Experimental Analysis of Approaches to Multidimensional Conditional Density Estimation*

Anna Berger[1][0000−0002−0268−2370] and Sergey Guda[1,2][0000−0002−2398−1847]

[1] Institute of Mathematics, Mechanics, and Computer Science,
Southern Federal University, Milchakova 8a, 344090 Rostov-on-Don, Russia
[2] The Smart Materials Research Center, Southern Federal University,
Sladkova Street 174/28, 344090 Rostov-on-Don, Russia
{anna.ig.berger, gudasergey}@gmail.com

Abstract. Recently several original methods for conditional density estimation (CDE) have been developed. The abundance of information comprised by the full conditional density of target variables is great when compared to the regression or quantile regression estimates. Still, there are only few independent experimental investigations of these methods, especially concerning a multidimensional target variable, and this paper aims to address this issue. We consider several approaches such as kernel density estimation, reduction to binary classification, Naïve Bayes, Bayesian Network, "varying coefficient“ approach, random forests and Approximate Bayesian Computation applied to a conditional density estimation problem. We examine these methods when applying to various datasets together with the dependency of the methods' performance on different parameters including the number of irrelevant covariates, smoothness, and flatness of the distribution. Considered datasets include artificial models with required properties and with the known exact value of CDE evaluation measure and a real-world dataset arisen from the problem of structure recognition by XANES spectra, which is reduced to a regression task with a complex multimodal probability distribution of the target variable. The special attention is paid to the computation of the evaluation measure as the methods based on the direct optimization of the loss employ its imprecise but fast approximation which results in the poor prediction quality for datasets with a small target variance.

Keywords: conditional density estimation, multidimensional regression, experimental analysis

1 Introduction

Conditional density estimation of a random variable $z \in \mathbb{R}$ can be considered a generalization of a regression problem. Standard regression returns point estimation of a target variable z. It minimizes the standard deviation leading to a

* This work was supported by the Russian Foundation for Basic Research, project 18-02-40029.

prediction of the target expectation. In case of a multimodal or skewed distribution, the expectation value has a smaller probability than the mode — the most probable value. But a researcher expects namely the latter to obtain as a result of the regression algorithm as it is implied by the word “expectation”.

To overcome the limitation of the point regression estimation, in many applications such as time series analysis, the quantile regression is employed. For the given probability, it determines the minimal interval having the form $(q, +\infty)$ and containing target value [1]. This approach makes it possible to predict the interval containing the target variable for the given confidence level. Another way of generalization is to consider regression function $X \to Z$ as a manifold in ambient space $X \times Z$. It enables the construction of prediction regions with a given confidence level [3,9,10].

Still, the distribution of the target variable can happen to be too complex to be well estimated via quantile regression: for instance, if multimodality or a significant skew of the response is present in the data. In this case, the standard and quantile regression may be insufficient for the proper data analysis and solving the problem, while the estimator of full conditional density provides a more comprehensive accounting of the target variable.

Several recent works utilize the CDE of the full probability distribution in various application domains and, by doing so, achieve substantial improvements. The approach proves itself especially in settings with complicated sources of errors which are widespread in physics in general [2] — and in Cosmology [14], in particular.

The other possible scope of application for CDE is multitask learning. If the objective function can be decomposed into several autonomous parts (e.g. error squares for different target coordinates), then the multitask problem splits itself into an independent problem for each target component. However, estimation of some non-decomposable target, such as the mode of the target variable, is a different matter. In this case, one cannot optimize the components separately as the combination of univariate target component modes is, generally speaking, not the mode of the multivariate target. The mode regression was developed for univariate case (see [8], [15]). Nevertheless, to the best of the authors' knowledge, currently, there are no methods for multi-target mode regression, except using CDE. While namely these methods, for example, are needed to solve the problem of predicting the molecular structure by a given XANES spectrum (see [4] Section 3.5.2). Due to the independence of the XANES spectrum on symmetry transformations of molecule geometry and some geometry parameters, the molecule geometry probability distribution has multiple modes, which are hard to predict with ordinary regression techniques.

As of today, there is a lack of comparison of different CDE methods in literature. This paper aims to fill this gap and provides the experimental overview and comparison of these methods when applying to the data from various distributions. We demonstrate the soundness of such methods as kernel density estimation, reducing to binary classification, Naïve Bayes, Bayesian Network, "varying coefficient“ approach, random forests, Approximate Bayesian Computation and study the dependency of the methods performance quality on volatility degree when x changes, correlation between z components, the number of irrelevant covariates, flatness of the underlying distribution.

