Information and Inference: A Journal of the IMA (2023) **00**, 1–29
https://doi.org/DOI HERE

Adaptive Kernel Conditional Density Estimation

WENJUN ZHAO*
Division of Applied Mathematics, Brown University, 182 George Street, 02906, Rhode Island, USA

AND

ESTEBAN G. TABAK
Courant Intstitute of Mathematical Sciences, New York University, 251 Mercer Street, New York, 10012, New York, USA
*Corresponding author: wenjun_zhao@brown.edu

[Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year]

A methodology is proposed for the determination of factor-dependent bandwidths for the kernel-based estimation of the conditional density $\rho(x|z)$ underlying a set of observations. The adaptive determination of the bandwidths is based on a z-dependent effective number of samples and variance. The procedure extends to categorical factors, where a nontrivial "bandwidth" can be designed that optimally uses across-class information while capturing class-specific traits. A hierarchy of algorithms is developed, and their effectiveness is demonstrated on synthetic and real-world data.

Keywords: kernel density estimation; conditional density estimation; bandwidth selection.

1. Introduction

1.1. *Kernel conditional density estimation*

A central problem in data science is to estimate the probability distribution underlying a set of independent observations $\{x_i\}$ of a random variable x. For concreteness, we will assume that x is a d-dimensional real vector, with a probability distribution representable through a smooth density $\rho(x)$.

A widely used procedure for estimating $\rho(x)$ is kernel density estimation,

$$\rho^{est}(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x,x_i), \tag{1}$$

where the kernel function K_h is a non-negative, real-valued function that integrates to one over its first argument, and $h = \{h_l\}$ is a set of scaling parameters, the *bandwidths*, typically one per dimension. An example is the multivariate Gaussian kernel:

$$K^G_{\{h_l\}}(x_1,x_2) = \prod_{l=1}^{d} \frac{1}{(2\pi)^{\frac{1}{2}} h_l} \exp\left(-\frac{\left(x_1^l - x_2^l\right)^2}{2{h_l}^2}\right). \tag{2}$$

Kernel density estimation generalizes the notion of a histogram: each sample adds to the local value of the estimated density not within the histogram's cell where it falls but within a range scaling with the bandwidths h. These bandwidths, the only free parameters of the procedure, should scale with the spread of the samples $\{x_i\}$ and correlate negatively with their number N, as a larger number of samples allows one to tune the estimated distribution more finely. A common choice for h is given by the rule-of-thumb

$$h_l = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \frac{\sigma_l}{N^{\frac{1}{d+4}}}, \tag{3}$$

where σ_l is the empirical standard deviation of x^l. This choice can be shown to be optimal in the integrated square error sense when both the kernels and the underlying density are Gaussian [16].

The next level of complexity has the variable x depend on other variables z, the *factors* or *covariates*. The corresponding object to estimate from sample pairs $\{x_i, z_i\}$ is the conditional density $\rho(x|z)$. One can reduce this problem to the previous one by writing

$$\rho(x|z) = \frac{\pi(x,z)}{\mu(z)}, \tag{4}$$

where π and μ are the densities of the joint distribution of x and z and the marginal distribution of z respectively. Estimating each through Kernel density estimation automatically produces an estimate for $\rho(x|z)$. Yet this has at least two problems: on the one hand, it requires the estimation of a density π in the extended, higher-dimensional space (x,z). On the other, if $\mu^{est}(z)$ does not agree exactly with the marginal of $\pi^{est}(x,z)$, then $\rho^{est}(x|z)$ will not integrate to one over x for each value of z. The latter problem disappears if we select two kernel functions K^x and K^z for the individual variables x and z respectively, and build a joint kernel K^{xz} for (x,z) through

$$K^{xz}\left[(x,z),(x_*,z_*)\right] = K^x(x,x_*)\,K^z(z,z_*).$$

The corresponding conditional density estimation is given by the Nadaraya-Waston's formula [4, 5]

$$\rho^{est}(x|z) = \sum_{i=1}^{N} w(z,z_i) K^x_{h_x}(x,x_i), \quad w(z,z_i) = \frac{K^z_{h_z}(z,z_i)}{\sum_{j=1}^{N} K^z_{h_z}(z,z_j)}, \tag{5}$$

depending on the two sets of bandwidths h_x and h_z. Because the $w(z,z_i)$ add up to one over i and each K^x integrates to one over x, the $\rho^{est}(x|z)$ from (5) are valid probability densities for all values of z.

We will find it convenient –and conceptually more sound– to think of (5) not as deriving from (4), but as an extension of (1). In this conceptualization, given z, one performs regular density estimation for x, but weights the various samples x_i by their relevance, measured by $w(z,z_i)$, which quantifies the closeness between z_i and the value of z of current interest. Thus

we think of $\rho(x|z)$ as a family of distributions in x indexed by the factors z, a more natural characterization than (4), which takes into consideration the distribution of z, immaterial to the conditional distribution. We emphasize this conceptual distinction because it allows us to adopt a different set of bandwidths $\{h_x, h_z\}$ for each value of z, as each corresponds to a different estimation problem. This leads to a much more robust conditional density estimation than the regular one that uses (5) with fixed bandwidths.

Determining good values for the bandwidths is crucial to a good kernel conditional density estimation. This article proposes a methodology whereby values of the bandwidths h_x and h_z, depending on the target value z^* of z for which the estimation is sought, are determined jointly through the maximization of the leave-one-out likelihood of the data.

1.2. *Prior work*

Kernel density estimation goes back to the work of Rosenblatt and Parzen [1, 2]. An early comparative study of kernel-based density estimation methods can be found in [3]. The Nadaraya-Watson formula for kernel conditional regression was proposed in [4, 5], and used for conditional density estimation in [6]. Subsequent work on bandwidth selection in [7] used as main criterion the minimization of the integrated square error, and combined various approaches: reference rules based on assumptions on the form of the distribution to estimate, a bootstrap methodology based on a parametric model, and a regression-based approach. The integrated square error is also used in [9] for bandwidth selection based on leave-one-out cross-validation, while the work in [10] uses likelihood cross-validation –which is also our tool of choice–, with performance accelerated through dual trees. By contrast with this article's proposal, the bandwidths estimated in all these approaches are uniform in x and z.

Alternative, quite different approaches to conditional density estimation include the work in [11], based on Gaussian processes, the work in [12], based on least-squares importance fitting, the work in [13], which develops LinCDE, based on gradient boosting and Lindsey's method, and the conditional density simulation through the distributional barycenter problem in [14] and [15].

1.3. *Contribution*

The main contribution of this article is a methodology for determining variable bandwidths $h_z(z^*)$, $h_x(z^*)$, where z^* is the target value of the covariate z for which the conditional density estimation $\rho(x|z)$ is sought through Nadaraya-Waston's formula (5). In addition, the article extends the use of Kernel-based weights to discrete, categorical covariates z, for which no notion of distance between points exist, except for the distinction between equal and different values of z.

The article is structured as follows. Section 2 describes some measures of the accuracy of an estimate, with an emphasis on the relative entropy, which has natural data-based counterparts that we will use as objective functions for the determination of the bandwidths. Section 3 develops the methodology proposed for this determination, in various stages. First

some options for the determination of uniform bandwidths are reviewed, particularly the joint determination of h_x and h_z through the maximization of the leave-one-out log-likelihood of the data. Then a methodology is proposed whereby z-dependent bandwidths h_x are slaved to a global choice of h_z, which determines the effective number of samples $\{x_i\}$ available for each value of z, as well as the corresponding local variance. The bandwidth h_z and a set of parameters α quantifying deviation from Gaussianity are determined through the maximization of a leave-one-out log-likelihood function, where weights can be used to tailor it to the estimation of $\rho(x|z)$ for the target value $z = z^*$. Finally, the procedure is extended to categorical factors z, where how much to weight the x_i with $z_i \neq z^*$ –balancing the gains from a larger data-pool with the dilution of z-specific characteristics– is determined, much as h_z in the continuous case, through the maximization of the leave-one-out log-likelihood. The article concludes with some final remarks and connections to other lines of work.

2. Measuring the effectiveness of an estimate

A robust estimate of the discrepancy between the true distribution $\rho(x|z)$ and its estimate $\rho^{est}(x|z)$ serves two main purposes: to compare the performance of different estimation methodologies and as a criterion for the choice of a procedure's parameters, particularly its bandwidths.

In cases where the true density $\rho(x|z)$ is known –such as when testing the procedure on synthetic data– many alternative measures of the discrepancy between $\rho(x|z)$ and $\rho^{est}(x|z)$ are available. Frequently used choices are the *Hellinger distance*

$$H^2(\rho^{est},\rho)[z] = \frac{1}{2}\int \left(\sqrt{\rho^{est}(x|z)} - \sqrt{\rho(x|z)}\right)^2 dx = 1 - \int \sqrt{\rho^{est}(x|z)\rho(x|z)}\, dx,$$

ranging between 0 and 1, the *Integrated mean square error*

$$ISE\left(\rho^{est},\rho\right)[z] = \int \left(\rho^{est}(x|z) - \rho(x|z)\right)^2 dx$$

and the *relative entropy* or Kullback-Leibler divergence

$$KL\left(\rho \| \rho^{est}\right)[z] = \int \log\left(\frac{\rho(x|z)}{\rho^{est}(x|z)}\right)\rho(x|z)\, dx,$$

a non-negative quantity that only vanishes when $\rho^{est}(x|z) = \rho(x|z)$ almost everywhere.

