V}_{F_i})$ denotes a set of those direct parents of the core variables in $\mathcal{F}_i$ that are not already members of $\mathcal{F}_i$. The fragment $\mathcal{F}_1$ formed from variable X_1 in Fig. 1.b has $\mathcal{A}_1 = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$ while the fragment $\mathcal{F}_3$ formed from variable X_3 in Fig. 1.c has $\mathcal{A}_3 = \{X_2, Y_2, Y_3, Y_4, Y_5\}$. The local antecedent set $\mathcal{A}_i$ can contain the relevant context variables from the set $\mathcal{C}_\mathcal{O}$, as is the case in in Fig. 1.b, but it can contain also core variables. $\mathcal{A}_i$ can be determined with a simple procedure shown in Algorithm 1.

Algorithm 1 Find Fragment's Local Antecedent Set

Data: $QMDGP$
$\mathcal{A}_i \leftarrow \emptyset$
for *each* $X_p \in \pi^{core}(X_i)$ **do**
  $\mathcal{A}_i \leftarrow \mathcal{A}_i \cup \pi(X_p)$
end
$\mathcal{A}_i \leftarrow \mathcal{A}_i \cup (\pi(X_i) \setminus \pi^{core}(X_i))$
return $\mathcal{A}_i$

With the help of these concepts we can identify a smaller bound on the complexity of the model than Equation (2). The number of parameters required to define a probability distribution for fragment $\mathcal{F}_i$ is:

$$n_{\mathcal{F}_i} = |X_i| \prod_{Z_k \in \pi^{core}(X_i)} |Z_k| \prod_{Y_j \in \mathcal{A}_i} |Y_j|, \tag{3}$$

[2] We consider relaxing this assumption later.

[3] The number of parameters for this gate equals $(|X_i|-1)\prod_{Y_j \in \pi(X_i)} |Y_j|$

where $|X_i| \prod_{Z_k \in \pi^{core}(X_i)} |Z_k|$ expresses the combinations of states of the variables in fragment F_i.

Figure 2 shows different QM-DGP examples that will be used to illustrate the application of Equation (3). In this figure all context variables are labeled with C1, C2, ..., Cn, while the core variables forming the core and the fragments therein are labeled with V1, V2, ..., Vn. The model in Figure 2.a has the largest fragment $\mathcal{F}_{max} = \{V1, V2, V3\}$ with antecedents $\mathcal{A}_{max} = \{C1, C2, C3\}$. If all context and core variables have 3 states, then Equation (3) yields $n_{\mathcal{F}_{max}} = 729$. Figure 2.b, shows a model with the largest fragment $\mathcal{F}_{max} = \{V4, V5, V6\}$ and local antecedent set $\mathcal{A}_{max} = \{C2, C3, C4\}$. By assuming that all context variables have 3 states, while the core variables V4, V5, and V6 have 5 states, Equation (3) yields $n_{\mathcal{F}_{max}} = 3375$. Finally, Figure 2.c depicts a QM-DGP that features the the largest fragment $\mathcal{F}_{max} = \{V2, V3, V4, V5\}$ with the antecedent set $\mathcal{A}_{max} = \{C2, C3, C4, C5, V1\}$. By assuming that all antecedent and core variables have 3 states we obtain $n_{\mathcal{F}_{max}} = 19683$.

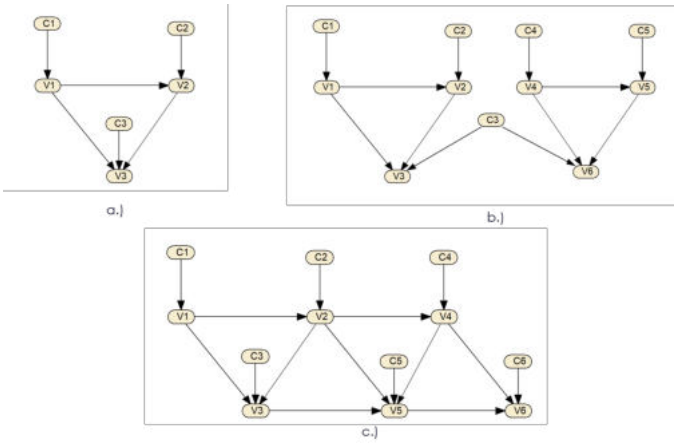

Fig. 2. Examples of QM-DGPs

III. Estimating the Required Size of Training Data

The QM-DGP based analysis in the preceding section shows that the number of possible combinations of states of variables $\mathcal{V}$ representing phenomena in a DGP corresponds to the DGP complexity. However, this analysis ignores the fact that the complexity is influenced also by the nature of the distributions over the states of variables $X_i \in \mathcal{V}$ of a DGP. To be able to estimate the sufficient size of the training set for a given DGP, the results of a QM-DGP based analysis must be combined with a sound approach to determining the size of data sets for the estimation of the distributions over any variable $X_i \in \mathcal{V}$. Section III.A discusses a theoretically sound approach to determine the data quantities required for the estimation of multinomial distributions for a single variable $X_i \in \mathcal{V}$, without any simplifying assumption on the distribution. Section III.B provides a formal procedure to determine the training data quantity required for learning a good quality model $\mathcal{M}$ reflecting the underlying DGP.

