Simultaneous interval regression for K-nearest neighbor

Mohammad Ghasemi Hamed, Mathieu Serrurier, Nicolas Durand

Simultaneous Interval Regression for K-Nearest Neighbor

Mohammad Ghasemi Hamed [1,2], Mathieu Serrurier[1], and Nicolas Durand[1,2]

[1] IRIT - Université Paul Sabatier
118 route de Narbonne 31062, Toulouse Cedex 9, France
[2] ENAC/MAIAA - 7 avenue Edouard Belin 31055 Toulouse, France

Abstract. In some regression problems, it may be more reasonable to predict intervals rather than precise values. We are interested in finding intervals which simultaneously for all input instances $x \in \mathcal{X}$ contain a β proportion of the response values. We name this problem simultaneous interval regression. This is similar to simultaneous tolerance intervals for regression with a high confidence level $\gamma \approx 1$ and several authors have already treated this problem for linear regression. Such intervals could be seen as a form of confidence envelop for the prediction variable given any value of predictor variables in their domain. Tolerance intervals and simultaneous tolerance intervals have not yet been treated for the K-nearest neighbor (KNN) regression method. The goal of this paper is to consider the simultaneous interval regression problem for KNN and this is done without the homoscedasticity assumption. In this scope, we propose a new interval regression method based on KNN which takes advantage of tolerance intervals in order to choose, for each instance, the value of the hyper-parameter K which will be a good trade-off between the precision and the uncertainty due to the limited sample size of the neighborhood around each instance. In the experiment part, our proposed interval construction method is compared with a more conventional interval approximation method on six benchmark regression data sets.

1 Introduction

When dealing with regression problems, it may be risky to predict a point which may be illusionary precise. Due to the existence of learning biases, especially the limited amount of available data and the necessarily incomplete language used for describing them, the obtained model does not describe exactly the true unknown model. In situations with lack of sufficient observations to obtain precise results or when the considered model is too complex, one may rather want to find intervals which are the most likely to contain a desired proportion of the population of the response values. Such intervals are mostly used in application demanding a high level of confidence, like aircraft trajectory prediction, security systems. The most common approach to estimate these prediction intervals is to use statistical inference to calculate confidence intervals on the error's variable.

However, this corresponds to a confidence interval about the error and not about the real prediction value. Another disadvantage of this approach is to assume that the prediction interval sizes are constant and so independent of the considered instance.

It is known that taking a quantile of an estimated distribution using a sample set having a limited number of instance is not a statistically approved way to infer an interval that must contain a desired proportion of the true unknown distribution that might have generated this sample set. This is because the estimation procedure did not take into account the uncertainty related due to the limited sample size used for estimating the distribution. In statistics, there are already many ways to build confidence intervals around a prediction variable like prediction intervals, tolerance intervals, confidence intervals of quantiles and simultaneous confidence intervals of quantiles, etc. and each of them has its own properties. In this work, we are interested in intervals which have similar properties to the simultaneous tolerance intervals for regression. Simultaneous interval regression, introduced in this paper, could be seen as a form of confidence envelop for the real value of the prediction variable Y given any value of predictor variables in their domain $x \in \mathcal{X}$. This concept is similar to simultaneous tolerance interval for regression with an high confidence level $\gamma \approx 1$. This type of approach is different to quantile regression introduced by Koenker and Bassett (1978) [8] in which conditional quantile of the response variable Y given the predictor values $X = x$ is calculated. Quantile regression is much more flexible than least square regression when dealing with heterogeneous conditional distributions, because it makes no distributional assumption about the error term in the model and just provide a conditional distribution of the prediction given the predictor values [9]. Quantile regression focus on the direct estimation of regression quantile and ignores the uncertainty related to the limited number of observations. Some authors have already treated the problem of confidence intervals for regression quantiles [7, 2, 6], however they always focus to find confidence interval on the regression parameters and not on the prediction variable. One might think to use the confidence interval on the regression parameters in quantile regression to derive intervals on the conditional quantile. However, The major difference of these derived confidence intervals with tolerance intervals in least square regression is that they are not two-sided, but just a one-sided confidence interval on the prediction variable. Another fundamental difference with simultaneous tolerance interval in least square regression is that they are confidence intervals of regression quantiles given $X = x$ and not simultaneous on the entire domain of independent variables.

