Modelling of nonlinear filtering Poisson time series

Modelling of nonlinear filtering Poisson time series

Vladimir V Bochkarev and Inna A Belashova

Kazan Federal University, Russia

E-mail: inkin91-91@mail.ru

Abstract. In this article, algorithms of non-linear filtering of Poisson time series are tested using statistical modelling. The objective is to find a representation of a time series as a wavelet series with a small number of non-linear coefficients, which allows distinguishing statistically significant details. There are well-known efficient algorithms of non-linear wavelet filtering for the case when the values of a time series have a normal distribution. However, if the distribution is not normal, good results can be expected using the maximum likelihood estimations. The filtration is studied according to the criterion of maximum likelihood by the example of Poisson time series. For direct optimisation of the likelihood function, different stochastic (genetic algorithms, annealing method) and deterministic optimization algorithms are used. Testing of the algorithm using both simulated series and empirical data (series of rare words frequencies according to the Google Books Ngram data were used) showed that filtering based on the criterion of maximum likelihood has a great advantage over well-known algorithms for the case of Poisson series. Also, the most perspective methods of optimisation were selected for this problem.

1. Introduction

Widespread methods of time series filtering, especially wavelet thresholding, involve usage of data in which probability of fluctuations distribution is subject to the normal law [1]. Algorithms of wavelet thresholding are based on data dimension reduction strategy. It is assumed that to describe information contained in the series, it's enough to have several nonzero wavelet coefficients, which number is considerably smaller than the length of the series.

The minimum square error criterion is used for series with a normal distribution of variations. This results in a well-established strategy of data filtering. It implies that wavelet coefficients, which modulus values are below a certain threshold, are set to nil. If the distribution of variations differs significantly from the normal one, this strategy cannot produce good results. In this case, more general criterion of maximum likelihood is normally used to select non-zero coefficients, as well as their numerical values.

For data, which distribution differs from the normal one, the likelihood function depends on the parameters in a complicated way. Therefore, to find the parameters, we use numerical optimization methods such as genetic algorithms, simulated annealing method and deterministic algorithms (method confidence regions) [2,3].

In this article, we examined filtration of time series based on the criterion of maximum likelihood by the example of the series, which distribution is governed by Poisson's law. Poisson's law is introduced as a universal distribution for a number of independent rare events and is widely used for solving practical problems. The amount of network treatment, the number of device failures, the

number of cases of rare diseases, the frequency of rare words and many other values are distributed by Poisson's law.

One of the interesting practical applications can be visualization and analysis of time series of word usage frequencies obtained using diachronic text corpora, for example, Google Books Ngram. The Google Books Ngram (https://books.google.com/ngrams) corpus, based on the largest electronic collection of texts Google Books, allows obtaining time series of word usage frequencies for 8 languages for the last five centuries. The project focuses primarily on the study of changes in the language and culture [5]. The online service Google Books Ngram Viewer is available on the project page. It allows visualization of time series of the selected words frequencies, using anti-aliasing by means of the moving average for better visual representation. Using the moving average can result in a distorted view of rapid variations of frequencies, jumps and bursts associated with certain historical events. Best results are obtained using nonlinear wavelet filters. However, to apply them to the series of rare words frequencies, they need to be adapted to the Poisson distribution.

This work has two objectives:

- To check how far significant benefit can be obtained from filtering of Poisson time series based on the criterion of maximum likelihood;
- To select the most appropriate algorithms to optimize the likelihood function.

To assess the quality of filtering, we use both simulated test Poisson series with known parameters and actual series of rare words frequencies. The quality of filtration for the series of empirical data can be evaluated on the basis of qualitative considerations that take into account the Poisson distribution properties.

2. Method

Likelihood function algorithm is used as a quantitative measure of approximation quality for a time series with the Poisson distribution [4]:

$$llh = const + \sum_t \{x_t \log \lambda_t - \lambda_t\}$$

Here λ_t is a mathematical expectation of the process at time t. The objective is to reduce dimension of the investigated time series by presenting the sequence λ_t as a wavelet series with a small number of items. It can be shown as follows:

$$\lambda_t \approx \sum_{(j,k)\in I} c_{jk}\psi_{jk}(t)$$

At that, the number of items in the formula (the content of set *I*) should be significantly smaller than the length of the series.

Two problems are to be solved: 1) to select the most suitable set of non-zero coefficients *I*; 2) to determine the optimal numerical values for these coefficients. The first task is more difficult than the second one because discrete optimisation should be performed with a great number of options. In this article, we tried to solve this problem using stochastic search algorithms suspenders (simulated annealing, genetic algorithms), as well as the greedy algorithm. The second problem can be easily solved using any known algorithm of non-linear optimisation.

Thus, the filtration algorithm includes the following steps:

- Selection of the initial set of non-zero coefficients;
- Defining a new set of non-zero coefficients;
- Varying the values of nonzero coefficients in order to minimize the quality function – the maximum likelihood function;
- Checking the termination condition and if it is not satisfied – going to the second step.