A. Estimating Data Distributions

This section discusses the quantities of data that are required to obtain sufficiently accurate estimates of the multinomial probability distributions over the k states of a discrete random variable X, given certain conditions, i.e. the states of X's parents are instantiated to specific values $\pi(X) = \epsilon$. In [1] (section II.D) a heuristic was proposed to determine the quantity of data. In the current section, we propose to use a statistical approach with the desired accuracy as a parameter. Different approaches exist in the literature to estimate the sample size needed to learn a multinomial distribution[4]. For instance, Tortora [5] proposes a method to compute the sample size necessary to achieve a desired confidence level α and a desired confidence interval width W_i when an approximation of the proportion p_i of the multinomial distribution is available. The approach relies on the approximation of simultaneous confidence intervals proposed by Goodman [6].

$$\hat{p}_i^{\pm} = \frac{1}{2(N+B)}(B + 2n_i \pm \sqrt{B(B + 4p_i(N - n_i))}\ ), \quad (4)$$

where B is the upper α/k percentile of the chi-square distribution:

$$B = \chi^2_{1,1-\alpha/k}$$

Sison [7] proposes a method leading to even narrower confidence interval. The required proportion $p_i = n_i/N$ can be provided by a small scale study or by domain experts when feasible. Otherwise, a conservative choice can be made, for instance choosing a reasonable fraction of the value that p_i would have for a uniform distribution over the k states for instance $0.05/k$. Similarly, the desired confidence interval width can be chosen as a fraction of corresponding proportion, for instance $0.05 * p_i$. Finally the confidence level can be chosen within generally accepted values (e.g. $\alpha = 0.05$). Table I presents possible values of n_i and N computed from Equation (4) for different numbers of states k, expected fractions of probability, and choice of interval width.

states	prob. ratio (lowest p_i)	ratio CI width to lowest p_i	n_i (total) $\alpha = 0.05$	n_i (κ_{X_i}) $\alpha = 0.01$
3	1 (0.333)	1 (0.333)	11 (33)	14 (42)
3	0.5 (0.166)	1 (0.166)	15 (90)	19 (114)
5	1 (0.2)	1 (0.01)	17 (85)	21 (105)
5	0.5 (0.1)	1 (0.1)	20 (200)	25 (250)
10	1 (0.1)	1 (0.1)	25 (250)	29 (290)
10	0.5 (0.05)	1 (0.05)	25 (250)	32 (640)

TABLE I
NUMBER OF DATA POINTS PER STATE OF A DISCRETE VARIABLE AND κ_{X_i} COMPUTED FROM EQUATION (4).

B. Training Complexity

In this paper the required quantities of training data $|\mathcal{D}_{train}|$ are estimated by combining (i) the results from the QM-DGP

[4]We consider here the general case. When the variable X is a discretized (in k bins) representation of a normally distributed continuous variable, simpler approaches exist (for instance adapting methods designed to compute an optimal bin width such as Sturge's Rule) which also appropriately result in smaller required sample size.

analysis and (ii) the approach to determining the quantities of data needed for the estimation of multinomial distributions from section III-A.

Equation (2) and the methods described in Section III-A enable simple estimation of $|\mathcal{D}_{train}|$ that support learning of a good quality model $\mathcal{M}$:

$$|\mathcal{D}_{train}| = \kappa_{max} \cdot n_{\mathcal{V}_\mathcal{M}}, \tag{5}$$

where κ_{max} is given by the appropriate entry in Table I for the variable $X_i \in \mathcal{V}_k$ with the largest number of states and $n_{\mathcal{V}_\mathcal{M}}$ is obtained with Equation (2).

Conditionally independent fragments introduced in Section II-B support improved estimation of the required data quantities, given that the machine learning method can exploit the structure of the QMDGP. Algorithm 2 implements a procedure that estimates the complexity of each fragment in a QM-DGP and returns $|\mathcal{D}_{train}|$ corresponding to the most complex fragment.

