Iterated Conditional Expectation Algorithm on DAGs and Regression Graphs

Máté Baranyi[a], Marianna Bolla[a]

[a]*Department of Stochastics, Institute of Mathematics,*
Budapest University of Technology and Economics

Abstract

Constructions for regression graphs and verification of the statistical model via linear, linearized, and logistic regressions along them have recently been intensively studied by Wermuth, Sadeghi, and Cox. In the present paper an iterative regression method, using local averaging estimators, is introduced for prediction, based on a complete training sample. The method makes it possible to perform nonparametric regressions recursively, irrespective of the type of the context and response variables. As a consequence, predictions for the multiple response variables of a test sample are performed in the possession of their context variables only. Consistency is proved if the joint distribution is Markov compatible with a DAG or with a regression graph. In the latter case, the prediction goes on from chain component to chain component, with vector valued smoothers, mainly of product kernel types in the implementation. Practical considerations and application to randomly generated and real-world data are also presented.

Keywords: graphical models, nonparametric regression, iterated conditional expectation, mean-square consistency, product kernel

1. Introduction

A regression graph is a refined chain graph that gives rise to a collection of so-called traceable regressions along the directions of its edges, see [1, 2]. It contains both directed and undirected edges in the following way. If we keep only the undirected edges, the graph falls apart into connected chain components, and if we keep only the directed ones, we have a DAG (Directed Acyclic Graph). The components are numbered so that the last ones (with highest indices) correspond to the so-called context variables that are given in the context of the experiment. Context variables of the same component are connected with undirected edges based on the concentration graph on them. From the context variables arrows point to variables in the lower indexed components, which are primary, secondary, etc. responses, and they can also be connected with directed edges. Between the joint response variables of the same connected component, there may be dashed lines (sometimes denoted as bidirected edges),

Email addresses: baranyim@math.bme.hu (Máté Baranyi), marib@math.bme.hu (Marianna Bolla)

URL: math.bme.hu/~baranyim (Máté Baranyi), math.bme.hu/~marib (Marianna Bolla)

which indicate dependences on conditional covariance base. Joint response variables are also called to be on equal standing, and they usually correspond to similar aspects of an experiment, without direct causality between them. For every directed edge $j \to i$ in the DAG part, the relation $i < j$ holds (this so-called topological ordering is referred to as 'j is the parent of i'). So the 'youngest' vertex has label 1, and the 'older' a vertex, the larger its label is (we can think of labels as ages). The arrows point from the right (past) to the left (future), see [3] and Figs. 4,7, and 9 as examples. Let $C_1, C_2, \dots, C_\ell$ denote the chain components ($\ell < d$, where d is the number of variables). Here C_ℓ contains the context variables, and the chain components are also indexed so that lower index components contain 'younger' vertices. With the shorthand $C_{>m} := C_{m+1} \cup \cdots \cup C_\ell$, we can make the following conditional independence statements (denoted by $\perp\!\!\!\perp$) that uniquely define the missing edge positions, whatever their type is:

- for $i < j$ such that $X_i \in C_m$ and $X_j \in C_{>m}$ are in different chain components, there is no $j \to i$ arrow if and only if $X_i \perp\!\!\!\perp X_j \mid C_{>m} \setminus \{X_j\}$;
- for $X_i, X_j \in C_\ell$ (context component), there is no solid line between them if and only if $X_i \perp\!\!\!\perp X_j \mid C_\ell \setminus \{X_i, X_j\}$;
- for X_i and X_j in the same C_m (non-context component, $m < \ell$), there is no dashed line between them if and only if $X_i \perp\!\!\!\perp X_j \mid C_{>m}$.

In Theorem 1 of [2], the authors state that two regression graphs are Markov equivalent if and only if they have the same skeleton (the graph resulting by making every edge undirected) and the same set of collision Vs, irrespective of the type of edges, see Fig. 1. In their Theorem 2, they prove that a regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a DAG on the same skeleton if and only if it does not contain any chordless collision path in four nodes. These so-called *forbidden quadruple*s are shown in Fig. 2. When we prove consistency of our algorithm along a DAG, the starting object can be this DAG, with the above labeling of the vertices given in Algorithm 1 of [2] that is applicable if there is no forbidden quadruple in the underlying chain graph. The possible dashed line based arrows, though they are taken into consideration in the iteration, do not add any extra information to the prediction of joint responses along the arrows of the original regression graph. This is supported by the Cochran's theorem, and will be further discussed in Section 4.2. Note that this DAG is not unique, but it is important that the actual topological ordering of its vertices should comply with the ordering of the chain components. Whenever the regression graph does not contain any forbidden quadruple, the DAG version of our algorithm, processed on the Markov equivalent DAG, gives the same result as the version that uses parallel regressions along the chain components. Further, the DAG version of our algorithm is applicable to other directed graphical models with a DAG representation, and without any relation to regression graphs.

Even in the presence of a forbidden quadruple, if our joint distribution is Markov compatible with a regression graph, we can prove consistency along the chain components with vector valued smoothers. By using parallel regressions, we disregard the dashed edges of the response components during the iteration. It is justified by the fact that in this type of chain graphs the conditioning set

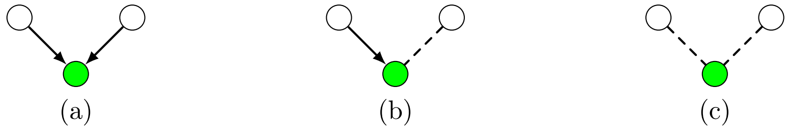

Figure 1: Possible collision Vs of non-context nodes in a regression graph

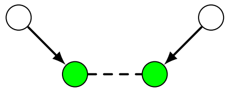

Figure 2: Chordless collision paths on four nodes

for a response variable does not include other responses within the same chain component.

In an ordering of the vertices compatible with the chain graph, instead of linear, linearized, or logistic regression we take conditional expectation in a nonparametric way, like the ACE (Alternating Conditional Expectation) algorithm of Breiman and Fiedman [4]. The ACE is applicable to an additive model that finds a unit variance measurable function of the dependent variable and those of the independent ones, the sum of which approximates best the former in L^2 norm. In the convergence proof of the ACE, the emphasis is on the optimal function of the single regressor, when predicting the current phase (additive function) of all the other variables. On the contrary, we find the measurable function of all the predictors recursively, and the finite iteration ends up with predicting the very target (targets). Though, in the consistency proof we compare the conditional expectation to the smoother, it differs from the ACE setup. Here we need not alternate, we just take directed conditional expectations consecutively, with the nonparametric regression concepts of [5]. Therefore, we call our recursive algorithm ICE (Iterated Conditional Expectation).

In comparison to former results in the literature, our algorithm uses the theory of chain graphs and regression graphs, also the conditional expectation concept of the ACE and EM algorithms, however we have no infinite iteration and maximization step as we do not have a parametric likelihood, but work in a nonparametric setup. Our contribution is that we extend the application area of regression graph models (and probabilistic DAGs), by applying nonparametric smoothings recursively along the directed part of the graph. The algorithm is a regression technique with multiple outputs, and it is applicable in environments with missing data as well, since we only need the values of the context variables for the prediction. The selection of the smoothing functions can be automated according to the types of the variables in the nonparametric regression, as discussed in our implementation. In this way, the ICE algorithm can be the building block of an artificial intelligence, where the first step is the construction of the regression graph. The basis for the construction is a training sample, expert knowledge, and the statistical validation of the edges, see [2]. This far not trivial task is not considered in the present paper, here the focus is on the prediction of multiple responses, given the graph. The soundness of the ICE algorithm is also supported by the proof of its consistency, meaning that with large training samples we get near optimal estimates (in L^2 sense) for the test samples.

The mean-square consistency of the individual smoothers, based on local averaging, is guaranteed under the conditions of the C. J. Stone's theorem, see [5] and Theorem A.1 of AppendixA. For example, k-nearest neighbor (k-NN), partitioning, and kernel regression estimates are mean-square consistent under particular conditions (see again [5]). Actually, we mainly use the product kernel regression approach of [6] which adopts well to different types of variables. It is asymptotically normal, and in turn, mean-square consistent too, see AppendixA. For comparison, we include results obtained with other smoothers too, in our applications.

The organization of the paper is as follows. In Section 2, a detailed description of the ICE algorithm is given, while in Section 3 its consistency is proved both for DAGs and for regression graphs. In Section 4, the implementation of the kernel smoothing is discussed together with kernel and bandwidth selection according to the types and sample characteristics of the variables and other practical considerations. In Section 5, the algorithm is applied to randomly generated and real-world demographic and educational data. Comparison of the product kernel based smoothers with other local averaging methods is also presented. In Section 6, conclusions are drawn and possible further perspectives, e.g., extension to undirected decomposable graphical models, are suggested. In AppendixA and AppendixB, some background material on nonparametric regression, in particular, on kernel regression is provided.

2. Description of the ICE algorithm

2.1. ICE on a DAG

Let the joint distribution of $X_1, \dots, X_d$ be (in this ordering) Markov compatible with a DAG. Say, at least the values of the variables $X_{k+1}, \dots, X_d$ are known; these include the ones without parents in the DAG. Then we first predict X_k, then X_{k-1}, ..., finally X_1, based on an n-element complete training sample $x_1^{(i)}, \dots, x_d^{(i)}$, $i = 1, \dots, n$. We also introduce some notation: $\mathbf{X}_{>j} := \{X_{j+1}, \dots, X_d\}$, $\mathbf{x}_{>j} := \{x_{j+1}, \dots, x_d\}$ and $\mathbf{X}_{\mathrm{pa}(j)}$ denotes the set of the parents of X_j, from where directed edges point to it.

For $j = k, k-1, \dots, 1$ we define

$$P_j(\mathbf{x}_{>j}) := \mathbb{E}(X_j \,|\, \mathbf{X}_{>j} = \mathbf{x}_{>j}) = \mathbb{E}(X_j \,|\, \mathbf{X}_{\mathrm{pa}(j)} = \mathbf{x}_{\mathrm{pa}(j)}) \tag{1}$$

as the exact solution of the regression task; it is a projection and a measurable function of its arguments. The last equality follows by Markovity.

If our data are from multivariate Gaussian distribution, then the above conditional expectations are linear functions of the variables in the condition, and are obtainable by linear regression, where the coefficients are estimated from the training data. Otherwise, we take the conditional expectation in a nonparametric way, by the smoothing procedures discussed in [4, 5]. Therefore, we define

$$S_j^{(n)}(\mathbf{x}_{>j}) := \sum_{i=1}^{n} x_j^{(i)} W_j^{(n,i)}(\mathbf{x}_{>j}) = \sum_{i=1}^{n} x_j^{(i)} W_j^{(n,i)}(\mathbf{x}_{\mathrm{pa}(j)}) \tag{2}$$

as a smoother, based on local averaging and mapping into the convex hull of the sample space for X_j. By Markovity, instead of the set $\mathbf{x}_{>j}$, the parent set $\mathbf{x}_{\mathrm{pa}(j)}$ is considered.

Here the non-negative weights $W_j^{(n,i)}(\mathbf{x}_{>j})$ are calculated from the n-element training sample, and are usually normalized such that

$$\sum_{i=1}^{n} W_j^{(n,i)}(\mathbf{x}_{>j}) = 1.$$

The smoother $S_j^{(n)}$ itself depends on the segment $x_j^{(i)}, x_{j+1}^{(i)}, \dots, x_d^{(i)}$ for $i = 1, \dots, n$ of the training sample, and is also a measurable function of its arguments. The smoothers we implemented are mainly Nadarya–Watson type kernel estimators, where

$$W_j^{(n,i)}(\mathbf{x}_{>j}) = \frac{K(\mathbf{x}_{>j}, \mathbf{x}_{>j}^{(i)})}{\sum_{i'=1}^{n} K(\mathbf{x}_{>j}, \mathbf{x}_{>j}^{(i')})}.$$

For possible kernels K, see the product kernel approach and other local averaging estimators, discussed in Section 4, AppendixA, and AppendixB.

Then the iterated conditional expectations

$$P_j(P_{j+1}(P_{j+2}(\dots(P_k(\mathbf{X}_{>k}), \dots), \mathbf{X}_{>k}), \mathbf{X}_{>k}), \mathbf{X}_{>k})$$

are imitated with the iterated smoothings

$$S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\mathbf{X}_{>k}), \dots), \mathbf{X}_{>k}), \mathbf{X}_{>k}), \mathbf{X}_{>k}).$$

In Section 3, it is proved that the mean-square difference between the iterated smoothings and the iterated conditional expectations tends to 0 as $n \to \infty$, for $j = k-1, \dots, 1$. Therefore, the successive estimates are close to the optimal (in L^2 sense) that means the consistency of the algorithm. For detailed explanation, see Section 3.1.

