Computational Probabilistic Analysis of Distributional Time Series

B S Dobronets and O A Popova
Institute of Space and Information Technology, Siberian Federal University, Kirenskogo 26, Krasnoyarsk, 660074 Russia
E-mail: B Dobronets@yandex.ru

Abstract. The article considers a new approach to forecasting of distributional time series based on computational probabilistic analysis. Functional data analysis and symbolic data analysis are currently used to study such data. A comparison of these approaches is given. Computational probabilistic analysis to forecasting of distributional time series uses special numerical operations on probability density functions. The article provides numerical examples of the analysis of distributional time series.

1. Introduction

With the development of big data, new opportunities have emerged in the development of methods for numerical data processing. An important place is occupied by the problem of representing data in the form of mathematical models for the purpose of their further application for problems of numerical modeling and forecasting. The essence of this approach is that data can be organized into objects that provide information that goes beyond a simple scalar capability.

We will consider the issues of forecasting distributional time series. Let the distribution functions be known at times \( t = 1, \ldots, n \), it is necessary to forecast distribution functions for times \( t = n + 1, \ldots \). Note that distributional time series are studied in functional data analysis as functional time series and in symbolic data analysis (SDA) \([4, 5]\). The FDA uses functional principal component analysis to explore various sources of the non stationarity of functional time series \([3]\).

We consider Functional time series consisting of probability density functions \( f_i(x), x \in D \subset R \). These probability density functions (PDF) will be assumed from the Hilbert space, for example, \( L^2 \). SDA \([5]\), also provides new research techniques in this direction.

Our approach is based on data aggregation procedures for input parameters, and the use of computational probability analysis (CPA) \([7, 8, 9]\). The transition to a more generalized representation using aggregation is necessary for several reasons. First, aggregation can significantly reduce the amount of data. Secondly, big data weakly reflect the general trends and properties of the studied set and is very variable due to the influence of various random factors and the spread of values. In this case aggregation allows to see the existing trends and patterns.

An example of a distributional time series is the histogram time series as a set of time-ordered histograms \([1]\). Developing this approach, we can talk about the distributional time series \([3]\). To construct a forecasting distributional time series, we proposed numerical operations on PDF
and applied Richardson extrapolation to increase the accuracy of constructing piecewise polynomial distribution functions [11].

The basis of CPA is the use of numeric operations over PDF. We use various types of PDF representation, including piecewise polynomial functions, which are set by grids of dimension \( n \) and values of functions at nodes [6, 7, 8].

The PDF of random variables \( x, y, z \) will be denoted by bold font \( \mathbf{x}, \mathbf{y}, \mathbf{z} \). Let us identify through \( \mathbf{R} \) the set of all probability density functions.

2. Aggregation
Consider a time series \((t_i, y_i)\) \( i = 1, 2, \ldots, N \). Further, regarding the random variables \( y_i, i = 1, 2, \ldots, N \), we assume that they have probability density functions \( b_{my_i} \) and \( y_i \), continuously depends on \( t \). Consider the construction of an approximation of the family \( y_i, t \in [t_0, t_N] \) and functional regression

\[
y(t) = \sum_{l=1}^{m} a_l \phi_l(t) + \varepsilon(t), t \in [t_0, t_N],
\]

where \( a_l \) are independent random variables, \( \varepsilon(t) \) are independent and identically distributed variables with zero mean and limited variance.

For some \( t_0 \in [t_0, t_N] \) and \( h > 0 \) on the set \( D_h = \{(t_i, y_i)|t_i \in [t_0 - h, t_0 + h]\} \) build a regression \( r(t), t \in [t_0 - h, t_0 + h] \).

Next, using nonparametric estimates for \( Z_h = \{z_i = y_i - r(t_i)|t_i \in [x_0 - h, x_0 + h]\} \), we construct an approximation \( \hat{y}^h(t_0) \approx y(t_0) \) [13]. The estimate \( \hat{y}^h(t_0)(\xi) : \)

\[
\hat{y}^h(t_0)(\xi) = \frac{1}{2h} \int_{t_0-h}^{t_0+h} y(t)(\xi)dt.
\]

In [11], the issues of increasing the accuracy of estimates are considered.

Note that when constructing the probability density functions \( y_{t_0} \), data with different probability density functions got into and

\[
y_{t_0}(\xi) = \hat{y}^h_{t_0}(\xi) + C h^2 + O(h^4),
\]

where \( C \) is constant independent of \( h \). To improve the accuracy, using Richardson extrapolation, we construct estimates for \( h \) and \( 2h \): \( \hat{y}^h \) and \( \hat{y}^{2h} \). Further

\[
y_{t_0}(\xi) = \frac{4}{3} \hat{y}^h_{t_0}(\xi) - \frac{1}{3} \hat{y}^{2h}_{t_0}(\xi) + O(h^4).
\]

If

\[
\hat{y}^h(t_0)(\xi) - \hat{y}^h(t_0)(\xi) \approx O(h^4)
\]

then

\[
y(t_0)(\xi) = \frac{4}{3} \hat{y}^h(t_0)(\xi) - \frac{1}{3} \hat{y}^{2h}(t_0)(\xi) + O(h^4).
\]

Consider the use of piecewise polynomial functions to approximate \( y_{t_0} \). The easiest way is to use piecewise linear functions.