For measuring quality performance CDE loss is employed as the most commonly used measure for this type of problem. It implies calculating several multivariate integrals, which are not always computed precisely especially in the algorithms, which directly optimize the CDE loss. That motivates us to study the error of computing the CDE loss by different methods separately.

There are several techniques, which can be adopted for conditional density estimation. Kernel density estimation is one of the simplest of them though it heavily suffers from the curse of dimensionality in the multivariate setting. Another view on a CDE problem can be reformulating it in terms of a binary classification problem and further utilizing existing powerful binary classifiers. The assumption of conditional independence of response variables results in the Naïve Bayes approach, and, as in many cases it is not fulfilled, but the relationship among the target variables is known, the Bayesian Network as well can be built to model the outcome. A different view of the problem is presented by a group of "varying coefficient" approaches which include FlexCode [7] and RFCDE [13] and exploit expanding the conditional density function into orthogonal series. The combination of CDE and Approximate Bayesian Computation (ABC) incorporates the advantages of both approaches and leads to better density estimates [6].

The remainder of the paper is structured as follows. We discuss the existing approaches to CDE in more detail in Section 2. Section 3 addresses the chosen performance measure and its approximations employed in several methods. In Section 4 the datasets under consideration are described. In Section 5 we present the results and the reflections on our findings. We conclude the research and present some suggestions for future work in Section 6.

2 CDE approaches

Assume we observe the finite sample of data $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^{n}$ with multidimensional covariates $\boldsymbol{x}_i \in \mathbb{R}^m$ and a multidimensional response $\boldsymbol{z}_i \in \mathbb{R}^\ell$. The goal of conditional density estimator methods is to reconstruct the full conditional probability density function $p(\boldsymbol{z}|\boldsymbol{x})$ as, in general, it provides us with a more comprehensive understanding of the underlying probability distribution than point estimations of conditional mean and variance. The following paragraphs give a brief description of the methods which can be employed for conditional density estimation.

Kernel Density Estimation. The first approach to this problem is estimating $p(\boldsymbol{x}, \boldsymbol{z})$ and $p(\boldsymbol{x})$ separately and then blending them as $p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z})}$. The estimation of a joint probability function $p(z, x)$ and a marginal probability function $p(\boldsymbol{x})$ can be performed with kernel density estimators (KDE).

Binary Classification Approach. One more approach to conditional density estimation is based on reformulating the problem of density estimation as a binary classification problem. To accomplish it, we assign class $c = 1$ to all points of the given dataset, sample new points from a uniform distribution in a bound rectangle of our data sample and then treat the latter as the elements of the class $c = 2$. Applying then any existing classifier such as logistic regression or decision trees, we calculate probabilities $p(c|\boldsymbol{x}, \boldsymbol{z})$ and $p(c|\boldsymbol{x})$.

The number of sampled points is the same as in the given data sample to keep the dataset balanced. This approach allows us to estimate again $p(\boldsymbol{x}, \boldsymbol{z})$ and $p(\boldsymbol{x})$ separately and, after that, compute $p(\boldsymbol{z}|\boldsymbol{x})$:

$$p(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{V_{xz}} \frac{p(0|\boldsymbol{x}, \boldsymbol{z})}{p(1|\boldsymbol{x}, \boldsymbol{z})}, \quad p(\boldsymbol{x}) = \frac{1}{V_x} \frac{p(0|\boldsymbol{x})}{p(1|\boldsymbol{x})}, \quad p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z})},$$

where V_{xz} and V_x are the volumes of bound rectangles of $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^n$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ samples correspondingly.

Naïve Bayes Approach. Naïve Bayes approach can be applied under the assumption of conditional independence of $\boldsymbol{z}$ coordinates — response variables $z^{(i)}$. It enables application of univariate methods to the problems with multidimensional $\boldsymbol{z}$.

Bayesian Network Approach. The extension of the previous estimator is the approach based on a Bayesian network built on the response variables $z^{(i)}$. The parent-child dependencies of response variables $z^{(i)}$ should be defined at the training step and should represented via a directed acyclic graph. Then, the conditional density estimation is performed as follows:

$$p(\boldsymbol{z}|\boldsymbol{x}) = p(z^{(1)}, z^{(2)}, \ldots, z^{(\ell)}|\boldsymbol{x}) = \prod_{i=1}^{\ell} p(z^{(i)}|\text{Parents}(z^{(i)}), \boldsymbol{x}).$$

FlexCode approach proposed in the paper [7] involves expanding the conditional density function into a series in which each coefficient can be estimated directly via regression if the chosen basis is orthonormal. This approximation reduces the multidimensional conditional density estimation problem to point estimation problem which is more straightforward to fulfill.