These quantities measure the discrepancy between $\rho^{est}(x|z)$ and $\rho(x|z)$ for each value of z; global measures of discrepancy use their expected value over z. The relative entropy admits a

natural data-based version:

$$\int KL\left(\rho\|\rho^{est}\right)\mu(z)\,dz = \int \log\left(\frac{\rho(x|z)}{\rho^{est}(x|z)}\right)\rho(x|z)\,\mu(z)\,dx\,dz \approx E(\rho) - \frac{1}{N}\sum_i \log\rho^{est}(x_i|z_i),$$

where

$$E(\rho) = \int \log\left(\rho(x|z)\right)\rho(x|z)\,\mu(z)\,dx\,dz$$

is the expected value over z of the entropy of $\rho(x|z)$. Even though $E(\rho)$ is not known when we have only samples $x_i \sim \rho$ available, it is a number that does not depend on the estimation $\rho^{est}(x)$. Then the value of the *log-likelihood*

$$L(\rho) = \frac{1}{N}\sum_i \log\rho^{est}(x_i|z_i) \tag{6}$$

can be used to compare the in-sample accuracy of different estimates ρ.

Yet (6) is not a good quantifier of performance when the same set of data $\{x_i, z_i\}$ is used both to estimate $\rho(x|z)$ and to assess ρ^{est} through (6). For instance, a set of bandwidths approaching zero yield an unbounded value for L, as ρ^{est} reduces to a sum of delta-functions over the $\{x_i\}$, a severe case of over-fitting. To remedy this, we can divide the available samples into a training set used to estimate ρ and a testing set where to asses ρ^{est}. Alternatively, in order to maximize the use of the samples available for training, one can resort to a leave-one-out procedure, replacing (6) by

$$L^1(\rho) = \frac{1}{N}\sum_i \log\rho^{est}_{-i}(x_i|z_i), \tag{7}$$

where each $\rho^{est}_{-i}(x|z)$ is estimated using all sample pairs except (x_i, z_i) itself.

3. Bandwidths for conditional density estimation

We propose a methodology for adaptive kernel conditional density estimation, using Nadaraya-Waston's formula (5), with bandwidths determined automatically from the data, which depend on the value of z for which $\rho^{est}(x|z)$ is sought. Rather than detailing from the outset this article's ultimate proposal, this section proceeds through a hierarchy of approaches to determining the bandwidths h_z and h_x.

3.1. *Fixed bandwidths*

A straightforward application of Nadaraya-Waston's formula (5) uses uniform values for the sets of bandwidths h_z and h_x. Some natural options for their choice are:

1. **Bandwidths chosen independently.** Since we have two kernel functions, one in x and one in z-space, a naive approach would choose them independently, solely based on their

individual distributions, as conveyed by their samples $\{x_i\}$ and $\{z_i\}$. The simplest recipe would use the rule of thumb (3) for each. A more sophisticated choice would pick each as the maximizer of the corresponding leave-one-out log-likelihood. For the bandwidths h_x, we would solve the problem

$$\max_{h_x} L_x^1 = \frac{1}{N}\sum_i \log\left(\rho_{-1}^{est}(x_i,h_x)\right), \quad \rho_{-1}^{est}(x_i,h_x) = \frac{1}{N-1}\sum_{j\neq i} K_h(x,x_j) \tag{8}$$

and similarly for h_z. This choice has two main advantages: it works for any kernel function and it does not depend on the assumption that the underlying distributions are Gaussian. Nonetheless, it still suffers from the fundamental flaw of disregarding the relation between the variables x and z.

2. **Jointly chosen bandwidths.** There are a number of reasons why the choice of the two sets of bandwidths h_x and h_z must be linked. Key among them is that the choice of h_z determines, for each value of z, an effective number of samples that contribute to the estimation of $\rho(x|z)$ –quantified in the next subsection– and the optimal bandwidths h_x depend on the statistics of this number. Thus larger values of h_z would yield a larger effective number of samples and hence smaller values for the optimal h_x, converging to the one for regular density estimation as $h_z \to \infty$. A natural way to choose the two sets of bandwidths is through the joint maximization of the leave-one-out likelihood

$$h_x, h_z = \arg\max_{a,b} L = \frac{1}{N}\sum_i \log \rho_{-i}^{est}(x_i|z_i,a,b),$$

where

$$\rho_{-i}^{est}(x|z,a,b) = \sum_{j\neq i} w_b^i(z,z_j) K_a^x(x,x_j), \quad w_b^i(z,z_j) = \frac{K_b^z(z,z_j)}{\sum_{l\neq i} K_b^z(z,z_l)}. \tag{9}$$

For other alternatives considered in the literature, see subsection 1.2.

3.2. *Variable bandwidths: $h_x(x_i)$ vs. $h_x(z)$*

The optimal bandwidths depend on the number of samples: as more become available, one can afford the finer resolution provided by smaller bandwidths. Yet this raises the question of whether a better estimate could be obtained by letting h_x be non-uniform, so as to provide a finer resolution in areas where the density of samples is larger. Consider first the possibility that h_x be a function of the corresponding kernel's center x_i. To simplify matters, we can analyze this case in the context of regular density estimation, with no factor z. The corresponding kernel density estimation would adopt the form

$$\rho^{est}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{h_i}(x,x_i),$$

with the h_i selected so as to be smaller in areas of larger density $\rho(x)$. Picking each h_i independently could lead to severe over-fitting, but maybe a procedure could be devised whereby h_i becomes a smooth function of x_i, possibly specified by a small number of parameters?

Rather than proposing a methodology for selecting h_i, let us explore the plausibility of such proposal with a simple example. Consider the one-dimensional normal distribution

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

and draw from it $N = 1000$ samples $\{x_i\}$. The optimal uniform bandwidth, according to the rule of thumb, is given by

$$\bar{h} = \left(\frac{4}{3}\right)^{\frac{1}{5}} \frac{1}{N^{\frac{1}{5}}} = 0.2661.$$

Figure 1 shows (in blue) the resulting density estimation, a very good fit, since the rule of thumb was developed precisely for the Gaussian case. In order to experiment with making h finer where ρ is larger, propose instead

$$h_i = \frac{a}{\rho(x_i) + \varepsilon}, \quad a = \bar{h} \frac{N}{\sum_i \frac{1}{\rho(x_i)+\varepsilon}}, \quad \varepsilon = \frac{1}{N},$$

which makes the $\{h_i\}$ inversely proportional to the actual density while preserving their mean value $\bar{h}$. The resulting estimation, displayed in figure 1, over-estimates the density in places where it is already large, and under-estimates it in places where it is small.

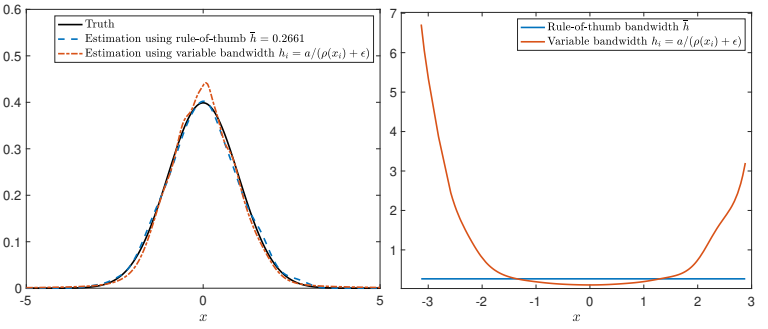


FIG. 1. Estimation of a standard normal distribution using variable bandwidths. Left: estimated and true density. Right: bandwidth as a function of x.

The reason for this unwanted effect is not hard to elucidate. By making the bandwidth smaller in areas of high density, we have made the corresponding kernels more concentrated,

so each of them adds more mass to the density estimation for points nearby, and less mass to the density estimation for points far away. Reciprocally, the larger bandwidth in areas of low density makes each individual kernel expand, hence contributing less to the estimation of the local density and more to that of points far away. Both of these effects conspire to over-estimate hight densities and under-estimate low ones, as observed in the example.

We conclude from this argument that proposing a variable h is not a good idea for regular kernel density estimation. We point this fact out because, as we discuss next, the same principle does not hold for kernel conditional density estimation. It is still true that the bandwidths should not depend on x, but they can –and should– depend on z for a more accurate estimate

A conditional distribution $\rho(x|z)$ is a family of regular distributions in x parameterized by z. We can conceptualize the Nadaraya-Waston's formula (5) along these lines, as a regular kernel density estimation where the sample points x_i are weighted by their relevance for the value of z under consideration, a relevance measured by their corresponding weights

$$w_i(z) = \frac{K^z_{h_z}(z,z_i)}{\sum_{j=1}^N K^z_{h_z}(z,z_j)}.$$

Under this conceptualization, there is no reason why the bandwidths used for the density estimation for different values of z should be the same. The density $\rho(x|z)$ may have larger variance for some values of z than for others and some areas in z-space may be better represented than others, giving rise to a larger effective number of samples for the estimation of the corresponding $\rho(x|z)$. Thus it is natural to propose a set of bandwidths that depends on z.

