Improving reliability of aggregation, numerical simulation and analysis of complex systems by empirical data

Boris S Dobronets and Olga A Popova

Institute of Space and Information Technology, Siberian Federal University,
Kirenskogo 26, Krasnoyarsk, 660074, Russia

E-mail: address BDobronets@yandex.ru

Abstract. The paper considers a new approach of regression modeling that uses aggregated data presented in the form of density functions. Approaches to Improving the reliability of aggregation of empirical data are considered: improving accuracy and estimating errors. We discuss the procedures of data aggregation as a preprocessing stage for subsequent to regression modeling. An important feature of study is demonstration of the way how represent the aggregated data. It is proposed to use piecewise polynomial models, including spline aggregate functions. We show that the proposed approach to data aggregation can be interpreted as the frequency distribution. To study its properties density function concept is used.Various types of mathematical models of data aggregation are discussed. For the construction of regression models, it is proposed to use data representation procedures based on piecewise polynomial models. New approaches to modeling functional dependencies based on spline aggregations are proposed.

1. Introduction

The linear regression analysis is often used to discovery dependencies of the empirical data. Properties of empirical data including the type and level of uncertainty significantly affect on the results of the simulation. It is well-known, that the random uncertainty concept addressed by probability theory plays a fundamental role to study data uncertainty [1, 2].

Aggregation is quite a popular method of converting big data [3-6]. For example, the application of the histogram allows to reduce dimension of the data set and level of uncertainty and to significantly increase the efficiency of numerical calculations. It is important to note that the histograms are examples of the symbolic data using in the Symbolic data analysis [1, 4, 7].

Symbolic Data Analysis and Data Mining use the histograms to study a variety of different processes and are applied to modeling the variability of quantitative characteristics.

Histogram data models and histogram regression models based on the Symbolic analysis is a new important direction to discover knowledge in a data base. Billard L., Diday E. proposed the symbolic data type named histogram-valued variables to employ for regression modeling [1, 7].

This problem becomes more complicated if large amounts of data are processed. In this case it is useful to look at the empirical data in an aggregated form. Aggregation is a popular method of converting data. For example, the application of the histogram allows to reduce dimension of data set and level of uncertainty and to significantly increase the efficiency of numerical calculations. It is important to note that histograms are the examples of symbolic data used in the Symbolic data analysis

[7]. Symbolic Data Analysis and Data Mining use histograms to study a variety of different processes and are applied for modeling the variability of quantitative characteristics. Histogram data models and histogram regression models based on the Symbolic analysis is a new important direction to discover knowledge in a data base.

In our study we consider a new approach to regression modeling using input data aggregation. To develop our approach for performing efficient aggregation we employ piecewise polynomial aggregation functions, including piecewise linear functions and piecewise constant functions. Histogram is a good example of piecewise constant functions of which are perfectly employed in our study.

To examine the structure of data aggregation we use the probability density functions (PDF). The concept of the mathematical aggregation functions is used to the regression modeling. To illustrate this we will regard the spline aggregation function in more detail. This approach will allow to employ the density function models as input and output data. It is of further importance that, the data uncertainty is studied to identify the relationship between the input and output characteristics when the input probability density functions are unknown. Thus, in order to describe any specific PDF we need to consider their spline interpolation.

In this work we propose a new linear regression model named a PDF-valued variable regression. The abbreviated form of the regression model is called a Distributions Regression. If we use a spline aggregation model it is named a PDF-spline valued variable Regression Model and a Distributions Regression in shot. The following statements confirm the justification of PDF-spline models. The application of the spline procedures allows big data aggregating reduce the level of uncertainty and to significantly increase the efficiency of numerical calculations. These splines allow considerably accurate representions of the arbitrary distribution.

To demonstrate the degree of the relevance of the proposed methods to reality, we developed a theoretical study and provided numerical examples to illustrate it. With this we propose a conclusive discussion of this approach applicability to the uncertainty treatment and big data processing. The comparison of NPA and Monte Carlo method showed good agreement of the results. At the same time, numerical experiments demonstrate that the PDF arithmetic is more than hundred times faster than the Monte Carlo method [8-10]. As a result, the NPA approach can be successfully applied for solving computational and engineering problems [9,10,11].