**Piecewise linear functions** (frequency polygons). Let \( s \) be a piecewise linear approximation of \( y_{t_0} \). Then

\[
s(\xi) = \sum_{l=0}^{m} y_l \phi_l(\xi),
\]
where \( y_l = y_{\ell_0}(\xi_l) \), \( \varphi_l, l = 0, \ldots, m \) is a piecewise linear basis. Note that the following conditions are necessary

\[
s(\xi) \geq 0, \quad \xi_m \in \mathbb{R}, \quad \int_{\xi_0}^{\xi_m} s(\xi) d\xi = 1.\tag{1}
\]

The first condition (1) is satisfied by virtue of \( y_l \geq 0 \). If the condition (2) is not satisfied

\[
\int_{\xi_0}^{\xi_m} s(\xi) d\xi = r \neq 1,
\]

then

\[
\hat{s}(\xi) = \sum_{l=1}^{m} \frac{y_l}{\varphi_l(\xi)}.
\]

**Hermite spline.** Cubic Hermite spline, defined as

\[
s(x) = \sum_{i=0}^{n} \varphi_0((x - x_i)/h)f(x_i) + \varphi_1((x - x_i)/h)f'(x_i),
\]

where \( \omega = \{x_i = hi, i = 0, \ldots, n\} \) is uniform mesh with step \( h \).

A quintic Hermite spline, defined as

\[
s(x) = \sum_{i=0}^{n} \phi_0((x - x_i)/h)f(x_i) + \psi_1((x - x_i)/h)f'(x_i) + \psi_2((x - x_i)/h)f''(x_i),
\]

where \( \phi_i \in C^2 \) are basis functions and if \( |x| \geq 1 \) than \( \phi_i(x) \equiv 0 \). If \( |x| < 1 \) than

\[
\phi_0(x) = (1 - |x|)^3(6x^2 + 3|x| + 1),
\]

\[
\phi_1(x) = (1 - |x|)^3|x|(3|x| + 1),
\]

\[
\phi_2(x) = (1 - |x|)^3x^2.
\]

Restrictions on the spline coefficients for fulfilling the condition (1) are given in [2].

Let us consider an example of constructing an approximation of the probability density function \( f \) of the annual temperature in the city of Krasnoyarsk using quintic Hermite splines. To do this, in the first stage kernel estimator with the parameter \( h = 1 \) were used. The approximation \( \hat{f}_i \) of the probability density function was constructed at the grid nodes \( \{x_i|l = 0, \ldots, N\} \). We assume that the support \( (f) = (x_0, x_N) \). The boundary conditions \( s(x_0) = 0, s'(x_0) = 0, s(x_N) = 0, s'(x_N) = 0 \). To construct a spline, we define the grid \( \{\xi_i \in (x_0, x_N), i = 0, \ldots, n\} \), we find the unknown values \( f_i, f'_i, f''_i \), using the least squares method

\[
\sum_{l=1}^{N} (s(x_l) - \hat{f}_l)^2 \rightarrow \min.
\]

Note that the number and location of nodes is regularization.

Figure 1 shows the construction of an approximation of the probability density function of temperature for 1960. Points are kernel estimator of the probability density function, solid line is Quintic Hermitian spline \( s \) with spline grid nodes \( \omega = \{-40, -10, -1, 23, 36\} \), boundary conditions are

\[
s(-40) = 0, s(36) = 0,
\]

\[
s'(-40) = 0, s'(36) = 0.
\]
3. Operations on probability density functions

The issues of forecasting time series of distributions were considered in works [1, 12]. In [1] used the barycenter method for forecasting. Figure 2 compares the barycenter and probabilistic operations on two triangular probability density functions. In [12] shows the need for special operations development. In particular, the sum \( tf_1 + (1-t)f_2 \) of two triangular probability density functions must be triangular as shown in Figure 3.

![Figure 2](image2.png)

**Figure 2.** Comparison of barycenter (1) and probabilistic (2) operations

Next we study the time series of the distributions \( y_i, i = 1, 2, \ldots, n \), where the probability density functions \( y_i \) are piecewise polynomial functions with supports \([a_i, b_i]\). We will build the grids based on the features of the probability density function: we will take into account the boundaries of carriers, points of local extrema, etc

\[
\omega_i = \{x_i^0 = a_i, \ldots, x_i^m = b_i\}.
\]

On the sets \( X_j = \{x_j^i, i = 1, 2, \ldots, n\} \), we construct the regression curves \( r_j(t) \). Further, let \( S(t) \) be the transformation of the set \( \{r_j(t), j = 0, 1, \ldots, m\} \) into \( \omega = \{x_0 = a, \ldots, x_m = b\} \)

Piecewise polynomial functions \( y_i \) and grids \( \omega_i \) we transform using \( S(t) \) in the function \( s(t) \) on the grid \( \omega \). Consider, for example, the use of fifth degree Hermitian splines. Then for each node \( x_l \in \omega \) the values \( s_{li}, s'_{li}, s''_{li} \) are known and we can construct the regression curves \( s_l(t), s'_l(t), s''_l(t) \). Using these regression curves, it is possible to construct piecewise polynomial functions \( s(t) \) on the grid \( \omega \) for any \( t \). Then, using the inverse transformation \( S^{-1}(t) \), construct the value \( \bar{y}(t) \) for any \( t \). Note that in this case it is necessary to check the conditions (1), (2).