2.2. ICE on a regression graph

Here the joint density of $X_1, \dots X_d$ can be factorized along the chain components $C_1, \dots, C_\ell$ like

$$f_{C_\ell} \prod_{k=1}^{\ell-1} f_{C_{\ell-k} \,|\, C_{>\ell-k}}, \tag{3}$$

so the conditional expectations are taken from chain component to chain component, starting from the last one, C_ℓ. We postulate that at least the context variables, included in C_ℓ, are known. The above formula is in terms of the joint densities and conditional densities of the chain components in their given reversed ordering, however it easily adopts to discrete variables (with probability mass functions instead of the densities).

In the general step, for $k = \ell-1, \ell-2, \dots, 2, 1$, let

$$\mathbf{P}_k(\mathbf{x}_{C_{>k}}) := \mathbb{E}(\mathbf{X}_{C_k} | \mathbf{X}_{C_{>k}} = \mathbf{x}_{C_{>k}}) \tag{4}$$

denote the exact solution of the regression task. The result is a vector of components

$$P_{k,j}(\mathbf{x}_{C_{>k}}) := \mathbb{E}(X_j | \mathbf{X}_{C_{>k}} = \mathbf{x}_{C_{>k}}) = \mathbb{E}(X_j | \mathbf{X}_{\mathrm{pa}(j)} = \mathbf{x}_{\mathrm{pa}(j)}), \quad X_j \in C_k. \tag{5}$$

This is the instance of parallel regressions, where we utilized that all the parents of $X_j \in C_k$ are in $C_{>k}$. We also used that the conditional expectation of a random vector on another one is the vector of components that are the conditional expectations of its components on the same conditioning set. By the Markov properties of the regression graphs, even if we condition on the same set of random variables, the individual components may have different parent sets to which the conditioning sets are limited.

The smoothers are also defined component-wise:

$$S_{k,j}^{(n)}(\mathbf{x}_{C_{>k}}) := \sum_{i=1}^{n} x_j^{(i)} W_{k,j}^{(n,i)}(\mathbf{x}_{C_{>k}}) = \sum_{i=1}^{n} x_j^{(i)} W_{k,j}^{(n,i)}(\mathbf{x}_{\mathrm{pa}(j)}), \quad X_j \in C_k \tag{6}$$

is the approximate solution using a smoother, based on local averaging. The smoother $\mathbf{S}_k^{(n)}(\mathbf{x}_{C_{>k}})$ consists of coordinates $S_{k,j}^{(n)}(\mathbf{x}_{C_{>k}})$ for $X_j \in C_k$.

Then, by the forthcoming Section 3.2, our estimates are close to the optimal if the mean-square difference between the iterated component-wise conditional expectations

$$\mathbf{P}_k(\mathbf{P}_{k+1}(\mathbf{P}_{k+2}(\dots(\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})$$

and the analogous smoothers

$$\mathbf{S}_k^{(n)}(\mathbf{S}_{k+1}^{(n)}(\mathbf{S}_{k+2}^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})$$

tends to 0 as $n \to \infty$, for $k = \ell-1, \ell-2, \dots, 2, 1$.

3. Consistency of the ICE algorithm

3.1. Consistency along a DAG

3.1.1. When the target variable is absolutely continuous

Now assume that $X_1, \dots, X_d$ (in the topological DAG ordering) have finite variances and obey an absolutely continuous joint distribution. The preparation for the proof of the mean-square consistency of the ICE algorithm follows. Since

$$\mathbb{E}[S_k^{(n)}(\mathbf{X}_{>k}) - X_k]^2 = \mathbb{E}[S_k^{(n)}(\mathbf{X}_{>k}) - P_k(\mathbf{X}_{>k})]^2 + \mathbb{E}[P_k(\mathbf{X}_{>k}) - X_k]^2, \tag{7}$$

and the second term on the right depends only on the nature of the underlying distribution (the larger the coefficient of determination, the smaller this regression error is), it is the first term on the right that can be directly decreased by the sampling and smoothing procedure. (Note that the coefficient of determination is closely related to the maximal correlation of Rényi [7], and to the multiple correlation in the Gaussian case.) Therefore, the above objective function on the left of (7) is close to the optimal value if the first term on the right is close to 0. In the spirit of [5], we call our nonparametric estimate mean-square consistent if

$$\mathbb{E}\int_{\mathbb{R}^{d-k}} [S_k^{(n)}(\mathbf{x}_{>k}) - P_k(\mathbf{x}_{>k})]^2 \, \mu(d\mathbf{x}_{>k}) \to 0, \quad n \to \infty,$$

where μ denotes the joint distribution of the variables behind it. Here both the n-element sample for $X_k, X_{k+1}, \dots, X_d$, used by the smoother, and the values of the variables $X_{k+1}, \dots, X_d$ of a new-coming observation are random, this is why we separated these two concepts, when taking the expectation. Briefly,

$$\mathbb{E}[S_k^{(n)}(\mathbf{X}_{>k}) - P_k(\mathbf{X}_{>k})]^2 \to 0, \quad n \to \infty,$$

where $\mathbb{E}$ now takes the expectation with respect to both of the above concepts of randomness.

First we prove consistency in the first step of the iteration (see Lemma 3.1), then we prove the consistency of the whole ICE on DAG process by induction (see Theorem 3.1). Going backward, to predict X_{k-1}, we have to prove the mean-square consistency of the estimate

$$S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k}).$$

Here, because of the orthogonality properties of the conditional expectation operator (as a projection), we get that

$$\begin{aligned}
&\mathbb{E}[S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - X_{k-1}]^2 \\
&= \mathbb{E}[S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k}))]^2 \\
&+ \mathbb{E}[P_{k-1}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(P_k(\mathbf{X}_{>k}), \mathbf{X}_{>k}))]^2 \\
&+ \mathbb{E}[P_{k-1}(P_k(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(X_k, \mathbf{X}_{>k}))]^2 \\
&+ \mathbb{E}[P_{k-1}(X_k, \mathbf{X}_{>k})) - X_{k-1}]^2.
\end{aligned}$$

The last two terms are the regression errors that are cumulated in the consecutive steps. Therefore, it suffices to estimate the first two terms, the sum of which also estimates the mean-square difference

$$\begin{aligned}
&\mathbb{E}[S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(P_k(\mathbf{X}_{>k}), \mathbf{X}_{>k}))]^2 \\
&\leq 2\mathbb{E}[S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k}))]^2 \qquad (8) \\
&+ 2\mathbb{E}[P_{k-1}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(P_k(\mathbf{X}_{>k}), \mathbf{X}_{>k}))]^2
\end{aligned}$$

between the iterated conditional expectation and the iterated smoothing. We will use this form in our forthcoming Lemma 3.1 and Theorem 3.1.

Lemma 3.1. *With the notation of Equations (1) and (2), let $P_k(\mathbf{x}_{>k})$ and $S_k^{(n)}(\mathbf{x}_{>k})$ be the exact solution of the regression task and a smoother (based on local averaging and mapping into the convex hull of the sample space of X_k), respectively. Assume that the smoother gives a mean-square consistent estimate for the conditional expectation, i.e.,*

$$\mathbb{E}[S_k^{(n)}(\mathbf{X}_{>k}) - P_k(\mathbf{X}_{>k})]^2 \to 0, \quad n \to \infty.$$

Then the one-step-ahead smoothing is also mean-square consistent, i.e.,

$$\mathbb{E}[S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{X}_{>k}), \mathbf{X}_{>k})) - P_{k-1}(P_k(\mathbf{X}_{>k}), \mathbf{X}_{>k}))]^2 \to 0, \quad n \to \infty.$$

Proof. For the first term on the right of (8),

$$\mathbb{E}\int_{\mathbb{R}^{d-k}}[S_{k-1}^{(n)}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k})-P_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2\,\mu(d\mathbf{x}_{>k})\to 0$$

as the $\mathbb{R}^{d-k}$-dimensional integral is the integral of a non-negative function under some restrictions, where for the unrestricted integral

$$\mathbb{E}\int_{\mathbb{R}^{d-k+1}}[S_{k-1}^{(n)}(\mathbf{x}_{>k-1})-P_{k-1}(\mathbf{x}_{>k-1})]^2\,\mu(d\mathbf{x}_{>k-1})\to 0,\quad n\to\infty$$

holds by the assumed mean-square consistency. It is important that $S_k^{(n)}(\mathbf{x}_{>k})$ is within the convex hull of the sample space for X_k.

For the second term on the right of (8), we use that P_{k-1} is an $L^2(\mu)$ operator, and therefore, it can be approximated (with any small precision) with a continuous function of compact support, that is also uniformly continuous (see Theorem A.1. of [5]). So to any $\varepsilon > 0$ there is a $\tilde{P}_{k-1}$ which is uniformly continuous and

$$\int_{\mathbb{R}^{d-k+1}}[P_{k-1}(\mathbf{x}_{>k-1})-\tilde{P}_{k-1}(\mathbf{x}_{>k-1})]^2\,\mu(d\mathbf{x}_{>k-1})\le\varepsilon. \tag{9}$$

With this,

$$\begin{aligned}
&[P_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))-P_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2\\
&\le 3[\tilde{P}_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))-\tilde{P}_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2\\
&+3[P_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))-\tilde{P}_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))]^2+\\
&+3[P_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})-\tilde{P}_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2\\
&\le 3K^2[(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k})-(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2+\\
&+3[P_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))-\tilde{P}_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))]^2+\\
&+3[P_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})-\tilde{P}_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2\\
&=3K^2[S_k^{(n)}(\mathbf{x}_{>k})-P_k(\mathbf{x}_{>k})]^2+\\
&+3[P_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))-\tilde{P}_{k-1}(S_k^{(n)}(\mathbf{x}_{>k}),\mathbf{x}_{>k}))]^2+\\
&+3[P_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})-\tilde{P}_{k-1}(P_k(\mathbf{x}_{>k}),\mathbf{x}_{>k})]^2,
\end{aligned} \tag{10}$$

where K is the constant coming from the uniform continuity of $\tilde{P}_{k-1}$. As

$$\mathbb{E}[S_k^{(n)}(\mathbf{X}_{>k})-P_k(\mathbf{X}_{>k})]^2\to 0,\quad n\to\infty$$

by the consistency assumption,

$$\begin{aligned}
&\mathbb{E}[\tilde{P}_{k-1}(S_k^{(n)}(\mathbf{X}_{>k}),\mathbf{X}_{>k}))-\tilde{P}_{k-1}(P_k(\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2\\
&\le K^2\mathbb{E}[S_k^{(n)}(\mathbf{X}_{>k})-P_k(\mathbf{X}_{>k})]^2\to 0,\quad n\to\infty.
\end{aligned}$$

Finally, the integral (over $\mathbb{R}^{d-k+1}$) of the last two terms of (10), apart from the constant 3, is bounded from above with ε, due to (9). Taking the expectation of this 2ε error term with respect to the sampling procedure will result in an error $2D\varepsilon$, where D is a positive constant (valid for all n) coming from the fact that the sample space is bounded. □

Now we are able to formulate the general theorem about the mean-square consistency of the ICE on a DAG.

Theorem 3.1. *Let $X_1, \dots, X_d$ obey an absolutely continuous joint distribution with finite variances that is Markov compatible with a DAG (in this ordering). They are indexed such that $\mathbf{X}_{>k}$ are the variables without parents in the DAG and the others are to be predicted based on an n-element sample. With the notation of Equations (1) and (2), for $j = k, k-1, \dots, 1$ let $P_j(\mathbf{x}_{>j})$ and $S_j^{(n)}(\mathbf{x}_{>j})$ be the exact solution of the regression task and a smoother (based on local averaging and mapping into the convex hull of the sample space of X_j), respectively. Assume that the individual smoothers give a mean-square consistent estimate for the conditional expectations, i.e.,*

$$\mathbb{E}[S_j^{(n)}(\mathbf{X}_{>j}) - P_j(\mathbf{X}_{>j})]^2 \to 0, \quad n \to \infty, \quad j = k, k-1, \dots, 1.$$

Then the iterated smoothings are also mean-square consistent, i.e.,

$$\begin{aligned}&\mathbb{E}[S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\ &- P_j(P_{j+1}(P_{j+2}(\dots(P_k(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2 \to 0, \quad n \to \infty\end{aligned}$$

for $j = k, k-1, \dots, 1$.

Note that the parameters of our mainly kernel-based smoothers and of the product kernels will be adopted to the data; other than kernel-based local averaging estimators are also considered in Section 5. Assumptions, under which the individual smoothers give a mean-square consistent estimate, are discussed in Section 4, AppendixA, and AppendixB. Also, by Markovity, instead of the set $\mathbf{X}_{>j}$, the parent set $\mathbf{X}_{\mathrm{pa}(j)}$ is considered during the algorithm.