Simultaneous tolerance intervals have already been treated by several authors [11, 19, 12] for linear regression. These works are based on three assumptions. First, the error follows a normal distribution. Second, the mean is linear with respect to the input variable. Finally, the standard deviation around the mean is constant and independent with respect to the input variables (homoscedastic-

ity assumption). This paper tries to overcome the last two limitations by using tolerance interval and a non parametric regression method, however we still assume that the error distribution is normal. Thus, we define simultaneous interval regression for K-nearest neighbor. In simultaneous interval regression for KNN, the local tolerance interval is used in order to find the value of the parameter K that has the better trade-off between the precision and the uncertainty. Given a dataset and a desired proportion of response value β, the goal is to find the optimal combination of hyper-parameters (MIN_K, MAX_K and γ), for which the simultaneous condition on the obtained intervals of the underlying data set is satisfied. The interval construction approach is proposed for KNN in general. This method exploits the local density of the neighborhood to find the most appropriate intervals to contain the desired proportion of response values, so the proposed interval construction method may be more effective with heterogeneous data set with heteroscedastic error.

This paper is organized as follows: section 2 is a brief introduction of tolerance interval and simultaneous tolerance interval for least square regression. Section 3 is devoted to the description of our approach with KNN and in the last section, we apply our method on six benchmark regression databases.

2 Tolerance interval and least square regression

2.1 Tolerance interval

Let $X_1, \cdots, X_n$ denote a random sample from a continuous probability distribution and let $\mathbb{X} = (X_1, \cdots, X_n)$. A tolerance interval is an interval which guarantees with a specified confidence level γ, to contain a specified proportion β of the population. The $I^T_{\gamma,\beta}$ sign, is used to refer to a β-content γ-coverage tolerance interval [1]. Then, we have:

$$\forall \beta \in (0,1),\ P_{\mathbb{X}}(P(X \in I^T_{\gamma,\beta}|\mathbb{X}) \geq \beta) = \gamma \tag{1}$$

By making the assumption that our sample set comes from a univariate normal distribution, then the lower and the upper bound of the tolerance interval $I^T_{\gamma,\beta} = [X_l, X_u]$ for a sample of size n is obtained as follows :

$$X_l = \hat{\theta} - \mathbf{k}\hat{\sigma}, X_u = \hat{\theta} + \mathbf{k}\hat{\sigma} \tag{2}$$

$$\mathbf{k} = \sqrt{\frac{(n-1)(1+\frac{1}{n})Z^2_{1-\frac{1-\beta}{2}}}{\chi^2_{1-\gamma,n-1}}} \tag{3}$$

where $\hat{\theta}$ is the sample mean of the distribution, $\hat{\sigma}$ its sample standard deviation, $\chi^2_{1-\gamma,n-1}$ represents the $1-\gamma$ quantile of the chi-square distribution with $n-1$ degree of freedom and $Z^2_{1-\frac{1-\beta}{2}}$ is the squared of $(1-\frac{1-\beta}{2})$ quantile value of the standard normal distribution [5]. For more details on tolerance intervals see [1] .

Regression is the process of creating an approximating function that looks for capturing relevant patterns in the data. In a classical fixed design regression problem (parametric or nonparametric), there are m pairs $(x_i, y(x_i))$ of observation where x_i is a vector of input variables and $y(x_i)$ is the observed value of the response variable. It is usually supposed that the mean of the random variable $Y(x_i)$ follows an unknown function f^* with a random error term ε_i defined as:

$$Y(x_i) = f^*(x_i) + \varepsilon_i, \text{ where } E(\varepsilon_i) = 0. \tag{4}$$