Algorithm 2 QM-DGP Complexity Scan

$|\mathcal{D}_{train}| \leftarrow 0$;
for *each variable* $X_i \in \mathcal{V}_\mathcal{M}$ **do**
  Find $\mathcal{F}_i$ and $\mathcal{A}_i$ using algorithm (1);
  Compute $n_{\mathcal{F}_i}$ using Equation (3);
  Determine κ_{X_i} for X_i using Table I;
  if $\kappa_{X_i} \cdot n_{\mathcal{F}_i} > |\mathcal{D}_{train}|$ **then**
    $|\mathcal{D}_{train}| \leftarrow \kappa_{X_i} \cdot n_{\mathcal{F}_i}$
  end
end
return $|\mathcal{D}_{train}|$

Note that this approach considers only variables in each fragment $\mathcal{F}_i$ and their local antecedent set $\mathcal{A}_i$. This is appropriate because variables influencing $\mathcal{A}_i$ have only indirect influence on the distributions over the variables in $\mathcal{F}_i$.

C. Refined Analysis Through Auxiliary Models

Equation (5) and Algorithm 2 provide very crude estimates of the required $|\mathcal{D}_{train}|$ as they do not consider the strength of correlations in $P(\mathcal{V})$ from which the DGP samples. It fails to account for the fact that some combinations of states might never materialize while other could be so rare that the chance of capturing a sufficient number of cases in the training data is very low. In this paper we revise the concept of Auxiliary Models introduced in [1]. If some amount of training data is already available at the time of the analysis an auxiliary PGM $\mathcal{M}_{aux}$ can be trained. Such an M_{aux} has the topology corresponding to the reduced QM-DGP, as for example the graph in figure 1.b. It is trained on the currently available data using simple Maximum Likelihood estimation based on state counting. The learned $\mathcal{M}_{aux}$ encodes the frequencies of the state observations of all variables in the model.

By inspecting the estimated distributions in $\mathcal{M}_{aux}$ we can find for each variable $Y_j \in \mathcal{A}_i$ the state with the smallest marginal probability $P_{min}(Y_j)$ and determine the probability of the least likely combination of states in $\mathcal{A}_i$:

$$P_{min}(\mathcal{A}_i) = \prod_{Y_j \in \mathcal{A}_i} P_{min}(Y_j). \tag{6}$$

With $P_{min}(\mathcal{A}_i)$ and $n_{\mathcal{F}_i}$, obtained with Equation (3), we can compute factor $\gamma(\mathcal{F}_i)$, a ratio between the number of required training cases to sufficiently cover all combinations of the states of the core model of $\mathcal{F}_i$ during learning and the expected number of occurrences of the least likely combination of $\mathcal{A}_i$'s states in $\mathcal{D}_{train}$:

$$\gamma(\mathcal{F}_i) = n_{\mathcal{F}_i} \cdot \kappa_{X_i} / (|\mathcal{D}_{train}| \cdot P_{min}(\mathcal{A}_i)). \tag{7}$$

$\gamma(\mathcal{F}_i)$ is a factor by which the initially estimated $|\mathcal{D}_{train}|$ should be increased to obtain a sufficient expected number of training cases to reliably estimate parameters describing the relations between variables $\mathcal{V}_{\mathcal{F}_i}$ of fragment $\mathcal{F}_i$.

Algorithm 3 uses equations (6) and (7) to find an improved estimate of the required training set size $|\mathcal{D}'_{train}|$ that considers initial estimates of the relations between the variables in M_{aux}.

Algorithm 3 Refined QM-DGP Complexity Scan

Compute $|\mathcal{D}_{train}|$ using algorithm 2;
Train auxiliary model $\mathcal{M}_{aux}$;
$\gamma_{max} \leftarrow 1$;
for *each variable* $X_i \in \mathcal{V}_\mathcal{M}$ **do**
  Find $\mathcal{F}_i$ and $\mathcal{A}_i$ using algorithm 1;
  For each $Y_j \in \mathcal{A}_i$ determine $P_{min}(Y_j)$;
  Compute $P_{min}(\mathcal{A}_i)$ using Equation (6) and $P_{min}(Y_j)$;
  Compute $n_{\mathcal{F}_i}$ using Equation (3);
  Determine κ_{X_i} for X_i using Table I;
  Compute $\gamma(\mathcal{F}_i)$ using Equation (7);
  if $\gamma(\mathcal{F}_i) > \gamma_{max}$ **then**
    $\gamma_{max} \leftarrow \gamma(\mathcal{F}_i)$;
  end
end
$|\mathcal{D}'_{train}| \leftarrow \gamma_{max} \cdot |\mathcal{D}_{train}|$;
return $|\mathcal{D}'_{train}|$;

IV. Experimental Results

A. Evaluation Method

The empirical evaluation relied on two types of metrics, the classification accuracy p_{err} and the log likelihood of the model $\mathcal{L}_{\mathcal{D}_{test},\mathcal{M}}$=$log(p(\mathcal{D}_{test}|\mathcal{M}))$ [2].