Proof. Using Lemma 3.1 for the one-step-ahead prediction ($j = k$), with induction we proceed backward. Assume that we have proved the consistency until some $j \in \{k, k-1, \dots, 2\}$, i.e., we have proved that

$$\begin{aligned}&\mathbb{E}[S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\ &- P_j(P_{j+1}(P_{j+2}(\dots(P_k(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2 \to 0, \quad n \to \infty.\end{aligned} \tag{11}$$

Now we will prove the same for $j-1$. Indeed,

$$\begin{aligned}&\mathbb{E}[S_{j-1}^{(n)}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\ &- P_{j-1}(P_j(P_{j+1}(\dots,(P_k(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2 \\ &\le 2[\mathbb{E}[S_{j-1}^{(n)}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\ &- P_{j-1}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2+ \\ &2\mathbb{E}[P_{j-1}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\ &- P_{j-1}(P_j(P_{j+1}(\dots,(P_k(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2.\end{aligned}$$

The first term on the right of the inequality tends to 0 (as $n \to \infty$) by the mean-square consistency assumption. For the second term, we use that P_{j-1} is an $L^2(\mu)$ operator, and therefore, it can be approximated (with any small precision ε) with a continuous function $\tilde{P}_{j-1}$ of compact support, that is also uniformly continuous (with constant K). So, the same argument as in the proof of Lemma 3.1 yields

$$\begin{aligned}
&2\mathbb{E}[\tilde{P}_{j-1}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\
&-\tilde{P}_{j-1}(P_j(P_{j+1}(\dots,(P_k(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2 \\
&\le K^2\mathbb{E}[S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})- \\
&-P_j(P_{j+1}(P_{j+2}(\dots(P_k(\mathbf{X}_{>k}),\dots),\mathbf{X}_{>k}),\mathbf{X}_{>k}),\mathbf{X}_{>k})]^2,
\end{aligned}$$

that tends to 0 ($n \to \infty$) by the induction hypothesis (11). The integral (over $\mathbb{R}^{d-j+1}$) of the two terms, containing the differences between the functions of P_{j-1} and $\tilde{P}_{j-1}$ at the same arguments, is bounded from above with ε, due to an estimate analogous to (9). Taking the expectation of this 2ε error term with respect to the sampling procedure will again result in an error $2D\varepsilon$ that can be arbitrarily small. This finishes the proof. □

It is important that the target variable has finite variance and it is absolutely continuous. Intermediate responses should be continuous, but discrete ordered variables can work quite well if they have a sufficiently large sample space. If the target variable is discrete, categorical, then in that step we use the method to be introduced in the forthcoming Section 3.1.2.

3.1.2. When the target variable is categorical

When X_k is categorical, taking on c different values, then the prediction is based on the Bayes rule via mode maximization, see [5]:

$$P_k(\mathbf{x}_{>k}) := \operatorname*{argmax}_{1\le i\le c} \mathbb{P}(X_k = i \,|\, \mathbf{X}_{>k} = \mathbf{x}_{>k}).$$

With introducing the binary variables $\mathbb{I}_{\{X_k=i\}}$ ($i = 1,\dots,c$), the posterior probabilities are the conditional expectations, i.e., regression functions:

$$P_{k,i}(\mathbf{x}_{>k}) := \mathbb{P}(X_k = i \,|\, \mathbf{X}_{>k} = \mathbf{x}_{>k}) = \mathbb{E}(\mathbb{I}_{\{X_k=i\}} \,|\, \mathbf{X}_{>k} = \mathbf{x}_{>k}).$$

Given the training data for $\mathbb{I}_{\{X_k=i\}}, \mathbf{X}_{>k}$, we construct the (smoothing) estimate $S_{k,i}^{(n)}(\mathbf{x}_{>k})$ for each $i = 1,\dots,c$, and the plug-in estimate

$$S_k^{(n)}(\mathbf{x}_{>k}) := \operatorname*{argmax}_{1\le i\le c} S_{k,i}^{(n)}(\mathbf{x}_{>k}).$$

Now we show that the error of the plug-in estimate is close to the optimal error of $P_k(\mathbf{x}_{>k})$:

$$\begin{aligned}
0 &\le \mathbb{P}(S_k^{(n)}(\mathbf{X}_{>k}) \ne X_k) - \mathbb{P}(P_k(\mathbf{X}_{>k}) \ne X_k) \\
&= \mathbb{P}(P_k(\mathbf{X}_{>k}) = X_k) - \mathbb{P}(S_k^{(n)}(\mathbf{X}_{>k}) = X_k) \\
&= \sum_{i=1}^{c} [\mathbb{P}(P_k(\mathbf{X}_{>k}) = X_k \,|\, X_k = i) - \mathbb{P}(S_k^{(n)}(\mathbf{X}_{>k}) = X_k \,|\, X_k = i)]\mathbb{P}(X_k = i) \\
&\le \sum_{i=1}^{c} \int |P_{k,i}(\mathbf{x}_{>k}) - S_{k,i}^{(n)}(\mathbf{x}_{>k})| \, \mu(d\mathbf{x}_{>k}) \\
&\le \left[\sum_{i=1}^{c} \int |S_{k,i}^{(n)}(\mathbf{x}_{>k}) - P_{k,i}(\mathbf{x}_{>k})|^2 \, \mu(d\mathbf{x}_{>k}) \right]^{1/2} .
\end{aligned}$$

This is true for any data in the smoothing within the n-element sample space for $X_k, X_{k+1}, \dots, X_d$. Then taking the expectation with respect to the sample,

$$\begin{aligned}
&\mathbb{E}[\mathbb{P}(S_k^{(n)}(\mathbf{X}_{>k}) \ne X_k) - \mathbb{P}(P_k(\mathbf{X}_{>k}) \ne X_k)] \\
&\le D \left[\sum_{i=1}^{c} \int [S_{k,i}^{(n)}(\mathbf{x}_{>k}) - P_{k,i}(\mathbf{x}_{>k})]^2 \, \mu(d\mathbf{x}_{>k}) \right]^{1/2} \to 0,
\end{aligned}$$

as $n \to \infty$, by the mean-square consistency assumption, where D is a positive constant (independent of n) resulting from the boundedness of the sample space. The same holds for $j \in \{k-1, \dots, 2\}$. We also refer to the Bayes risk consistency in discrimination, see [8].

3.2. Consistency along a regression graph

Now the mean-square consistency is formulated in terms of the mean-square consistency of the vectors that in turn reduces to the the mean-square consistency of their coordinates. Since

$$\begin{aligned}
&\mathbb{E}\|\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}) - \mathbf{X}_{C_{\ell-1}}\|^2 \\
&= \mathbb{E}\|\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}) - \mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell})\|^2 + \mathbb{E}\|\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}) - \mathbf{X}_{C_{\ell-1}}\|^2 \\
&= \sum_{j:\, X_j \in C_{\ell-1}} \mathbb{E}|S_{\ell-1,j}^{(n)}(\mathbf{X}_{C_\ell}) - P_{\ell-1,j}(\mathbf{X}_{C_\ell})|^2 + \mathbb{E}\|\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}) - \mathbf{X}_{C_{\ell-1}}\|^2,
\end{aligned}$$

where the last term depends only on the nature of the underlying distribution, it is the first term that can be directly decreased by the sampling and smoothing procedure. It will be proved that the finitely many terms after the summation tend to 0 as $n \to \infty$, whenever the mean-square consistency of the individual smoothers is guaranteed. In this way, the following consistency theorem is stated under similar conditions as Theorem 3.1.

Theorem 3.2. *Let $X_1, \dots, X_d$ have an absolutely continuous joint distribution with finite variances that is Markov compatible with a regression graph. The variables are organized into chain components $C_1, \dots, C_\ell$ so that their joint distribution obeys the factorization (3); further, they are indexed such that the* *last chain component C_ℓ contains the context variables and the others are to be predicted backward, from component to component, based on an n-element sample.*

With the vector extensions (5) and (6) of the conditional expectations and the smoothers, and assuming that the individual smoothers component-wise give a mean-square consistent estimate for the conditional expectations, the iterated smoothings are also mean-square consistent, i.e.,

$$\begin{aligned}&\mathbb{E}\|\mathbf{S}_k^{(n)}(\mathbf{S}_{k+1}^{(n)}(\mathbf{S}_{k+2}^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})-\\ &-\mathbf{P}_k(\mathbf{P}_{k+1}(\mathbf{P}_{k+2}(\dots(\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\|^2\to 0,\quad n\to\infty,\end{aligned}$$

for $k=\ell-1,\ell-2,\dots,2,1$.

Proof. In the first step,

$$\mathbb{E}\|\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell})-\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell})\|^2=\sum_{j:\,X_j\in C_{\ell-1}}\mathbb{E}\left[S_{\ell-1,j}^{(n)}(\mathbf{X}_{C_\ell})-P_{\ell-1,j}(\mathbf{X}_{C_\ell})\right]^2,$$

where the finitely many terms after the summation tend to 0 (as $n\to\infty$), whenever the mean-square consistency of the individual smoothers is guaranteed.

Analogously to the considerations of Lemma 3.1 and Theorem 3.1, by induction, we proceed backward. Assume that we have proved the consistency until some k, and now we prove it for $k-1$:

$$\begin{aligned}&\mathbb{E}\|\mathbf{S}_{k-1}^{(n)}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})-\\ &-\mathbf{P}_{k-1}(\mathbf{P}_k(\dots(\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\|^2\\ &\le 2\mathbb{E}\|\mathbf{S}_{k-1}^{(n)}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})-\\ &-\mathbf{P}_{k-1}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\|^2\\ &+2\mathbb{E}\|\mathbf{P}_{k-1}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\\ &-\mathbf{P}_{k-1}(\mathbf{P}_k(\dots(\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\|^2.\end{aligned}$$

The first term (apart from the multiplier 2) on the right of the inequality is

$$\begin{aligned}&\mathbb{E}\|\mathbf{S}_{k-1}^{(n)}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})-\\ &-\mathbf{P}_{k-1}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\|^2\\ &=\sum_{j:\,X_j\in C_{k-1}}\mathbb{E}[S_{k-1,j}^{(n)}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})-\\ &-P_{k-1,j}(\mathbf{S}_k^{(n)}(\dots(\mathbf{S}_{\ell-1}^{(n)}(\mathbf{X}_{C_\ell}),\dots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})]^2,\end{aligned}$$

where the individual terms after the summation tend to 0 (as $n\to\infty$) by the mean-square consistency of the individual smoothers.

For the second term we use that for all j, such that $X_j\in C_{k-1}$, $P_{k-1,j}$ is an $L^2(\mu)$ operator, and therefore, it can be approximated (with any small precision ε_j) with a continuous function $\tilde{P}_{k-1,j}$ of compact support, that is also uniformly continuous (with constant K_j). With it,

$$\begin{aligned}
&\mathbb{E}[\tilde{P}^{(n)}_{k-1,j}(\mathbf{S}^{(n)}_k(\ldots(\mathbf{S}^{(n)}_{\ell-1}(\mathbf{X}_{C_\ell}),\ldots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\\
&-\tilde{P}_{k-1,j}(\mathbf{P}_k(\ldots(\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}),\ldots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})]^2\\
&\le K_j^2\mathbb{E}\|\mathbf{S}^{(n)}_k(\ldots(\mathbf{S}^{(n)}_{\ell-1}(\mathbf{X}_{C_\ell}),\ldots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\\
&-\mathbf{P}_k(\ldots(\mathbf{P}_{\ell-1}(\mathbf{X}_{C_\ell}),\ldots),\mathbf{X}_{C_\ell}),\mathbf{X}_{C_\ell})\|^2
\end{aligned}$$

that tends to 0 ($n \to \infty$) by the induction hypothesis. The integral over $\mathbb{R}^{|C_k|+\cdots+|C_\ell|}$ of the two terms, containing the differences between the functions of $P_{k-1,j}$ and $\tilde{P}_{k-1,j}$ at the same arguments, is bounded from above with ε_j, due to an estimate analogous to (9). Taking the expectation of this ε_j error term with respect to the sampling procedure will again result in an error $D\varepsilon_j$ that can be arbitrarily small. Here D is a positive constant (independent of n), coming from the fact that the sample space is bounded. Applying this for all j, such that $X_j \in C_{k-1}$, the sum of these finitely many terms will also tend to 0. This finishes the proof. □

4. The numerical algorithm

4.1. Implementation of the smoothings

Note that in Theorem 3.1 and Theorem 3.2 we utilize that the smoothing is mean-square consistent in every step. Again, there are different smoothers, for example, k-NN, partitioning, and kernel regression estimation. These are all local averaging techniques of Equation (A.1), each with different conditions regarding their mean-square consistency (see [5]), usually derived from Theorem A.1. In the realization of our iterated method, we mainly use kernel-based smoothing, more precisely the Nadaraya–Watson (NW) kernel regression estimation [9, 10]. Some theoretical background on kernel regression is provided in AppendixA.