One could naively think that, given a value of z for which $\rho(x|z)$ is sought, we could simply determine an optimal pair (h_z,h_x) along the same lines as for the regular density estimation. A problem immediately appears though: if z is a continuous variable, for any given value of z we typically have none or at most one sample of $\rho(x|z)$. We cannot perform an optimization over bandwidths based on just one sample –and even less when we base it on a leave-one-out procedure! Thus we cannot determine (h_z,h_x) just for one z: we need instead the whole set of $\{h_z(z_i),h_x(z_i)\}$, so that all $\{\rho^{est}_{-i}(x_i|z_i)\}$ are available for likelihood maximization.

We develop a hierarchy or proposals for determining variable bandwidths. The first proposal maximizes the leave-one-out likelihood over global bandwidths $\{h_z\}$,

$$h_z = \arg\max_b L = \frac{1}{N}\sum_i \log \rho^{est}_{-i}\left(x_i|z_i,b\right), \tag{10}$$

where

$$\rho^{est}_{-i}\left(x_i|z_i,b\right) = \sum_{j\neq i} w^i_b(z,z_j) K^x_{h_x(z,b)}(x,x_j), \quad w^i_b(z,z_j) = \frac{K^z_b(z,z_j)}{\sum_{l\neq i} K^z_b(z,z_l)}, \tag{11}$$

with the variable bandwidths $h_x(z,b)$ *slaved* to the choice of $h_z = b$ through an extension of the rule of thumb, as specified below:

1. We first determine the *effective number of samples*

$$N_e(z,h_z) = \sum_i \frac{K^z_{h_z}(z,z_i)}{K^z_{h_z}(z_i,z_i)} \tag{12}$$

for each value of z under a global choice of h_z. In regular density estimation, all samples x_i have equal weight in the determination of $\rho(x)$. This is not the case in conditional density estimation $\rho(x|z)$, where sample points x_i with corresponding factors z_i near z have more significance than those with z_i far away. The value of the kernel $K^z_{h_z}(z,z_i)$ provides a measure of that significance. Notice that this value is normalized differently for the calculation of $N_e(z,h_z)$ than for the determination of the weights $w_{h_z}(z,z_j)$. For the latter, the normalization was such that the sum of the weights added up to one. Such normalization would not make sense for the former, as it is precisely this sum that we are after, the effective number of samples that contribute to the estimation of $\rho(x|z)$. We need instead a normalization factor that gives sample x_i a weight equal to one when $z = z_i$, hence the $K^z_{h_z}(z_i,z_i)$ in the denominator of (12).
2. We determine next the empirical conditional mean and standard deviation of each component l of x under $\rho(x|z)$,

$$\hat{\mu}_l(z,h_z) = \sum_j \frac{K_{h_z}(z,z_j)}{\sum_l K_{h_z}(z,z_l)} x^l_j, \quad \hat{\sigma}_l(z,h_z)^2 = \sum_j \frac{K_{h_z}(z,z_j)}{\sum_k K_{h_z}(z,z_k)} \left(x^l_j - \hat{\mu}_l(z,h_z)\right)^2. \tag{13}$$

3. Now we have all the ingredients required to apply a local rule of thumb:

$$h_{x,l}(z,h_z) = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \frac{\sigma_l(z,h_z)}{N_e(z,h_z)^{\frac{1}{d+4}}}, \tag{14}$$

which takes into account the effective variance and number of samples corresponding to the value of z under consideration.

When applying this procedure to estimate ρ^{est}_{-i} in (11), the steps above use all sample pairs except (x_i,z_i).

We used the rule of thumb because we cannot estimate each $h_x(z_i)$ independently. Yet one may want to avoid using a rule grounded on a Gaussian assumption, which may be unrealistic for the data at hand. For any distribution $\rho(x)$ of fixed shape, the optimal bandwidths for kernel density estimation adopt the form

$$h_l = F_l(N)\sigma_l.$$

The linear dependence of h_l on σ_l follows simply from a scaling argument: if we change the units in which we measure x^l, both the standard deviation σ_l and the bandwidth h_l must change accordingly.

For conditional kernel density estimation, ρ, N and σ depend on z and h_z, so we must have

$$h_{x,l}(z) = F_l(z, N_e(z,h_z))\sigma_l(z,h_z),$$

which leaves us with the task of estimating the functions F_l. The simplest approach is the one proposed above, where F is completely determined by the rule of thumb. At the other end of the spectrum, we could perform a full non parametric estimation of F, based on the maximization of the likelihood. In between, practical approaches can establish a compromise between accuracy and complexity. The simplest among these, which we adopted for the experiments in section 4, uses

$$h_{x,l}(z,h_z) = \alpha_l \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \frac{\sigma_l(z,h_z)}{N(z,h_z)^{\frac{1}{d+4}}}, \tag{15}$$

a modified rule of thumb where the d additional parameters $\{\alpha_l\}$ account for global deviations from Gaussianity.

A further refinement of the procedure just described tailors the objective function itself to the value $z = z^*$ for which $\rho^{est}(x|z)$ is sought, replacing (10) by

$$h_z(z^*) = \arg\max_b L^*(z^*) = \sum_i \gamma(z^*, z_i) \log \rho^{est}_{-i}(x_i|z_i, b), \tag{16}$$

where $\gamma(z^*, z_i)$ is a weight that quantifies our prior belief on the degree of similarity between $\rho(x|z^*)$ and $\rho(x|z_i)$. Under this refinement, both h_x and h_z depend on the target value z^* of z. For each target z^*, the (weighted) likelihood in (16) is maximized over h_z and the $\{\alpha_l\}$ in (15), yielding an h_x that is a function of z_i. Thus the factor z appears twice in (16): as the target z^*, of which h_z and the $\{\alpha_l\}$ depend, and as the z_i of which h_x depends through the rule of thumb for each value of z^*, so we have three functions: $h_z(z^*)$, $\alpha(z^*)$ and $h_x(z^*, z)$. Even though the final estimation uses only $h_x(z^*, z^*)$, all of the $h_x(z^*, z_i)$ are used for optimizing $h_z(z^*)$ and $\alpha(z^*)$.

The optimization problems in (16) and (10) have the same computational cost when $\rho^{est}(x|z)$ is sought for a single value z^* of z. When more values are required, all the way to one for each sample point z_i, as required for the computation of the in-sample log-likelihood, the cost of (16) grows accordingly. Using (16) is worth the cost though, since tailoring both h_x and h_z to the target sought increases the accuracy of the estimation. Notice that the changes in $h_x(z)$ derive from two sources: the change in the optimal h_z, which affects $N_e(z)$ and $\sigma_x(z)$, and the optimization of the parameters $\{\alpha_l\}$, which are also tailored to z through the weights $\gamma(z, z_i)$.

A natural choice for the weighting function $\gamma(z, z_i)$ in (16) is a kernel function in z space, for which an additional bandwidth $\hat{h}_z$ must be selected. In the examples below, we simply hand-picked a uniform and relatively large bandwidth $\hat{h}_z$, rather than optimizing it through cross-validation. Notice that the extreme choice $\hat{h}_z \to \infty$ reduces (16) to (10).

3.3. *Categorical factors*

We have considered so far factors z that are continuous variables, or have at least a notion of distance between pairs of points, so that a kernel $K_z(z_i, z_j)$ can be built. Yet one often encounters discrete factors $z \in (z^1, \ldots, z^k)$ with only k possible values, which are moreover categorical, so that no natural notion of distance between them exists. The conventional wisdom in such cases is to think of the $\rho(x|z^l)$ as k distinct distributions $\rho_l(x)$, to be estimated independently. This corresponds to using a kernel

$$K(z_i, z_j) = \mathbf{1}_{z_i = z_j}$$

which assigns zero distance to members of the same class and infinite distance otherwise.

This is a sensible choice when the number of samples available within each class is enough for a robust estimation. When this is not the case, however, treating each class separately fails to use any possible commonality among classes, whose estimation may benefit from the larger total pool of available observations.

A natural extension addressing this shortcoming assigns a weight a to the samples within the class whose estimation is sought, and $1 - a$ to the others:

$$K(z_i, z_j) = a\mathbf{1}_{z_i = z_j} + (1 - a)\mathbf{1}_{z_i \neq z_j},$$

where $a \in [\frac{1}{2}, 1]$. The reason for the lower bound at $a = \frac{1}{2}$ is that the kernel should achieve its maximum value at $z_i = z_j$: no value of z is more relevant to the estimation of $\rho(x|z)$ than z itself. In particular, the choice $a = \frac{1}{2}$ is optimal when the distribution $\rho(x|z)$ does not depend on z at all, while the conventional procedure of estimating the density in each class independently corresponds to the upper bound $a = 1$. Optimizing the leave-one-out likelihood L^1 over a determines automatically how much the benefits of using across-class information compensate for the corresponding dilution of class-specific traits.