RFCDE: Random Forests for Conditional Density Estimation. The paper [13] further develops the idea of a series expansion and focuses on building ensembles of regression trees suggesting the method called Random Forests for Density Estimation (RFCDE). Its main contribution is the way of choosing the partition splits in the nodes of the trees: instead of minimizing traditional mean-squared loss, they optimize the loss specific to CDE (which will be discussed in Section 3). Several simplifications allow them to keep the optimization process computationally feasible.

NNKCDE: Approximate Bayesian Computation Method (ABC-CDE method). The last paper to consider in the scope of this research is [6], which suggests an efficient methodology for non-parametric conditional density estimation for the problems with inaccessible or intractable likelihood and available though limited data simulations. It aims for estimating the posterior density with the means of Approximate Bayesian Computation (ABC). The suggested framework combines the advantages of both ABC and CDE approaches and proposes not only the way of estimating the posterior density upon observing high-dimensional data but also the way of comparing the performance of ABC and other related methods and choosing the optimal summary statistics.

During the first step of the algorithm, the training set is constructed via a simple ABC rejection sampling algorithm for choosing the sample points $\boldsymbol{x}$ close to the given $\boldsymbol{x}_0$ according to some pre-defined distance function $d(\boldsymbol{x}, \boldsymbol{x}_0)$ and δ such as $d(\boldsymbol{x}, \boldsymbol{x}_0) < \delta$. This training set is further employed for building the conditional density estimator. For this purpose the authors of the paper [6] adopt the FlexCode estimator from their previous paper [7], nevertheless mentioning the flexibility of the estimator choice. All the advantages mentioned above of the FlexCode estimator apply to this problem as well.

3 CDE measure

One of the most straightforward approaches to assessing the quality of the obtained density function is evaluating some of its point estimates such as median, mean or any appropriate raw or central n-th moments of the distribution. The problem is that in many cases, such as multimodality, asymmetry and heteroscedastic noise, the point estimates do not fully describe the underlying data structure and therefore, cannot be considered representative for estimation quality assessment.

In order to quantify the distribution in a more comprehensive way, we need to measure the difference between the estimation and the true values in all data points of the sample. The most commonly used metric for estimating the discrepancy between exact $p(\boldsymbol{z}|\boldsymbol{x})$ and approximate $\hat{p}(\boldsymbol{z}|\boldsymbol{x})$ is the mean integrated square error (MISE):

$$\text{MISE} = \mathbb{E}_x \left(\int (\hat{p}(\boldsymbol{z}|\boldsymbol{x}) - p(\boldsymbol{z}|\boldsymbol{x}))^2 d\boldsymbol{z} \right).$$

As it is hard to calculate, the reduced measure is used:

$$\begin{aligned} L(p, \hat{p}) &= \iint \hat{p}^2(\boldsymbol{z}|\boldsymbol{x}) dF(\boldsymbol{x}) d\boldsymbol{z} - 2 \iint \hat{p}(\boldsymbol{z}|\boldsymbol{x}) p(\boldsymbol{z}|\boldsymbol{x}) dF(\boldsymbol{x}) d\boldsymbol{z} = \\ &= \iint \hat{p}^2(\boldsymbol{z}|\boldsymbol{x}) dF(\boldsymbol{x}) d\boldsymbol{z} - 2 \iint \hat{p}(\boldsymbol{z}|\boldsymbol{x}) dF(\boldsymbol{x}, \boldsymbol{z}) \end{aligned} \tag{1}$$

where $F(x)$ is a marginal cumulative distribution function. $L(p, \hat{p})$ differs from MISE by a constant, which doesn't depend on the estimator $\hat{p}$. MISE generalizes the mean squared error and controls the overall MSE of the entire density function. It is closely related to the L_2 error of estimating a function.