2. Reliable wstimates of the accuracy of aggregation

In those cases where data can be interpreted as frequency distributions, it is advisable to use methods for constructing the probability density function for aggregation. It is remarkable that the histogram stood as the only nonparametric density estimator until the 1950s, when substantial and simultaneous progress was made in density estimation and in spectral density estimation. During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt, Parzen, and Cencov [12].

Next, consider using Richardson's extrapolation to improve the accuracy of the kernel estimator [13].

The basic kernel estimator may be written compactly as [12]

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x-\xi_i}{h}) = \frac{1}{Nh}\sum_{i=1}^{N} K_h(x-\xi_i),$$

where $K_h(t) = K(t/h)/h$.

Note

$$K_h(x,\xi_i) = K(\frac{x-\xi_i}{h}).$$

where ξ is a random variable with probability density function $f(x)$.
Then expected value

$$\mathrm{E}[\hat{f}(x)] = \mathrm{E}[K_h(x,\xi)]$$

and variability

$$\sigma_N = \mathrm{Var}[\hat{f}(x)] = \frac{1}{N}\mathrm{Var}[K_h(x,\xi)].$$

Suppose that the kernel K satisfies the requirements

$$\int_{-\infty}^{\infty} K(\eta)d\eta = 1, \int_{-\infty}^{\infty} \eta K(\eta)d\eta = 0$$

and

$$\int_{-\infty}^{\infty} \eta^3 K(\eta)d\eta = 0.$$

Denote

$$\int_{-\infty}^{\infty} \eta^2 K(\eta)d\eta = \sigma^2.$$

Define $f^h(x)$ as follows

$$f^h(x) = \mathrm{E}[\hat{f}_h(x)] = f(x) + \sigma^2 h^2 f''(x)/2 + O(h^4). \tag{1}$$

and $f^{2h}(x)$ as

$$f^{2h}(x) = \mathrm{E}[\hat{f}_{2h}(x)] = f(x) + 4\sigma^2 h^2 f''(x)/2 + O(h^4). \tag{2}$$

Let we apply the Richardson's extrapolation to $f^h(x)$ and $f^{2h}(x)$ [7]. In the next stage, we multiply (1) on 1/4 to subtract the result from (2). Excluding $\sigma^2 h^2 f''(x)/2$ from (4) and (5), we get

$$f(x) = \frac{4}{3} f^h(x) - \frac{1}{3} f^{2h}(x) + O(h^4).$$

Noting that we have constructed the approximation to the function $f(x)$

$$f_{cor}^h(x) = \frac{4}{3}\hat{f}^h(x) - \frac{1}{3}\hat{f}^{2h}(x). \tag{3}$$

with the accuracy $O(h^4)$.

In figure 1 we represent numerical example. The solid line is exact probability density function *f(x),* a − kernel estimator of probability density function, b − correction of kernel estimator function by Richardson's extrapolation.

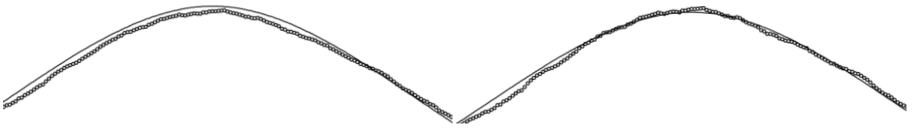
(a) (b)

Figure 1. Improving the accuracy of the probability density function estimation.

Thus, successively setting $z \in \omega$, we obtain the values $f_{cor}^{h}(x_i) = f(x_i) + O(h^4)$. Further, using the obtained values, we can construct systems of linear algebraic equations for constructing a cubic spline [14].

On the other hand, applying the Runge rule, we can obtain the estimate [11]

$$f''(x) = 2(f^h(x) - f^{2h}(x))/(3\sigma h^2) + O(h^2)$$

or

$$\| \hat{f}'' \| = \frac{2}{3\sigma h^2} \| \hat{f}^h - \hat{f}^{2h} \| \tag{4}$$

Thus, a posteriori estimate is constructed for the second derivative of the density function. This allows one to obtain an estimate for the accuracy of the approximations constructed.

3. Spline aggregation

A spline is a sufficiently smooth polynomial function that is piecewise-defined, and possesses a high degree of smoothness at the places where the polynomial pieces connect (which are known as knots). We will consider the probability density of the random variables as an approximated spline.

Prior to consideration of spline aggregation, we propose study of mathematical models applicable for the representation of splines and will discuss the interpolation questions with their application.