4. Numerical examples

This section compares the performance of the various procedures for conditional density estimation discussed before, using both synthetic and real data.

4.1. *Synthetic examples*

Testing first the various procedures on simulated datasets, for which we can compare the exact conditional densities with their estimation, we repeat each experiment 20 times, each with newly drawn samples from the corresponding distributions $\mu(z)$ and $\rho(x|z)$. Then we compute the average over those 20 realizations of the following metrics:

- Mean squared Hellinger distance:

$$\bar{H}^2 = \frac{1}{N_z^g} \sum_{z_i^g} H^2(\rho^{true}(\cdot|z_i^g), \rho^{est}(\cdot|z_i^g)),$$

where z^g is a grid in z-space with N_z^g points. When z is discrete, the grid is just the list of all of its possible values; when z is continuous, we adopt a uniform grid with 100 equidistant points.

The Hellinger distance for each value of z is estimated via numerical integration using the trapezoidal rule. For one-dimensional variables x, we adopt a grid of 4000 points in $[\min x - 6\sigma_x, \max x + 6\sigma_x]$ and, for two-dimensional x, a grid with 400×400 points.

- In-sample mean log-likelihood:

$$\frac{1}{N} \sum_i \log \rho^{est}(x_i|z_i).$$

- Mean leave-one-out log-likelihood:

$$L^1 = \frac{1}{N} \sum_i \log \rho_{-i}^{est}(x_i|z_i),$$

where the estimation of the conditional density at each sample pair does not use that individual sample. This is also the objective function that we optimize over bandwidths in some of the procedures.

- Mean out-of-sample log-likelihood,

$$L^o = \frac{1}{N_{test}} \sum_{i \in I_{test}} \log \rho^{est}(x_i|z_i),$$

evaluated on additional testing data. While the in-sample points are used for bandwidth selection and constructing estimator, the mean out-of-sample likelihood quantifies how well the estimator can be generalized to points that are not included in the density estimation.

We use these metrics to compare the following six approaches to kernel conditional density estimation:

1. Rule-of-thumb for both x and z:

$$h_{z,l} = \left(\frac{4}{N(D_z+2)}\right)^{\frac{1}{D_z+4}} \sigma_{z,l}, \quad l = 1,...,D_z, \quad h_{x,d} = \left(\frac{4}{N(D_x+2)}\right)^{\frac{1}{D_x+4}} \sigma_{x,d}, \quad d = 1,...,D_x.$$

Here D_z and D_x are the dimensions of z and x respectively. In all synthetic examples, $D_z = 1$, and either $D_x = 1$ or $D_x = 2$.

2. Rule-of-thumb for both x and z, but with z-dependent parameters for x:

$$h_{z,l} = \left(\frac{4}{N(D_z+2)}\right)^{\frac{1}{D_z+4}} \sigma_{z,l}, \quad h_{x,d}(z,h_z) = \left(\frac{4}{N_e(z,h_z)(D_x+2)}\right)^{\frac{1}{D_x+4}} \sigma_{x,d}(z,h_z).$$

   The difference with the prior approach is that N and σ are replaced by their effective values given z, so the bandwidths h_x depend on z.
3. Optimization of the leave-one-out likelihood L^1 over h_z, with $h_x(z)$ slaved to h_z as above.
4. Joint optimization of L^1 over uniform values of both h_z and h_x, a total of $D_z + D_x$ parameters, one bandwidth for each dimension of z and x.
5. Similar to 3., but with L^1 optimized over both h_z and the parameters $\{\alpha_l\}$ in (15). For computational efficiency, in all numerical examples, $\alpha_l = \alpha$ is fixed as constant for all dimensions of x. When $\alpha = 1$, approach 5 reduces to 3.
6. Same as 5., but optimizing $L^*(z)$ from (16) instead of L^1. When applied to estimate all $\rho^{est}(x_i|z_i)$, this procedure is far more expensive, as it requires solving an optimization problem for each $h_z(z_i)$. Yet it is not so costly in real practice, where it is applied only to those values of z where the estimation is sought.

We have selected a small set of examples suited to assess the functionality of the various algorithms, with x either one or two-dimensional, and a single factor z, either real or categorical. In all cases, 500 pairs (x_i, z_i) are generated and used for the algorithms, and 250 additional pairs are generated from the same joint distribution, to be used for the evaluation of the out-of-sample log-likelihood. The simulations chosen are the following:

1. Gaussian distribution with uniform mean and z-dependent variance:

$$z \sim U(0,1), \quad x \sim N(0, \left(0.5\cos(4\pi z + \frac{\pi}{4}) + 0.55\right)^2).$$

2. Non-Gaussian unimodal distribution:

$$z = Beta(2,3), \quad y \sim N(-\sin(2\pi z), 0.2^2), \quad x = y^3.$$

3. Bimodal distribution:

$$z \sim U(0,1), \quad x \sim 0.3N(0,(0.2z+0.05)^2) + 0.7N(2z+0.5,(0.2z+0.05)^2).$$

4. Categorical factor z:

$$z \sim \frac{1}{10}\delta_0 + \frac{3}{10}\delta_{1/4} + \frac{2}{10}\delta_{1/2} + \frac{3}{10}\delta_{3/4} + \frac{1}{10}\delta_1,$$

$$x \sim \frac{1}{8}N\left(\frac{5}{6}(z+2)(1-z), (\frac{1}{5}(z+2))^2\right) + \frac{1}{8}N\left(-\frac{5}{6}(z+2)(1-z), (\frac{1}{5}(z+2))^2\right) + \frac{3}{4}N\left(0, (\frac{1}{3}(z+2))^2\right).$$

5. A two-dimensional Gaussian:

$$z \sim N(1/2,(1/6)^2), \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} -5\sin(2\pi z) \\ 5(1-z)^2 \end{bmatrix}, (0.6\cos(2\pi z + \frac{\pi}{4}) + 0.8)^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

6. A two-dimensional, highly non-Gaussian, unimodal distribution:

$$z \sim Beta(2,3), \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} (-\sin(2\pi z) + 0.2\varepsilon_1)^3 \\ (4(z-1/2) + 0.2\varepsilon_2)^3 \end{bmatrix}, \quad \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

7. Two-dimensional, multimodal distribution with x_1 and x_2 conditionally independent given $z \sim U(0,1)$:

$$x_1 \sim \frac{1}{2} N(4z(1-z)\cos(2\pi z + \frac{\pi}{4}), (\frac{z}{2} + \frac{1}{32})^2) + \frac{1}{2} N(-4z(1-z)\cos(2\pi z + \frac{\pi}{4}), (\frac{1-z}{2} + \frac{1}{32})^2),$$

$$x_2 \sim \frac{1}{2} N(4z(1-z)\sin(2\pi z + \frac{\pi}{4}), (\frac{1-z}{2} + \frac{1}{32})^2) + \frac{1}{2} N(-4z(1-z)\sin(2\pi z + \frac{\pi}{4}), (\frac{1-z}{2} + \frac{1}{32})^2).$$

The results of applying the alternative procedures to these simulations are summarized in the tables and plots below. The tables list the mean value and standard deviation (in brackets) of each measure of accuracy over the 20 realizations of each experiment. The procedures with best mean performance are singled out in blue font, and all equivalent approaches under a two-sided, paired t-test at significance level 5% in light blue. In the figures, the plots of the various estimated densities (for one random realization) are distinguished by their color, with the true distribution displayed in solid black.

For the Gaussian of the first experiment (table 1 and figure 2), the third approach is consistently the best. The strong dependence of the variance of x on z necessitates a z-dependent bandwidth, and the fact that all $\rho(x|z)$ are Gaussian justifies the use of the rule of thumb. The quite comparable results of the fifth and sixth approaches –though only better when measured in-sample– yield values of the parameter α very close to one, as beholds a Gaussian distribution.

The situation changes in the second experiment, where the distributions are far from Gaussian. The results in table 2 and figure 3 show the fifth and sixth approaches as clear winners, with values of α around 0.6, far from the default $\alpha = 1$ of the third approach, which this time has the third best performance. From the results of the two cases so far, one would choose the fifth approach, with variable bandwidth and adjustment factor for departures from Gaussianity, as it performs almost as well as the third approach even when the underlying distribution is indeed Gaussian, and is less computationally expensive than the sixth.