If the “varying coefficient” approach [7] is employed, then the MISE measure can be rewritten in a form more convenient for optimization:

$$\widehat{L}(p, \hat{p}) = \sum_{i=1}^{I} \frac{1}{n} \sum_{k=1}^{n} \hat{\beta}_i^2(\boldsymbol{x}_k) - 2\frac{1}{n} \sum_{k=1}^{n} \hat{p}(\boldsymbol{z}_k|\boldsymbol{x}_k). \tag{2}$$

This is also the loss employed in [13] for defining the optimal split in tree nodes. Its properties allow the execution to be performed in a parallel manner which makes the optimizational process relatively fast.

For our experiments, we employ two approaches to computing the MISE: straightforward (1) which involves computing the multiple integral with rectangle rule in the bounded rectangle of the training sample and the approximate one (2) suggested in [7]. We also examine these MISE losses by comparing them with the best possible loss computed for the datasets with the pre-defined distribution and study the errors gained.

4 Datasets

We evaluate the approaches discussed in Section 2 on several datasets to reveal the strengths and weaknesses of the methods and to assess the problems they could help to overcome. They include artificial datasets with required properties and with the known exact value of CDE evaluation measure and practical significant multimodal dataset [11,12].

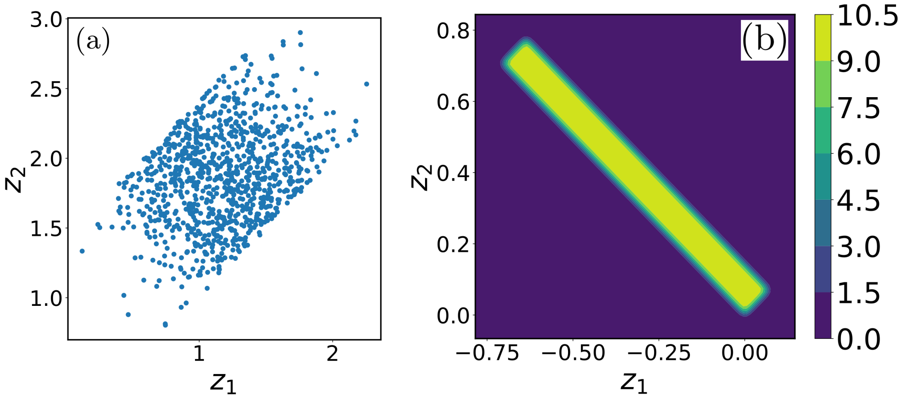

Fig. 1: (a) SURD($n_x = 10$, $n_y = 1$, $n_z = 2$, $\boldsymbol{d} = [0.1, 1]$, $\sigma = 0.01$, $\boldsymbol{\alpha} = [0.3, 0.3]$, $\varphi = \frac{\pi}{4}$), (b) density of the distribution given in (a) at $\boldsymbol{x} = \boldsymbol{0}$.

The artificial datasets are generated from one general scheme of sampling by fixing different parameters. We consider the smoothed uniform rotated distribution SURD(n_x, n_y, n_z, $\boldsymbol{d}$, σ, $\boldsymbol{\alpha}$, φ):

$$\begin{gathered} x^{(i)},\ y^{(j)} \sim \text{Uniform}(0,1),\ i = 1..n_x,\ j = 1..n_y, \\ \widetilde{z}^{(k)} \sim \text{Uniform}(0, d^{(k)}) + \text{Normal}(0,\sigma), k = 1..n_z, \\ \boldsymbol{z} = Q\widetilde{\boldsymbol{z}} + \boldsymbol{\alpha}\sum_{i=1}^{n_x} x_i. \end{gathered} \tag{3}$$

Here $x^{(i)}$ are relevant covariates, $y^{(j)}$ are irrelevant covariates, Q stands for the orthogonal rotation matrix which for every pair of coordinates $(2k-1, 2k)$, $k = 1..\frac{n_z}{2}$ rotates the point by the degree of rotation φ. The variance of the distribution along axes is determined by the vector $\boldsymbol{d}$ while the $\boldsymbol{\alpha}$ parameter explicitly controls the dependence of $\boldsymbol{z}$ on the values of $\boldsymbol{x}$. The σ parameter regulates the smoothness of the distribution. Figure 1 shows the example of sampling according to the described scheme together with the density of the distribution at the fixed point $\boldsymbol{x} = \boldsymbol{0}$.