Let $\omega = \{x_0 < x_1 < x_2 < \ldots < x_n\}$ be mesh and interpolation conditions

$$s(x_i) = f(x_i), i = 0, \ldots, n.$$

Where boundary conditions are given as

$$s'(x_0) = 0, s'(x_n) = 0.$$

The cubic spline on a mesh $\{x_i\}$ with step $\bar{h} = \max(x_{i+1} - x_i), i = 0, \ldots, n$ satisfies the estimate

$$\| f^{\nu} - s^{\nu} \| \le \bar{h}^{4-\nu} \| f^{(4)} \|, \nu = 0, 1, 2. \tag{5}$$

The task of the spline construction is reduced for solving a system of linear algebraic equations with a tridiagonal matrix [14]

$$\begin{aligned} &\lambda_j m_{j-1} + 2m_j + \mu_j m_{j+1} = d_j, \\ &2m_0 + m_1 = 3(f_1 - f_0)/h_1 - h_1 z_0^2/2, \\ &2m_n + m_{n-1} = 3(f_n - f_{n-1})/h_n + h_n z_n^2/2, \\ &d_j = 3\lambda_j (f_j - f_{j-1})/h_j + 3\mu_j (f_{j+1} - f_j)/h_{j+1}, \quad j = 1, \ldots, N-1. \end{aligned} \tag{6}$$

where $m_i = s'(x_i)$. As a result, a cubic spline on the intervals $[x_{j-1}, x_j], j = 1, \ldots, N$ will have the following representation [8]:

$$s(x) = m_{j-1}(x_j - x)^2 (x - x_{j-1})/h_j^2 - m_j (x - x_{j-1})^2 (x_j - x)/h_j^2 +$$

$$+f_{j-1}(x_j-x)^2(2(x-x_{j-1})+h_j)/h_j^3++f_j(x-x_{j-1})^2(2(x_j-x)+h_j)/h_j^3,$$

Let consider the spline approach to build regression model with the Distributions-valued variables. This approach is useful due to the following reasons. Underlying of this approach is the notion of the spline. The spline can be regarded as a mathematical object that is easy to describe and calculate the mathematical procedures and operations, in the process of maintaining the essence of data frequency distribution.

Since the spline is a piecewise polynomial function then it can be regarded as a data aggregation function in aggregation issues. Aggregation function performs numerical calculations on a data set and returns the spline values. Splines are useful for data uncertainty analysis due to fact that they adequately represent the random distribution of random variables.

Despite its simplicity, the spline also covers all possible ranges of probability density function estimation. Simple and flexible spline structure greatly simplifies their use in numerical calculations and it has a clear visual image, which is useful for analytical conclusions. It is important to note that the construction of regression models with aggregated inputs require the use of appropriate numerical procedures. To this end, we consider numerical probabilistic analysis. We propose to use of the numerical probabilistic analysis to compute the arithmetic operations for the aggregated data and to apply for regression modeling.

Consider the problem of aggregating data by splines. For this purpose, we construct a spline *s* approximating the density function *f,* so that the estimate

$$\| f-s \| \leq Ch^4.$$

Thus, successively assuming $z \in \omega$, we obtain the values $f_{cor}^h(x_i)=f(x_i)+O(h^4)$ and the system of linear algebraic equations (8) for constructing a cubic spline. To improve the reconstruction accuracy of probability density at the point z we use the combination of kernel assessments with the parameters h and $2h$.

As an evidence we refer to Schweizer who states, that "distributions are the numbers of the future" [15]. Thus, instead of simplifying them, it seems better to propose methods which deal with distributions directly. In order to do this, one has to determine how to represent the observed distributions.

In our study we propose to represent them by using a piecewise polynomial aggregation function, as long as it offers a good tradeoff between simplicity and accuracy.

4. Distributions Regression

Consider a linear model

$$Y=a_0+\sum_{i=1}^{n}a_iX_i+\varepsilon,$$

where X_i, $i=1,\dots,n$ are independent predictor variables, Y is a dependent variable, ε is an error. From the observed values of Y_j $X_{i,j}$ after the aggregation of the density Y, X_i are represented by splines: S_y, S_i.