Independently of the metric, in each experiment a ground truth model ${\mathcal{M}^i}_G$ was used for sampling of a large test set ${\mathcal{D}^i}_{test}$ containing 10^6 cases. The same model was then used to sample different sizes of training sets. For each training set size $\mathcal{S}_l$, hundred new training sets ${\mathcal{D}^i}_{train}$ were sampled. For each training set a model $\mathcal{M}^i$ was trained and tested with $\mathcal{D}^i_{test}$. In the case of classification accuracy p_{err}, the resulting mean error rate $\mu_{p_{err}}$ and the corresponding 90% quantile $q_{err,90\%}$ were recorded for each size and plotted. Moreover, optimal error rate $p_{err}^{optimal}$ is obtained by testing the ground truth model ${\mathcal{M}^i}_G$ with ${\mathcal{D}^i}_{test}$. $p_{err}^{optimal}$ is approximately the Bayes Error [8]. For each ${\mathcal{M}^i}_G$, the horizontal line corresponding

to $p^k_{err} = (1 + k/100) \cdot p^{optimal}_{err}$ (see Figure 3), where factor $k >= 0$ determines the percentage by which the test error deviates from $p^{optimal}_{err}$ provides a estimate of closeness of the trained model with the Bayes error and is used as a convergence threshold in terms of classification error p_{err}. The projection of the intersection between the horizontal line p^k_{err} and the $q_{err,90\%}$ curve on the horizontal axis representing $|\mathcal{D}^i_{train}|$ provides the size of the training data set for which the classification error p_{err} with the trained model $\mathcal{M}^i$ will not exceed p^k_{err} error rate in more than 90% of the cases. Hence $|\mathcal{D}^{k\%}_{train}|$ denotes the experimentally determined size of the data set for which this is true in the case of a $k\%$ deviation from Bayes Error $p^{optimal}_{err}$. For the data likelihood metric $\mathcal{L}_{\mathcal{D}^i{}_{test},\mathcal{M}^i{}_k}$, the same procedure is used, except that $\mu_{\mathcal{L}_{\mathcal{D}^i{}_{test},\mathcal{M}^i}}$ and the 10% quantile $q_{\mathcal{L},10\%}$ are used (see Figure 4). In the experiments we measured $|\mathcal{D}^{1\%}_{train}|$, $|\mathcal{D}^{0.5\%}_{train}|$ and $|\mathcal{D}^{0.3\%}_{train}|$ for 1%, 0.5% and 0.3% deviations from the optimal results, respectively.

B. Experiments description

Six experiments were carried out using synthetic data generated from different ground truth models $\mathcal{M}^i_G$ with the topologies depicted in Figure 2 .

In Experiment 1 to 3, the ground truth models $\mathcal{M}^1_G$, $\mathcal{M}^2_G$ and $\mathcal{M}^3_G$ had the topology shown in Figure 2.a. The core variables in the identified fragments of $\mathcal{M}^1_G$, $\mathcal{M}^2_G$ and $\mathcal{M}^3_G$ had 3, 5 and 10 states respectively. The conditional probabilities in $\mathcal{M}^i_G$ were obtained through random sampling, which resulted in complex correlations. Variable V_2 was used as the classification variable to compute the error as defined in section IV-A.

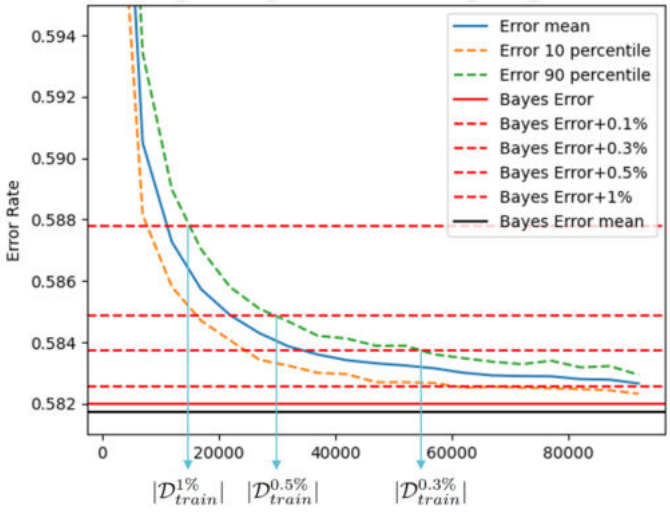

Fig. 3. Convergence of the error rate for $\mathcal{M}_2$.