We have implemented our method based on the Python package of **Statsmodels** [11]. This package uses a product kernel setup in the multivariate case, see Equation (A.3), thus works with a vector $\mathbf{h}$ of bandwidths for the kernel regression estimation. The source for this is mainly the works of Racine and Li [6, 12], who have refined the product kernel approach with the usage of mixed (continuous and discrete) regressors and a proper cross-validated choice of bandwidths. For the theoretical background behind the selection of the kernels and bandwidths, see [6] and AppendixA.

To compare the product kernel approach with other methods, we include results obtained by using a full bandwidth matrix $\mathbf{H}$ (see AppendixA), by the local linear version of the product kernel approach, and by the k-NN regression (see AppendixB).

Kernel selection of Racine and Li. In Equation (A.3), each univariate kernel and the corresponding bandwidth parameter is chosen according to the regressor variables. This implementation needs the type of the regressors as an input. They can be of three types: (a) continuous, (b) ordered (discrete), (c) unordered (discrete, categorical). Based on this, the three default univariate kernels are as follows:

(a) Gaussian kernel in the continuous case:

$$K_h^{(G)}\left(x, x^{(i)}\right) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{\left(x-x^{(i)}\right)^2}{2h^2}};$$

(b) Wang–Ryzin kernel in the ordered case:

$$K_h^{(WR)}\left(x, x^{(i)}\right) = (1-h) \times \mathbb{I}_{\{x=x^{(i)}\}} + \frac{1}{2}(1-h)h^{|x-x^{(i)}|} \times \mathbb{I}_{\{x\neq x^{(i)}\}};$$

(c) Aitchison–Aitken kernel in the unordered case if the categorical variable X admits c different values:

$$K_h^{(AA)}\left(x, x^{(i)}\right) = (1-h) \times \mathbb{I}_{\{x=x^{(i)}\}} + \frac{h}{c-1} \times \mathbb{I}_{\{x\neq x^{(i)}\}}.$$

In this way, the multivariate product kernel becomes

$$K_{\mathbf{h}_n}\left(\mathbf{x}, \mathbf{x}^{(i)}\right) = \prod_{j=1}^{r_c} K_{h_{n,j}}^{(G)}\left(x_j, x_j^{(i)}\right) \prod_{k=1}^{r_o} K_{h_{n,k}}^{(WR)}\left(x_k, x_k^{(i)}\right) \prod_{l=1}^{r_u} K_{h_{n,l}}^{(AA)}\left(x_l, x_l^{(i)}\right),$$

where r_c, r_o, r_u are the number of continuous, discrete ordered, and discrete unordered regressors, respectively. The choice of the kernel is of little importance compared to the bandwidths, and the conditions on the kernel can be relaxed as pointed out by Racine and Li:

- the Gaussian kernel can be replaced with any compactly supported kernel function that is Hölder continuous, e.g., the Epanechnikov kernel;
- the Wang–Ryzin kernel can be replaced with the Aitchison–Aitken kernel, which further simplifies things but performs slightly worse.

Bandwidth selection. The bandwidth is chosen in every iteration step independently, with two possible validation methods or with the forthcoming rule-of-thumb. The first is the leave-one-out cross-validation, by minimizing

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - S_Y^{(n)\setminus\{i\}}(\mathbf{x}^{(i)})\right)^2 M(\mathbf{x}^{(i)}), \tag{12}$$

where $S_Y^{(n)\setminus\{i\}}$ is the estimator in which the ith sample entry is left out from the prediction. $M(\cdot)$ is an optional weight function that trims out boundary observations. In [6], the asymptotic normality results are proved for this choice of bandwidth.

Alternatively, one can use the minimization of the corrected AIC Hurvich criteria, see [13]. These two methods produce asymptotically the same results, [14]. The optimization is performed by the Nelder–Mead downhill simplex method in the underlying Python implementation.

Scott's rule-of-thumb [15] may also work well for the selection of the bandwidth:

$$h_{n,j} \approx 1.06 \times \hat{\sigma}_j \times n^{-\frac{1}{4+r}},$$

where $\hat{\sigma}_j$ is the sample standard deviation of X_j and r is the number of regressors in the particular step. Note that this is close to optimal for Gaussian variables, and also used as the starting point by the code in the aforementioned cross-validation optimization processes.

The aforementioned bandwidth selections are data driven, their asymptotic behavior for the NW (local constant) and the local linear kernel regression is studied in [6, 13, 14, 16]. In [6], the authors prove that under the conditions, listed in Assumptions A.1 and A.2, the estimate (with the leave-one-out cross-validated choice of bandwidths) is asymptotically normal, see Theorem A.2. Their proof uses the same $\hat{h_c}$ for all continuous, and another same $\hat{h_d}$ for all discrete regressors, also they use the Aitchison–Aitken kernel for all discrete variables. They anticipate that similar results can be proved for distinct parameters and using the Wang–Ryzen kernel for the discrete ordered regressors. Asymptotic normality does not always imply mean-square consistency, but under some extra conditions we are able to prove mean-square consistency of their estimator, see Lemma A.1.

Using a full bandwidth matrix. Again, we include results of the NW estimate with a full bandwidth matrix $\mathbf{H}$ as well, see Equation (A.2). We use the multivariate version of the Gaussian kernel,

$$K_{\mathbf{H}}^{(G)}\left(\mathbf{x},\mathbf{x}^{(i)}\right) = \frac{1}{|\mathbf{H}|\sqrt{2\pi}^{r}}\, e^{-\frac{1}{2}\left[\mathbf{H}^{-1}\left(\mathbf{x}-\mathbf{x}^{(i)}\right)\right]^{T}\left[\mathbf{H}^{-1}\left(\mathbf{x}-\mathbf{x}^{(i)}\right)\right]},$$

and the multivariate version of Scott's rule [17] for the selection of a full bandwidth matrix:

$$\mathbf{H}_n \approx 1.06 \times \hat{\Sigma}^{\frac{1}{2}} \times n^{-\frac{1}{4+r}},$$

where $\hat{\Sigma}$ is the sample covariance matrix of $\mathbf{X}$, and r is the number of regressors in a particular step. As for the mean-square consistency of this estimator, we have not found any results in the literature yet.

4.2. Practical considerations

As for the realization of our algorithm, some remarks are in order.

- In case of a discrete unordered categorical intermediate response, the kernel regression method leads to a Bayes classification method via dummy indicator variables, see Section 3.1.2. Then we continue with the original variable in a consecutive step on the regression graph, not with its indicators. In case of a discrete ordered intermediate response, the regression estimation produces a value between two possible consecutive values of the variable. One can leave these predicted values as they are, or substitute a rounded value for them to have a classification instead of regression.

- In case of long chains, our iterative algorithm may result in overtly smooth predictions compared to a direct prediction, because the smoothing will effect not only the target, but all previously predicted variables too.

- In lack of a forbidden quadruple, we can find a Markov equivalent DAG to a regression graph. If we proceed from chain component to chain component along the directed edges, and predict the joint response variables

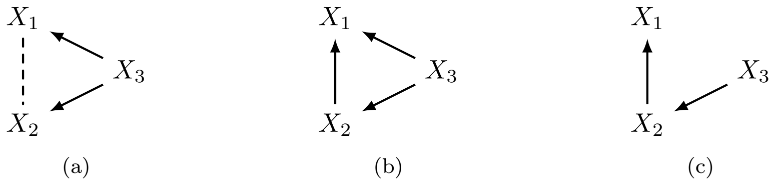

Figure 3: Examples for Cochran's formula

in the same chain component in parallel, we get the same result thanks to the Markov equivalence of the original chain graph and the DAG.

In Fig. 3, we illustrate the equivalence of the predictions in the two setups with three variables. Let X_1, X_2 be on equal standing, connected with a dashed line in the same chain component, whereas their joint parent X_3 is in a chain component of their past (see Fig. 3a). So our iteration predicts X_1 only with X_3, with regression coefficient $\beta_{1|3}$. After constructing a DAG, then say, there is a directed edge $X_2 \to X_1$ (replacing the former dashed line), see Fig. 3b. Along this DAG, our algorithm predicts first X_2 with X_3, then X_1 with X_3 (old parent) and with X_2 (new parent). But in this case, the direct and indirect effect of X_3 is added together as $\beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3}$, where $\beta_{1|3.2}$ is the partial regression coefficient of X_3 when regressing X_1, given also X_2 as regressor for X_1. By the Cochran's formula (see also Wermuth–Sadeghi [2] and Cox–Wermuth [18]):

$$\beta_{1|3} = \beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3}, \tag{13}$$

so we get the same result if our variables are Gaussian, or we confine ourselves to the second moments. As the smoothings imitate the conditional expectations, the predictions would be the same if we used the Markov equivalent DAG. The above equivalence extends to several variables, actually the seminal paper of Wright [19] discusses it with a more complicated notation.

From Equation (13) we can see that in the special case when $X_3 \to X_2 \to X_1$ form a Markov chain, i.e., $X_1 \perp\!\!\!\perp X_3 | X_2$ (see Fig. 3c), $\beta_{1|3.2} = 0$ and $\beta_{1|2.3} = \beta_{1|2}$. Therefore, $\beta_{1|3} = \beta_{1|2}\beta_{2|3}$. If our variables have unit variances, the above product rule extends to the correlation coefficients, and so, $r_{13} = r_{12}r_{23}$. Consequently, for the regression errors we have $1 - r_{12}^2 \le 1 - r_{13}^2$, showing that the prediction error of X_1 by X_2 is smaller than that by X_3, via X_2. In the course of the algorithm, there are indirect and direct effects, so the prediction errors combine in the above way. The main concepts extend to absolutely continuous distributions, other than Gaussian, if we confine ourselves to the second moments.

5. Application

Three examples of application are included. The first is based on randomly generated data; the theoretical regression function can be well controlled in this environment. The second deals with the 2014's Egypt Demographic and

Health Survey (EDHS 2014) data and examines the effect of background characteristics on the ideal number of children a woman thinks manageable to have. The third considers data collected at the Budapest University of Technology and Economics (BME) to study students' academic achievements in the first two semesters, related to their background characteristics and university application scores. In the two real-life examples the regression graphs are built by expert knowledge and the R package **gRchain** [20] which implements the methods described in [18] for multivariate regression chain graphs. In these two examples 90%–10% training–test data split is applied, and the results refer to the test data. The following smoothings are selected:

- the NW estimate of Racine and Li with Scott's bandwidths (NW_{sc});
- the same NW estimate with cross-validated bandwidths (NW_{cv});
- the NW estimate but with the full bandwidth matrix of Scott (NW_H);
- the local linear estimate of Racine and Li with Scott's bandwidths (LL_{sc});
- the k-NN estimate with cross-validated choices for k.

5.1. Randomly generated regression graph

Consider the following randomly generated example. The context variables (X_5, X_6, X_7, X_8) obey a zero-centered 4-dimensional normal distribution with covariance and concentration matrices as follows.

$$\mathbf{\Sigma} = \begin{bmatrix} 0.8 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1.2 & 0.8 & 0.4 \\ 0.4 & 0.8 & 1.2 & 0.6 \\ 0.2 & 0.4 & 0.6 & 0.8 \end{bmatrix}, \qquad \mathbf{\Sigma}^{-1} = \begin{bmatrix} -2 & -1 & 0 & 0 \\ -1 & -2 & -1 & 0 \\ 0 & -1 & -2 & -1 \\ 0 & 0 & -1 & -2 \end{bmatrix}.$$

The size of the sample is 1000. Response variables are generated with the equations:

$$\begin{aligned} X_4 &:= X_5^2 + \varepsilon_4, \\ X_3 &:= X_8^3 + \varepsilon_3, \\ X_2 &:= X_6 X_7 + \varepsilon_2, \\ X_1 &:= X_2 + X_3 + X_4 + \varepsilon_1, \end{aligned}$$

where ε_is are independent, standard normal error terms. This naturally results in the regression graph of Fig. 4.

The context variables X_5, X_6, X_7, X_8 are considered to be known, and we predict X_1 through the intermediate responses X_2, X_3, X_4. In this artificial environment, the regression functions are known, therefore obviously, the prediction is accurate enough, see Fig. 5.

