A Nonparametric Kernel Approach to Interval-Valued Data Analysis

YONGHO JEON
Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea

JEONGYOUN AHN, CHEOLWOO PARK
Department of Statistics, University of Georgia, Athens, GA 30602, USA

September 8, 2014

Abstract

This paper concerns data sets in which variables are in the form of intervals, which are obtained by aggregating information about variables from a larger data set. We propose to view the observed set of hyper-rectangles as an empirical histogram, and to use a Gaussian kernel type estimator to approximate its underlying distribution in a nonparametric way. We apply this idea to both univariate density estimation and regression problems. Unlike many existing methods used in regression analysis, the proposed method can estimate the conditional distribution of the response variable for any given set of predictors even when some of them are not interval-valued. Empirical studies show that the proposed approach has a great flexibility in various scenarios with complex relationships between the location and width of intervals of the response and predictor variables.

Key words: Conditional distribution; Interval prediction; Nonparametric density estimation; Symbolic data.
1 Introduction

This paper concerns statistical analysis of data in which the variables are observed in the form of an interval, with lower and upper bounds. If a data set contains $p$ variables that are interval-valued, each observation is represented by a $p$-dimensional hypercube in $\mathbb{R}^p$. Traditional methods for vectorial data in density estimation and multiple regression cannot be readily applied to this type of data. Interval-valued data are a special case that fall into a much broader class of data called symbolic data (Billard and Diday, 2007). Symbolic data analysis uses a hypercube as the unit of measurement.

Where would these interval-valued data be observed? In some cases, the true value of a variable can be hidden within an interval. Questions regarding a person’s age or income in most demographic surveys give several choices of ranges to choose from, e.g. 25 – 30 years old or $50,000 – $75,000. Some variables are intrinsically observed as intervals. For example, the daily body temperature of a person naturally has lower and upper bounds. The actual selling price of a car model has a range, rather than being fixed at some value. In what follows we argue that in general a single-valued data set may be transformed into an interval-valued one. Due to the ever-increasing ability to collect data automatically, valuable information is being accumulated with minimal effort. Examples include web traffic and search data. Analyzing such big data in their original form is often not practical since it requires enormous computing resources and data storage. Consequently it has become a common practice to aggregate numerous observations into a few values using summary statistics such as sample mean, median, or, in our case, range.

Choosing a proper method for interval-valued data depends on the innate property of the data set. When each interval belongs to one independent unit, such as a person’s income in a survey, we can regard the observed interval as the absolute bounds that contain the true unobserved value. On the contrary, if the intervals are obtained from aggregating observations, we can consider the bounds as “soft”, which means that they are sample minima and maxima, thus might not include all possible values. In this paper we focus on this more general latter scenario. Note that after the aggregation, the sample size of the
resulting interval-valued data set may no longer be large.

Analyzing interval-valued data has been an interest to many researchers, especially for the regression problem. A common approach is to extract numerical attributes from intervals and use them to fit an ordinary regression model. Billard and Diday (2000) proposed the center method (CM) that fits a regular regression model with the centers of the intervals. Billard and Diday (2002) fitted two separate models with lower and upper bounds respectively. Lima Neto and de Carvalho (2008) proposed the center and range method (CRM) that fits two respective regression models for the centers and the ranges of the intervals. Billard and Diday (2007) regressed the center and range of the response variable separately, on a combined set of predictors that includes both centers and ranges of the predictor variables. Their model has twice as many predictors compared to most other methods and this over-parametrization has been found to have a negative impact on the out-of-sample prediction (Ahn et al., 2012).

Several attempts have been made to deal with a hyper-rectangle as the unit of the analysis, rather than converting the problem to the familiar domain of vectorial data. Blanco-Fernández et al. (2011), Blanco-Fernández et al. (2012), Blanco-Fernández et al. (2013), González-Rodríguez et al. (2007), Sinova et al. (2012) derived regression estimators for convex compact random sets using a set arithmetic approach. Xu (2010) proposed the symbolic covariance method (SCM) that applied a symbolic sample covariance (Billard, 2007) to find the least squares solution. Lima Neto et al. (2009, 2011) proposed the bivariate symbolic regression model based on a generalized linear model framework with a bivariate exponential family of distributions. Silva et al. (2011) developed a copula-based regression model and studied statistical inference using bivariate joint distributions. Ahn et al. (2012) studied a resampling method for which a large number of conventional data sets are generated by randomly selecting a point within each interval. Recently, Giordani (2014) proposed a lasso type regression estimator for interval-valued data.

In this work, we view the observed set of hyper-rectangles as a joint empirical histogram, from which we can estimate the joint probability distribution of the variables. We propose
to use locally weighted Gaussian kernel for the density estimation and to use neighboring intervals to obtain the weights. We can then make various statistical inferences based on the estimated joint distribution. For example, in the regression setting, we can calculate the conditional distribution of the response variable for given predictors. From this, we can obtain prediction intervals as well as a point prediction with conditional mean.

The rest of the paper is organized as follows. In Section 2, we demonstrate the proposed approach in the context of the univariate density estimation. We then approach the regression problem with conditional distribution of the response given predictor values. In Sections 3 and 4, we apply the proposed method to simulated and real examples, respectively, and compare it with some existing methods. The paper ends with discussion in Section 5.

2 Proposed Method

Let \( X_1, \ldots, X_p \) be \( p \) real single-valued random variables with a joint density function \( f \). Suppose that we do not observe their realizations directly, but instead observe them in intervals: \( x_{ij}^I = [x_{ij}^L, x_{ij}^U] \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, p \). The objective is to estimate \( f \) or the conditional density of a variable given others’ observed intervals. Let \( f_n \) be the empirical joint density based on the observed intervals \( x_{ij}^I \):

\[
f_n(x_1, \ldots, x_p) = \sum_{i=1}^{n} w_i \prod_{j=1}^{p} \frac{1}{x_{ij}^U - x_{ij}^L} I(x_{ij}^L \leq x_j \leq x_{ij}^U),
\]

where \( w_i \)'s are positive weights with a typical choice \( w_i = n^{-1} \). One can directly use \( f_n \) as an estimate of \( f \). The estimate \( f_n \) is piecewise constant analogous to histograms of single-valued data, but there is no need to choose the widths or end points of the bins. In a regression problem, it is straightforward to obtain the conditional distribution of a variable given all the other variables with this estimate. Also, any conditional measure of interest such as conditional mean can be estimated based on this empirical conditional distribution. However, there are possible drawbacks of this naive estimate: a) a non-smooth density function is not appropriate for many continuous random variables, and b) it may not be
defined in regions where the hyper-rectangles are sparsely observed.