Let`s consider these steps in detail. Initially, the set of non-zero coefficients and the values of these coefficients can be selected in accordance with a known wavelet thresholding algorithms (in this

paper, we used the fixed form threshold). The final result of applying such approach will certainly be no worse (for the value of the likelihood function) than using the standard method.

The set of non-zero coefficients is given as a binary vector, where 1 corresponds to non-zero coefficient and 0 corresponds to zero coefficient. In each of the above algorithms, the rules of varying this vector include saving the number of units (i.e., non-zero wavelet coefficients). When using the genetic algorithms, the vector turns to be a chromosome, to which mutation and crossover operations are applied at each step. After that, numerical values are specified for all the obtained coefficients, the log-likelihood function is calculated and the selection is performed. The desired final result is a set of non-zero coefficients at which the value of the likelihood function is maximum.

We tested the following filtering methods:

- Standard wavelet-thresholding;
- Method with optimization of the log-likelihood function by simulated annealing;
- Method with finding the optimal values of coefficients determined by the search (the dog leg algorithm was used). At that, the selection of non-zero coefficients was not changed.
- Method with optimization of the log-likelihood using genetic algorithms. At that, the genetic algorithms were used both for selection a set of non-zero coefficients and location their optimal values;
- Method which is a combination of the two previous methods; In this case, the selection of the set of nonzero coefficients is performed using genetic algorithms and the numerical values are calculated using the dog leg algorithm (hereinafter, this option is called a "combined method");
- Method with optimization of log-likelihood using "greedy" algorithm.

Let`s consider the last method. As the below examples show, the best results are obtained by using the genetic algorithms and the combined method. However, if the data size is great, it will take a lot of time to determine the best set of non-linear coefficients. In order to speed up the filtering, the "greedy" algorithm was used. It's realized in the following way: first, we assess how far the log-likelihood function is changed as a result of zeroing of one or another coefficient. Then, at each step, we set to nil the coefficient, the removal of which changes the objective function least of all and update the values of the coefficients. The algorithm is realized when a predetermined number of non-zero coefficients is left.

3. Results and discussion

A comparative testing of different filtering algorithms was performed using statistical modelling. Poisson series with time-dependent parameter $\lambda_t$were simulated. In most cases, filtering using the criterion of maximum likelihood shows significantly better results compared with the standard wavelet thresholding. If various optimization algorithms are compared, the best results are obtained if the combined algorithm is used (the set I is selected by the genetic algorithm, and the values of the coefficients are found by the determined search). Figure 1 shows an example of the filtering result for the test case «Noisy Bumps» from the Wavelet Toolbox of the Matlab package. It should be noted that different filtering methods are compared using the same number of non-zero coefficients. As can be seen from this figure, filtering methods based on the maximum likelihood criterion described both broad and narrow peaks doing the task better than the standard wavelet thresholding. Special attention should be paid to the intervals 120-160 and 600-700. There are two relatively narrow peaks in the initial series λ_t. These peaks are not seen as two peaks but as a single broad peak when the standard wavelet-thresholding is used. However, these peaks are distinguished as two peaks if the combined filtering method is used. This can be important for real practice. For quantitative comparison, Table 1 shows values of log-likelihood for the given example using different filtering methods.

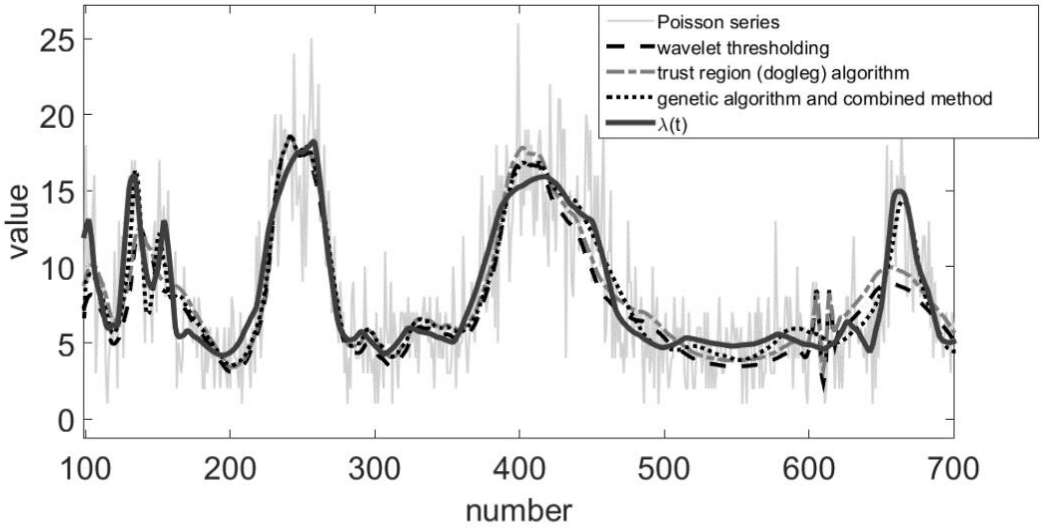

Figure 1 Comparison of the filtering results with different methods of the Poisson series (the example of "Noisy Bumps" from the Wavelet Toolbox package MATLAB)