Thus, the goal of regression is to learn from data a function f that is as close as possible to the unknown function f^*. In least square regression, it results to find the function that minimize the mean square of the error (MSE), i.e. find f that minimize :

$$MSE(f) = \frac{1}{m}\sum_{1}^{m}(y(x_i) - f(x_i))^2 \tag{5}$$

In the following, we will always assume that the error follows a normal distribution. A conventional approach employed by some practitioners is to assume that f is a non-biased estimator of f^* with $Var(\varepsilon_i) = \sigma^2$ being constant which means that it does not depends of the input vector (homoscedasticity assumption), and to use the MSE of the found f as an estimation of σ^2 (i.e. $MSE(f) = \hat{\sigma}^2$). Thus the conventional approach assumes that the error distribution normal and homoscedastic. In this approach inter-quantiles of the estimated normal distribution are used, as an approximate solution to find intervals that contain a chosen proportion of the underlying distribution for a given value of dependent variables or intervals that contain a chosen amount of underlying distribution for all the possible values of dependent variables, (respectively similar to tolerance intervals and simultaneous tolerance intervals). For instance, the 0.95 inter-quantile $[f(x) - 1.96\hat{\sigma}, f(x) + 1.96\hat{\sigma}]$ is often used as an interval that will contain 95% of the distribution of $Y(x)$ (i. e. as a regression tolerance interval). As shown by Wallis [17], this statement is not true since $\hat{\sigma}$ and $f(x)$ are only estimations of the true standard deviation σ and the true mean function f^*. These estimations are usually made on a finite sample and are then pervaded with uncertainty. Thus, tolerance intervals for least square regression have been introduced in order to take into account this uncertainty. These intervals are described formally by (6). We name such intervals, β-content γ-coverage regression tolerance intervals and they are represented by $I(x)^T_{\gamma,\beta}$.

$$\forall x,\ P\left(\int_{L_{\beta,\gamma}(x)}^{U_{\beta,\gamma}(x)} p_x(t)dt \geq \beta\right) \geq \gamma \text{ where } Y(x) = f^*(x) + \varepsilon_i \tag{6}$$

Where $p_x(t)$ represents the the probability density function of $Y(x)$ for an specified value of predictor variable x. It is important to observe that tolerance intervals in regression are defined separately for each input vector x. Therefore, for two different input vectors x_1 and x_2, $I(x_1)^T_{\gamma,\beta}$ and $I(x_2)^T_{\gamma,\beta}$ are different and the event of $Y(x_1) \in I(x_1)^T_{\gamma,\beta}$ is independent of $Y(x_2) \in I(x_2)^T_{\gamma,\beta}$.

2.2 Difference between tolerance and prediction intervals

One might think to use prediction intervals instead of tolerance intervals. Note that in terms of prediction, tolerance intervals are not the same as prediction intervals. For a given x, tolerance intervals contain at least $100\beta\%$ of the population of $Y(x)$, however a β prediction interval contains in mean $100\beta\%$ of the distribution of $Y(x)$. In other words, the expected percentage of the population of $Y(x)$ contained in its β prediction interval $I(x)_{\beta}^{Pred}$ is β. This is stated formally as follows:

$$\forall x, E(P(Y(x) \in I(x)_{\beta}^{Pred})) = \beta \ \textit{where}\ Y(x) = f^*(x) + \varepsilon_i \tag{7}$$

For a detailed discussion about the differences between prediction and tolerance intervals, the reader can find more in [4].

2.3 Simultaneous tolerance intervals for least square regression

As seen above, tolerance intervals for least square regression are point-wise intervals which are obtained separately for each vector of x. Lieberman and Miller [11] extended the Wallis [17] idea to the simultaneous case. Simultaneous tolerance intervals are constructed so that with confidence level γ, simultaneously for all possible values of input vector x, at least β proportion of the whole population of the response variable Y is contained in the obtained intervals. In fact simultaneous tolerance interval for least square regression $[LS_{\beta,\gamma}(x), US_{\beta,\gamma}(x)]$ create an envelope around the entire mean regression function $f(\cdot)$ such that for all $x \in \mathcal{X}$, the probability that $Y(x)$ is contained in $[LS_{\beta,\gamma}(x), US_{\beta,\gamma}(x)]$ is simultaneously β, and this coverage is guaranteed with a confidence level γ. We name such intervals, β-content γ-coverage simultaneous regression tolerance intervals, we represent them by $I(x)_{\gamma,\beta}^{TS}$ and they are described formally by Equation (8), where $p_x(t)$ represents the the probability density function of $Y(x)$ for an specified value of predictor variable x.