We can improve the naive estimate (1) by imposing some structures on the distribution of the data. We start by fitting the single joint normal density \( \phi(x_1, \ldots, x_p | \hat{\mu}, \hat{\Sigma}) \) with mean \( \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p) \) and covariance \( \hat{\Sigma} = \{\hat{\sigma}_{jj'}\}_{j,j'=1,\ldots,p} \), by matching the first and second moments. Under the empirical distribution (1) with \( w_i = 1/n \), we have

\[
\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} m_{ij}, \quad \hat{\sigma}_{jj} = \frac{1}{n} \sum_{i=1}^{n} v_{ij}, \quad \hat{\sigma}_{jj'} = \frac{1}{n} \sum_{i=1}^{n} m_{ij}m_{ij'} - \hat{\mu}_j\hat{\mu}_{j'},
\]

where \( m_{ij} = (x_{L_{ij}} + x_{U_{ij}})/2 \) is the center and \( v_{ij} = (x_{U_{ij}} - x_{L_{ij}})^2/12 \) is a function of the ranges of the intervals. Note that \( m_{ij} \) and \( v_{ij} \) are the mean and variance of Uniform distribution \( U(x_{L_{ij}}, x_{U_{ij}}) \), respectively. However in practice, assuming that a single multivariate normal density approximates the true density is perhaps too restrictive because the target density might be asymmetric and/or have a local structure. We propose to use a Gaussian kernel approach since it is able to reflect local data structures and also has some nice properties such as continuity and smoothness. The proposed estimator with \( n \) mixture multivariate normal densities is the following:

\[
f_n(x_1, \ldots, x_p) = \frac{1}{n} \sum_{k=1}^{n} \phi(x_1, \ldots, x_p | \hat{\mu}_k, \hat{\Sigma}_k),
\]

where \( \phi(\cdot | \hat{\mu}_k, \hat{\Sigma}_k) \) is the multivariate normal density with mean \( \hat{\mu}_k = (\hat{\mu}_{k,1}, \ldots, \hat{\mu}_{k,p}) \) and covariance \( \hat{\Sigma}_k = \{\hat{\sigma}_{k,jj'}\}_{j,j'=1,\ldots,p} \). The mean and covariance are estimated by

\[
\hat{\mu}_{k,j} = \sum_{i=1}^{n} w_{ki}m_{ij}, \quad \hat{\sigma}_{k, jj} = \sum_{i=1}^{n} w_{ki}v_{ij}, \quad \hat{\sigma}_{k,jj'} = \sum_{i=1}^{n} w_{ki}m_{ij}m_{ij'} - \hat{\mu}_{k,j}\hat{\mu}_{k,j'}
\]

for \( j, j' = 1, \ldots, p \ (j \neq j') \), where we use local weights \( w_{ki}, i = 1, \ldots, n \), with \( \sum_i w_{ki} = 1 \). The \( w_{ki} \) are determined by considering the distances between the \( i \)th and \( k \)th rectangles, such that neighboring observations are more heavily weighted than distant ones. More details on the estimation of the weights will be provided in Section 2.1.

The proposed approach falls in the category of nonparametric function estimation because we do not assume a particular form on the population density. Although the density
estimator in (3) contains many parameters, the only parameter that should be carefully estimated from the data is the tuning parameter that determines the local weights $w_{ki}$, which will be discussed in detail in Sections 2.1 and 2.2. In this sense, the proposed estimator can be viewed as an extension of a multivariate kernel density estimator to interval-valued data.

Figure 1 illustrates the proposed approach with two toy data sets with size $n = 50$ each. The interval data in the upper panels are generated as follows. The centers $x_i^C$ of $X$ are generated from $N(5, 2^2)$, and the centers $y_i^C$ of $Y$ are generated by $y_i^C = x_i^C + \epsilon_i$, where $\epsilon_i \sim N(0, 0.75^2)$. In the bottom panels, the centers are generated by $x_i^C \sim N(5, 2^2)$ and $y_i^C = 6 + 4 \sin(0.25\pi x_i^C) + \epsilon_i$, where $\epsilon_i \sim N(0, 1.5^2)$. The half-ranges $x_i^R$ and $y_i^R$ are independently generated from $U(0.5, 1.5)$, and the lower and upper bounds of the intervals are obtained as $x_i^L = x_i^C - x_i^R$, $x_i^U = x_i^C + x_i^R$, $y_i^L = y_i^C - y_i^R$, and $y_i^U = y_i^C + y_i^R$.

In the left panels in Figure 1, random samples of size 1000 are generated from the bivariate normal with the mean and covariance estimated by (2) and plotted in red circles. In the right panels random samples of size 1000 from (3) utilizing the local component means and covariances are plotted. It can be seen that the data in the upper panels are well approximated by a single bivariate normal distribution. However, the distribution of rectangles in the bottom panels exhibits more complicated local structures and cannot be properly estimated by a single bivariate normal density. In contrast, the proposed approach in (3) can incorporate the local structures successfully.

In the following two subsections, we discuss the proposed method under the settings of univariate density estimation in Section 2.1 and regression in Section 2.2.