The fourth, fifth and sixth approaches perform best in the third experiment, with categorical factors, followed closely by the third approach, see table 3 and figure 4. These four outperform the others because they can play with the parameter a, our surrogate for the bandwidth h_z for categorical factors. With too few samples available for each value of z to assess each

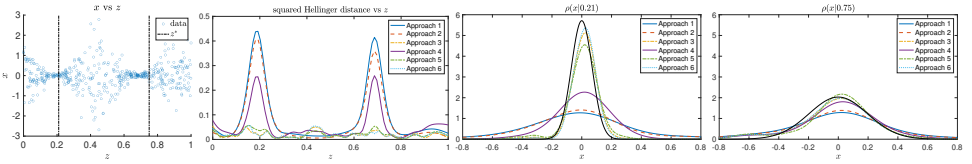
Gaussian 1D						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.09(0.00)	0.08(0.00)	0.02(0.01)	0.06(0.01)	0.03(0.03)	0.04(0.02)
Mean log-likelihood (higher is better)						
In-sample	-0.67(0.06)	-0.66(0.06)	-0.39(0.09)	-0.49(0.07)	-0.37(0.10)	-0.34(0.07)
In-sample,leave-one-out	-0.76(0.05)	-0.70(0.05)	-0.53(0.06)	-0.68(0.06)	-0.52(0.06)	-0.50(0.06)
Out-of-sample	-0.76(0.10)	-0.71(0.07)	-0.55(0.09)	-0.70(0.15)	-0.61(0.16)	-0.68(0.23)

TABLE 1 *Performance when the distribution of x is Gaussian with z-dependent variance (Experiment 1), with the best values indicated in blue. The rule-of-thumb is almost designed for this example, and the approaches with variable bandwidth in x outperform the others. In approach 5, the mean of α is 1.10 with standard deviation 0.23, consistent with the optimality of the rule-of-thumb for conditional Gaussians.*

Unimodal and skewed 1D						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.22(0.01)	0.20(0.01)	0.14(0.01)	0.14(0.01)	0.11(0.01)	0.12(0.02)
Mean log-likelihood (higher is better)						
In-sample	-0.37(0.06)	-0.20(0.08)	0.13(0.11)	0.05(0.11)	0.25(0.10)	0.31(0.10)
In-sample, leave-one-out	-0.41(0.07)	-0.24(0.08)	-0.01(0.10)	-0.19(0.09)	0.04(0.09)	0.11(0.10)
Out-of-sample	-0.40(0.04)	-0.24(0.05)	-0.07(0.11)	-0.16(0.14)	-0.01(0.14)	-0.02(0.29)

TABLE 2 *Experiment 2. When all conditional densities are skewed –though still unimodal– , while approach 3 performs better than approach 4 with optimal constant bandwidths, approach 5 is consistently the best, with the parameter α correcting the rule-of-thumb. The average value of α is 0.56 with standard deviation 0.14.*

conditional distribution accurately on its own, the algorithm has found it worth to weight in information from the distributions corresponding to other values of z, to leverage the features that they share. In addition, approaches 3, 5 and 6 adapt the bandwidth h_x to the different



(a) Data (b) Squared Hellinger distance (c) Estimation at $z = 0.21$ (d) Estimation at $z = 0.75$

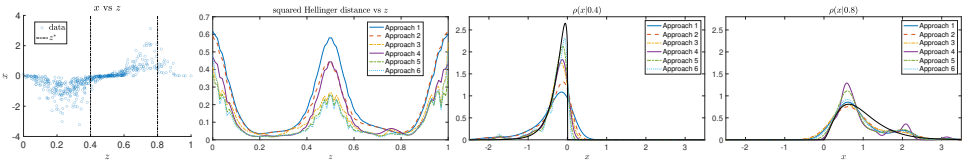
FIG. 2. Experiment 1. When the conditional distribution of x is a Gaussian, approaches 3, 5 and 6 give a consistently good performance across a broad range of z-dependent variances.

standard deviation and sample size of each conditional distribution, while the fourth approach gains from its independence from any Gaussian assumption.

Categorical z						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.02(0.00)	0.05(0.00)	0.02(0.01)	0.01(0.00)	0.01(0.00)	0.01(0.00)
Mean log-likelihood (higher is better)						
In-sample	-2.08(0.02)	-2.15(0.02)	-2.10(0.02)	-2.00(0.03)	-2.00(0.03)	-2.00(0.03)
In-sample,leave-one-out	-2.11(0.02)	-2.18(0.02)	-2.11(0.02)	-2.05(0.02)	-2.05(0.02)	-2.05(0.02)
Out-of-sample	-2.11(0.04)	-2.18(0.03)	-2.10(0.03)	-2.05(0.04)	-2.05(0.04)	-2.05(0.04)

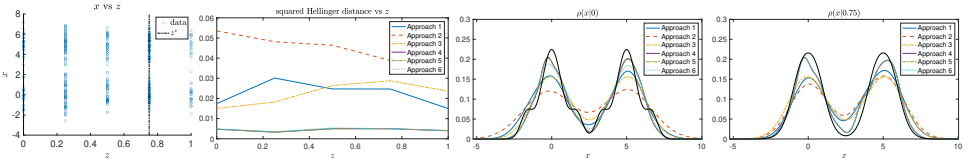
TABLE 3 *Experiment 3, categorical factor and bimodal distribution. With a categorical factor z, approach 2 would be the conventional benchmark: to treat each value of z separately, performing five density estimates. Instead, approaches 3, 4, 5 and 6 outperform the others by allowing samples corresponding to different values of z to inform each other. The average value of a for approach 3 is 0.65 with standard deviation 0.15, and the average of α for approach 5 is 0.33 with standard deviation 0.07.*

The advantage of the fifth and sixth approaches becomes far more pronounced in example 4, non-Gaussian and bimodal, displayed in table 4 and figure 5. They outperform all others



(a) Data (b) Squared Hellinger distance (c) Estimation at $z = 0.4$ (d) Estimation at $z = 0.8$

FIG. 3. Experiment 2. With skewed distributions, the fifth/sixth approaches perform best under all measures of accuracy, followed by the third.



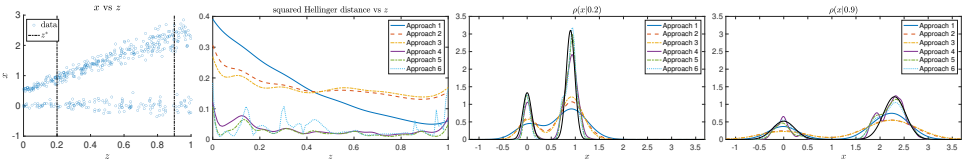
(a) Data (b) Squared Hellinger distance (c) Estimation at $z = 0$ (d) Estimation at $z = 0.75$

FIG. 4. Experiment 3. Approaches 3, 4 and 5 provide better and smother estimates for different values of the categorical factor z.

quite significantly, though the fourth approach performs much better than the third, as the capability to address non-Gaussianity –in this case though the optimization of a global bandwidth h_x– becomes more significant than incorporating a variable bandwidth.

Bimodal 1D						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.02(0.00)	0.05(0.00)	0.02(0.00)	0.01(0.00)	0.00(0.00)	0.01(0.00)
Mean log-likelihood (higher is better)						
In-sample	-2.08(0.02)	-2.15(0.02)	-2.10(0.02)	-2.00(0.03)	-2.00(0.03)	-2.00(0.03)
In-sample,leave-one-out	-2.11(0.02)	-2.18(0.02)	-2.11(0.02)	-2.05(0.02)	-2.05(0.02)	-2.05(0.02)
Out-of-sample	-2.11(0.04)	-2.18(0.03)	-2.10(0.03)	-2.05(0.04)	-2.05(0.04)	-2.05(0.04)

TABLE 4 *Example 4. A bimodal case provides one clear instance where the rule-of-thumb fails. Nonetheless, the results show that, enriching the rule-of-thumb with the extra* α *parameter in approaches 5 and 6, the resulting variable bandwidths still outperform a freer optimization of a fixed bandwidth. For reference, the average value of* α *in approach 5 was* 0.24*, with a standard deviation of* 0.02*.*



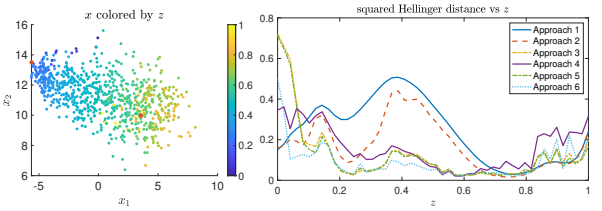
(a) Data (b) Squared Hellinger distance (c) Estimation at $z = 0.2$ (d) Estimation at $z = 0.9$

FIG. 5. Example 4. The fifth and sixth approaches, followed by the fourth, clearly outperform all others when applied to this bimodal, z-dependent distribution.

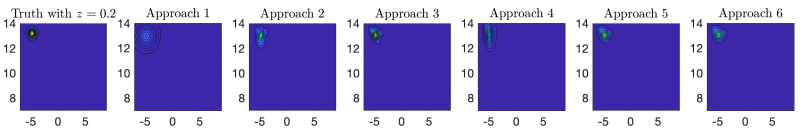
The situation does not change much when x becomes two-dimensional. The results for the fifth example (Table 5, figure 6), an isotropic Gaussian with z-dependent mean and variance, again show the third approach performing best, followed closely by the fifth, with a value of the parameter α not far from 1. The fourth approach, with fixed but optimal bandwidths, is not far behind, the reason being that the variance of x depends only mildly on z. For the sixth example, quite skewed but unimodal (Table 6, figure 7), the sixth approach works predictably best, followed by the fifth, third and fourth. The advantage of the fifth and sixth approaches over the third gets again magnified in the multimodal scenario of example 7 (Table 7, figure 8), where α differs significantly from one.