In this example, Gaussian kernel was used for each variable. In Figs. 5 and 6, the response variable (X_1) is plotted versus the four context variables. The data points, their iterated theoretical regression, and the smoothed estimate by ICE are included; a leave-one-out cross-validation, see Equation (12), was applied in each step of the iteration. Error scores (R^2 and MSE) versus the true value of X_1 and the theoretical regression function, are shown in Tab. 1 with respect to

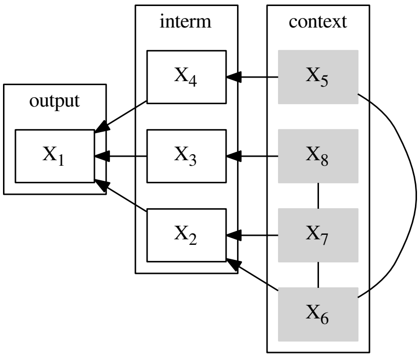

Figure 4: Regression graph of the generated data

the different bandwidth choices and smoothers listed in Section 4.1. With the relatively small sample size of 1000, the cross-validated methods were able to achieve better results than the rule-of-thumb. The scores in the first row are calculated according to

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - x_1^{\{i\}}\right]^2, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - x_1^{\{i\}}\right]^2}{\sum_{i=1}^{n}\left[\overline{x_1} - x_1^{\{i\}}\right]^2}; \quad (14)$$

while, in the second row,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - P_1^{\{i\}}(\cdot)\right]^2, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - P_1^{\{i\}}(\cdot)\right]^2}{\sum_{i=1}^{n}\left[\overline{P_1(\cdot)} - P_1^{\{i\}}(\cdot)\right]^2},$$

where $S_1(\cdot)$ and $P_1(\cdot)$ are as in our Theorem 3.1, and the shorthand $(\cdot)$ for the arguments corresponds to the smoothings and conditional expectations of the previous steps, respectively. The superscript $^{\{i\}}$ indicates the ith test sample entry. These results reassure that in the continuous case and when a well-compatible regression graph is available, the estimate is accurate enough. In this simulated case, the different smoothers give similar results. The cross-validation improved the results of the NW estimate but the local linear regression, even without cross-validation, outperformed the other smoothers. Using a full bandwidth matrix had little if any impact on the results.

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
X_1	4.385	3.982	4.402	3.790	4.543
$P_1(\cdot)$	0.578	0.304	0.629	0.194	1.164

(a) Mean-square errors

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
X_1	0.682	0.696	0.684	0.705	0.675
$P_1(\cdot)$	0.965	0.971	0.962	0.978	0.895

(b) R^2 scores

Table 1: Error scores of the generated example: versus the true value of X_1 (first rows) and the theoretical regression function (second rows)

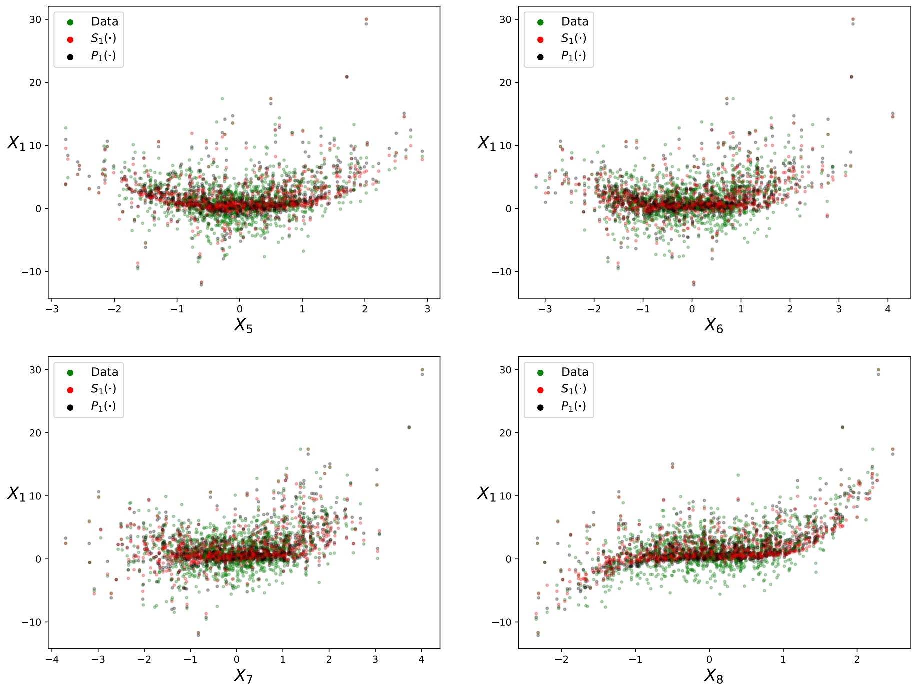

Figure 5: ICE on the generated example: the response variable (X_1) is plotted versus each of the four context variables (X_5, X_6, X_7, X_8); the data points are in green, the iterated theoretical regression is in black, and the smoothed estimate by ICE is in orange

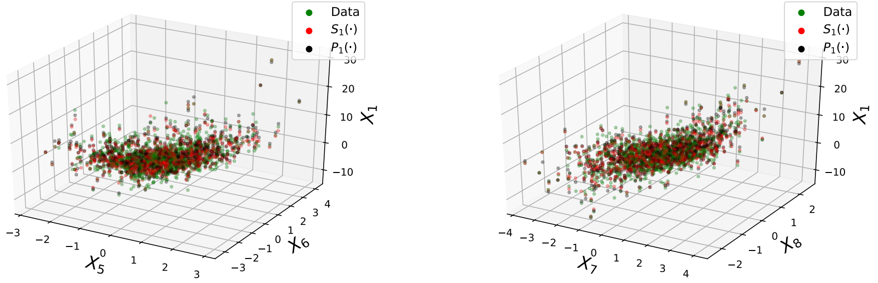

Figure 6: ICE on the generated example (3D): the response variable (X_1) is plotted versus the two pairs of context variables (X_5, X_6 and X_7, X_8); the data points are in green, the iterated theoretical regression is in black, and the smoothed estimate by ICE is in orange

5.2. The EDHS 2014 data

Based on data from the EDHS 2014, we examined the effect of background characteristics on the ideal number of children a woman thinks manageable to have. The research question is: to what extent do age and education level of married couples affect the conceivable ideal number of children, through intermediate variables. The focus is on a selected random sample of more than 17000 urban married women aged 20-49 years. Fig. 7 shows the regression graph, built on the variables listed below. The grouping of the variables is based on the relationships between the variables suggested by the experts.

For some descriptive statistics of the included variables, see Tab. 2. As shown in the regression graph, Fig. 7, the far right box includes the relevant context variables in the model:

- Husband's and Wife's Education level in years;
- Husband's and Wife's Age.

The next box from the right contains two intermediate variables:

- Woman's age at the first marriage;
- Wealth index of the family (ordered variable from 1 to 5).

The graph shows that some variables are considered explanatory for some variables and responses to others. Moving to the next box, the secondary responses are presented:

- Number of years the woman has been using any Contraception method;
- Number of births.

The first box on the left is the primary response variable, the ideal number of children the woman thinks to be optimal.

The choice of kernels is based on considering the age/year-related variables to be continuous, and the others to be ordered discrete. The variables were min-max scaled to the $[0, 1]$ interval before the application, but were rescaled afterwards for the sake of the figures. Our results show an overtly smooth regression surface after the iteration. In Fig. 8, one can see the iterated estimates along the graph. A row of scatter plots belongs to every non-context variable.

	count	mean	std	min	max
AgeWoman	17686	32.79	7.89	20	49
WealthIndex	17686	3.23	1.42	1	5
NumOfBornChildren	17686	2.68	1.62	0	15
AgeWomanAtFirstMar	17686	20.55	4.05	9	49
AgeHusband	17686	39.16	9.56	14	86
IdealNumOfChildren	17686	3.08	1.37	0	24
SchoolYearsWoman	17686	9.34	5.55	0	23
SchoolYearsHusband	17686	9.82	5.28	0	23
ContraceptionYears	17686	5.07	6.16	0	32

Table 2: Descriptive statistics of the variables in the EDHS example

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
IdealNumOfChildren	0.003	0.003	0.003	0.003	0.003
NumOfBornChildren	0.007	0.007	0.007	0.007	0.007
ContraceptionYears	0.030	0.030	0.030	0.030	0.030
WealthIndex	0.100	0.095	0.094	0.094	0.094
AgeWomanAtFirstMar	0.007	0.006	0.006	0.006	0.006

(a) Mean-square errors

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
IdealNumOfChildren	0.028	0.029	0.031	0.031	0.023
NumOfBornChildren	0.378	0.390	0.393	0.396	0.386
ContraceptionYears	0.201	0.203	0.190	0.198	0.195
WealthIndex	0.244	0.251	0.254	0.253	0.252
AgeWomanAtFirstMar	0.309	0.313	0.322	0.325	0.322

(b) R^2 scores

Table 3: Error scores per variable along the iteration in the EDHS example

In each subplot, a non-context variable is presented versus its parent variables, one by one. These are cross-sections of the multidimensional space. Red dots correspond to the actual data points, blue and yellow dots belong to the estimates given by the ICE versus the estimated value of the parent and the actual value of it, respectively. In a subplot, the blue and yellow dots obviously coincide if the parent is a context node or a node with known values. Note that some jittering is added to the data points to have more comprehensible figures. For example, we can see that, regardless of the background characteristics, the iterated regression estimate for the ideal number of children is around 3. In our iterated regime we can calculate error scores for each non-context variable. The calculated R^2-scores and the mean-square errors, Equation (14), are presented in Tab. 3 for all non-context variables with respect to the different smoothings listed in Section 4.1. With the relatively large sample, the other kernel methods were unable to achieve much better results than the rule-of-thumb NW estimate. The k-NN provided similar results.

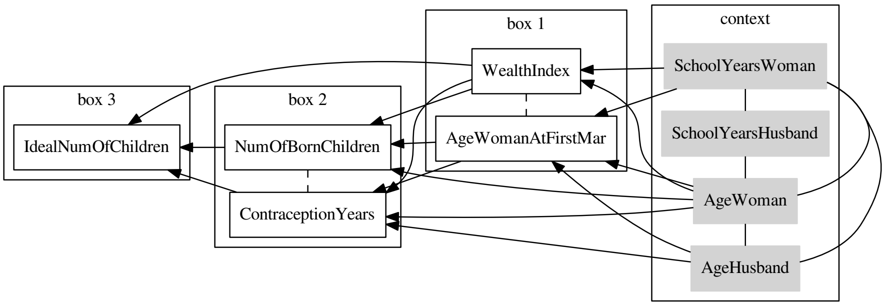

Figure 7: Regression graph of the EDHS example

5.3. On students academic achievements

The Central Academic Office of BME (short: KTH) registers the data of all students entering into the institution. We were interested in the iterative prediction of some characteristics of the students, based on variables which are already available at their application into the university.

The involved data consists of records for over 12000 students with the limitation that it is their first time entering into our university and they wrote a so-called ‘zeroth midterm’ of mathematics.

In the regression graph of Fig. 9, the boxes are formed by the chronological ordering of the included variables. The far right box includes the relevant context variables in the model, which are already available during the application process into the university:

- Matura points (secondary school exit examination);
- Study points, calculated from the grades of the core secondary school subjects;
- Extra points, earned in the application process;
- Gender and Age when starting the first semester.

For a detailed description of the Hungarian university admission system, see [21]. The next box from the right contains an intermediate variable, which is the points earned at the aforementioned ’zeroth midterm’. It is a mandatory entry examination written by first year students of the BME in mathematics. It is taken on the first week of the semester by almost all entering students.

Moving to the next box, the secondary responses are:

- Grade of the first semester mathematics (calculus) course (discrete ordered variable from 1 to 5);
- Credit weighted average result of the first semester without the result of the aforementioned calculus course (continuous).

The first box on the left is the primary response variable, the cumulative credit weighted average result of the first two semesters without the result of the first semester calculus course. Some descriptive statistics of the involved variables are in Tab. 4.