### 2.1 Univariate Density Estimation

Let $X$ be a real-valued random variable and $x_i^I = [x_i^L, x_i^U]$ be its interval-valued observation for $i = 1, \ldots, n$. Let $f$ be the population density of $X$, which can be approximated by the following Gaussian kernel estimator:

$$f_n(x) = \frac{1}{n} \sum_{k=1}^{n} \phi(x|\hat{\mu}_k, \hat{\sigma}_k),$$

where $\phi(x|\mu, \sigma)$ is the Gaussian density function with mean $\mu$ and variance $\sigma^2$. This estimator is given by (5).
Figure 1: Illustration of the proposed approach using two toy data sets with linear (upper panels) and nonlinear (bottom panels) relationship between two variables. A single bivariate normal density is used for left panels and locally weighted normal densities are used for right panels. In the upper panels, the interval data are well approximated by both estimators. On the other hand, the Gaussian kernel approach in the lower panels can incorporate the local structures well.
where $\phi(\cdot | \hat{\mu}_k, \hat{\sigma}_k)$ is the univariate normal density with mean $\hat{\mu}_k$ and standard deviation $\hat{\sigma}_k$ computed by (4) with local weights $w_{ki}$, $i = 1, \ldots, n$. The local weights $w_{ki}$ are determined as follows. Using the centers of the intervals, we first calculate Euclidean distances between $k$th and $i$th intervals and sort the distances. Note that one can also use the endpoints to define the distance between intervals. Let $R_{ki}$ be the rank of the $i$th observation in the sorted distances to the $k$th interval such that $R_{kk} = 1$. Then, the weights are determined such that

$$w_{ki} \propto \frac{1}{h} K \left( \frac{R_{ki} - 1}{h} \right)$$

and

$$\sum_{i=1}^{n} w_{ki} = 1,$$

(6)

where $K$ is a symmetric kernel function (we use the standard normal density in this work) and $h$ is the bandwidth. The definition in (6) makes intervals that are close to the $k$th interval carry more weight than the distant ones. The role of the bandwidth $h$ is to control how many neighboring intervals carry nontrivial weights. If $h$ is large, each interval contributes equally; i.e., all the weights $w_{ki}$ become close to $n^{-1}$ and the Gaussian kernel estimator in (5) approaches a single normal density. If $h$ is small, then $w_{kk} \approx 1$ and $w_{ki} \approx 0$ for $i \neq k$; i.e., the $k$th interval alone determines the corresponding density component $\phi(x | \hat{\mu}_k, \hat{\sigma}_k)$ in (5).

The choice of $h$ plays an important role; thus, we propose to control the number of observations that hold nontrivial weights through a smoothing parameter $\alpha$. Specifically, for a given $\alpha$ in the range of $(1/n, 1]$, we determine $h$ such that the $[n\alpha]$-th nearest interval has the weight $([n\alpha] - 1)/h = 2$, where $[a]$ is the largest integer less than or equal to $a$. Because we use the standard normal density for the kernel, the nearest $[n\alpha]$ intervals to the $k$th interval approximately carry 95.4% of the weights collectively. For the purpose of density estimation, it is desirable to use a small $\alpha$ when the data exhibit some local structures while a large $\alpha$ should be used when the data are close to a single Gaussian structure.

In what follows we consider two different distributional settings to demonstrate the method. As discussed in Section 1, the true nature of the bounds of interval-valued data is generally unknown, so it is not straightforward to generate random interval-valued data from a simple underlying population. In order to create a population of intervals, we generate a
Figure 2: A simulated population for interval data. The centers are generated from $N(0, 2^2)$ and the half-ranges from $U(0.1, 3)$. Random heights are used for $y$-axis for the purpose of visualization. The population density is obtained by the empirical histogram of these intervals.
large number, say $N$, of intervals from which the population density $f$ can be obtained from the histogram of $N$ such intervals.

![Figure 3: Univariate density estimation for (a) normal and (b) normal mixture with interval-valued data with varying smoothing parameter $\alpha$. In each plot, the solid line represents the population density $f$ and the other lines represent the proposed estimators with different $\alpha$ values.](image)

(a) Normal

(b) Normal mixture

We consider two different settings. In the first setting, the centers are generated from $N(0, 2^2)$ and the half-ranges from $U(0.1, 3)$. Figure 2 displays 500 of those intervals, in which random heights are used for $y$-axis for convenient visual separation (Cleveland, 1993). In the second setting, we use a normal mixture for centers, $0.3 \times N(-0.6, 0.2^2) + 0.7 \times N(0.6, 0.3^2)$, and the half-ranges are generated from $U(0.25, 0.5)$. For each setting, we simulate the population with $N = 10,000$ intervals, and we randomly sample $n = 100$ and estimate the density with the procedure described above. Figure 3 displays the true population densities (solid lines) and the proposed Gaussian kernel type estimates with varying $\alpha$. In Figure 3(a), note
that the population density has thicker tails than a normal density, due to the range effect. Note also that lower values of $\alpha$ fit the data better at the tails where the data are sparse, and relatively higher values such as 0.3 work better at the center where the data are dense. In Figure 3(b), where the true density is bimodal, lower values of $\alpha$ provide a better estimate of the true mode.

As demonstrated in the examples above, choosing $\alpha$ is an important problem. For an optimal selection of $\alpha$, we propose to use the Kullback-Leibler loss (Kullback and Leibler, 1951),

$$-\int f(x) \log f_n(x)dx,$$

where $f(x)$ is a histogram type estimator using the interval-valued data,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{I(x_i^L \leq x \leq x_i^U)}{x_i^U - x_i^L},$$

which is a univariate version of (1). Then, we choose $\alpha$ that minimizes the empirical Kullback-Leibler cross-validation

$$CV(\alpha) = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i^U - x_i^L} \int_{x_i^L}^{x_i^U} \log f_n^{(i)}(x)dx,$$

where $f_n^{(i)}(x)$ represents the density estimate based on the Gaussian kernel estimator in (5) with the $i$th interval left out. For the two simulated data in Figure 3, the proposed criterion selects $\alpha = 0.115$ and 0.011, respectively. We note that, as Figure 3 suggests, it might be better to use different values of $\alpha$ depending on the region. Extending the criterion (7) to the localized selection is left as a future study.