In Experiment 4, the ground truth model $\mathcal{M}^4_G$ had topology shown in Figure 2.b which features two conditionally independent fragments $\mathcal{F}_1$ and $\mathcal{F}_2$. All context variables had 3 states, the core variables in $\mathcal{F}_1$ had 3 states while the core variables in $\mathcal{F}_2$ had 5 states. Variable C_3 was used as the classification variable to compute the error as defined in section IV-A.

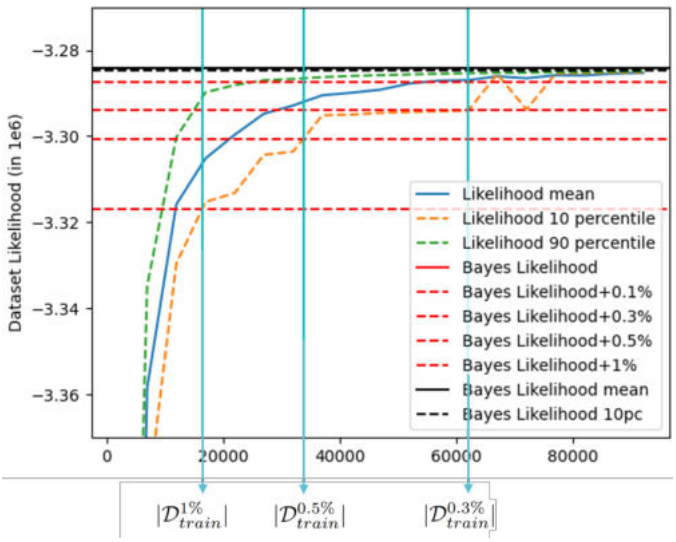

Fig. 4. Convergence of the data likelihood for $\mathcal{M}_2$.

In Experiment 5, the ground truth model $\mathcal{M}^5_G$ had the topology shown in Figure 2.c, which features three tightly coupled fragments $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$. All context variables and core variables had 3 states. The complexity of this model is determined by fragment $\mathcal{F}_2$. Variable V_5 was used as the classification variable to compute the error as defined in section IV-A.

Experiment 6 investigates the effectiveness of the auxiliary model method. It can happen that the less likely contexts during data collection process are the most relevant for the operation. To simulate this situation, the training data was sampled from $\mathcal{M}^6_G$ having the topology shown in Figure 2.a with core and context variables having 3 states. Compared to Experiment 1, the context variable had very sharp distributions, the least likely context had probability of 0.2 %, about 18 times smaller than in the case the distributions over the context variables were uniform. The test data $\mathcal{D}_{test}$, however, was sampled only in this context. Variable T_i was used as the classification variable to compute the error as defined in section IV-A.

Experiment 7 used an existing model $\mathcal{M}^7_G$ that describes relations between diabetes and various causes and symptoms [5] shown in Figure 5. $\mathcal{M}^7_G$ was considered as ground truth DGP used for the sampling of testing and training data. The data is synthetic, however, it was characterized through realistic correlations. Variable $Diabetes$ was used as the classification variable to compute the error as defined in section IV-A. The largest fragment was $\mathcal{F}_{max} = \{Diabetes, Diastolic\ Blood\ Pressure, 2hour\ serum\ insulin\ test, Plasma\ Glucose\ Concentration\}$ and the local an-

[5]©Saverio Manago, available from Norsys (https://www.norsys.com/netlibrary/index.htm?net=Diabetes%20Learned.htm, retrieved on February 28th, 2024) based on a dataset from University of California at Irvine (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)

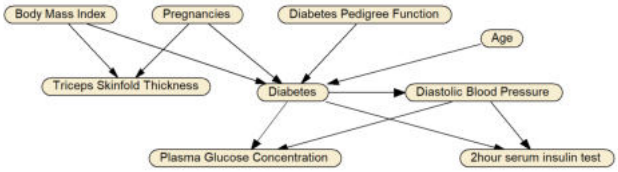

Fig. 5. The QM-DGP of the Diabetes model $\mathcal{M}_6$.

Exp	$\lvert\mathcal{D}^{CS}_{train}\rvert$	$\lvert\mathcal{D}^{RCS}_{train}\rvert$	$\lvert\mathcal{D}^{1\%}_{train}\rvert$	$\lvert\mathcal{D}^{0.5\%}_{train}\rvert$	$\lvert\mathcal{D}^{0.3\%}_{train}\rvert$
1	8000	8000	$\approx$ 5200	$\approx$ 9600	$\approx$ 15700
2	57375	57375	$\approx$ 15000	$\approx$ 29000	$\approx$ 55000
3	675000	675000	$\approx$ 67000	$\approx$ 140000	$\approx$ 250000
4	57375	57375	$\approx$ 18000	$\approx$ 38000	$\approx$ 68000
5	216513	216513	$\approx$ 18000	$\approx$ 45000	$\approx$ 78000
6	8000	195000	$\approx$ 20000	$\approx$ 84000	$\approx$ 110000
7	7128	22600	$\approx$ 10000	$\approx$ 22000	$\approx$ 44000
8	42800	154000	$\approx$ 60000	$\approx$ 130000	-