The artificial univariate multimodal dataset Fork(n_x, n_y, m, σ) (considered in the RFCDE paper [13]) is generated by the following scheme to verify our findings from the multivariate case: $x^{(i)}$, $y^{(j)} \sim \text{Uniform}(0,1)$, $i = 1..n_x$, $j = 1..n_y$, $\boldsymbol{s} \sim \text{Multinomial}\left(1, \frac{1}{m}, ..., \frac{1}{m}\right)$, $\boldsymbol{v} = (v_1, ...v_m)$, $v_i = -1 + \frac{2(i-1)}{m-1}$, $k = \langle \boldsymbol{s}, \boldsymbol{v} \rangle$, $z \sim \text{Normal}\left(k \sum_{i=1}^{n_x} x^{(i)}, \sigma\right)$. Here $x^{(i)}$ are relevant covariates, $y^{(j)}$ are irrelevant covariates, $\boldsymbol{s}$ is an unobserved one-hot binary vector covariate which introduces multimodality in the conditional densities, $m \geq 1$ is the parameter controlling the number of peaks in the multimodal distribution. Angle brackets $\langle \cdot, \cdot \rangle$ in the equation for k denote the dot product.

The practically significant multimodal dataset containing examples for the PyFitIt software [11,12] was built by calculating XANES spectra for various geometry modifications of $[Fe(terpy)_2]^{2+}$ complex by FDMNES [5]. Then the spectra were smoothed to get the same shape as the experimental one. The argument $\boldsymbol{x}$ here is a XANES spectrum (dimension = 86), the target variable $\boldsymbol{z}$ is a 6-dimensional vector of geometry parameters. The dataset was constructed in such a way that the partial probability distribution $p(\boldsymbol{z})$ of the target $\boldsymbol{z}$ is uniform in a rectangle $[-0.3, 0.5]^6$. To check the results, we use the dataset Feterpy_combined with reduced number of geometry parameters: 3-dimensional $\boldsymbol{z}$. Feterpy dataset contains 729 spectra, Feterpy_combined — 500.

5 Experiments

During our experiments we determine the relative error of the method as

$$\frac{|L_{cur} - L_{best}|}{|L_{best}|}, \tag{4}$$

where L is defined in (1). All KDE estimators are run with a Gaussian kernel and a bandwidth h set to 0.4. The base classifier for the Binary Classification Approach is the LGBM classifier from the LightGBM framework that uses tree-based learning algorithms with *n_estimators = 200* and *learning_rate = 0.005.* The base one-dimensional estimator for the Naïve Bayes approach is set to KDE with the same parameters. The RFCDE model is employed with the following set of parameters: *n_trees = 1000, mtry = 4, n_basis = 15, node_size = 20.* The base estimator for Bayesian Network is the RFCDE model with the abovementioned parameters. We utilize the NNKCDE model and set its parameters to *k = 10, bandwidth = 0.2.* Any other parameters in the algorithms under consideration are set to default.

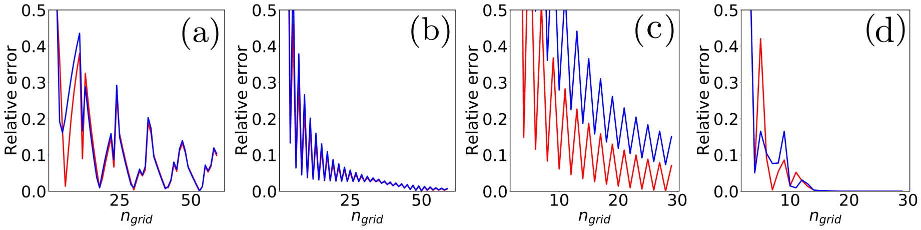

Fig. 2: Relative errors for straightforward (red) and approximate (blue) approaches for computing the MISE loss for
(a) SURD($n_x = 2$, $n_y = 1$, $n_z = 2$, $\boldsymbol{d} = [0.1, 1]$, $\sigma = 0$, $\boldsymbol{\alpha} = [0, 0]$, $\varphi = \frac{\pi}{4}$),
(b) SURD($n_x = 2$, $n_y = 1$, $n_z = 2$, $\boldsymbol{d} = [1, 1]$, $\sigma = 0$, $\boldsymbol{\alpha} = [0, 0]$, $\varphi = \frac{\pi}{4}$),
(c) SURD($n_x = 2$, $n_y = 1$, $n_z = 3$, $\boldsymbol{d} = [1, 1, 1]$, $\sigma = 0$, $\boldsymbol{\alpha} = [0, 0, 0]$, $\varphi = \frac{\pi}{4}$),
(d) Fork($n_x = 10$, $n_y = 1$, $m = 2$, $\sigma = 1$).