2.2 Regression with Conditional Distribution

For convenience let $Y = X_1$ be the response variable and $Z = (X_2, \ldots, X_p)'$ be the predictor variables in a regression problem. We propose to estimate the joint density of $Y$ and $Z$ with a Gaussian kernel estimator in (3)

$$f_n(y, z) = \frac{1}{n} \sum_{k=1}^{n} \phi_k(y, z),$$
where \( \phi_k(y, z) = \phi \left( y, z | \hat{\mu}_k, \hat{\Sigma}_k \right) \). By decomposing the mean vector \( \hat{\mu}_k = (\hat{\mu}_{k(y)}, \hat{\mu}_{k(z)}) \) and the covariance matrix \( \hat{\Sigma}_k = [\hat{\Sigma}_{k(yy)} \hat{\Sigma}_{k(yz)}; \hat{\Sigma}_{k(zy)} \hat{\Sigma}_{k(zz)}] \), each component in the estimator can be expressed as a product of the conditional density of \( Y \) given \( Z = z \) and the marginal of \( Z \):

\[
\phi \left( y, z | \hat{\mu}_k, \hat{\Sigma}_k \right) = \phi \left( y | \hat{\mu}_{k(y)}, \hat{\Sigma}_{k(yy)} \right) \phi \left( z | \hat{\mu}_{k(z)}, \hat{\Sigma}_{k(zz)} \right),
\]

where \( \hat{\mu}_{k(y)} = \hat{\mu}_{k(y)} + \hat{\Sigma}_{k(yz)} \hat{\Sigma}_{k(zz)}^{-1} (z - \hat{\mu}_{k(z)}) \) and \( \hat{\Sigma}_{k(yz)} = \hat{\Sigma}_{k(yy)} - \hat{\Sigma}_{k(yz)} \hat{\Sigma}_{k(zz)}^{-1} \hat{\Sigma}_{k(zy)} \). Then, the conditional density of \( Y \) given \( Z \) is

\[
f_n(y|z) = \frac{\sum_{k=1}^n \phi \left( y | \hat{\mu}_{k(y)}, \hat{\Sigma}_{k(yy)} \right) \phi \left( z | \hat{\mu}_{k(z)}, \hat{\Sigma}_{k(zz)} \right)}{\sum_{k=1}^n \phi \left( z | \hat{\mu}_{k(z)}, \hat{\Sigma}_{k(zz)} \right)}, \tag{8}
\]

where \( v_k(z) = \phi \left( z | \hat{\mu}_{k(z)}, \hat{\Sigma}_{k(zz)} \right) / \sum_{i=1}^n \phi \left( z | \hat{\mu}_{i(z)}, \hat{\Sigma}_{i(zz)} \right) \). The conditional expectation in (9) gives a natural point prediction for the response variable. Other types of inference regarding the response can be done in a straightforward way. For example, one can obtain a confidence interval for the point estimation, and also quantiles of the distribution.

We propose to choose \( \alpha \) that minimizes the leave-one-out sum of squared errors for predicting the centers of the observed data

\[
CV_{reg}(\alpha) = \frac{1}{n} \sum_{i=1}^n \left\{ y_i^C - \hat{\mu}^{(i)}(y|z_i^C) \right\}^2, \tag{10}
\]

where \( y_i^C \) and \( z_i^C \) are the center points for each interval, and \( \hat{\mu}^{(i)}(y|z_i^C) \) represents the conditional expectation (9) given the center point of the \( i \)th interval for \( z \) based on the Gaussian kernel estimate in (3) with the \( i \)th observation left out. This proposed criterion (10) shows good empirical performance, as demonstrated in Section 3. Nevertheless, we note that a different type of cross-validation criterion can also be considered and suggest it as a future study.
2.3 Prediction with Interval-valued Predictors

When performing regression with interval-valued data, it is important the method used should be capable of making predictions on the response variable. The prediction can be for centers of the response interval, which we interpret as conditional mean, and also can be for the two bounds of the response interval. In what follows we illustrate how the center and bounds can be predicted in a sensible manner under the unified framework of our approach.

Given a predictor hyper-rectangle \( R \in \mathbb{R}^{p-1} \), the proposed approach obtains the conditional cumulative distribution function from (8):

\[
P(Y \leq y' | \mathbf{Z} \in R) = \frac{\sum_k \int_{z \in R} \int_{-\infty}^{y'} \phi \left( y, z | \mu_{(k)}, \Sigma_{(k)} \right) dy dz}{\sum_k \int_{z \in R} \phi \left( z | \mu_{(k)}, \Sigma_{(k)} \right) dz},
\]

with which one can also calculate the conditional mean or the center of the response interval.

Predicting the lower and upper bounds is not as straightforward as the center prediction, since in practice the interpretation of the bounds is ambiguous. In this work we use the observed data to predict the bounds by assuming that the observed intervals are sample versions of the \( q_L \)th and \( q_U \)th quantiles where the \( q_L \) and \( q_U \) values may depend on the location of the predictor hyper-rectangle. For the \( i \)th observation with the response interval \([y_{iL}, y_{iU}]\) and the predictor hyper-rectangle \( R_i \), one can obtain the cumulative probabilities \( p_{iL} = P(Y \leq y_{iL} | \mathbf{Z} \in R_i) \) and \( p_{iU} = P(Y \leq y_{iU} | \mathbf{Z} \in R_i) \) with the conditional distribution function in (11). Given the predictor hyper-rectangle \( R \) for a new observation, we identify the index set \( I \) of the \( m \) nearest hyper-rectangle observations in terms of their center locations, and let \( \hat{q}_L = \text{median}_{i \in I} \{ p_{iL} \} \) and \( \hat{q}_U = \text{median}_{i \in I} \{ p_{iU} \} \). Then, the interval of the response variable for a new observation can be predicted by using the \( \hat{q}_L \)th and \( \hat{q}_U \)th quantiles of the estimated conditional distribution (11). We use \( m = n/5 \) in our empirical study.

Figure 4 illustrates the proposed regression method with a simulated data set. We established a population with 10,000 rectangles generated from some underlying structure. The smooth red curves show the 5th and 95th population percentiles of the conditional distribution of \( Y \) given \( X \). We obtained a sample of size 50 by randomly selecting rectangles from the population, shown with the black rectangles. The dotted red lines are conditional
Figure 4: Illustration of regression with the conditional distribution. Rectangles represent the simulated data set with size 50. The smooth red curves, solid cyan lines, and red dotted lines display the conditional 5th and 95th percentiles that are true, percentiles that are directly estimated from the rectangles, and percentiles that are estimated by the proposed method, respectively.
percentiles calculated from the estimated conditional distribution fitted with the proposed method. The solid cyan lines represent conditional percentiles directly calculated from the sample rectangles. It is evident that the estimated quantiles from the proposed method are smoother and closer to the population quantiles than the ones from the sample rectangles.