$$P\left(\min_{x\in\mathcal{X}}\left(\int_{LS_{\beta,\gamma}(x)}^{US_{\beta,\gamma}(x)} p_x(t)dt\right) \geq \beta\right) \geq \gamma \ \textit{where}\ Y(x) = f^*(x) + \varepsilon_i \tag{8}$$

These intervals have been studied for the linear regression by several authors [11, 19, 12]. For an introduction to the subject, the reader can see Lieberman and Miller [11]. They explained the problem in details and presented four different methods to construct such intervals for linear regression. For more information about simultaneous inference, see [1, 13].

3 Simultaneous interval regression for K-Nearest Neighbor (KNN)

3.1 K-Nearest Neighbor (KNN)

Non-parametric regression is a type of regression analysis in which the response value is not a predefined function of the predictor variables and vector of param-

eter θ which must be estimated form data. In contrary to parametric regression, which is based on the construction on a model based on a training set, the prediction for a vector x is made by a local estimation inside the training set. The motivation of non-parametric methods is their utility when dealing with too complex models or when having non-linear or linear with heteroscedastic data. Therefore, in such situations, exploiting the neighborhood of the input data to estimate the local distribution of response value may be justified. KNN uses the distribution of response values in the neighborhood of the input vector x to find its unknown response value. In this work, we assume that the local distributions are normal, and we will use tolerance intervals of normal distribution to obtain the required intervals. This section, makes also the general assumptions of fixed regression design described in the previous section. With KNN, which are linear smoothers, the estimated function for the input vector x, $f(x)$ will be defined as:

$$f(x) = \sum_{i=1}^{n} l_i y(x_i). \tag{9}$$

where l_i, is the weight associated to the observation $y(x_i)$. The computation of these weights, requires an unknown hyper-parameter named as the bandwidth. The bandwidth is the size of the neighborhood (K) around the considered input vector which is used to compute these weights. Then, KNN which is a kernel smoother is defined as follows :

$$f(x) = \frac{\sum_{i=1}^{n} Ker_b(d(x, x_i))y(x_i)}{\sum_{i=1}^{n} Ker_b(d(x, x_i))} \tag{10}$$

where

$$Ker_b(u) = \frac{1}{b} Ker(\frac{u}{b}),$$

$Ker(\cdot)$ is an a kernel function, $d(\cdot)$ is a distance function and b is the distance between the input vector x and its furthest K-nearest neighbor. In fact KNN is a specialized form of Nadaraya-Watson (NW) [14, 18] kernel estimator in which the bandwidth b is not constant and depends on the distance between input vector x and its furthest K-nearest neighbor. Usually, The size of the neighborhood, K, has to be fixed before the learning phase and it will be constant for all the input vectors. In the following, the neighborhood of x is denoted as :

$$Kset_x = \{(x_i, y(x_i)), d(x, x_i) \leq b\}.$$

Some of the common kernel functions are defined as belows [10] where $I(\cdot)$ is the indicator function:
Tricube: $K(u) = \frac{70}{81}(1 - |u|^3)^3\, I(|u| \leq 1)$,
Gaussian : $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, I(|u| \leq 1)$,
Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)\, I(|u| \leq 1)$,
Uniform: $K(u) = \frac{1}{2}\, I(|u| \leq 1)$.