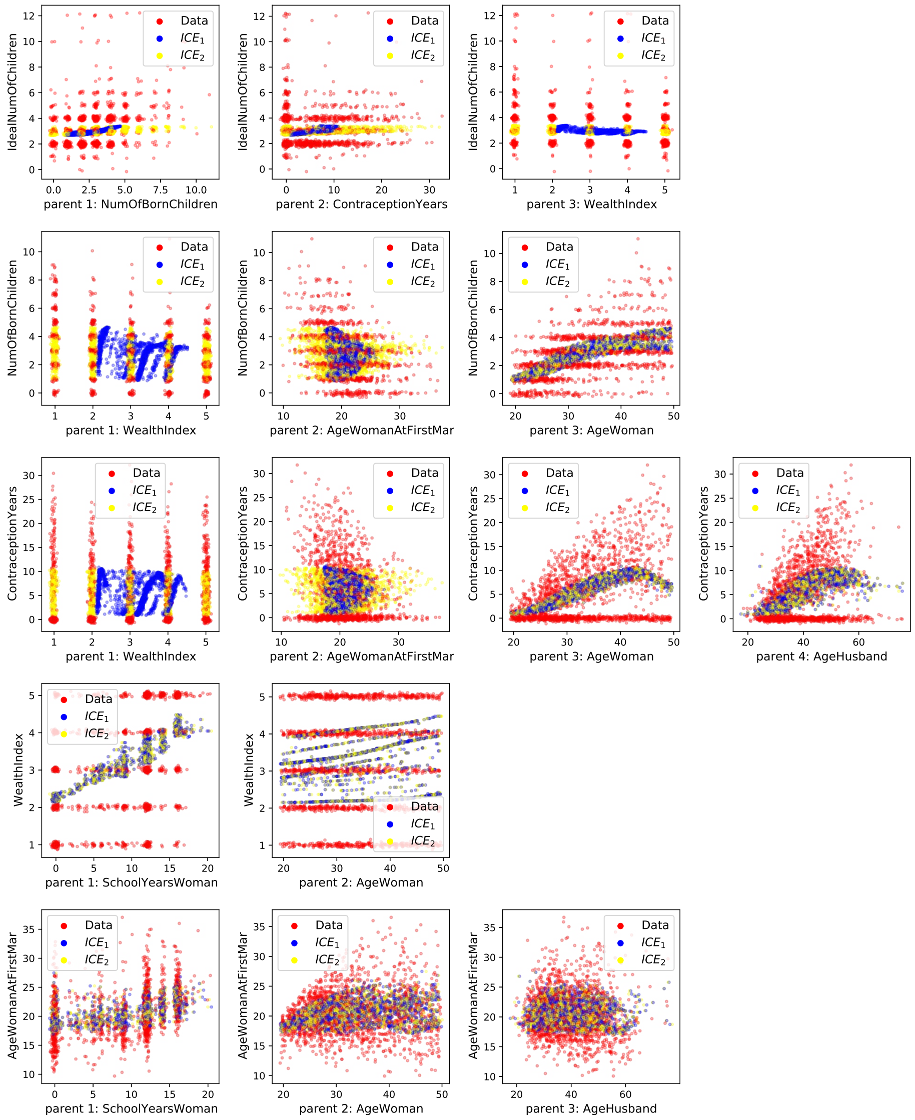

Figure 8: ICE on the EDHS data: to each non-context variable a row of the scatter plot belongs. In each subplot, a non-context variable is presented versus one of its parent variables. Red dots correspond to the actual data points, blue and yellow dots belong to the estimates given by the ICE versus the estimated value of the parent and the actual value of it, respectively.

	count	mean	std	min	max
ZeroExamPoints	11876	29.50	15.66	-13	60
MathsGrade	11876	2.45	1.25	1	5
Gender	11876	0.74	0.44	0	1
Age	11876	20.19	1.02	18	43
CWA w/o Maths	11876	1.86	1.15	0	5
CumCWA w/o Maths	11876	2.10	1.28	0	6
StudyPoints	11876	174.71	14.75	87	200
MaturaPoints	11876	167.66	17.96	81	200
ExtraPoints	11876	69.20	23.46	0	100

Table 4: Descriptive statistics of the variables in the educational example

The variables in this example are not all continuous, but for most of them, the sizes of the sample spaces are relatively large, so we can treat them as continuous variables. Finally, Gender is considered as binary, Age as discrete ordered (its sample space is too small), and the Grade of mathematics also as discrete ordered, every other variable is treated as continuous.

In Fig. 10, scatter plots of the iteration steps are presented, akin to the EDHS example, with the leave-one-out cross-validated choices of bandwidths. Error scores are shown in Tab. 5 for all non-context variables and with respect to the different smoothings listed in Section 4.1. Again, with the relatively large sample, the other kernel methods were unable to achieve much better results than the rule-of-thumb NW estimate, and the k-NN also provided similar numbers. It seems that, with a large enough sample, the techniques aimed at improving the results – like cross-validation over rule-of-thumb, full bandwidth matrix over independent kernels, local linear over NW estimation – have only small effect on the results when applying the ICE algorithm. However, the computational burden can be significant. The problem with k-NN is that, as the sample size grows, finding the close neighbors becomes intractable. Though, implementations handle this issue with clever algorithms, the choice of k remains a problem to be solved.

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
CumCWA w/o Maths	0.024	0.024	0.025	0.024	0.024
CWA w/o Maths	0.026	0.026	0.026	0.026	0.026
MathsGrade	0.072	0.072	0.074	0.070	0.071
ZeroExamPoints	0.031	0.031	0.032	0.033	0.031

(a) Mean-square errors

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
CumCWA w/o Maths	0.426	0.432	0.412	0.428	0.426
CWA w/o Maths	0.430	0.433	0.426	0.431	0.432
MathsGrade	0.368	0.356	0.326	0.347	0.333
ZeroExamPoints	0.331	0.333	0.318	0.300	0.338

(b) R^2 scores

Table 5: Error scores per variable along the iteration in the educational example

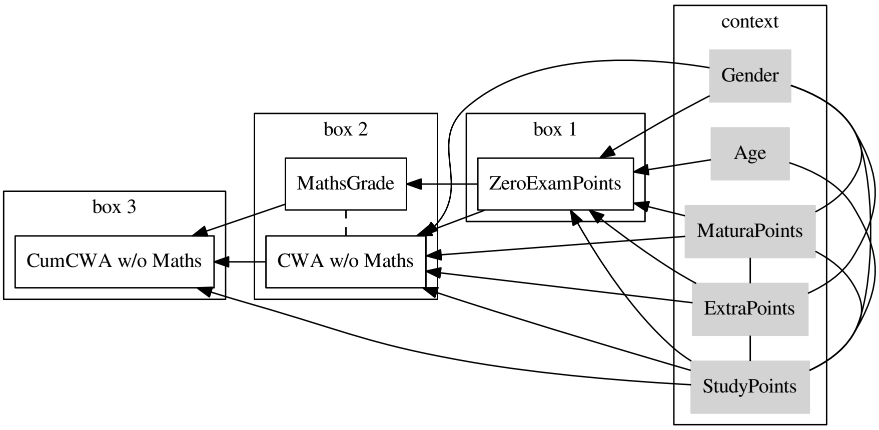

Figure 9: Regression graph of the educational example

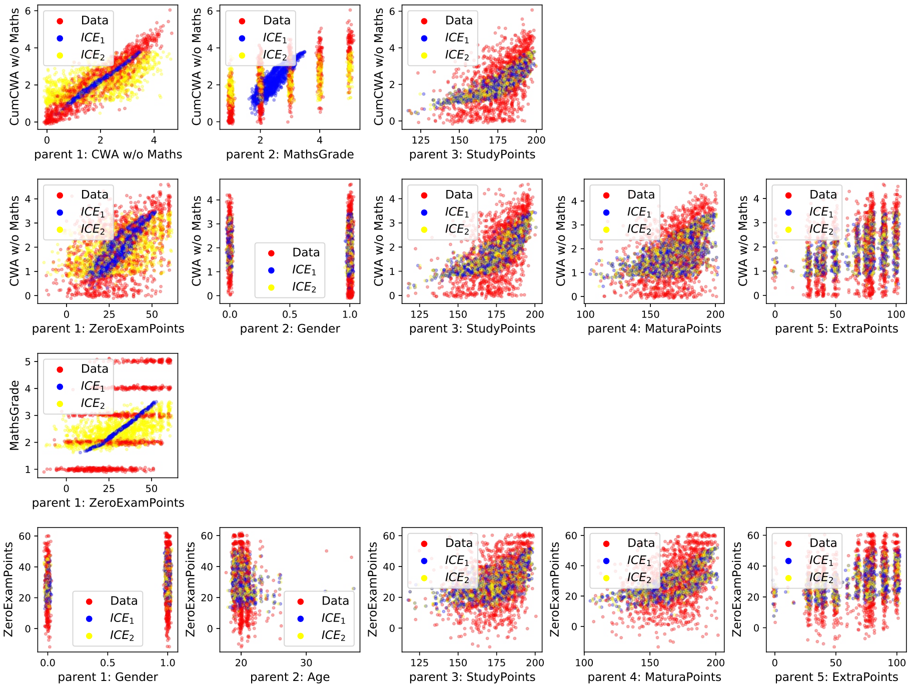

Figure 10: ICE on the educational data: to each non-context variable a row of the scatter plot belongs. In each subplot, a non-context variable is presented versus one of its parent variables. Red dots correspond to the actual data points, whereas, blue and yellow ones to the estimates given by the ICE versus the estimated value of the parent and the actual value of it, respectively.

6. Discussion

The ICE algorithm expands the application area of graphical models and regression graphs. Consistency of the algorithm is proved under generic conditions and useful observations are presented for future applications. The ICE seems versatile enough to be applied in graphical modeling frameworks other than DAGs and regression graphs. Prediction along chain graphs could be part of an artificial intelligence if graph construction is included too.

If no regression graph is known, but the undirected skeleton is triangulated, we can find a junction tree, and make predictions from separators to residuals according to the following factorization of the joint probability density or mass function

$$p(\mathbf{x}) = \prod_{i=1}^{k} p(\mathbf{x}_{R_j} | \mathbf{x}_{S_j}).$$

This resembles Equation (3), and the kernel smoothing usually applies to random vectors.

Consider the ordering of the cliques, obeying the running intersection property with cliques C_j, residuals R_j and separators S_j (indexed from the past to the future), $S_1 = \emptyset$ and $R_1 = C_1$. Assume that we have the coordinates of $\mathbf{x}$ corresponding to C_1. Then

$$\mathbf{x}_{R_j} = \mathbb{E}(\mathbf{X}_{R_j} \,|\, \mathbf{X}_{S_j} = \mathbf{x}_{S_j})$$

for $j = 2, \dots, k$, where k now denotes the number of cliques. Because of $C_j = R_j \cup S_j$, we so get $\mathbf{x}_{C_j}$ and via marginalizing, the new $\mathbf{x}_{S_{j+1}}$ is obtained. For a new-coming case, in the possession of an n-element training sample, we have the estimate

$$\widehat{\mathbf{x}_{R_j}} = \sum_{i=1}^{n} \mathbf{x}_{R_j}^{(i)} W_{R_j}^{(n,i)}(\mathbf{x}_{S_j}) \tag{15}$$

for $j = 2, \dots, k$. Here the same weights can be used for all the components of $\mathbf{x}_{R_j}$ as the variables in the same S_j have usually similar sample characteristics. Since both the separators S_j and the residuals R_j are complete subgraphs (see [3]), the prediction of the joint responses can be easily programmed as they have the same parents in the clique $C_j = R_j \cup S_j$. Hence, in case of a decomposable (equivalently, chordal or triangulated) graph we can proceed from separator to residual, then marginalize.

In case of a generic chain graph, the joint responses usually do not share the same parents. In addition, in the predictions the bandwidths are adopted to the (only) target variable, which technique is not quite straightforward in case of multiple targets with possibly diverse sampling statistics. In a future research we plan to formulate the consistency of the multiple response situations in terms of the error covariance matrix. The ICE algorithm is also applicable to more general constructs emerging in econometrics, like time series, where the directions of the arrows indicate not only causation but time sequence of the observations. Longitudinal data can be treated too. We also plan to involve SEM (Structural Equation Modeling) and PLS (Partial Least Squares) techniques by distinguishing between measurement and latent variables, see e.g., [22].

AppendixA. Background material on kernel regression

In this section we work with the usual (Y: target, $\mathbf{X}$: regressors) setup to keep the notation simple, with an n-element training sample $y^{(i)}, \mathbf{x}^{(i)}$, $i = 1, \ldots, n$. The number or regressors will be denoted by r, out of which the number of continuous ones is p, and $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_d)$ decomposes into continuous and discrete coordinates.

The supervised learning procedure of the kernel regression (see [9, 10]) is historically preceded by the unsupervised learning procedure of kernel density estimation. It is a local averaging technique of Equation (2), which reads here:

$$S_Y^{(n)}(\mathbf{x}) := \sum_{i=1}^{n} y^{(i)} W^{(n,i)}(\mathbf{x}). \tag{A.1}$$

The mean-square consistency of such methods is usually formulated in terms of the following theorem, from which specific conditions can be derived for the different types of local averaging techniques, see e.g., [5].