3 Simulation Study

In this section we compare the proposed Gaussian kernel method (GM) with some popular interval-valued regression methods including CM, CRM, and SCM. We also include CRM3, a modified nonlinear CRM, for which two respective cubic polynomial regression models are fitted for centers and ranges. We consider six different population settings. In order to create a population of interval-valued data, we generate 10,000 hyper-rectangles from each setting and then randomly select $n = 100$ with replacement to compose a sample for regression. We also generate a test data set with the same size to evaluate prediction errors. Each setting is discussed in the list below with a brief description.

- Setting 1: nonlinear center relationship, independent ranges
  The center $x^C_i$ is generated from $N(5, 2^2)$, and $y^C_i$ is generated by $y^C_i = 6 + 4 \sin(0.25\pi x^C_i) + \epsilon_i$, where $\epsilon_i \sim N(0, 1.5^2)$. The half-ranges $x^R_i$ and $y^R_i$ are independently generated from $U(0.5, 1.5)$.

- Setting 2: three clusters, no relationship within a cluster
  The 30% of the population are from $x^C_i \sim N(0, 1^2)$, $y^C_i \sim N(12, 3^2)$, $x^R_i \sim U(0.5, 1.5)$, $y^R_i \sim U(0.5, 1.5)$, 40% are from $x^C_i \sim N(5, 1^2)$, $y^C_i \sim N(6, 0.5^2)$, $x^R_i \sim U(0.5, 1.5)$, $y^R_i \sim U(0.15, 0.45)$, and the remaining 30% are from $x^C_i \sim N(10, 1^2)$, $y^C_i \sim N(0, 3^2)$, $x^R_i \sim U(0.5, 1.5)$, $y^R_i \sim U(0.5, 1.5)$.

- Setting 3: linear center relationship, heteroscedastic response
  Let $g_i$ be 1 and 2 with probability 0.5 and 0.5 respectively. When $g_i = j$ ($j = 1, 2$), we have the centers $x^C_i \sim N(2, 1^2)$, $y^C_i = 15 - x^C_i + \sigma_j \epsilon_i$, with $\sigma_1 = 0.75$, $\sigma_2 = 0.3$,
and $\epsilon_i \sim N(0, 1^2)$, and the half-ranges $x^R_{i}/0.5 \sim U(0.5, 1.5)$, $y^RL_{i}/b^L_{j} \sim U(0.5, 1.5)$, $y^{RU}_{i}/b^U_{j} \sim U(0.5, 1.5)$ with $b^L_{1} = 0.2, b^U_{1} = 2, b^L_{2} = 0.04$, and $b^U_{2} = 0.4$.

- Setting 4: two predictors, linear center relationship, ranges linearly dependent on centers

The $x^C_{i1}$ and $x^C_{i2}$ are from $N(5, 3^2)$ independently. The $y^C_{i}$ is generated by $y^C_{i} = 5 + x^C_{i1} - x^C_{i2} + \epsilon_i$, where $\epsilon_i \sim N(0, 1.5^2)$. The $x^R_{i1}$, $x^R_{i2}$ and $y^R_{i}$ are generated by $x^R_{i1} = 2 - 0.1x^C_{i1} + \tau_{i1}$, $x^R_{i2} = 1 + 0.1x^C_{i2} + \tau_{i2}$ and $y^R_{i} = 1 + 0.2y^C_{i} + \delta_i$, where $\tau_{i1}, \tau_{i2}$ are from $N(0, 0.1^2)$ and $\delta_i$ from $N(0, 0.2^2)$. We take absolute values for half-ranges.

- Setting 5: two predictors, nonlinear center relationship, ranges nonlinearly dependent on centers

The $x^C_{i1}$ and $x^C_{i2}$ are from $N(5, 3^2)$ independently. The $y^C_{i}$ is generated by $y^C_{i} = 4 - 2(x^C_{i1} - 5) \exp\{-0.05(x^C_{i1} - 5)^2\} - 100/(x^C_{i2} + 10) + \epsilon_i$, where $\epsilon_i \sim N(0, 1^2)$. For half-ranges, we first generate $v_{i1} = u_{i1} + \exp(-0.5x^C_{i1} + \tau_{i1})$, $v_{i2} = u_{i2} + \exp(-0.5x^C_{i2} + \tau_{i2})$ and $w_i = \exp(-3 + 0.2x^C_{i1} - 0.2x^C_{i2} + \delta_i)$ where $u_{i1}, u_{i2} \sim U(0, 0.5), \tau_{i1}, \tau_{i2} \sim N(0, 0.2^2)$, and $\delta_i \sim N(0, 0.5^2)$. Then, we let $x^R_{i1} = 2v_{i1}/(1 + v_{i1})$, $x^R_{i2} = 3v_{i2}/(1 + v_{i2})$ and $y^R_{i} = 3w_i/(1 + w_i)$.

- Setting 6: three predictors, additive (one linear and two nonlinear) center relationships, ranges nonlinearly dependent on centers

The $x^C_{i1}$, $x^C_{i2}$ and $x^C_{i3}$ are from $N(5, 3^2)$ independently. The $y^C_{i}$ is generated by $y^C_{i} = 4 - 2(x^C_{i1} - 5) \exp\{-0.05(x^C_{i1} - 5)^2\} - 100/(x^C_{i2} + 10) + x^C_{i3}/5 + \epsilon_i$, where $\epsilon_i \sim N(0, 1^2)$. For half-ranges, we first generate $v_{i1} = u_{i1} + \exp(-0.5x^C_{i1} + \tau_{i1})$, $v_{i2} = u_{i2} + \exp(-0.5x^C_{i2} + \tau_{i2})$, $v_{i3} = u_{i3} + \exp(-0.5x^C_{i3} + \tau_{i3})$ and $w_i = \exp(-3 + 0.2x^C_{i1} - 0.2x^C_{i2} - 0.2x^C_{i3} + \delta_i)$ where $u_{i1}, u_{i2}, u_{i3} \sim U(0, 0.5), \tau_{i1}, \tau_{i2}, \tau_{i3} \sim N(0, 0.2^2)$, and $\delta_i \sim N(0, 0.5^2)$. Then, we let $x^R_{i1} = 2v_{i1}/(1 + v_{i1})$, $x^R_{i2} = 3v_{i2}/(1 + v_{i2})$, $x^R_{i3} = 3v_{i3}/(1 + v_{i3})$ and $y^R_{i} = 3w_i/(1 + w_i)$.