3.2 KNN simultaneous interval regression

Our goal is to find intervals which contain simultaneously a proportion of the response values for all input instances $x \in \mathcal{X}$. We name the problem stated just above as simultaneous interval regression. This is similar to consider simultaneous tolerance interval for regression with a high confidence level $\gamma \approx 1$. The goal of this paper is to consider the simultaneous interval regression problem for KNN. The interval construction approach is proposed for KNN in general. This method exploits the local density of the neighborhood to find the most appropriate intervals to contain the desired proportion of response values, so the proposed interval construction method may be more effective with heterogeneous data set with heteroscedastic error. Note that, tolerance and simultaneous tolerance intervals have not yet been treated for non-parametric methods. Thus, given an input vector x, K, β, and γ, the tolerance interval for the response variable is computed by using Equation (2) with

$$\hat{\theta} = f(x), n = K$$

and

$$\sigma = (K-1)^{-1} \sum_{y(x_i) \in Kset_x} (y(x_i) - \bar{y})^2, \textit{ where } \bar{y} = K^{-1} \sum_{y(x_i) \in Kset_x} y(x_i).$$

Note that, in contrary to the sample mean, the sample standard deviation does not take into account the distance between the considered input vector and its neighbors. Indeed, if the weights l_i was embedded in the computation of σ, K would overestimate the amount of information used for the estimation of the standard derivation.

As pointed out previously, it is common to fix K and use a general KNN estimator. These settings are denoted as "KNN regression with fixed K". The fixed K idea in KNN regression comes from the assumption which suppose that the data are homogeneously distributed in the feature space. In this section tolerance intervals are used to find the "best" K for each input vector x. Let the sample set containing the response values of the K-nearest neighbors of x be $Kset_x$. For a fixed value of β, and for each input vector x, the computation begins by an initial value of K, then the β-content γ-coverage normal tolerance interval of $Kset_x$ is calculated. This process is repeated for the same input vector x but different values of $K, MIN_K \leq K \leq MAX_K$. Finally, for a given x, the interval having the smallest size between other tolerance intervals resulted by different $Kset_x$ for $MIN_K \leq K \leq MAX_K$ is chosen as the desired interval.

This leads us to choose the interval that has the best trade-off between the precision and the uncertainty to contain the response value. Indeed, when K decreases the neighborhood considered is more faithful but it increases the uncertainty of the estimation. On the contrary, when K increases, the neighborhood becomes less faithful but the size of the tolerance intervals decreases. In fact the mentioned intervals take into account the number of instances in the neighborhood, and their size reflects also the neighborhood's density . Thus, choosing the

K that minimizes a fixed β-content γ-coverage normal tolerance ensures to have the best trade off between the faithfulness of the neighborhood and the uncertainty of the prediction due to the sample size. This is summarized in Algorithm 1. For the case of the computational complexity, the computation process of KNN simultaneous interval regression is $(MAX_K - MIN_K)$ times higher than the complexity of KNN regression with fixed K. Because from the beginning to the $Kset_x$ finding step, everything remains the same for both of the regression methods, then in the interval calculation phase, KNN regression with fixed K computes just one interval and instead our method computes MAX_K ones. For more detail on the complexity of KNN see [15].

Algorithm 1 Simultaneous interval regression with KNN

for all $x \in testSet$ **do**
  $IntervalSize_{min} \leftarrow Inf$
  for all $i \in MIN_K, \ldots, MAX_K$ **do**
    $Kset_x \leftarrow$ response value of the K nearest instances to x
    $Interval \leftarrow \beta$-content γ-coverage normal tolerance interval of $Kset_x$
    if $size(Interval) \leq IntervalSize_{min}$ **then**
      $K \leftarrow i$
      $foundInterval \leftarrow Interval$
      $IntervalSize_{min} \leftarrow size(Interval)$
    end if
  end for
  $Interval_x \leftarrow foundInterval$
end for

3.3 Tunning MIN_K, MAX_K and γ

MIN_K and MAX_K are global limits that stop the search if the best K value is not before . This may occur when in some part of the data set, the local density of the response variable is relatively high. In practice, it is known that this kind of local density is not always a sign of similarity, therefore these two bounds serve to restrict the search process in a region where it is most likely to contain the best neighborhood of x. MIN_K ,MAX_K and γ are algorithms hyper-parameter and they can be found by evaluating the effectiveness of the algorithm on the training set.