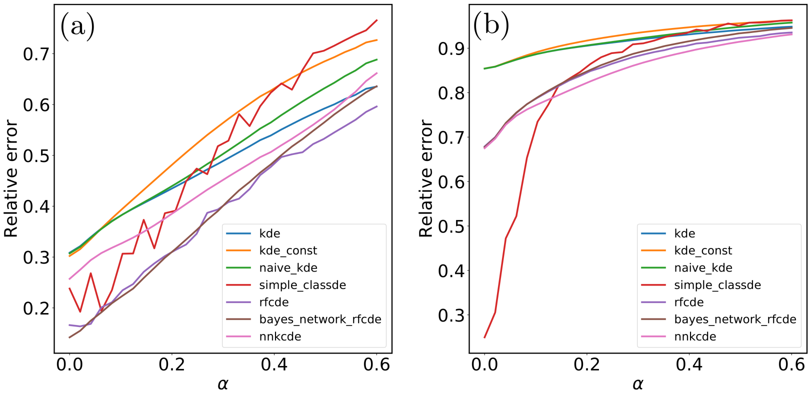

Fig. 3: Relative errors for CDE methods depending on $\boldsymbol{\alpha} = [\alpha, \alpha]$, $\alpha \in (0, 0.6)$ for (a) SURD($n_x = 10$, $n_y = 0$, $n_z = 2$, $\boldsymbol{d} = [1, 1]$, $\sigma = 0$, $\varphi = \frac{\pi}{4}$) and (b) same as (a) except $\boldsymbol{d} = [0.1, 1]$.

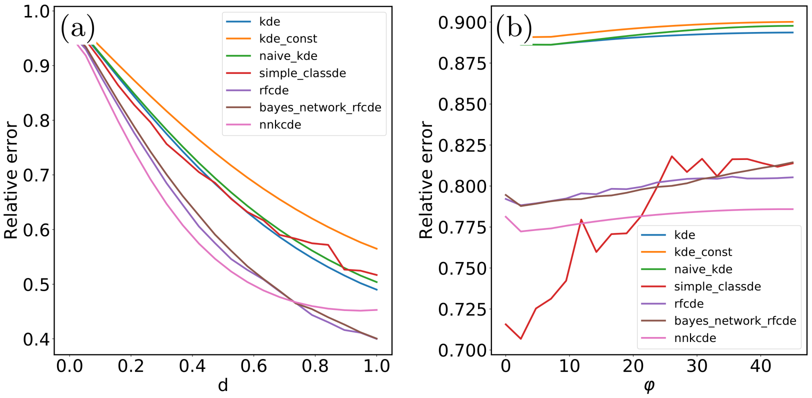

Fig. 4: Relative errors for CDE methods depending on (a) $\boldsymbol{d} = [d, 1]$, $d \in (0, 1)$ for SURD($n_x = 10$, $n_y = 0$, $n_z = 2$, $\boldsymbol{\alpha} = [0.3, 0.3]$, $\sigma = 0$, $\varphi = 0$) and (b) $\varphi \in \left(0, \frac{\pi}{4}\right)$ for SURD($n_x = 10$, $n_y = 0$, $n_z = 2$, $\boldsymbol{\alpha} = [0.1, 0.1]$, $\boldsymbol{d} = [0.1, 1]$, $\sigma = 0$).

We start our experiments by assessing the errors obtained by two approaches to calculating the MISE measure discussed in Section 3. The parameter under investigation is the number of grid points taken along each axis to compute the integrals in either straightforward (1) or approximate (2) manner. Figure 2 displays that the approximate approach needs a more frequent grid to achieve the competitive relative error rate in comparison with more computationally expensive straightforward numerical integration especially for $n_z = 3$. Another interesting observation is the fact that the data points flatness along one of the axes leads to the worse performance of both approaches which can be explained by the inability of both numerical integration and employed cosine basis to calculate integral for the flattened data. It can be also noted that increasing the dimensionality of the target variable n_z heavily influences the relative error rate for the thinner grid. In the univariate case, two approaches to estimating the MISE loss gained by the estimator appear to exhibit better results as it is shown in Figure 2d.

Henceforth, in our experiments, we employ the straightforward approach for evaluating the MISE loss as it requires fewer points of the grid to accomplish more true-to-life values obtained by the loss of the algorithm.