TABLE II
ESTIMATED TRAINING DATA VOLUMES $|\mathcal{D}_{train}|$ AND EMPIRICALLY DETERMINED CONVERGENCE POINTS FOR 90% QUANTILE ERRORS.

tecedent set $\mathcal{A}_{max}$ ={$Body\ Mass\ Index$, $Pregnancies$, $Diabetes\ Pedigree\ Function$, Age}.

In Experiment 8, the QM-DGP approach was tested on a Neural Network with four hidden layers and LeakyReLU activation functions using the Diabetes ground truth model $\mathcal{M}^7_G$ that was used also in Experiment 7. A large validation set taken outside of the training set was used to design the network and prevent over-fitting during training. The assumption was that a fully-connected NN cannot efficiently encode the sparse topology of the underlying QM-DGP. Consequently, the required training data size was estimated using Equation (5), by assuming that all core variables are fully connected.

In each experiment the required quantities of training data were estimated by using the Algorithm 2 and Algorithm 3, with the following intermediary values:

- **Exp 1:** $\mathcal{M}^1_G$, $n_{\mathcal{V}_1} = 729$, $\kappa_{max} = 11$.
- **Exp 2:** $\mathcal{M}^2_G$, $n_{\mathcal{V}_2} = 3375$, $\kappa_{max} = 17$.
- **Exp 3:** $\mathcal{M}^3_G$, $n_{\mathcal{V}_3} = 27000$, $\kappa_{max} = 25$.
- **Exp 4:** $\mathcal{M}^4_G$, $n_{\mathcal{V}_{F_{max}}} = 3375$, $\kappa_{max} = 17$.
- **Exp 5:** $\mathcal{M}^5_G$, $n_{\mathcal{V}_{F_{max}}} = 19683$, $\kappa_{max} = 11$.
- **Exp 6:** $\mathcal{M}^6_G$, $n_{\mathcal{V}_1} = 729$, $\kappa_{max} = 11$.
- **Exp 7:** $\mathcal{M}^7_G$, $n_{\mathcal{V}_{F_{max}}} = 648$, $\kappa_{max} = 11$.
- **Exp 8:** $\mathcal{M}^7_G$, $n_{\mathcal{V}_7} = 3888$, $\kappa_{max} = 11$.

C. Results and Discussion

Table II shows the experimental results, where $|\mathcal{D}^{CS}_{train}|$ and $|\mathcal{D}^{RCS}_{train}|$ denote the data size estimates obtained with Algorithm 2 and Algorithm 3, respectively. $|\mathcal{D}^{1\%}_{train}|$, $|\mathcal{D}^{0.5\%}_{train}|$ and $|\mathcal{D}^{0.3\%}_{train}|$, on the other hand, represent the empirically determined data quantities for 90%-quantile errors, as defined in section IV-A.

Table III shows the results of the same experiments, but the convergence points were determined with the curve for the 10%-quantile $q_{\mathcal{L},10\%}$ of the log likelihoods $\mathcal{L}_{\mathcal{D}^i{}_{test},\mathcal{M}^i}$.

Exp	$\lvert\mathcal{D}^{CS}_{train}\rvert$	$\lvert\mathcal{D}^{RCS}_{train}\rvert$	$\lvert\mathcal{D}^{1\%}_{train}\rvert$	$\lvert\mathcal{D}^{0.5\%}_{train}\rvert$	$\lvert\mathcal{D}^{0.3\%}_{train}\rvert$
1	8000	8000	$\approx$ 8100	$\approx$ 8600	$\approx$ 8900
2	57375	57375	$\approx$ 16000	$\approx$ 34000	$\approx$ 62000
3	675000	675000	$\approx$ 135000	$\approx$240000	$\approx$ 350000
4	57375	57375	$\approx$ 15000	$\approx$ 24000	$\approx$ 36000
5	216513	216513	$\approx$ 17500	$\approx$ 24000	$\approx$ 110000
6	8000	195000	$\approx$ 9400	$\approx$ 22000	$\approx$ 87000
7	7128	22600	$\approx$ 7000	$\approx$ 9000	$\approx$ 13000

TABLE III
ESTIMATED TRAINING DATA VOLUMES $|\mathcal{D}_{train}|$ AND EMPIRICALLY DETERMINED CONVERGENCE POINTS FOR 10%QUANTILE LOG LIKELIHOOD. EXPERIMENT IS OMITTED SINCE THE LOG LIKELIHOOD METRIC IS NOT AVAILABLE FOR EXPERIMENTS WITH NEURAL NETWORKS.