Our goal is to find an envelop that gives β proportion of all the predictions. In this scope β is chosen with respect to the user expectation. Given a KNN function f and a validation set that contains m pairs $(x_i, y(x_i))$ the proportion of data inside the tolerance envelope is computed by the MIP function (Mean Inclusion Percentage) :

$$MIP = \frac{\sum I(y(x_i) \in I^T_{f(x)})}{m} \tag{11}$$

where I is the indicator function and $I^T_{f(x)}$ is the interval found by the algorithm above. The process of finding the optimal value of γ is more tricky. Indeed,

The choice of a good value γ is crucial in order to have simultaneous tolerance intervals that guarantee the expected value of MIP (i.e. $MIP \geq \beta$). High values of γ will guarantee that $MIP \geq \beta$ but the computed intervals can become very large. Thus,we experimentally search for the smallest value of γ that makes $MIP \geq \beta$. Note that, with this approach, the value of γ can be much lower that β and this may happen when the local density of the response values is quite dense.

4 Experiments

4.1 The experiment's approach

We compare the effectiveness of our methods based on tolerance intervals with the conventional interval construction approach described in section 2.1 (represented by "Conv.") for a given β proportion of the response values . Thus, the conventional approach is the computation of inter-quantile of population based on the classical KNN algorithm with a fixed K. The goal is to find simultaneous β-content regression intervals where $\beta = 0.9, 0.95$ and 0.99. The motivation of this choice of β is that these inter-quantiles are the most used ones in machine-learning and statistical hypothesis-testing. Another reason justifying our choice is that they are harder to approximate.

When considering the simultaneous interval regression, it is expected for the fraction of prediction values inside the envelope, for each of the 10 models in cross validation, to be greater or equal to β. For example, for $\beta = 0.95$ in a 10-fold cross validation, it is expected for each of the 10 built model to have a Mean Inclusion Percentage (MIP) greater or equal to 0.95 ($MIP \geq \beta$). In our experiments part, we are interested to compare the obtained intervals by the mentioned methods regardless to any variable selection or outliers detection preprocessing. The mentioned results are the mean inclusion percentages and the Mean of Interval Size (MIS) in each of the 10-fold in the cross validation scheme. The MIP (see Equation (11)) and MIS over all the 10-fold cross validation is also contained in the results.

In a first attempt, data set is divided into two parts of $\frac{2}{3}n$ and $\frac{1}{3}n$, where n represents the data set size. The part containing $\frac{2}{3}$ of instances are used to tune the hyper-parameters. The hyper-parameters are MIN_K, MAX_K and γ for our proposed interval regression method and just K value for the Fixed KNN (denoted as Conv.). For the classical KNN, the fixed K maximizing the Root Mean Squared Error (RMSE) of response variable is chosen. For our proposed method, the hyper-parameters having the smallest MIS and also satisfying the simultaneous β-inclusion constraint (see Section 3.3) are selected. Finally, all of the instances will serve to validate the results using a 10-cross validation scheme.

4.2 Results

For this purpose the following six well known regression data sets are used : "Auto MPG" (Auto) [3] , "Concrete Compressive Strength" (Concrete) [3],"Concrete Slump Test" [3] (Slump),"Housing" [3],"Wine Quality" [3] (Wine) (the red wine) and "Parkinsons Telemonitoring" [16]. "Parkinsons Telemonitoring" data set contains two regression variable named as "motor_UPDRS" and "total_UPDRS", so we consider it as two distinct datasets named respectively as "Parkinson1" (containing "total_UPDRS" values without "motor_UPDRS") and "Parkinson2" (containing "motor_UPDRS" values without "total_UPDRS").
Table 1 summarizes the application of the algorithm 1 ("Var. K") and the conventional interval construction approach combined with KNN ("Conv.") to the seven datasets seen above. For each 10-fold cross validation scheme, the following quality measures are computed:

- MFIP: Mean Fold Inclusion Percentage (value of the MIP for one fold). It must be greater or equal to the desired β for each of the 10 models build in the cross validation phase.
- Min(MFIP): minimum value of MFIP between all the 10 models. If we have $min(MFIP) < \beta$, *that represents the failures of the approach to cover the required* β *proportion of the response values* .