Figure 3 shows the results of linear combinations

\[
f_s = tf_1 + (1-t)f_2, \quad t \in (0,1)
\]
Figure 3. Linear combinations of two triangular distribution

represented by lines 1, dashed lines 2 show the boundaries of supports and extrema of the probability density functions.

**Model example.** For $t_i = ih$ we generate random variables $y_i$:

$$y_i = ((4 - t_i) t_i / 8 + 1/4) + (x_1 + x_2 + x_3 + x_4)/(4 + 4t_i),$$

where $N = 80000$, $h = 1/N$, $x_i, i = 1, 2, 3, 4$ are uniform random variables. The results of modeling are shown in figure 4.

Figure 4. Initial data

Figure 5. Approximation of the family $y$

In figure 5 shades of gray are represented the approximations of the $y$ family. Red solid lines show the boundaries of the supports and the mathematical expectation of $y(x)$.

4. Numerical example

In figure 6 the blue lines 1 show the piecewise polynomial approximations of temperature probability density functions for 1960, 1970, 1980 and 1990. The red lines 2 are the grid of nodes $\omega = \{-40, -10, -1, 23, 36\}$. For approximations, quintic Hermite splines were used.

Thus, knowing the values of $s_i(t)$, $s'_i(t)$, $s''_i(t)$ of the constructed splines at the grid nodes, we can predicting probability densities.

Consider the construction of a time series of temperature probability densities by days. To do this, we will use the data on the temperature in the city of Krasnoyarsk for the last 70 years. To plot the temperature density function, for example, April 15, we select the temperature data
from April 1 to April 30. Further, using nonparametric estimates, we construct estimates for the probability density function. To construct a piecewise polynomial approximation, Hermitian splines of the fifth degree are used. In Figure 4, dashed line 1 shows the probability density functions of temperature from 15.04 to 15.06, blue line 2 shows the regression on the constructed probability density functions of temperature. Red dotted lines 3 show the boundaries of supports and maxima of the regression.

Using the regression, the forecast for June 7 and 15 is built. Figure 7 compares the forecast and the probability density function for June 15th.

The above studies of the use of distributional time series have shown the prospects for the development of this direction. The distributional time series are supposed to be used in future in those areas where the usual time series do not adequately describe the processes that are taking place, are not efficient, and require large computational costs. Such processes arise during the processing of Earth sounding information, in forecasts of hydrological series and econometrics.

References
[1] Arroyo J, Gonzalez-Rivera G and Maté C 2011 Smoothing Methods for Histogram-valued Time Series: An Application to Value-at-Risk Statistical Analysis and Data Mining. 4 216–228
[2] Dougherty R, Edelman A and Source J Nonnegativity-, Monotonicity-, or Convexity-Preserving Cubic and Quintic Hermite Interpolation Mathematics of Computation 52186 pp 471–494
[3] Chang Y, Kim C and Park J 2016 Nonstationarity in time series of state densities Journal of Econometrics 192 pp 152–167
[4] Ramsay J and Silverman B 2005 Functional Data Analysis (Springer 2nd ed., New York)
[5] Billard L and Diday E 2006 Symbolic Data Analysis: Conceptual Statistics and Data Mining (West Sussex)
[6] Dobronets B and Popova O 2017 Improving the accuracy of the probability density function estimation Journal of Siberian Federal University, Mathematics and Physics 10(1) pp 16–21.
[7] Dobronets B and Popova O 2018 Piecewise polynomial aggregation as preprocessing for data numerical modeling, IOP Conf. Series: Journal of Physics: Conf. Series 1015 032028
[8] Dobronets B and Popova O 2018 Improving reliability of aggregation, numerical simulation and analysis of complex systems by empirical data IOP Conf. Series: Materials Science and Engineering 354 012006
[9] Dobronets B and Popova O 2020 Computational probabilistic analysis: models and methods (Krasnoyarsk: Siberian Federal University) (In Russian)
[10] Popova O 2017 Application of numerical probabilistic analysis in interpolation problems Computational Technologies 222 pp 99–114 (In Russian)
[11] Popova O 2019 Using Richardson extrapolation to improve the accuracy of processing and analyzing empirical data Measurement Techniques 62(2)
[12] Popova O, Dobronets B and Merko A 2020 Distributional Time Series For Forecasting And Risk Assessment
Modeling and Analysis of Safety and Risk in Complex Systems: Proceedings of the Sixteenth International Scientific Conference MASR — 2020 (Saint-Petersburg, Russia) pp 122–127

[13] Scott D 2015 *Multivariate density estimation: theory, practice, and visualization* (Hoboken, New Jersey: John Wiley & Sons)