Experiments 1, 2 and 3 show that the estimated $|\mathcal{D}^{CS}_{train}|$ and $|\mathcal{D}^{RCS}_{train}|$ increase with the model complexity as expected. While these estimates are conservative, they have consistently the right order of magnitude.

Experiments 4 and 5 demonstrate that the most complex fragment $\mathcal{F}_2$ determines the overall training complexity. In Experiment 4, as $\mathcal{F}_2$ is identical to the single fragment in Experiment 2, the overall training complexity is in the same range as the one in Experiment 2. Experiment 5 also shows that the Algorithm 2 is suitable for complex models.

The empirical result in Experiment 6 show that close to optimal performance is achieved for the corrected estimates using the auxiliary model. In Experiment 7 close to optimal performance was achieved for the estimated $|\mathcal{D}_{train}|$ using as ground truth model a model learned from real data.

Experiment 8 showed that the neural network generally required lower data quantities than estimated in order to achieve near-optimal performance (within 1% of the Bayes error). It is assumed that this happens because of the neural network's capability to generalize well. A neural network with an optimal architecture (one which has no more or less parameters than is justified given the learning task) forces the function approximation of the NN to take on a more general form, which results in the neural network being able to make good predictions for unseen data. In the case of larger than optimal networks (networks with more depth and width than is justified given the problem), which is likely the case here given that the NN has four hidden layers, the inductive bias of neural networks may play a role in forcing the network to learn simpler function approximations (simplicity bias) [9]. Additionally, the use of a large validation set in combination with early stopping prevents larger neural networks from overfitting.

Given the classification problem and the NN architecture, the network is not able to get within 0.3% of the Bayes error. We hypothesize that this may be a result of the neural network's inability to exploit the causal structure of the QM-DGP. Further experimentation on the relationship between the QM-DGP and neural network architectures could provide support for this hypothesis.

In all experiments, the Algorithm 2 and Algorithm 3 yielded estimates that supported training of high quality models. In all cases, the discrepancy between the optimal values for the error

rates and the Bayes Error was less than 1% of the Bayes Error and the same was true for the log likelihood estimates. In some cases the estimates supported training of models for which the discrepancy between the Bayes Error and the experimentally determined errors was less than 0.1%.

All experiments confirm that Algorithm 2 and Algorithm 3 correctly capture the tendencies between the QM-DGP complexity and the required training data size.

For the used relative discrepancy between the Bayes Error and the measured error, the Algorithm 2 yields increasingly conservative estimates of the training size as the number of states of the core variables increases.

Moreover, as expected, the experiments confirm that the distributions over the context variables have a significant impact on the required size of the training data. Algorithm 3 seems to correctly factor in non-uniform distributions over the contexts. Contrary to Algorithm 2, however, Algorithm 3 requires certain amounts of real world data. Fortunately, this requires relatively small amounts of data. For the experiments in this paper, we first applied Algorithm 3 and then assumed that 5% percent of $|\mathcal{D}^{CS}_{train}|$ was sampled to obtain the coarse estimates of the distributions over the context variables that are needed in Algorithm 3. Thus, the procedure can be used early in the sampling process, allowing early corrections of the required size estimates.

V. Algorithm Independent Analysis

Learning good classifiers requires (i) training data sets that carry all the relevant information about the correlations between the chosen variables and (ii) adequate model architectures that can capture that information.

The experiments show that, given a DGP and its QM-DGP, we can train a classifier whose error rates are close to the **Bayes Error** [8] if (i) model $\mathcal{M}$ is a PGM, (ii) its DAG is identical to the QM-DGP and (iii) the size of the training data $|\mathcal{D}_{train}|$ is determined with the proposed approach. Namely, the error rates of the trained models are close to the error rates of the classifiers using the ground truth models (the models used as DGPs) which, by definition, are Optimal Bayes Classifiers that cannot be outperformed by any other classifier [8]. Since the Optimal Bayes Classifiers contain the maximum possible relevant information about the relations of the modelled variables $\mathcal{V}_{\mathcal{M}}$, the trained models that reached near Bayes error in our experiments must have been able to extract the relevant information from the dataset. Therefore, the experimental results suggest that by using the proposed approach to determining $|\mathcal{D}_{train}|$ also the training data $\mathcal{D}_{train}$ likely contains most of the relevant information about the relations between the modelled phenomena involved in the underlying data generation process.