Theorem A.1 (C. J. Stone [5, 23]). *Assume that the following conditions are satisfied for any distribution of* $\mathbf{X}$*:*

1. $\exists\, c$ *such that for every non-negative function* $g(\cdot)$ *with* $\mathbb{E}\left[g(\mathbf{X})\right] < \infty$ *and for any* n*:*
$$\mathbb{E}\left[\sum_{i=1}^{n} |W^{(n,i)}(\mathbf{X})| \cdot g(\mathbf{X}_i)\right] \leq c \cdot \mathbb{E}\left[g(\mathbf{X})\right],$$
2. $$\exists\, D \geq 1 \quad \textit{such that} \quad \mathbb{P}\left(\sum_{i=1}^{n} |W^{(n,i)}(\mathbf{X})| \leq D\right) = 1,$$
3. $$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} |W^{(n,i)}(\mathbf{X})| \cdot \mathbb{I}_{\{\|\mathbf{X}_i - \mathbf{X}\| > a\}}\right] = 0, \quad \forall\, a > 0,$$
4. $$\sum_{i=1}^{n} |W^{(n,i)}(\mathbf{X})| \to 1 \quad \textit{as } n \to \infty, \quad \textit{in probability},$$
5. *asymptotically all weights become small:*
$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} |W^{(n,i)}(\mathbf{X})|^2\right] = 0.$$

Then for all distribution of $(\mathbf{X}, Y)$ *with* $\mathbb{E}\left[Y^2\right] < \infty$*,*

$$\lim_{n\to\infty} \mathbb{E}\left[\int_{\mathbb{R}^d} |S_Y^{(n)}(\mathbf{x}) - P_Y(\mathbf{x})|^2 \mu(\mathrm{d}\mathbf{x})\right] = 0$$

holds.

The Nadaraya–Watson (NW) kernel estimate of the regression around an appropriate $\mathbf{x}$ is

$$S_Y^{(n)}(\mathbf{x}) = \frac{\sum_{i=1}^n K_{\mathbf{H}_n}\left(\mathbf{x}, \mathbf{x}^{(i)}\right) y^{(i)}}{\sum_{j=1}^n K_{\mathbf{H}_n}\left(\mathbf{x}, \mathbf{x}^{(j)}\right)}, \tag{A.2}$$

where $\mathbf{H}_n = \{h_{i,j}^{(n)}\}_{i=1,j=1}^{r,r}$ is the bandwidth matrix, which has to be positive definite, and

$$K_{\mathbf{H}}\left(\mathbf{x}, \tilde{\mathbf{x}}\right) := \left|\mathbf{H}\right|^{-1} \kappa\left(\mathbf{H}^{-1}\left|\mathbf{x} - \tilde{\mathbf{x}}\right|\right),$$

where κ is a kernel function and $|\cdot|$ refers to coordinatewise absolute value. It is a local averaging estimate of Equation (A.1) with weights

$$W^{(n,i)}(\mathbf{x}) = \frac{K_{\mathbf{H}_n}\left(\mathbf{x}, \mathbf{x}^{(i)}\right)}{\sum_{j=1}^n K_{\mathbf{H}_n}\left(\mathbf{x}, \mathbf{x}^{(j)}\right)}.$$

Choosing the right bandwidth matrix $\mathbf{H}$ is as crucial in the multivariate setup as the selection of the only bandwidth parameter h in the univariate case (only one regressor). However, it is a computationally demanding matrix optimization task, therefore practitioners often choose diagonal bandwidth matrices $\mathbf{H} = \mathrm{diag}(\mathbf{h})$, which reduces the complexity of the task significantly. Considering a full bandwidth matrix gives more flexibility, but notably increases the amount of bandwidth parameters that needs to be chosen. Off-diagonal entries of $\mathbf{H}$ correspond to some dependency between the regressors, but it is an understudied topic in the literature regarding regression problems. Note that in our graphical modeling framework we have to perform several regression estimations, the number of which is $\#\{\text{dimensions}\} - \#\{\text{context variables}\}$, one for each intermediate node; this further increases the computation time. It would make sense to use full bandwidth matrices in our case, because we have some knowledge (through edges between parents of a given node) about the dependencies between the regressors in every step. However, we wanted to decrease the computation time and make the process more automatic, therefore we also used diagonal bandwidth matrices by default.

Mean-square consistency of the NW estimator. For the mean-square consistency of the NW estimator, there are several results. In case of a single deterministic bandwidth parameter h_n, when $\mathbf{H} = diag(h_n, \ldots, h_n)$, and is deterministic in the sense that only depends on the sample size, the following result holds, see [5]. The NW estimator of this case is mean-square consistent if:

- K_h is a bounded non-negative function on $\mathbb{R}^r$ with compact support; further, $K_h \geq \beta \mathbb{I}_B$ with some $\beta > 0$ and some closed sphere B centered at the origin with positive radius ($\mathbb{I}$ is the indicator function);
- $h_n \to 0$ and $nh_n^r \to \infty$ as $n \to \infty$.

This type of kernel is called *boxed* kernel in [5]. In particular, a bounded kernel with compact support, such that it is bounded away from zero at the origin, satisfies this condition. For example, the Epanechnikov kernel is such, but the Gaussian is not. However, we can take truncated Gaussian kernels, the support of which contains the sample entries.

This result can be extended (see [5]) to product kernels (multivariate kernels defined as the product of univariate kernels) with a vector of bandwidth parameters $\mathbf{h}_n = \operatorname{diag}(h_{n,1}, \ldots, h_{n,r})$. In this way, the multivariate kernel is the product of univariate kernels:

$$K_{\mathbf{h}_n}\left(\mathbf{x}, \mathbf{x}^{(i)}\right) = \prod_{j=1}^{r} K_{h_{n,j}}\left(x_j, x_j^{(i)}\right). \tag{A.3}$$

The NW estimator constructed from a product kernel is mean-square consistent if:

- each univariate kernel in the product is *boxed* kernel in the previous sense;
- $h_{n,j} \to 0$ for $j = 1, \ldots, r$ and $n \prod_{j=1}^{r} h_{n,j} \to \infty$.

From the second condition $nh_{n,j} = \infty$ follows for $j = 1, \ldots, r$, but not vice versa.

These consistency results are universal in the sense that they do not depend on the underlying distributions. Practitioners often use the same kernel functions for discrete and continuous regressors, which is a questionable modeling choice. There are kernel functions designed especially for discrete regressors (see the Section 4.1), but their consistency in mean-square sense has not been extensively studied so far.

Product kernel estimator of Racine and Li. Product kernel estimation was further refined for mixed regressors by Racine and Li [6]. Here we list their conditions (with additional remarks) they used to prove their consistency results. Again, the number of continuous regressors will be denoted by p. For the sake of their proof, the same h_c, h_d bandwidths was chosen for all continuous and for all discrete regressors, respectively.

Assumption A.1 ([6])**.** *Conditions on the regression estimation.*

(i) (h_d, h_c) *lie in a shrinking set* $\mathcal{H}_d^{(n)} \times \mathcal{H}_c^{(n)}$, *where*

$$\mathcal{H}_d^{(n)} = \left[0, \min\left(1, \frac{C_0}{\log n}\right)\right] \qquad \mathcal{H}_c^{(n)} = \left[\frac{n^{\delta - \frac{1}{p}}}{C_1}, \frac{C_1}{n^{\delta}}\right]$$

for some $C_0, C_1, \delta > 0$*;*

- *Note that the first part of the condition is* $h_d \to 0$ *and for small* δ *the second part of the condition is virtually identical to* $h_c \to 0$ *and* $nh_c^p \to \infty$.

(ii) the univariate kernel function $K_c(\cdot)$ *used for the continuous variables is non-negative, symmetric, bounded,* $\int K_c(v)v^4 \,\mathrm{d}v < \infty$, *m-times differentiable and* $\int |K_c^{(s)}(v)v^s| \,\mathrm{d}v < \infty$ *for all* $s = 1, \ldots, m$, *where* $m > \max\{2 + 4/p, 1 + p/2\}$*;*

- *Note that the Gaussian kernel satisfies these conditions; further, this condition can be replaced with a compactly supported kernel function that is Hölder continuous, e.g., the Epanechnikov kernel.*

(iii) the weighting function $M(\cdot)$ *(optionally used in the cross-validation, see Equation* (12)*) is bounded and supported on a compact set with nonempty interior for all realization of the discrete variables;*

(iv) the joint density of the variables ($f(\cdot)$ from now on) is bounded from below on the support of $M(\cdot)$.

- *Note that conditions* (iii) *and* (iv) *are only needed to have a uniform convergence rate, $M(\cdot)$ can be omitted from the cross-validation.*

Assumption A.2 ([6])**.** *Conditions on the underlying distribution.*

(i) $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ are i.i.d. and $Y - P_Y(\mathbf{X})$ has finite fourth moment;

(ii) $\sigma(\cdot, \mathbf{x}_d)$, $P_Y(\cdot, \mathbf{x}_d)$, $f(\cdot, \mathbf{x}_d)$ are all twice differentiable, and their partial derivatives (up to the second order) are all bounded by some functions with finite fourth moment for all possible $\mathbf{x}_d$, where

$$\sigma^2(\mathbf{x}) := \mathbb{E}\left[(Y - P_Y(\mathbf{X}))^2 \,\middle|\, \mathbf{X} = \mathbf{x}\right]$$

and $f(\cdot)$ is the joint density of the variables;

(iii) lengthy technical condition, only needed to rule out the case when the regression function is independent of the continuous regressors.

Theorem A.2 ([6])**.** *Under the Assumptions A.1 and A.2, the product kernel regression estimate (with the leave-one-out cross-validated bandwidth choice of Racine and Li, see Equation* (12)*) is asymptotically normal:*

$$\sqrt{n\hat{h_c}^{\,p}}\left(S_Y(\mathbf{x}) - P_Y(\mathbf{x}) - \hat{B}(\hat{h_c}, \hat{h_d})\right) \to \mathcal{N}\left(0, \hat{\Omega}(\mathbf{x})\right),$$

where

$$\begin{aligned}
\hat{\Omega}(\mathbf{x}) &= \frac{\hat{\sigma}^2(\mathbf{x})}{\hat{f}(\mathbf{x})} \int \left(\prod_{i=1}^p K_c(v_i)\right)^2 \, \mathrm{d}\mathbf{v} \\
\hat{B}(\hat{h_c}, \hat{h_d}) &= \hat{h_c}^{\,2} \Phi_1(\mathbf{x}) \int K_c(v) v^2 \, \mathrm{d}v + \hat{h_d} \Phi_2(\mathbf{x}) \\
\Phi_1(\mathbf{x}) &= \frac{\nabla \hat{f}(\mathbf{x})^T \nabla S_Y(\mathbf{x})}{\hat{f}(\mathbf{x})} + \frac{\mathrm{tr}(\nabla^2 S_Y(\mathbf{x}))}{2} \\
\Phi_2(\mathbf{x}) &= \sum_{\tilde{\mathbf{x}}_d : d(\mathbf{x}_d, \tilde{\mathbf{x}}_d) = 1} [S_Y(\mathbf{x}_c, \tilde{\mathbf{x}}_d) - S_Y(\mathbf{x}_c, \mathbf{x}_d)] \frac{\hat{f}(\mathbf{x}_c, \tilde{\mathbf{x}}_d)}{\hat{f}(\mathbf{x}_c, \mathbf{x}_d)} \\
d(\mathbf{x}_d, \tilde{\mathbf{x}}_d) &= \sum_{i=1}^k \mathbb{I}_{\{x_{d,i} \neq \tilde{x}_{d,i}\}} \\
\hat{\sigma}^2(\mathbf{x}) &= \frac{1}{n\hat{f}(\mathbf{x})} \sum_{i=1}^n (Y_i - S_Y(\mathbf{X}_i))^2 K_{\mathbf{h}_n}\left(\mathbf{x}, \mathbf{x}^{(i)}\right) \\
\hat{f}(\mathbf{x}) &= \frac{1}{n} \sum_{i=1}^n K_{\mathbf{h}_n}\left(\mathbf{x}, \mathbf{x}^{(i)}\right).
\end{aligned}$$

Lemma A.1. *Under Assumptions A.1, A.2, further assuming that*

$$\limsup_{n\to\infty} \mathbb{E}\left|\sqrt{n\hat{h_c}^{\,p}}\left(S_Y(\mathbf{x}) - P_Y(\mathbf{x}) - \hat{B}(\hat{h_c}, \hat{h_d})\right)\right|^3 < \infty, \tag{A.4}$$

$$\mathbb{E}\left[\Phi_1(\mathbf{X})^2\right] < \infty \quad \textit{and} \quad \mathbb{E}\left[\Phi_2(\mathbf{X})^2\right] < \infty, \tag{A.5}$$

the estimate is mean-square consistent.