The next parameter to study is $\boldsymbol{\alpha}$ which regulates the dependence of $\boldsymbol{z}$ on the values of $\boldsymbol{x}$. For simplicity we set $\boldsymbol{\alpha} = [\alpha, \alpha]$. We evaluate all conditional density estimators discussed in Section 2 except FlexCode which implementation provided by the authors of the paper [7] does not support multidimensional datasets. According to Figure 3, all concerned CDE algorithms indicate the similar tendency: the stronger the dependence between $\boldsymbol{z}$ and $\boldsymbol{x}$, the more difficult is to provide the correct estimations for the fixed $\boldsymbol{x}$. Flatness of data along one

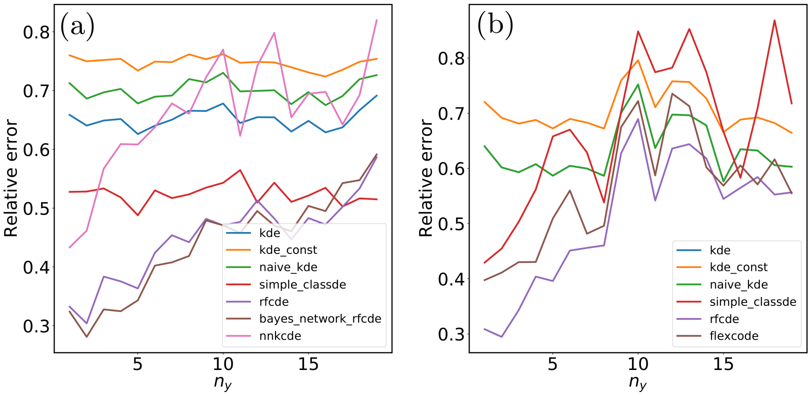

Fig. 5: Relative errors for CDE methods depending on $n_y \in [1, 20]$ for
(a) SURD($n_x = 1$, $n_z = 2$, $\boldsymbol{\alpha} = [2, 2]$, $\boldsymbol{d} = [1, 1]$, $\sigma = 0$, $\varphi = 0$)
(b) Fork($n_x = 1$, $m = 2$, $\sigma = 0.1$).

of the axes even stronger obstructs the correct conditional density estimation as can be seen when analyzing Figure 3a against Figure 3b.

Furthermore, we examine the dependency on data flatness along one of the axes separately as it proved to be critical for the quality of estimation in the previous experiments. Without loss of generality, we vary the first component of $\boldsymbol{d}$. Figure 4a confirms the idea expressed in the previous paragraphs: the closer the variances along axes to each other, the better the algorithms model the underlying distribution. Figure 4a indicates as well that the family of orthogonal series approaches shows the best performance among concerned methods though still suffering from an imbalanced variance of data along different axes.

Another interesting parameter to discuss is φ which determines the dependency between target variables $z^{(i)}$ by controlling the degree of the data points rotation Figure 4b. One can observe that the Binary Classification Approach approximation becomes noticeably inaccurate with amplification of the angle φ. The estimation appears to be imprecise as the volume of the bounded rectangles V_x and V_{xz} can be sufficiently larger than the area in which the true density values have fallen. This fact leads to the approximation error becoming larger and explains the deteriorated performance of this method with an increased φ.

Table 1: Values of CDE evaluation measure (1) for Feterpy datasets

	kde	kde_const	naive_kde	simple classde	rfcde	bayes_network rfcde	nnkcde
Feterpy	-0.6	-0.4	-0.6	3.9	-	-4.5	-7.7
Feterpy combined	-1.4	-0.8	-1.3	-1.8	-5.5	-4.2	-8.7

Finally, we evaluate the influence of the number of irrelevant covariates n_y on the performance of the CDE algorithms. Figure 5a indicates that while RFCDE, NNKCDE and Bayesian Network approach (employing RFCDE one-dimensional estimator internally) show the lowest relative error among all the methods, their performance deteriorates with growing n_y. A similar tendency can be observed for the family of orthogonal series approaches in the univariate case in Figure 5b.

Practically significant datasets have the drawback: the exact value of the CDE evaluation measure is not known. So, the relative error (4) can't be calculated and we have to be content with the value (1). We calculate the measure values for considered algorithms for the dataset with 6-dimensional z and for the dataset with combined geometry parameters (see Feterpy_combined in [11]), which has 3-dimensional z. The results are collected in Table 1. The NNKCDE algorithm outperforms the others both by quality and speed. RFCDE required more than 80 Gb of memory for Feterpy dataset and didn't finish calculation.