Moreover, it is plausible to assume that such a training data set should in principle enable training of a model $\mathcal{M}$ that is close to optimal by using any sound ML paradigm. This assumes, however, that the architecture of model $\mathcal{M}$ adequately captures the complexity of the data generation processes. Such an architecture corresponds to a set of parameters that can capture all the relevant information from $\mathcal{D}_{train}$ during the ML process. While the mapping between the QM-DGP and the model architecture is straightforward if PGMs are used, this is not the case with Neural Networks, SVMs, etc. In cases where the structure of the QM-DGP cannot be easily considered in the ML model architecture, we can assume fully connected core variables of the QM-DGP and estimate the training data set size using Equation (5). This is likely to be a conservative estimate. However, it is a safe bet, as it was demonstrated in the presented experiments (Experiment 8).

The QM-DGP based approach to estimating $|\mathcal{D}_{train}|$ by approximating the complexity of a DGP can be seen as an alternative to the well known approaches based on the PAC Learnability framework and the Vapnik-Chevronenkis (VC) dimension [8]. The PAC based approaches rely on the estimation of the hypothesis space $\mathcal{H}$, essentially the set of all possible models of a certain architecture. It is often difficult to determine $\mathcal{H}$ and the approach does not work in cases where $|\mathcal{H}| = \infty$. As noted in [10], also the VC-dimension might be difficult to obtain for many interesting models. Overall, these approaches focus on the learners complexity, while the presented QM-DGP-based approach focuses on the complexity of the underlying data generation processes.

VI. Conclusions

This paper presents an improved analysis of the complexity of machine learning tasks using the Qualitative Models of Data Generating Processes (QM-DGP) introduced in [1]. The focus is on the methods for determining the complexity of the data generating processes. This approach facilitates the design of model architectures that can absorb relevant information through machine learning and it is a basis for the determination of the quantities of training data required for learning of good quality models.

As stated in [1], the number of different contexts and the complexity of the data generating process determine the modelling complexity and the size of the training sets required for training of adequate models. However, this paper provides a more thorough treatment of the learning complexity and the evaluation of the effectiveness of the approach. Firstly, the process of determining the model complexity considers the dependencies captured by the QM-DGP in a more rigorous way and provides an improved approach to identifying the major parts of the QM-DGP contributing to the complexity. Secondly, an improved approach to determining the required training data volumes is proposed. It combines (i) a sound approach to determining the size of data sets for categorical variables and (ii) a method for determining the combinations of possible situations in which the distributions over variables must be estimated. Thirdly, more rigorous evaluation metrics were introduced, enabling preciser measurement of the effectiveness of the method for determining the required training data quantities. Especially useful is the evaluation based on the data likelihood.

The effectiveness of the proposed methods is experimentally evaluated in controlled settings. Different ground truth models

are used to sample test and training data. The experiments show that the required training data quantities are estimated, such that the trained models support classification close to Bayes Error. While the majority of the experiments are carried out on PGMs whose variables were fully observed during training, the approach is valid also for training of models with latent variables. The experiments were carried also with the help of Neural Networks, confirming that the QM-DGP is not limited to PGMs.

REFERENCES

[1] G. Pavlin, K. B. Laskey, F. Mignet, F. S. Slijkhuis, E. Blasch, V. Dragos, J. P. de Villiers, and L. Jansen, "Qualitative models of data generation processes: Facilitating data-intensive AI solutions," in *26th International Conference on Information Fusion, FUSION 2023, Charleston, SC, USA, June 27-30, 2023*. IEEE, 2023, pp. 1–8. [Online]. Available: https://doi.org/10.23919/FUSION52260.2023.10224126

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] J. Pearl and D. Mackenzie, *The Book of Why*. Basic Books, 2018.

[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[5] R. D. Tortora, "A note on sample size estimation for multinomial populations," *The American Statistician*, vol. 32, no. 3, pp. 100–102, 1978. [Online]. Available: http://www.jstor.org/stable/2683352

[6] L. A. Goodman, "On simultaneous confidence intervals for multinomial proportions," *Technometrics*, vol. 7, no. 2, pp. 247–254, 1965. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1965.10490252

[7] C. P. Sison and J. Glaz, "Simultaneous confidence intervals and sample size determination for multinomial proportions," *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 366–369, 1995. [Online]. Available: http://www.jstor.org/stable/2291162

[8] T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997, vol. 1, no. 9.

[9] G. Valle-Pérez, C. Q. Camargo, and A. A. Louis, "Deep learning generalizes because the parameter-function map is biased towards simple functions," Apr. 2019, arXiv:1805.08522 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1805.08522

[10] K. P. Murphy, "Machine learning - a probabilistic perspective," in *Adaptive computation and machine learning series*, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:17793133