Gaussian 2D						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.31(0.00)	0.24(0.00)	0.05(0.00)	0.10(0.00)	0.11(0.05)	0.08(0.02)
Mean log-likelihood (higher is better)						
In-sample	-2.97(0.03)	-2.60(0.05)	-1.69(0.05)	-1.91(0.07)	-1.65(0.11)	-1.71(0.06)
In-sample,leave-one-out	-3.00(0.04)	-2.77(0.04)	-2.17(0.06)	-2.35(0.06)	-2.14(0.05)	-2.11(0.05)
Out-of-sample	-3.01(0.04)	-2.78(0.06)	-2.23(0.09)	-2.37(0.08)	-2.48(0.24)	-2.39(0.17)

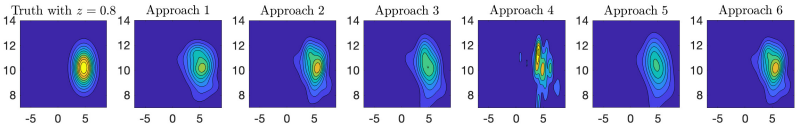
TABLE 5 *Performance of the various approaches on example 5, a two-dimensional isotropic Gaussian with z-dependent mean and variance. Here the third approach, with variable bandwidth and no adjustment for non-Gaussianity, performs best, followed by the sixth, fourth, and fifth, with parameter* α *with average value* 1.52 *and standard deviation* 0.21.



(a) Data (b) Squared Hellinger distance



(c) Truth and estimation at $z = 0.2$

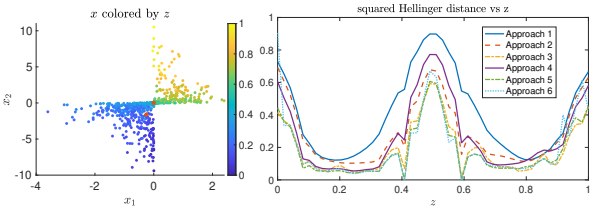


(d) Truth and estimation at $z = 0.8$

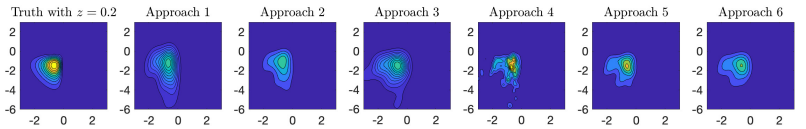
FIG. 6. Fifth example, a two-dimensional, z-dependent Gaussian. The third, fifth, and sixth approaches clearly outperform the others, with the fourth approach –fixed but optimal bandwidths– a not very distant fourth.

Unimodal and skewed 2D						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.40(0.01)	0.28(0.01)	0.20(0.01)	0.29(0.01)	0.19(0.01)	0.20(0.01)
Mean log-likelihood (higher is better)						
In-sample	-1.40(0.07)	-0.38(0.10)	0.15(0.10)	-0.37(0.15)	0.35(0.10)	0.41(0.11)
In-sample, leave-one-out	-1.49(0.08)	-0.51(0.10)	-0.14(0.09)	-0.93(0.15)	-0.10(0.09)	-0.02(0.10)
Out-of-sample	-1.48(0.06)	-0.49(0.14)	-0.16(0.17)	-0.89(0.16)	-0.11(0.18)	-0.08(0.22)

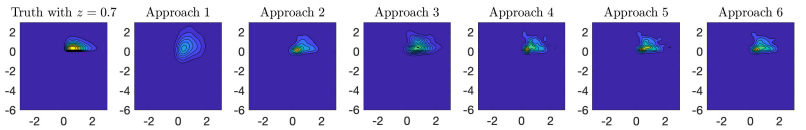
TABLE 6 *Performance on a skewed two-dimensional distribution (sixth example). The sixth approach works best, followed by the fifth, with parameter* α *away from* 1 *(average value* 0.72 *and standard deviation* 0.06*.)*



(a) Data (b) Squared Hellinger distance



(c) Truth and estimation at $z = 0.2$



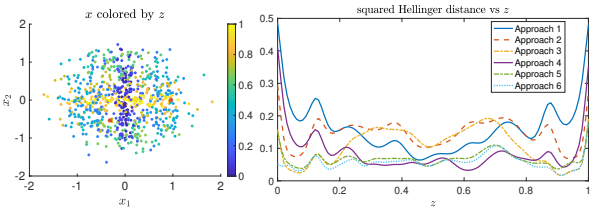
(d) Truth and estimation at $z = 0.7$

FIG. 7. Example 6 (skewed but unimodal, two-dimensional, z-dependent distributions.)

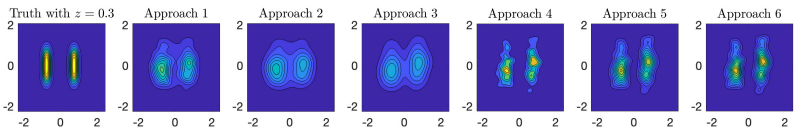
Multimodal 2D						
Approaches	1	2	3	4	5	6
Average squared Hellinger distance (lower is better)						
Averaged H^2	0.16(0.00)	0.15(0.00)	0.11(0.00)	0.09(0.00)	0.07(0.00)	0.06(0.00)
Mean log-likelihood (higher is better)						
In-sample	-1.12(0.03)	-1.13(0.03)	-0.95(0.03)	-0.62(0.06)	-0.60(0.06)	-0.56(0.06)
In-sample,leave-one-out	-1.24(0.03)	-1.21(0.03)	-1.11(0.03)	-1.04(0.04)	-1.01(0.04)	-0.95(0.04)
Out-of-sample	-1.25(0.05)	-1.22(0.05)	-1.13(0.06)	-1.08(0.06)	-1.04(0.07)	-1.00(0.07)

TABLE 7 *Example 7: multimodal two-dimensional distribution. The average value of α is 0.60, with a standard deviation of 0.02.*

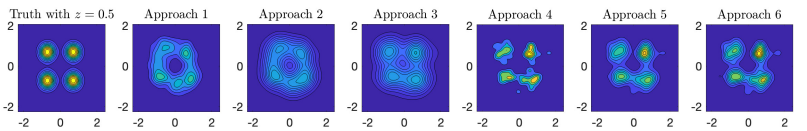
We conclude from all synthetic experiments that it is consistently best to adjust for non-Gaussianity (which the 4th approach does through the global optimization of h_x and the 5th



(a) Data (b) Squared Hellinger distance



(c) Truth and estimation at $z = 0.3$



(d) Truth and estimation at $z = 0.5$

FIG. 8. Example 7, a multimodal distribution where the Gaussian approximation clearly fails.

and 6th approaches through the parameter α) and for a z-dependent bandwidth (3rd, 5th and 6th approaches.) The 5th and 6th approaches, which combine both adjustments, perform the most robustly across cases, with the additional tailoring of the objective function to the target values of z in the 6th approach making it the most accurate.

5. Real-world examples

Next we test approaches 1 to 6 on real data from three quite different fields. Since the true distribution underlying the data is not available, our toolbox for comparing performance reduces to the log-likelihood in its three versions, in increasing order of reliability: in sample, leave-one-out and out-of-sample.

5.1. *Old faithful geyser*

A classical example for conditional density estimation is the Old Faithful geyser dataset [17]. The data, a total of 299 observations, consists of the waiting time between the start of successive eruptions and the duration of the subsequent eruption for the Old Faithful geyser in Yellowstone. We seek to estimate the distribution of the duration x of the subsequent eruption given the waiting time z. The scatter diagram of the data, displayed in figure 9, exhibits a switch from unimodal to bimodal at around $z = 70$ minutes.

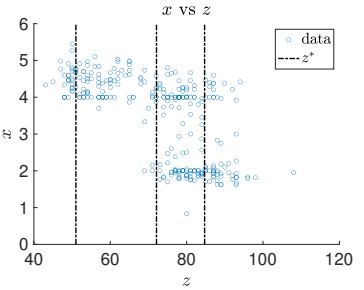


FIG. 9. Duration (in minutes) of subsequent eruptions x vs. waiting time z. The conditional density will be estimated at the values of z marked by the three dashed lines.

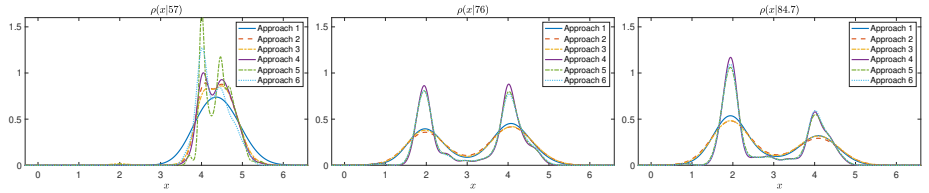
Figure 10 and table 8 summarize the results from the six procedures applied to this dataset. The three log-likelihood results have procedures 4, 5 and 6 as the best performers. This can also be seen in the plots of the estimated density for the three chosen values of z: only procedures 4, 5 and 6 divide the two cases with $z > 70$ into two clearly separated classes, and bound the support of ρ within its true range for $z = 57$.