The column MIS is the Mean of Interval Size for all the 10 models and σ_{is} is the standard deviation of the interval size over the whole dataset. Note that σ_{is} is not defined for the conventional method because its interval size is constant over the entire data set. The star * sign appears when Min(MFIP) satisfies the requirement (i.e. $Min(MFIP) \geq \beta$). When only one of the two compared methods satisfies this requirement, the result is put in bold.

For $\beta = 0.9$ $\beta = 0.95$ and for the data sets Parkinson1 and Parkinson2 in table 1, we can see that our method gives smaller intervals than the Conv. approach. However in contrary to the Conv. approach, the mentioned intervals contain the required β proportion of the response values. It is usually, a difficult task to satisfy the requirement for $\beta = 0.99$ and it becomes even harder for small data sets. Because each fold contains $\frac{n}{10}$ of total instances, so one percent is equals to $\frac{n}{1000}$. It means that the constructed intervals must miss at max $\frac{n}{1000}$ of total instances and this is a quite hard task for small and even medium data sets. But we can see that our method satisfies this condition for half of datasets and the mean of inferred intervals has not a big size compared to the required constraint. It is also interesting to note that our proposed method performs better in general for bigger datasets. This is because, our method is based on the local density of the data.

5 Conclusion

In this work, we have defined the idea of simultaneous interval regression for K-Nearest Neighbor (KNN). In simultaneous interval regression, the goal is to

Dataset	Algo.	90%		95%		99%	
		Min(MFIP)	$\overline{MIS}$ (σ_{is})	Min(MFIP)	$\overline{MIS}$ (σ_{is})	Min(MFIP)	$\overline{MIS}$ (σ_{is})
Parkinson1 (n=5875, p=25)	Conv.	94.54 *	6.62	94.55	7.88	95.4	10.36
	Var. K	90.98 *	5.01 (6.92)	**95.23 ***	6.38 (8.75)	**99.14 ***	11.19 (14.47)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (5, 40, 0.25)$		$(MIN_K, MAX_K, \gamma) = (5, 40, 0.35)$		$(MIN_K, MAX_K, \gamma) = (5, 40, 0.8)$	
Parkinson2 (n=5875, p=25)	Conv.	94.04 *	4.73	94.55	5.64	95.57	7.41
	Var. K	92.34 *	3.97 (5.37)	**95.23 ***	5.06 (6.77)	**99.14 ***	9.37 (11.85)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (5, 25, 0.3)$		$(MIN_K, MAX_K, \gamma) = (5, 25, 0.4)$		$(MIN_K, MAX_K, \gamma) = (5, 25, 0.87)$	
Wine (n=4898, p=12)	Conv.	78.93	1.84	90.59	2.19	93.46	2.88
	Var. K	**90.2 ***	2.5 (0.55)	**95.71 ***	3.51 (1.48)	98.77	5.04 (1.05)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20, 50, 0.9)$		$(MIN_K, MAX_K, \gamma) = (5, 25, 0.99)$		$(MIN_K, MAX_K, \gamma) = (20, 50, 0.999)$	
Concrete (n=1030, p=9)	Conv.	80.58	25.58	86.4	30.48	94.17	40.05
	Var. K	**91.26 ***	33.29 (11.86)	**95.14 ***	41.91 (14.8)	**99.02 ***	80.72 (26.47)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (10, 25, 0.6)$		$(MIN_K, MAX_K, \gamma) = (10, 25, 0.7)$		$(MIN_K, MAX_K, \gamma) = (10, 25, 0.99)$	
Auto (n=398, p=8)	Conv.	87.17	9.96	90	11.87	94.87	15.6
	Var. K	**94.87 ***	12.57 (6.48)	**95 ***	14.98 (7.72)	97.43	23.54 (11.98)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (7, 20, 0.95)$		$(MIN_K, MAX_K, \gamma) = (7, 20, 0.95)$		$(MIN_K, MAX_K, \gamma) = (7, 20, 0.99)$	
Housing (n=506, p=14)	Conv.	84.31	14.23	90.19	16.96	94	22.29
	Var. K	**92.15 ***	22.9 (13.09)	**96 ***	27.28 (15.6)	98	43.45 (24.44)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (10, 20, 0.99)$		$(MIN_K, MAX_K, \gamma) = (10, 20, 0.99)$		$(MIN_K, MAX_K, \gamma) = (10, 20, 0.999)$	
Slump (n=103, p=10)	Conv.	80	12.73	80	15.16	80	19.93
	Var. K	**90 ***	29.58 (9.83)	90	35.25 (11.71)	**100 ***	46.32 (15.4)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (5, 15, 0.99)$		$(MIN_K, MAX_K, \gamma) = (5, 15, 0.99)$		$(MIN_K, MAX_K, \gamma) = (5, 15, 0.99)$	

