SPRINGER BRIEFS IN PHYSICS

Călin Vamoş · Maria Crăciun

Automatic Trend Estimation

Springer

SpringerBriefs in Physics

For further volumes:
http://www.springer.com/series/8902

Călin Vamoş · Maria Crăciun

Automatic Trend Estimation

Springer

Călin Vamoş
"Tiberiu Popoviciu" Institute of Numerical Analysis
Romanian Academy
Cluj-Napoca
Romania

Maria Crăciun
"Tiberiu Popoviciu" Institute of Numerical Analysis
Romanian Academy
Cluj-Napoca
Romania

ISSN 2191-5423 ISSN 2191-5431 (electronic)
ISBN 978-94-007-4824-8 ISBN 978-94-007-4825-5 (eBook)
DOI 10.1007/978-94-007-4825-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012942924

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Huge amount of information is available as time series in many scientific fields: geophysics, astronomy, biophysics, quantitative finance, Internet traffic, etc. Processing so many time series is possible only by means of automatic algorithms usually designed in data mining. One of the critical tasks which has to be achieved by these algorithms is the automatic estimation of the trend contained in an arbitrary noisy time series. The aim of our book is to provide several automatic algorithms for nonmonotonic trend estimation. We do not intend to review the existing automatic trend estimation algorithms, but to present a thorough analysis for those presented in this book.

Obviously, an automatic algorithm is not able to work for all imaginable time series. By its automatic feature we mean that, without any subjective intervention, it efficiently processes time series of a well-defined type. The greater the diversity of the time series types, the more "automatic" the algorithm is. Therefore in designing a trend estimation algorithm an essential component is the method to evaluate its accuracy for a large diversity of time series. However, the algorithms are very often tested under unrealistic conditions and on too small number of time series. One reason for this situation is that the time series theory is dominated by stationary stochastic processes. The theoretical results for nonstationary time series containing a trend hold only under restrictive conditions, seldom satisfied by the real time series.

When the statistical theory is not applicable, Monte Carlo experiments can be used to evaluate the accuracy of the automatic algorithms. Even then the results are useful only if the members of the statistical ensemble have a diversity comparable with that of the real time series. The main difficulty is to generate realistic nonmonotonic trends. Usually, Monte Carlo simulations are performed on artificial time series much simpler than those encountered in practice, with monotonic (linear, power-law, exponential, and logarithmic) or periodic (sinusoidal) trends. The approach based on numerical Monte Carlo experiments in our book is much more general and the trends generated by our original algorithm are meaningful for real time series.

Chapter 1 contains fundamentals in probability theory, statistics, and time series theory which are used in the rest of the book. We analyze the autoregressive noise of order one denoted AR(1), which is a simple model depending only on two parameters: the variance and the constant of the serial correlation. Even for more complex noises an AR(1) model is a zero order approximation capturing their most important features. The noise serial correlation essentially influences the accuracy of the estimated trend because when it increases, the large-scale fluctuations of the noise cannot be distinguished from the trend variations.

In Chap. 2 we construct the statistical ensemble on which the Monte Carlo experiments are performed. There is no rigorous mathematical method to demonstrate that the variability of the obtained artificial time series is rich enough to simulate the variability of the real time series. In fact we construct an independent "numerical reality" on which we perform numerical experiments. Therefore, our approach is more typical to computational physics than to data mining or mathematical statistics. As examples of Monte Carlo experiments we evaluate the confidence interval for a method to estimate the serial correlation parameter of an AR(1) noise and we present a numerical method for testing if a time series is uncorrelated.

In Chaps. 3 and 4 we analyze in detail the accuracy of the classical algorithms of polynomial fitting and moving average in the case of arbitrary nonmonotonic trends. The quality of the estimated trend depends mainly on three parameters: the number of the time series values, the ratio between the amplitudes of the trend variations and the noise fluctuations, and the serial correlation of the noise. Our analysis shows that even in the case of the simplest trend estimation algorithms, due to the many parameters on which the artificial time series depend, a realistic evaluation of their performances is difficult and laborious.

In the last three chapters we present our original automatic algorithms for processing nonstationary time series containing a stationary noise superposed over a nonmonotonic trend. Their performances are tested by means of numerical experiments of the same type as those used in the previous chapters. The algorithms are designed to work on any time series, even if it has only a few values. Obviously, the best results are obtained for an AR(1) noise superposed over a deterministic trend with at least several hundreds of values. For other types of time series the outcomes of the algorithms have to be statistically analyzed by Monte Carlo experiments.

In Chap. 5 we design an automatic algorithm, called the averaged conditional displacement (ACD), to estimate a monotonic trend as a piecewise linear curve. The Monte Carlo experiments indicate that its accuracy is comparable with that of the classical methods, but it has the advantage to be automatic and to describe a much richer set of monotonic trend shapes. Applied to a time series with an arbitrary nonmonotonic trend, the ACD algorithm extracts one of the possible monotonic components which can be associated with the given trend. The probability that the estimated monotonic component is real can be estimated by a method based on surrogate time series.