5.2. *Human height and weight dataset*

We use partial census data for the Dobe area Kung San compiled from interviews conducted by Nancy Howell in the late 1960s (https://tspace.library.utoronto.ca/handle/1807/10395). This data set records the height (cm), weight (kg), age (years) and sex of 544 individuals. While there is no obvious difference between male and female height and weight at young ages, the conditional distribution becomes bimodal when the age increases beyond 25. A visualization of data is shown in figure 11. The separation between sexes on the upper right of figure 11 is made more apparent in the visualization in figure 12, which separates the data by sex and only displays individuals older than 25.

Figure 13 and table 9 show the results of the six procedures applied to the conditional estimation of $\rho(\text{height, weight}|\text{age})$, with plots for ages 10 and 50. The sixth approach has the best global performance in terms of likelihood, and it resolves both the unimodal distribution for age 10 and the bimodal one for age 50.

To verify whether the two peaks for age 50 are related to sex, we can alternatively estimate $\rho(\text{height, weight}|\text{age, male})$ and $\rho(\text{height, weight}|\text{age, female})$ separately. The results, displayed in figure 14, show how the last three approaches –particularly the sixth– resolve unimodal distributions for each sex, that combine naturally into a two-modal distribution for the total population.



(a) (b) (c)

FIG. 10. Conditional density estimation for the Old Faithful data, for three different values of z. The first three approaches, with no maximizing likelihood criteria on h_x, perform comparatively poorly, yielding too broad a support for $\rho(x|z=57)$ and two modes with significant overlap for $z=76$ and $z=84.7$. Adopting a variable $h_x(z)$, on the other hand, yields results comparable to those with with constant bandwidths. The value of α in approach 5 is 0.3547.

Approaches	1	2	3	4	5	6
In-sample	-0.84(0.02)	-0.83(0.02)	-0.83(0.02)	-0.49(0.08)	-0.51(0.05)	-0.50(0.05)
In-sample, leave-one-out	-0.87(0.02)	-0.87(0.02)	-0.86(0.02)	-0.66(0.05)	-0.68(0.03)	-0.68(0.04)
Out-of-sample	-0.88(0.06)	-0.90(0.06)	-0.89(0.06)	-0.84(0.67)	-0.76(0.26)	-0.76(0.27)

TABLE 8 *Log-likelihood values for the Old Faithful data, with 250 in-sample and 49 out-of-sample points. Approaches 4, 5 and 6 have comparable performances, superior to those of the first three.*

5.3. *Precipitation/Relative humidity and temperature*

With data from the US climate reference network (USCRN,https://www.ncei.noaa.gov/pub/data/uscrn/products/daily01/), daily averaged ground temperature ($^{\circ}C$), daily averaged

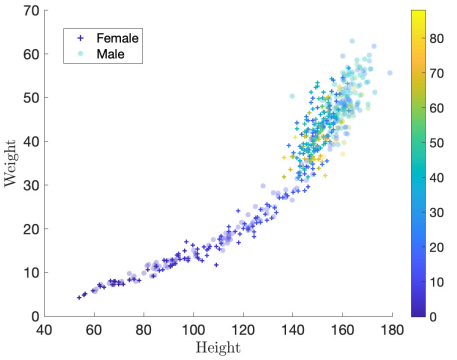


FIG. 11. Human height and weight data, with age indicated by color and gender by marker. Between ages 0-10+, there is no gap between sexes and the height increases with age. After that, height and weight change less sharply with age, and there is a separation between sexes (upper right).

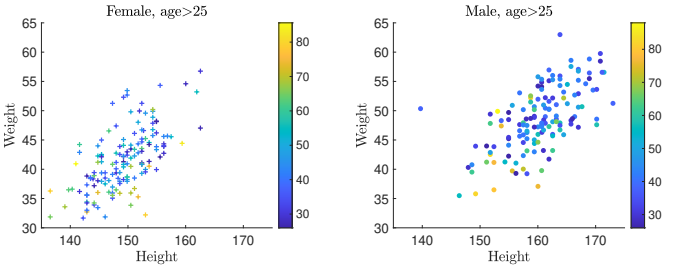
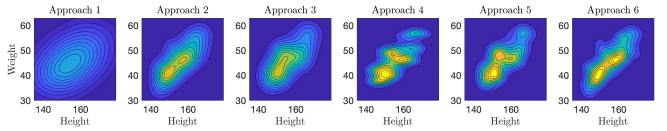


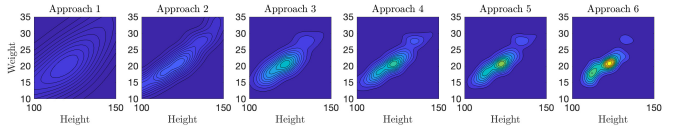
FIG. 12. Human height and weight data for individuals above 25. The distributions for both females and males are roughly football-shaped, Gaussian-like, with a mean difference of around 15cm in height and 10kg in weight.

Approach	1	2	3	4	5	6
In-sample	-6.96(0.03)	-6.35(0.03)	-5.93(0.05)	-5.70(0.04)	-5.71(0.04)	-5.65(0.07)
In-sample, leave-one-out	-6.99(0.03)	-6.46(0.03)	-6.20(0.03)	-6.23(0.03)	-6.15(0.03)	-6.04(0.04)
Out-of-sample	-6.98(0.03)	-6.47(0.04)	-6.25(0.06)	-6.26(0.07)	-6.20(0.08)	-6.10(0.07)

TABLE 9 *Height and weight data. Average log-likelihood values over 20 random ways of dividing the time series into 400 points for in-sample and 144 points for out-of-sample. Approach 6 has best likelihood for all cases, with no approaches with equivalent performance under a paired t-test.*

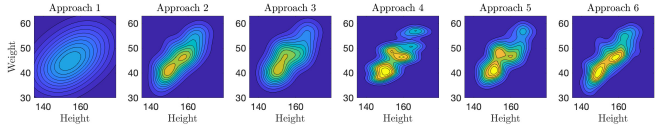


(a) Age 50, all gender

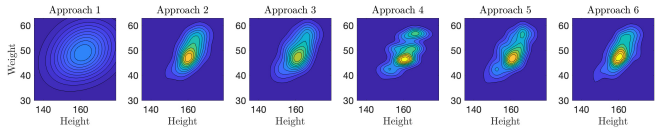


(b) Age 10, all gender

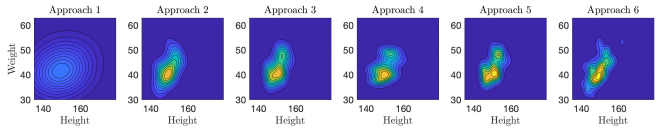
FIG. 13. Conditional densities estimated from the 6 procedures for age 50 and 10 given only age information. For age 50, approaches 4, 5, and 6 all capture two modes (with 5 and 6 giving smoother estimations), and for age 10, only one dominant peak arises in all estimations.



(a) Age 50, all gender



(b) Age 50, male



(c) Age 50, female

FIG. 14. Conditional densities estimated from the 6 procedures for age 50. The combined results for all sexes is well-represented as a mixture of (b) and (c), particularly for the last three approaches.

precipitation (mm), and daily averaged relative humidity (%) are jointly observed from 2018-01-01 to 2022-12-18 in Des Moines, IA, a total of 1731 observations, after excluding days with missing data.

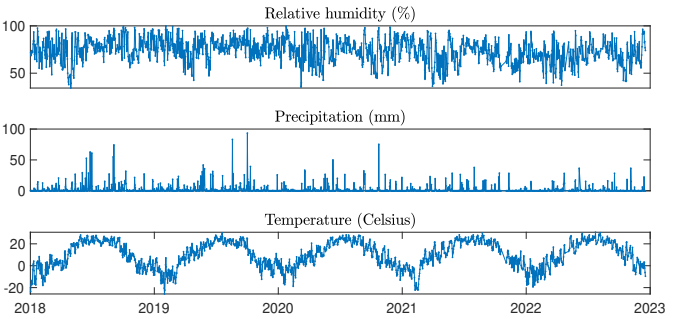


FIG. 15. Raw data with daily average temperature, relative humidity and total precipitation.

To investigate the relation between moisture and temperature, we quantify moisture through a nonlinear combination of relative humidity (RH) and precipitation (Precip):

$$x_1 = \frac{RH - \min(RH)}{\max(RH) - \min(RH)} + \sqrt{\frac{Precip - \min(Precip)}{\max(Precip) - \min(Precip)}}.$$

The dataset is split into a training and a testing set, with 1500 points used in-sample by our algorithm for bandwidth selection, and 231 points kept for the out-of-sample evaluation of the log-likelihood. We use a 3-dimensional z: a static factor: the day of year, representing the seasonal cycle, and two dynamical ones: the values of x_1 and x_2, as defined above, measured 24 hours before, to estimate the conditional density of the current x_1 and x_2. We enforce periodicity by replacing z_1 with the pair of periodic factors $\cos\left(2\pi\frac{z_1}{365.25}\right)$ and $\sin\left(2\pi\frac{z_1}{365.25}\right)$. An example of data after split is shown in figure 16.