Table 1 Comparison of the filtering results with different methods of the Poisson series (the example of "Noisy Bumps" from the Wavelet Toolbox package MATLAB)

Wavelet filtering	Genetic algorithms	Simulated annealing	Determined search	Combined method
$8.1178 \cdot 10^3$	$8.2146 \cdot 10^3$	$8.1178 \cdot 10^3$	$8.1643 \cdot 10^3$	$8.2147 \cdot 10^3$

If we take the largest and the smallest value of the log-likelihood function shown in the table and determine how many times one model is more probable than the other, it turns out to be a great value - 1.211 • 1042 times. Thus, using the likelihood function as a criterion for evaluation of filtering has significant advantage over the quadratic error functions for Poisson series. Good results of filtering the experimental data also confirm this.

As an example, Figure 2 shows the results of filtering of some frequencies for the English word «Karelian». It can be seen that the standard wavelet-thresholding cannot distinguish significant peaks associated with historical events. Even the biggest burst of 1939-1941 disappears (associated with the Soviet-Finnish war of 1939-1940 and 1941-1944). Attention should be paid to the unjustified extension of the transition interval in the second half of the 30s. Such disadvantages are not observed when filtering based on the criterion of maximum likelihood is used.

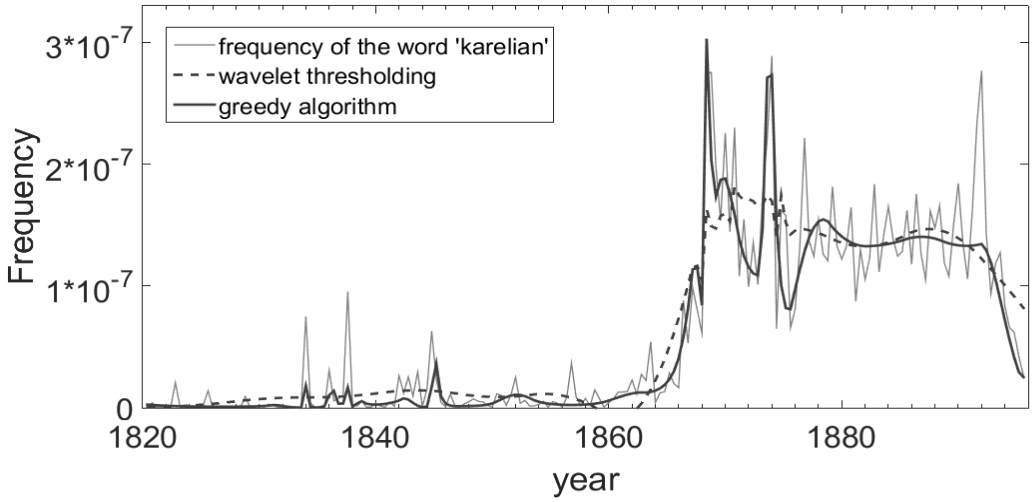

Figure 2 Comparison of the filtering results of the standard wavelet filtering and filtering based on the criterion of maximum likelihood (with optimization of the greedy algorithm). Frequency of use of the word «Karelian» is shown

Based on the qualitative assessment of the actual data filtering (frequency of use of rare words, the number of sunspots), it can be said that filtering based on the criterion of the maximum likelihood function (with optimization using a greedy algorithm) produces the best results, describing not only the trend, but also distinguishing the significant peaks.

4. Conclusion

The obtained results confirm that using the criterion of maximum likelihood has an advantage over minimum mean square error for the case of Poisson series filtering. Using genetic algorithms for selection of non-zero wavelet coefficients allows obtaining high-quality filtering, although it requires a significant amount of calculations. The use of the greedy algorithm can significantly reduce the calculation time with a slight deterioration in the quality.

5. Acknowledgments

This research was supported by the Russian Foundation for Basic Research (grant № 15-06-07402).

References

[1] S. Mallat A wavelet tour of signal processing. (Third Edition), Elsevier, 2009, 799 pp
[2] John R. Koza Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press, 1992
[3] W. Sun and Y.-x. Yuan, Optimization theory and methods : nonlinear programming. New York: Springer, 2006
[4] Yu.S. Maslennikova, V.V. Bochkarev, D.S. Voloskov Modelling of word usage frequency dynamics using artificial neural network // 2014 J. Phys.: Conf. Ser. 490 012180 (http://iopscience.iop.org/1742-6596/490/1/012180)
[5] Michel, J.-B., et al.: Quantitative analysis of culture using millions of digitized books. Science, 331,176–182 (2010)