Table 1. Comparing the interval construction approachs proposed to perform simultaneous interval regression for KNN.

find intervals which simultaneously contain a required proportion of the response values for all input instances. We have introduced one approach to build such intervals which can be applied to KNN. Since tolerance intervals take into account the neighborhood size, it allows us to automatically find the best value of K for each example rather than using a fixed K for all the test set. This can be useful in presence of heterogeneous data. In the experiments part, the introduced methods and its conventional versions are applied on six different regression data sets. The results show that our approach performs very well on dense datasets. In the case of dataset with small sample sizes compared to their number of variables, our method is less reliable, but it is still better than the conventional interval construction method in KNN.

Predicting simultaneous confidence intervals may be useful when we are interested in the combination of predictions. For instance, in the wine database,

the computation of simultaneous interval ensures that for a set of "m" bottles, "$m * \beta$" of the bottles will have simultaneously their score in their predicted interval. This can become more important in safety and security applications. As another example, let us take the aircraft trajectory prediction using a regression model. A warning occurs when the prediction intervals of two or more aircraft overlap. In this case, intervals found using simultaneous interval regression guarantee the safety measure of the collision detection approach.
As future work, we will focus on different non-parametric estimators such as Locally Weighed Scatter-plot Smoothing as well as regression cases where the error's distribution are not normal.

References

1. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley, 2009.
2. D. D. Boos. A new method for constructing approximate confidence intervals from m estimates. *J. Amer. Statistical Assoc.*, 75(369):pp. 142–145, 1980.
3. A. Frank and A. Asuncion. UCI machine learning repository, 2010.
4. G. J. Hahn and W. Q. Meeker. *Statistical Intervals: A Guide for Practitioners*. John Wiley and Sons, 1991.
5. W. G. Howe. Two-sided tolerance limits for normal populations, some improvements. *J. Amer. Statistical Assoc.*, 64(326):610–620, 1969.
6. M. Kocherginsky, X. He, and Y. Mu. Practical confidence intervals for regression quantiles. *J. Comp. and Graph. Stat.*, 14(1):pp. 41–55, 2005.
7. R. Koenker. Confidence intervals for regression quantiles. In *Proc. of 5th Symp. Asymp. Stat.*, 1994.
8. R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):pp. 33–50, 1978.
9. R. Koenker and K. Hallock. Quantile Regression: An Introduction. *J. Economic Perspectives*, 15(4):43–56, 2001.
10. Q. Li and J.S. Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.
11. G. J. Lieberman and Jr. Miller, R. G. Simultaneous tolerance intervals in regression. *Biometrika*, 50(1/2):pp. 155–168, 1963.
12. R. W. Mee, K. R. Eberhardt, and C. P. Reeve. Calibration and simultaneous tolerance intervals for regression. *Technometrics*, 33(2):pp. 211–219, 1991.
13. R. G. Miller. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 1991.
14. E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.
15. B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London ; New York, 1986.
16. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. 2009.
17. W. A. Wallis. Tolerance intervals for linear regression. In *Proc. Second Berkeley Symp. on Math. Statist. and Prob.*, pages pp. 43–51. Univ. of Calif. Press, 1951.
18. G. S. Watson. Smooth regression analysis. *Sankhyā Ser.*, 26:359–372, 1964.
19. A. L. Wilson. An approach to simultaneous tolerance intervals in regression. *Ann. of Math. Stat.*, 38(5):pp. 1536–1540, 1967.