The estimated conditional densities $\rho(x|z)$ for three dates are displayed in figure 17. The dates were selected so that two are in in summer, one humid and one dry, and the other in the winter. The conditional distribution of the latter shows a physical negative correlation between temperature and moisture: since the total amount of moisture available from the day before is captured by the dynamical factor z_3, an increase in temperature needs to be accompanied by a decrease in relative humidity. The likelihood values over 10 runs for the six procedures are shown in table 10, with the best values obtained for approaches 5 and 6, which are statistically equivalent in terms of out-of-sample log-likelihood.

6. Conclusions

This article developed a methodology for the determination of factor-dependent bandwidths for kernel conditional density estimation. Since a conditional distribution $\rho(x|z)$ is a family of x-distributions indexed by z, the parameters used for their estimation can –and should, for

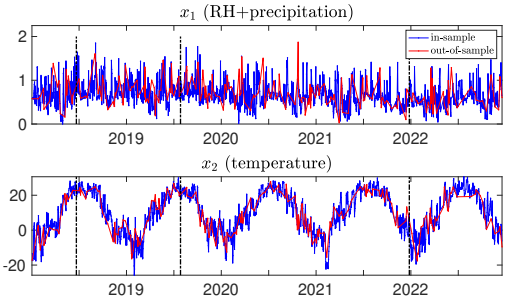
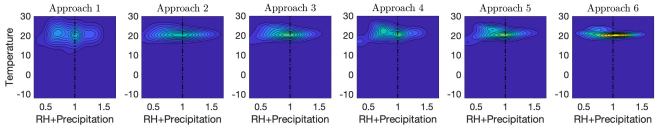
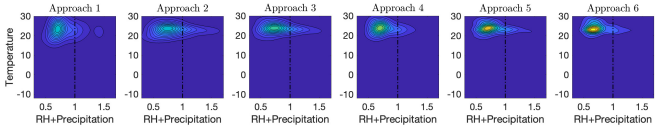


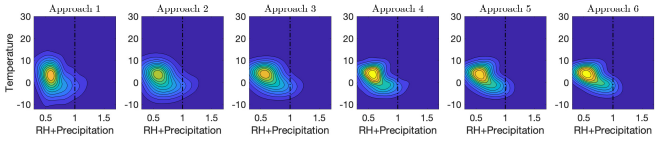
FIG. 16. An example of data after a random split, whereby the processed time series is randomly divided into in-sample and out-of-sample points.



(a) 2018-06-20



(b) 2019-07-28



(c) 2021-12-25

FIG. 17. Conditional density estimation for three given dates. The first two dates are in summer, with a high mean temperature component, one with nonzero precipitation and the other dry. The last date is in winter, where the variability in temperature is greater than that for summer, and the negative correlation between relative humidity and temperature becomes apparent.

accuracy and robustness– depend on the value of z for which the estimation is sought. For kernel conditional density estimation, these parameters are the sets of bandwidths $\{h_x(z)\}$ and $\{h_z(z)\}$. Yet the determination of the bandwidths cannot be made independently for each value of z, as the set of observations available for each z is typically very small, consisting of only one or no sample when z is a continuous variable. Even when z is discrete and categorical, so that a larger set of observations is potentially available for each value of z, it is wasteful not to exploit the possible commonality between the $\rho(x|z)$ for different z's.

In regular kernel density estimation, the bandwidths $\{h_x\}$ depend on the underlying distribution $\rho(x)$ and on the number of observations N available. It follows that, in conditional kernel density estimation, $\{h_x\}(z)$ must depend on the distribution $\rho(x|z)$ and the effective number of samples $N_e(z)$. The latter depends on the bandwidths $\{h_z(z)\}$, converging to N as the bandwidths grow to ∞ and to the number of z's nearest neighbors (typically one) as the bandwidths shrink to zero. A simple explicit formula determines $N_e(z)$ for each choice of $\{h_z\}$. Regarding the dependence of $\{h_x\}$ on $\rho(x|z)$, it is convenient to reduce it to simple functions of parameters of the distribution $\rho(x|z)$ that can be computed explicitly. The parameter most widely used in kernel density estimation is the standard deviation σ_l along each component l of x, whose conditional version depends on the choice of bandwidths $\{h_z\}$, much as N_e, and whose empirical counterpart can be computed explicitly. The bandwidths proposed in this article have the form $h_{x,l} = \alpha_l b_l$, where b_l is the bandwidth computed from the rule-of-thumb with parameters $N_e(z)$ and $\sigma_l(z)$, and α_l is a scaling parameter determined by leave-one-out validation, together with the $\{h_z(z)\}$, to account for deviations from the Gaussian hypothesis underlying the rule of thumb.

The bandwidths $\{h_z\}$ are global parameters, to which the $\{h_x(z)\}$ are slaved. One can refine the determination of the $\{h_z\}$, making them depend on the value $z = z_*$ for which the estimation is sought. This is achieved by modifying the leave-one-out log-likelihood that one maximizes over $\{h_z\}$, weighting the samples with $z = z_i$ according to their distance to z_*. The procedure that uses this refinement (the "sixth approach" in section 4) is systematically the most accurate, though also the most expensive when estimates of $\rho(x|z)$ for various values of z are required.

The numerical tests in section 4 show that the algorithms proposed can significantly improve the toolbox available for kernel conditional density estimation. Some of the

Approach	1	2	3	4	5	6
In-sample	-2.27(0.01)	-2.23(0.01)	-2.15(0.02)	-2.04(0.02)	-2.03(0.02)	-1.98(0.07)
In-sample, leave-one-out	-2.72(0.01)	-2.64(0.01)	-2.58(0.01)	-2.61(0.01)	-2.57(0.01)	-2.55(0.10)
Out-of-sample	-2.72(0.05)	-2.66(0.06)	-2.62(0.05)	-2.62(0.06)	-2.60(0.06)	-2.59(0.06)

TABLE 10 *Average log-likelihood values in-sample (1500 points) and out-of-sample (231 points) over 20 runs. In the periodic kernel example, the sixth approach has a handpicked combination of bandwidths for the weights* γ*:* $[4/3, 0.6, 10]$.

underlying ideas are also potentially useful for other tasks, such as the quantification of the dependence among sets of variables. One example is the simulation of conditional distributions through the data-driven optimal transport barycenter problem [15], which uses kernel conditional estimators to enforce the independence of the barycenter of the $\rho(x|z)$ from the factors z.

Data Availability Statement

The old faithful geyser dataset is derived from [17]. The human height and weight dataset is compiled from https://tspace.library.utoronto.ca/handle/1807/10395. The precipitation/relative humidity and temperature dataset can be accessed via US climate reference network (USCRN,https://www.ncei.noaa.gov/pub/data/uscrn/products/daily01/).

Funding

This work of E. G. Tabak is supported in part by the Office of Naval Research, through grants # N00014-19-1-2407 and # N00014-22-1-2192.

References

1. Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27, 832–837, 1956.
2. Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3),1065–1076, 1962.
3. Adrian W. Bowman. A comparative study of some kernel-based nonparametric density Estimators, *Journal of Statistical Computation and Simulation*, 21:3-4, 313-327, 1985.
4. Elizbar Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9, 141–142, 1964.
5. Geoffrey S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, Series A, 26, 359-372, 1964.
6. Murray Rosenblatt. Conditional probability density and regression estimates, *Multivariate Analysis II*, 1969
7. David M. Bashtannyk and Rob J. Hyndman. Bandwidth selection for kernel conditional density estimation, *Computational Statistics & Data Analysis*, 36, 2798, 2001.
8. Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2017.
9. Jianqing Fan and Tsz Ho Yim, A cross validation method for estimating conditional densities, *Biometrika*, 91, 8194, 2004.
10. Michael P. Holmes, Alexander G. Gray, and Charles Lee Isbell. Fast nonparametric conditional density estimation, *UAI*, 2007.
11. Vincent Dutordoir, Hugh Salimbeni, James Hensman, and Marc Deisenroth. Gaussian process conditional density estimation, *NeurIPS*, 2018.

12. Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional Density Estimation via Least-Squares Density Ratio Estimation, *PMLR*, 9:781-788, 2010.
13. Zijun Gao and Trevor Hastie. LinCDE: conditional density estimation via Lindsey method, *Journal of Machine Learning Research*, 23, 1–55, 2022.
14. Esteban G. Tabak, Giulio Trigila, Wenjun Zhao Conditional density estimation and simulation through optimal transport, *Machine Learning*, 2020.
15. Esteban G. Tabak, Giulio Trigila, Wenjun Zhao Distributional barycenter problem through data-driven flows, *Pattern Recognition*, 2022.
16. Bernard W. Silverman. *Density estimation for Statistics and Data Analysis*, London: Chapman & Hall/CRC, 1986
17. Adelchi Azzalini and Adrian W Bowman. A look at some data on the old faithful geyser, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1990