Table 1 displays the results of the methods in terms of out-of-sample prediction errors for bounds, center, and range, along with the corresponding standard errors from 100 repetitions. We note that there exists a simple relationship between the {center, range} prediction and
<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>lower</th>
<th>upper</th>
<th>lower + upper</th>
<th>center</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GM</td>
<td>105.10 (2.82)</td>
<td>93.20 (1.93)</td>
<td>198.30 (4.39)</td>
<td>84.48 (2.17)</td>
<td>58.70 (1.26)</td>
</tr>
<tr>
<td></td>
<td>CM</td>
<td>645.05 (11.45)</td>
<td>656.77 (12.49)</td>
<td>1301.81 (22.97)</td>
<td>622.59 (11.18)</td>
<td>113.25 (5.34)</td>
</tr>
<tr>
<td></td>
<td>CRM</td>
<td>628.37 (11.52)</td>
<td>633.98 (11.03)</td>
<td>1262.35 (22.36)</td>
<td>622.59 (11.18)</td>
<td>34.32 (0.35)</td>
</tr>
<tr>
<td></td>
<td>CRM3</td>
<td>336.00 (37.79)</td>
<td>333.21 (36.92)</td>
<td>669.21 (74.68)</td>
<td>325.92 (37.33)</td>
<td>34.73 (0.36)</td>
</tr>
<tr>
<td></td>
<td>SCM</td>
<td>660.26 (10.66)</td>
<td>673.37 (12.09)</td>
<td>1333.63 (21.32)</td>
<td>629.32 (10.02)</td>
<td>149.99 (6.33)</td>
</tr>
<tr>
<td>2</td>
<td>GM</td>
<td>628.71 (12.92)</td>
<td>641.31 (13.81)</td>
<td>1270.02 (26.23)</td>
<td>621.02 (13.00)</td>
<td>55.97 (2.69)</td>
</tr>
<tr>
<td></td>
<td>CM</td>
<td>746.95 (15.13)</td>
<td>749.64 (14.44)</td>
<td>1496.59 (28.59)</td>
<td>705.62 (14.33)</td>
<td>170.67 (3.58)</td>
</tr>
<tr>
<td></td>
<td>CRM</td>
<td>720.67 (14.64)</td>
<td>725.66 (14.35)</td>
<td>1446.33 (28.64)</td>
<td>705.62 (14.33)</td>
<td>70.17 (0.69)</td>
</tr>
<tr>
<td></td>
<td>CRM3</td>
<td>708.52 (13.12)</td>
<td>708.77 (13.02)</td>
<td>1417.29 (25.56)</td>
<td>690.80 (12.79)</td>
<td>71.38 (0.74)</td>
</tr>
<tr>
<td></td>
<td>SCM</td>
<td>746.35 (15.11)</td>
<td>749.45 (14.43)</td>
<td>1495.80 (28.70)</td>
<td>708.89 (14.43)</td>
<td>156.03 (3.17)</td>
</tr>
<tr>
<td>3</td>
<td>GM</td>
<td>37.36 (0.75)</td>
<td>56.45 (1.18)</td>
<td>93.81 (1.79)</td>
<td>41.68 (0.87)</td>
<td>20.89 (0.41)</td>
</tr>
<tr>
<td></td>
<td>CM</td>
<td>55.73 (1.07)</td>
<td>79.53 (1.58)</td>
<td>135.26 (1.77)</td>
<td>40.11 (0.76)</td>
<td>110.06 (1.24)</td>
</tr>
<tr>
<td></td>
<td>CRM</td>
<td>52.79 (0.97)</td>
<td>76.28 (1.39)</td>
<td>129.07 (1.64)</td>
<td>40.11 (0.76)</td>
<td>97.69 (0.90)</td>
</tr>
<tr>
<td></td>
<td>CRM3</td>
<td>55.06 (1.02)</td>
<td>75.13 (1.40)</td>
<td>130.19 (1.70)</td>
<td>40.15 (0.78)</td>
<td>99.78 (1.04)</td>
</tr>
<tr>
<td></td>
<td>SCM</td>
<td>50.70 (0.98)</td>
<td>85.73 (1.69)</td>
<td>136.43 (1.83)</td>
<td>40.54 (0.78)</td>
<td>110.71 (1.27)</td>
</tr>
<tr>
<td>4</td>
<td>GM</td>
<td>151.13 (3.58)</td>
<td>353.84 (8.41)</td>
<td>504.97 (11.61)</td>
<td>230.18 (5.37)</td>
<td>89.21 (2.26)</td>
</tr>
<tr>
<td></td>
<td>CM</td>
<td>378.13 (12.68)</td>
<td>506.47 (13.02)</td>
<td>884.61 (21.63)</td>
<td>204.74 (4.82)</td>
<td>950.26 (28.70)</td>
</tr>
<tr>
<td></td>
<td>CRM</td>
<td>145.01 (3.40)</td>
<td>300.02 (7.14)</td>
<td>445.04 (10.34)</td>
<td>204.74 (4.82)</td>
<td>71.11 (1.75)</td>
</tr>
<tr>
<td></td>
<td>CRM3</td>
<td>160.37 (3.52)</td>
<td>325.84 (7.38)</td>
<td>486.21 (10.65)</td>
<td>224.29 (5.00)</td>
<td>75.28 (1.79)</td>
</tr>
<tr>
<td></td>
<td>SCM</td>
<td>478.76 (12.62)</td>
<td>597.08 (14.66)</td>
<td>1075.85 (22.73)</td>
<td>219.97 (5.34)</td>
<td>1271.83 (28.66)</td>
</tr>
<tr>
<td>5</td>
<td>GM</td>
<td>172.42 (9.45)</td>
<td>166.11 (7.48)</td>
<td>338.53 (16.87)</td>
<td>165.91 (8.39)</td>
<td>13.44 (2.12)</td>
</tr>
<tr>
<td></td>
<td>CM</td>
<td>417.89 (8.35)</td>
<td>459.34 (11.80)</td>
<td>877.23 (17.86)</td>
<td>380.02 (8.61)</td>
<td>234.37 (5.52)</td>
</tr>
<tr>
<td></td>
<td>CRM</td>
<td>380.96 (8.81)</td>
<td>385.25 (8.64)</td>
<td>766.21 (17.29)</td>
<td>380.02 (8.61)</td>
<td>12.34 (0.41)</td>
</tr>
<tr>
<td></td>
<td>CRM3</td>
<td>227.43 (14.45)</td>
<td>220.12 (12.94)</td>
<td>447.55 (27.29)</td>
<td>220.45 (13.62)</td>
<td>13.29 (0.57)</td>
</tr>
<tr>
<td></td>
<td>SCM</td>
<td>368.29 (8.34)</td>
<td>405.10 (9.26)</td>
<td>773.39 (17.15)</td>
<td>379.42 (8.52)</td>
<td>29.11 (0.90)</td>
</tr>
<tr>
<td>6</td>
<td>GM</td>
<td>149.67 (4.89)</td>
<td>147.79 (5.01)</td>
<td>297.45 (9.83)</td>
<td>147.13 (4.80)</td>
<td>6.39 (3.02)</td>
</tr>
<tr>
<td></td>
<td>CM</td>
<td>256.91 (7.40)</td>
<td>265.28 (7.74)</td>
<td>522.18 (12.03)</td>
<td>207.54 (5.39)</td>
<td>214.20 (6.27)</td>
</tr>
<tr>
<td></td>
<td>CRM</td>
<td>207.63 (5.26)</td>
<td>208.71 (5.59)</td>
<td>416.34 (10.80)</td>
<td>207.54 (5.39)</td>
<td>2.52 (0.16)</td>
</tr>
<tr>
<td></td>
<td>CRM3</td>
<td>159.93 (19.70)</td>
<td>155.70 (18.78)</td>
<td>315.64 (38.41)</td>
<td>156.00 (19.14)</td>
<td>7.28 (2.64)</td>
</tr>
<tr>
<td></td>
<td>SCM</td>
<td>202.14 (5.27)</td>
<td>220.68 (5.92)</td>
<td>422.82 (10.84)</td>
<td>206.58 (5.36)</td>
<td>19.31 (1.08)</td>
</tr>
</tbody>
</table>