We shall seek the unknown parameters a_i, $i=0,1,\dots,n$ starting from the minimum of the functional

$$\Phi(a_0, a_1, \dots, a_n) = \| S_y - (a_0 + \sum_{i=1}^{n} a_i X_i) \|_2 \to \min .$$

By virtue of the independence of X_i, numerical operations on density functions can be used to calculate the functional $\Phi(a_0, a_1, \dots, a_n)$. The minimization of the functional $\Phi(a_0, a_1, \dots, a_n)$ can be carried out by the method of steepest descent.

Let us consider model problem n=2

$$Y = a_0 + \sum_{i=1}^{n} a_i X_i + \varepsilon,$$

For numerical realization, X_1, X_2 were generated as sums of random variables with an Irvine-Hall distribution n = 3 and shifted by 1 and 2, respectively, ε with probability density function $(|2x|-1)^2(2|2x|+1)$ with support $[-0.5, 0.5]$.

The variable Y was constructed as follows $Y = X_1 + X_2 + \varepsilon$.

The minimization of the functional $\Phi(a_0, a_1, a_2)$ was carried out by the method of steepest descent. For $a_0 = -0.089$, $a_1 = 1.031$, $a_2 = 1.029$, the value $\Phi(a_0, a_1, a_2)$ did not exceed the value $0.3 \cdot 10^{-3}$.

Thus, a numerical example showed the possibility of using distributions regression.

5. Conclusion

Although there are many ways of data aggregation, including simple average, we argue that the use of piecewise linear and piecewise polynomial aggregation functions will offer a more informative representation of the variability in the data, than other forms of data aggregation. To prove their thesis, we considered the aggregation procedure based on the histogram time series. Using these types of data aggregation for preprocessing and regression modeling you contribute to the reliability of the study of natural systems and processes. The spatial and time aggregation procedures help to reduce the amount of computation in data processing and are an important basis for the extraction of useful knowledge from large volumes of data. Developed methods reduce the level of uncertainty in the information flow; significantly reduce the processing time and the implementation of numerical procedures. This approach allows to the mode of interactive visual modeling to provide the necessary data for operational decision making under remote surveillance techniques and distributed object systems. In concluding the discussing about the applicability of this approach to practice we say about the advantage for uncertainty treatment and big data processing. Using the proposed model, applications with real and simulated data are presented.

References

[1] Dias S and Brito P 2015 *Linear Regression Model with Histogram-Valued Variables*. (wileyonlinelibrary.com). DOI:10.1002/sam.11260

[2] Koenker R 2005 *Quantile regression. Cambridge university press*

[3] Färe R, Grosskopf S and Primont D 2007 *Aggregation, Efficiency, and Measurement* (Springer, New York)

[4] Arroyoa J and Maté C 2009 *International Journal of Forecasting* 192–207

[5] Nava J 2012 *Journal of Uncertain Systems* vol 6 **2** 84–85

[6] Tchangani A 2013 *Journal of Uncertain Systems* **7(2)** 138–151

[7] Billard L and Diday E 2006 S*ymbolic Data Analysis: Conceptual Statistics and Data Mining.* (Wiley)

[8] Dobronets B S, Krantsevich A M and Krantsevich N M 2013 *J. Sib. Fed. Univ. Mathematics & Physics* **6(2)** 168–173

[9] Dobronets B and Popova O 2016 *Numerical Probabilistic Approach for Optimization Problems. Scientific Computing, Computer Arithmetic, and Validated Numerics. Lecture Notes in Computer Science* 9553 (Springer International Publishing) pp 43–53

[10] Dobronets B S and Popova O A 2014 *Reliable Computing* vol 19 274–289

[11] Dobronets B S and Popova O A *Journal of Siberian Federal University – Engineering & Technologies* **9**(7) 960-971

[12] Scott R W 2015 *Multivariate Density Estimation: Theory, Practice, and Visualization* (New York: John Wiley & Sons)

[13] Dobronets B S and Popova O A 2017 *Journal of Siberian Federal University – Mathematics & Physics* **10** (1) 16–21

[14] Ahlberg J H, Nilson E N and Walsh J L 1967 *The theory of Splines and Their Applications* (New York: Academic Press)

[15] Schweizer B 1984 *Distributions Are the Numbers of the Future, in Proceedings of The Mathematics of Fuzzy Systems Meeting* ed A di Nola and A Ventre, (Italy: University of Naples) pp 137–149.