Proof. We need that

$$\mathbb{E}\left[(S_Y(\mathbf{X}) - P_Y(\mathbf{X}))^2\right] = \tilde{\mathbb{E}}\left[\int (S_Y(\mathbf{x}) - P_Y(\mathbf{x}))^2 \mu(\mathrm{d}\mathbf{x})\right] \to 0 \text{ as } n \to \infty,$$

where $\tilde{\mathbb{E}}$ refers to the expectation over the distribution of the sample. We use the $\hat{B}()$ term of Theorem A.2:

$$\begin{aligned}&\mathbb{E}\left[(S_Y(\mathbf{X}) - P_Y(\mathbf{X}))^2\right] = \\ &\mathbb{E}\left[(S_Y(\mathbf{X}) - P_Y(\mathbf{X}) - \hat{B}(\hat{h_c}, \hat{h_d}) + \hat{B}(\hat{h_c}, \hat{h_d}))^2\right] \leq \\ &2\mathbb{E}\left[(S_Y(\mathbf{X}) - P_Y(\mathbf{X}) - \hat{B}(\hat{h_c}, \hat{h_d}))^2\right] + 2\mathbb{E}\left[(\hat{B}(\hat{h_c}, \hat{h_d}))^2\right]\end{aligned}$$

For the first term:

$$S_Y(\mathbf{x}) - P_Y(\mathbf{x}) - \hat{B}\left(\hat{h_c}, \hat{h_d}\right) \asymp \mathcal{N}\left(0, \frac{\hat{\Omega}(\mathbf{x})}{\sqrt{n\hat{h_c}^p}}\right)$$

asymptotically as $n \to \infty$, by Theorem A.2. Further,

$$\lim_{n\to\infty} \int \mathbb{E}\left[(S_Y(\mathbf{x}) - P_Y(\mathbf{x}) - \hat{B}(\hat{h_c}, \hat{h_d}))^2\right] \mu(\mathrm{d}\mathbf{x}) = \lim_{n\to\infty} \int \frac{\hat{\Omega}(\mathbf{x})}{\sqrt{n\hat{h_c}^p}} \mu(\mathrm{d}\mathbf{x}) = 0,$$

because Equation (A.4) ensures the convergence of moments (see Example 2.21 of [24]), and $\hat{\Omega}(\mathbf{x})$ is essentially bounded:

$$\hat{\Omega}(\mathbf{x}) = \frac{\hat{\sigma}^2(\mathbf{x})}{\hat{f}(\mathbf{x})} C_K \leq \frac{D^2}{\hat{f}(\mathbf{x})} C_K \leq \frac{D^2}{\varepsilon} C_K,$$

where $D = Y_{max} - Y_{min}$ is the range of Y in the sample, $\varepsilon = \inf\{\hat{f}(\mathbf{x}) | \hat{f}(\mathbf{x}) > 0\}$ and C_K is the finite integral in the definition of $\hat{\Omega}(\mathbf{x})$. (Note that $\hat{f}(\mathbf{x}) = 0$ only for those $\mathbf{x}$s that are "too far away" from the sample, resulting in undefined $S_Y(\mathbf{x})$ and $\hat{\Omega}(\mathbf{x})$. The probability of this event goes to zero as $n \to \infty$, so the bound will hold almost surely.) Therefore, we can use the dominated convergence theorem and that $nh_c^p \to \infty$ by Assumption A.1(i) to get the zero limit. For the second term:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\hat{B}(\hat{h_c}, \hat{h_d})^2\right] =&\mathbb{E}\left[\lim_{n\to\infty} \hat{B}(\hat{h_c}, \hat{h_d})^2\right] = \\ &\mathbb{E}\left[\lim_{n\to\infty}\left(\hat{h_c}^2 \Phi_1(\mathbf{X}) \int K_c(v) v^2 \ \mathrm{d}v + \hat{h_d} \Phi_2(\mathbf{X})\right)^2\right] = 0,\end{aligned}$$

where $\hat{h_c} \to 0$, $\hat{h_d} \to 0$ and $\int K_c(v) v^2 \ \mathrm{d}v < \infty$, by Assumption A.1. The limit and the expectation can be interchanged by the dominated convergence theorem:

$$\left|\hat{h_c}^2\Phi_1(\mathbf{X})\int K_c(v)v^2\ \mathrm{d}v+\hat{h_d}\Phi_2(\mathbf{X})\right|<C_1\Phi_1(\mathbf{X})\int K_c(v)v^2\ \mathrm{d}v+C_0\Phi_2(\mathbf{X}),$$

where C_0, C_1 are from Assumption A.1, and by Equation (A.5) the second moment of the right hand side is finite. □

AppendixB. Background material on other smoothers

Local linear regression. The NW estimate is also called local constant estimate, since it is actually the minimizer of

$$S_Y^{(n)}(\mathbf{x})=\underset{\alpha(\mathbf{x})}{\operatorname{argmin}}\sum_{i=1}^{n}K_{\mathbf{H}_n}\left(\mathbf{x},\mathbf{x}^{(i)}\right)\left[y^{(i)}-\alpha(\mathbf{x})\right]^2=\hat{\alpha}(\mathbf{x}),$$

where $\hat{\alpha}(\mathbf{x})$ is the optimal constant estimate around $\mathbf{x}$. Likewise, local linear and polynomial estimates can be constructed by substituting a polynomial (with coefficients depending on $\mathbf{x}$) for $\alpha(\mathbf{x})$. In the linear case:

$$\left\{\hat{\alpha}(\mathbf{x}),\hat{\beta}(\mathbf{x})\right\}=\underset{\alpha(\mathbf{x}),\beta(\mathbf{x})}{\operatorname{argmin}}\sum_{i=1}^{n}K_{\mathbf{H}_n}\left(\mathbf{x},\mathbf{x}^{(i)}\right)\left[y^{(i)}-\left(\alpha(\mathbf{x})-\beta(\mathbf{x})^T\mathbf{x}^{(i)}\right)\right]^2,$$

and the estimate from this is:

$$S_Y^{(n)}(\mathbf{x})=\hat{\alpha}(\mathbf{x})+\hat{\beta}(\mathbf{x})^T\mathbf{x}.$$

Local linear estimators usually work better in practice, for example, they do not suffer from bias at the boundary of the regressors as the NW estimator does. The local linear version of Racine and Li's kernel estimate is discussed in details in [14]. The mean-square consistency in case of continuous regressors is proved in [25], albeit under restrictive conditions on the coefficients of the polynomials.

The k-nearest neighbor regression. This so-called k-NN regression is another type of a local averaging regression estimate, see Equation (A.1). While fixing an $\mathbf{x}\in\mathbb{R}^r$, reorder the training data according to the increasing values of $\|\mathbf{X}_i-\mathbf{x}\|$:

$$\left\{\left(\mathbf{X}_{1,n}(\mathbf{x}),Y_{1,n}(\mathbf{x})\right),\left(\mathbf{X}_{2,n}(\mathbf{x}),Y_{2,n}(\mathbf{x})\right),\ldots\left(\mathbf{X}_{n,n}(\mathbf{x}),Y_{n,n}(\mathbf{x})\right)\right\}.$$

Then

$$S_Y^{(n)}(\mathbf{x}):=\frac{1}{k_n}\sum_{i=1}^{k_n}Y_{i,n}(\mathbf{x})$$

is the k_n-NN regression estimate of Y.

In case of continuous regressors, the following theorem can be stated on the mean-square consistency.

Theorem B.1 (Theorem 6.1 of [5]). *If* $\lim_{n\to\infty}k_n=\infty$ *and* $\lim_{n\to\infty}\frac{k_n}{n}=0$, *then the* k_n*-NN estimate is mean-square consistent for all distribution of* $(\mathbf{X},Y)$ *such that* $\mathbb{E}\left[Y^2\right]<\infty$ *and for each* $\mathbf{x}\in\mathbb{R}^r$*, the random variable* $\|\mathbf{X}-\mathbf{x}\|^2$ *is absolutely continuous.*

The last condition is the no-tie condition, which always can be satisfied by the inclusion of a new component Z attached to $\mathbf{X}$: $((\mathbf{X}, Z), Y)$, where Z is uniformly distributed on $[0, 1]$ and independent of $(\mathbf{X}, Y)$. Of course, the training sample is also supplemented with components $Z_1, Z_2, \ldots, Z_n$.

Acknowledgements

The authors are indebted to Nanny Wermuth for her valuable explanations, to Fatma Abdelkhalek for her expert knowledge regarding the EDHS data, to the Central Academic Office of BME for providing the educational data, further to Tamás Szabados and László Györfi for their helpful comments. We also thank the anonymous reviewers of the paper for their useful suggestions. The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology and Economics (BME-FIKP-MI/SC). The research is also supported by the EFOP-3.6.2-16-2017-00015-HU-MATHS-IN Hungarian Industrial and Innovation project, under the auspices of the Széchenyi István University, Győr.

References

[1] N. Wermuth, Traceable regressions, International Statistical Review 80 (3) (2012) 415–438.

[2] N. Wermuth, K. Sadeghi, Sequences of regressions and their independences, TEST 21 (2012) 215–279.

[3] M. Bolla, F. Abdelkhalek, M. Baranyi, Graphical models, regression graphs, and recursive linear regression in a unified way, Acta Scientiarum Mathematicarum 85 (12) (2019) 9–57. doi:10.14232/actasm-018-331-4.

[4] L. Breiman, J. H. Friedman, Estimating optimal transformations for multiple regression and correlation, Journal of the American Statistical Association 80 (1985) 580–619.

[5] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Series in Statistics, Springer, 2002.

[6] J. Racine, Q. Li, Nonparametric estimation of regression functions with both categorical and continuous data, Journal of Econometrics 119 (1) (2004) 99–130.

[7] A. Rényi, On measures of dependence, Acta Mathematica Academiae Scientiarum Hungarica 10 (3) (1959) 441–451. doi:10.1007/BF02024507.

[8] L. P. Devroye, T. J. Wagner, Distribution-free consistency results in nonparametric discrimination and regression function estimation, The Annals of Statistics 8 (2) (1980) 231–239.

[9] E. A. Nadaraya, On non-parametric estimates of density functions and regression curves, Theory of Probability & Its Applications 10 (1) (1965) 186–190.

[10] G. S. Watson, Smooth regression analysis, Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 26 (4) (1964) 359–372.

[11] S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with python, in: S. van der Walt, J. Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 57 – 61.

[12] J. S. Racine, Nonparametric Econometrics: A Primer, Foundations and Trends® in Econometrics 3 (1) (2008) 1–88.

[13] C. M. Hurvich, J. S. Simonoff, C.-L. Tsai, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60 (2) (1998) 271–293.

[14] Q. Li, J. Racine, Cross-validated local linear nonparametric regression, Statistica Sinica 14 (2) (2004) 485–512.

[15] D. W. Scott (Ed.), Multivariate Density Estimation, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1992.

[16] W. Hardle, P. Hall, J. S. Marron, How far are automatically chosen regression smoothing parameters from their optimum?, Journal of the American Statistical Association 83 (401) (1988) 86–95. doi:10.2307/2288922.

[17] W. K. Härdle, M. Müller, S. Sperlich, A. Werwatz, Nonparametric and semiparametric models, Springer Science & Business Media, 2004.

[18] D. R. Cox, N. Wermuth, Multivariate Dependencies: Models, Analysis and Interpretation, Vol. 67, CRC Press, 1996.

[19] S. Wright, The method of path coefficients, The Annals of Mathematical Statistics 5 (3) (1934) 161–215. doi:10.1214/aoms/1177732676.

[20] T. Rusch, M. Wurzer, R. Hatzinger, Chain Graph Models in R: Implementing the Cox-Wermuth Procedure, presented at the Psychoco International Workshop on Psychometric Computing, Zürich (Feb. 2013).

[21] M. Nagy, R. Molontay, Predicting dropout in higher education based on secondary school performance, in: 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), IEEE, 2018, p. 000389–000394.

[22] M. Tenenhaus, V. E. Vinzi, Y.-M. Chatelin, C. Lauro, PLS path modeling, Computational Statistics & Data Analysis 48 (1) (2005) 159–205.

[23] C. J. Stone, Consistent nonparametric regression, The Annals of Statistics 5 (4) (1977) 595–620. doi:10.1214/aos/1176343886.

[24] A. W. van der Vaart, Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998. doi:10.1017/CBO9780511802256.

[25] M. Kohler, Universal consistency of local polynomial kernel regression estimates, Annals of the Institute of Statistical Mathematics 54 (4) (2002) 879–899.