Table 1: Out-of-sample sum of squared errors for lower bounds, upper bounds, both lower and upper bounds, centers, and ranges are shown with the corresponding standard errors in parentheses.
the \{lower, upper\} bounds prediction, since \(\sum_{i=1}^{n}\{(\hat{y}_{Li} - y_{Li})^2 + (\hat{y}_{Ui} - y_{Ui})^2\} = 2\sum_{i=1}^{n}\{(\hat{y}_{Ci} - y_{Ci})^2 + (\hat{y}_{Ri} - y_{Ri})^2\}\). Despite this redundancy, we present both prediction results because they provide different insights in prediction ability. In general, we expect CM, CRM, and SCM to work well for settings with linear relationship between centers, such as Settings 3 and 4. On the other hand, we expect GM and CRM3 to capture nonlinear relationships well. Note that when ranges are dependent on centers, regardless of predictors or response, they are also related with ranges in the opposite side as well because of the center relationships. In these cases, such as Settings 4–6, we expect CRM or CRM3 to perform well. We expect the proposed method GM to perform well in various settings. Note that CM always has the same center prediction results as CRM.

In Setting 1, the proposed GM is noticeably better for bounds and center prediction, but both CRM and CRM3 predict ranges better than other methods because they specifically model ranges. The results for Settings 2 and 3 are similar in the sense that GM is best overall except for the center prediction in Setting 3, where all methods show similar results. Setting 4 is designed to be favorable to CRM, and as expected it is the best for this setting. The CRM3 and GM are the close second and third respectively in terms of bounds and range prediction. In Settings 5 and 6 where there are multiple predictors, GM is evidently superior for centers and bounds, but for the range, its performance is similar or slightly inferior to those of CRM and CRM3. We note that a good bounds prediction mostly implies a good prediction for both center and range, and vice versa. However, a good range prediction does not necessarily imply a good bounds or center prediction.

In Figure 5, we investigate the robustness of the criterion \(CV_{reg}(\alpha)\) in (10) with respect to repetitions using Setting 5, one of the more complicated simulation examples. The left panel displays the median values of \(CV_{reg}(\alpha)\) for a range of \(\alpha\) over 100 repetitions, along with quantiles. A sharp valley near \(\alpha = .10\) indicates that the criterion is not only sensible and but also consistently chooses similar values throughout the repetitions. The latter point is also supported by the histogram of the 100 chosen \(\alpha\) shown in the right panel.
Figure 5: Tuning parameter selection in regression. Quantiles of $CV_{reg}(\alpha)$ with varying smoothing parameter $\alpha$ based on 100 repetitions of Setting 5 (left panel) and a histogram of the selected $\alpha$ values (right panel).
In general a high-dimensional density function is known to be difficult to estimate due to the curse of dimensionality. The proposed method is based on a local method in density estimation, which might be vulnerable to this problem. We also examined the computational cost of the proposed method for Settings 5 (with two predictors) and 6 (with three predictors). The mean (and standard deviation) computing time of 10 independent runs including the prediction process is 224.44s (3.98) for Setting 5 and 376.88s (10.49) for Setting 6, which shows increasing difficulty in computation as the number of predictors increases. We further discuss this issue and suggest an alternative for a large number of predictors in Section 5.

4 Real Data Examples

Stock Indices Data

In this example, we investigate the relationship between two stock indices using the data analyzed in Cipollini et al. (2013). Among possible issues of interest, we concentrate on the realized variance, which is one of volatility measures in their paper, of two stock indices Dow Jones Industrials (DJ30) and Spanish IBEX (IBEX35). The data are collected between January 1996 and February 2009, which results in 3,411 daily realized variance observations. From these single-valued data we create 158 interval-valued observations by calculating Q1 and Q3 of each month for both stock indices. Figure 6 displays the data on a logarithm scale using the aggregated intervals. It shows a roughly linear relationship between two measures. We estimate the conditional density of IBEX35 given DJ30, and overlay the median and 1st, 5th, 95th, and 99th percentiles of the distribution. This reveals that there exists a nonlinear relationship between the two variables.