6 Conclusion

This research aims to investigate several approaches to conditional density estimation since point estimations or quantile regression do not suffice, for example, if a multimodal or skewed distribution is considered. Specifically, we perform the comparative study of several methods employing inherently different techniques: Kernel Density Estimators, Binary Classification Approach, Bayesian Networks, the family of orthogonal series approaches and one of the most recent models — NNKCDE which implements the ABC-rejection scheme. We provide an overview of these methods and varying the parameter of two synthetic datasets (for multivariate and univariate cases) and practically significant multimodal dataset, we demonstrate the strengths and the weaknesses of the algorithms under consideration. Not only concern our experiments the algorithms themselves, but they also tackle the problem of the quality of the loss computation performed with two approaches.

There are multiple directions in which this research may progress. During the experiments, we did not manage to examine datasets with a high dimensional target variable z. The problem is that the straightforward (1) way of calculating MISE measure is computationally exhaustive while the second (2) produces inaccurate estimation. So we have to either consider another measure or invent new algorithms of the MISE optimization suitable for a high-dimensional target.

References

1. Angrist, J.D., Pischke, J.S.: Mostly harmless econometrics: An empiricist's companion. Princeton university press (2008)
2. Bohm, G., Zech, G.: Introduction to statistics and data analysis for physicists, vol. 1. Desy Hamburg (2010)
3. Burnaev, E., Nazarov, I.: Conformalized kernel ridge regression. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). pp. 45–52. IEEE (2016). https://doi.org/10.1109/ICMLA.2016.0017

4. Guda, A.A., Guda, S.A., Lomachenko, K.A., et al.: Quantitative structural determination of active sites from in situ and operando xanes spectra: From standard ab initio simulations to chemometric and machine learning approaches. Catalysis Today (2018). https://doi.org/10.1016/j.cattod.2018.10.071
5. Guda, S.A., Guda, A.A., Soldatov, M.A., et al.: Optimized finite difference method for the full-potential xanes simulations: Application to molecular adsorption geometries in mofs and metal-ligand intersystem crossing transients. Journal of Chemical Theory and Computation **11**(9), 4512–4521 (Sep 2015). https://doi.org/10.1021/acs.jctc.5b00327
6. Izbicki, R., Lee, A.B., Pospisil, T.: Abc–cde: Toward approximate bayesian computation with complex high-dimensional data and limited simulations. Journal of Computational and Graphical Statistics pp. 1–20 (2019). https://doi.org/10.1080/10618600.2018.1546594
7. Izbicki, R., Lee, A.B., et al.: Converting high-dimensional regression to high-dimensional conditional density estimation. Electronic Journal of Statistics **11**(2), 2800–2831 (2017). https://doi.org/10.1214/17-EJS1302
8. Kemp, G.C., Silva, J.S.: Regression towards the mode. Journal of Econometrics **170**(1), 92–101 (2012). https://doi.org/10.1016/j.jeconom.2012.03.002
9. Kuleshov, A.P., Bernstein, A., Burnaev, E.: Conformal prediction in manifold learning. In: 7th Symposium on Conformal and Probabilistic Prediction and Applications, COPA 2018, 11-13 June 2018, Maastricht, The Netherlands. pp. 234–253 (2018)
10. Kuleshov, A.P., Bernstein, A., Burnaev, E.: Kernel regression on manifold valued data. In: 5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018. pp. 120–129 (2018). https://doi.org/10.1109/DSAA.2018.00022
11. Martini, A., Guda, S.A., Guda, A.A., et al.: Pyfitit: the software for quantitative analysis of xanes spectra using machine-learning algorithms. Mendeley Data (2019). https://doi.org/http://dx.doi.org/10.17632/dwrb56xrx6.1
12. Martini, A., Guda, S.A., Guda, A.A., et al.: Pyfitit: the software for quantitative analysis of xanes spectra using machine-learning algorithms. Computer Physics Communications (to appear)
13. Pospisil, T., Lee, A.B.: Rfcde: Random forests for conditional density estimation. arXiv preprint arXiv:1804.05753 (2018)
14. Rau, M.M., Seitz, S., Brimioulle, F., et al.: Accurate photometric redshift probability density estimation–method comparison and application. Monthly Notices of the Royal Astronomical Society **452**(4), 3710–3725 (2015). https://doi.org/10.1093/mnras/stv1567
15. Yao, W., Li, L.: A new regression model: modal linear regression. Scandinavian Journal of Statistics **41**(3), 656–671 (2014). https://doi.org/10.1111/sjos.12054