As in Section 3, we compare five methods, GM, CM, CRM, CRM3, and SCM. We randomly split the data into training and testing data with equal sample sizes, and repeat this procedure 100 times to calculate prediction errors. Table 2 reports the mean of prediction errors and their standard errors from the repetitions. In terms of center and bounds prediction, all the methods except CRM3 perform similarly, and the proposed GM yields the
Figure 6: Display of interval-valued data and pointwise conditional quantiles based on the proposed approach for stock indices data.
smallest prediction errors. Not surprisingly, CRM-based methods can predict the range most accurately, and GM performs better than SCM and CM.

Table 2: Result of the stock indices example. The mean of prediction errors based on the random split validation and the corresponding standard errors are shown in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>lower+upper</th>
<th>center</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>12.36 (0.27)</td>
<td>15.38 (0.30)</td>
<td>27.74 (0.52)</td>
<td>12.66 (0.26)</td>
<td>4.86 (0.19)</td>
</tr>
<tr>
<td>CM</td>
<td>13.09 (0.25)</td>
<td>17.54 (0.41)</td>
<td>30.62 (0.59)</td>
<td>13.85 (0.27)</td>
<td>5.84 (0.20)</td>
</tr>
<tr>
<td>CRM</td>
<td>13.28 (0.27)</td>
<td>16.35 (0.32)</td>
<td>29.63 (0.56)</td>
<td>13.85 (0.27)</td>
<td>3.84 (0.14)</td>
</tr>
<tr>
<td>CRM3</td>
<td>15.82 (1.39)</td>
<td>19.14 (1.35)</td>
<td>34.96 (2.73)</td>
<td>16.53 (1.36)</td>
<td>3.78 (0.18)</td>
</tr>
<tr>
<td>SCM</td>
<td>13.07 (0.24)</td>
<td>17.46 (0.39)</td>
<td>30.53 (0.57)</td>
<td>13.82 (0.27)</td>
<td>5.78 (0.19)</td>
</tr>
</tbody>
</table>

**Hawaiian Climate Data**

We analyze the Hawaiian climate data with two predictors, publicly available from the National Climatic Data Center at [http://www.ncdc.noaa.gov/](http://www.ncdc.noaa.gov/). A regression model is used to predict the daily sea level pressure based on the daily temperature and wind speed in Hawaii. A total of 5,856 single-valued observations are collected in the year 2012 from 16 stations and they are converted to interval-valued data using the Q1 and Q3 of the 16 stations each day for temperature, sea level pressure and wind speed. Figure 7 displays interval-valued scatterplots of the response variable against each predictor. Figure 7(a) shows that the sea level pressure seems to decrease as temperature increases and Figure 7(b) indicates the positive linear relationship between sea level pressure and wind speed. The conditional medians and quantiles obtained from the propose method are overlaid in the two plots and confirm these trends.

Table 3 displays the mean prediction errors of GM, CM, CRM, CRM3, and SCM based on the leave-out-one cross validation and the corresponding standard errors in parentheses. The GM noticeably performs better than the other four methods in terms of bounds and center
Figure 7: Display of interval-valued data and pointwise conditional quantiles based on the proposed approach for Hawaiian climate data.

prediction while the CRM-based methods are, as expected, accurate in range prediction. This example again demonstrates the usefulness of the proposed method in prediction of interval-valued data with multiple predictors.

5 Discussion

We propose to fit a joint distribution of observed interval-valued variables with a Gaussian kernel type estimator. The core idea is that we regard the collection of observed hyper-rectangles as an empirical multivariate histogram. Unlike many existing methods, the proposed method possesses flexibility of making various types of prediction. Utilizing the whole conditional distribution, one can make a point or interval prediction whether the value of the predictor variable is given as a point or as an interval. The proposed method is also capable of obtaining conditional median or quantiles as well as conditional mean.

The proposed approach can be readily applied to many other problems than regression.
Table 3: Result of the Hawaiian climate data. The mean of prediction errors based on the leave-out-one cross validation and their standard errors in parentheses are shown.

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>lower + upper</th>
<th>center</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>571.17 (3.39)</td>
<td>543.80 (3.26)</td>
<td>1114.97 (6.58)</td>
<td>552.96 (3.29)</td>
<td>18.11 (0.25)</td>
</tr>
<tr>
<td>CM</td>
<td>1020.25 (6.18)</td>
<td>965.27 (7.92)</td>
<td>1985.52 (11.29)</td>
<td>582.35 (3.29)</td>
<td>1641.64 (17.83)</td>
</tr>
<tr>
<td>CRM</td>
<td>603.22 (3.40)</td>
<td>567.86 (3.22)</td>
<td>1171.08 (6.58)</td>
<td>582.35 (3.29)</td>
<td>12.75 (0.07)</td>
</tr>
<tr>
<td>CRM3</td>
<td>607.76 (4.08)</td>
<td>579.06 (4.12)</td>
<td>1186.81 (8.16)</td>
<td>590.16 (4.08)</td>
<td>12.98 (0.07)</td>
</tr>
<tr>
<td>SCM</td>
<td>696.46 (3.67)</td>
<td>583.75 (3.27)</td>
<td>1280.20 (6.53)</td>
<td>586.66 (3.35)</td>
<td>213.77 (3.00)</td>
</tr>
</tbody>
</table>

Taking binary classification as an example, one can estimate a density for each class using the proposed approach and derive a classification rule based on the densities. Similarly, clustering can also be done by finding local modes of the Gaussian kernel estimator.

As seen in Section 3, application of the proposed approach to the problem with a large number of predictor variables may not be straightforward. One possible approach is to give a functional ANOVA structure to the log-density function as in Gu (2002), and utilize the fact that, under some functional log-density ANOVA models, the joint density is completely determined by the corresponding marginal densities to the function components in the model (Jeon, 2012). For example, one may estimate all the two-dimensional marginal densities first, then build the high-dimensional joint density by using the estimated marginal densities as building blocks. This merits further investigations in the future study.

References