In Chap. 6 we define the timescale of a local extremum of a time series such that it allows a classification of the local extrema with respect to their importance for the global shape of the time series. The local extrema with scales greater than a given value provide a partition of a noisy time series in segments which approximate the monotonic parts of the trend from a time series. The quality of this approximation is improved by first applying a moving average to the noisy time series. We use the monotonic component estimated by the ACD algorithm as a reference to measure the magnitude of the nonmonotonic variations of a time series. In this way we can build a criterion to stop the partition of a time series when the resulting segments may be considered monotonic.

In the last chapter we give an automatic form to the repeated central moving average (RCMA) analyzed in Chap. 4. In order to adjust the parameters of the RCMA algorithm to the characteristics of the processed time series, we have designed two simple statistical methods to estimate the noise serial correlation and the ratio between the amplitudes of the trend variations and of the noise fluctuations. The partitioning algorithm presented in Chap. 6 is used now to determine the local extrema of the estimated trend which corresponds to the real trend and not to the smoothed noise.

We illustrate the functioning of the analyzed algorithms by processing time series from astrophysics, finance, biophysics, and paleoclimatology. The examples of real time series are typical to the complex situations encountered in practice: data missing from the time series, superposition of several types of noises, long time series with tens of thousands of values, non-Gaussian probability distributions with fat tails, repeated values of the time series, additional conditions imposed on time series by the physical laws governing the studied phenomenon.

Our analysis is restricted to AR(1) noises superposed over nonmonotonic trends, but our methods can be applied to study other noise models. Such new applications could be: autoregressive noise of higher orders, long-range correlated noises, unevenly sampled time series, asymmetric probability distribution of the time series values. Obviously, the number of parameters could increase and the analysis of the accuracy of the estimated trend would become more burdensome.

We have limited our analysis to four methods of trend estimation: two classical (polynomial fitting and moving average) and two original and automatic (one for monotonic trends and the other for arbitrary nonmonotonic trends). Other trend estimation methods can be analyzed using the same type of Monte Carlo experiments. In order to obtain significant results, it is essential to use a statistical ensemble of artificial time series with a variety of trend shapes at least as rich as that generated by our algorithm presented in Chap. 2.

Even if the main definitions and theorems used in the book are briefly presented, nevertheless it is recommended that the reader has the knowledge of basic notions in probability, mathematical statistics, and time series theory. This book is of interest for researchers who need to process nonstationary time series. Detailed descriptions of all the numerical methods presented in the book allow the reader to reproduce the original automatic algorithms for trend estimation and time series partitioning. In addition, the source codes in MATLAB of the computer programs implementing them are freely available on the web so that the researchers who merely apply trend estimation algorithms could successfully use them.

Cluj-Napoca, April 2012

Călin Vamoş
Maria Crăciun

Contents

1 Introduction . . . 1
1.1 Discrete Stochastic Processes and Time Series . . . 1
1.2 Trend Definition and Estimation . . . 4
1.3 AR(1) Stochastic Process . . . 7
References . . . 12

2 Monte Carlo Experiments . . . 15
2.1 Monte Carlo Statistical Ensembles . . . 15
2.2 Numerical Generation of Trends . . . 18
2.3 Numerical Generation of Noisy Time Series . . . 21
2.4 Statistical Hypothesis Testing . . . 26
References . . . 30

3 Polynomial Fitting . . . 31
3.1 Polynomial Fitting . . . 31
3.2 Polynomial Fitting of Artificial Time Series . . . 34
3.3 An Astrophysical Example . . . 38
References . . . 42

4 Noise Smoothing . . . 43
4.1 Repeated Central Moving Average . . . 43
4.2 Smoothing of Artificial Time Series . . . 49
4.3 A Financial Example . . . 53
References . . . 58

5 Automatic Estimation of Monotonic Trends . . . 61
5.1 Average Conditional Displacement (ACD) Algorithm . . . 61
5.2 Automatic ACD Algorithm . . . 65
5.3 Evaluation of the ACD Algorithm . . . 70
5.4 A Climatological Example . . . 74

5.5 Monotonic Components of Nonmonotonic Trends 76
References 80

6 Estimation of Monotonic Trend Segments from a Noisy Time Series 81
6.1 Time Scale of Local Extrema 81
6.2 Local Extrema of Noisy Time Series 85
6.3 Local Extrema of RCMA Trends. 89
6.4 Significant Local Extrema of a Real Time Series 92
References 97

7 Automatic Estimation of Arbitrary Trends 99
7.1 Automatic RCMA (AutRCMA). 99
7.2 Statistical Significance of the Local Extrema of the AutRCMA Trend 105
References 110

Appendix A: Statistical Properties of the Linear Regression 111

Appendix B: Spurious Serial Correlation Induced by MA. 113

Appendix C: Continuous Analogue of the ACD Algorithm 117

Appendix D: Standard Deviation of a Noise Superposed over a Monotonic Trend 121

Appendix E: Construction of a Partition of Scale Δn 127

Appendix F: Estimation of the Ratio Between the Trend and Noise Magnitudes 129

Chapter 1
Introduction

A complete presentation of the theory of stochastic processes can be found in any treatise on the probability theory, e.g., [18] and for time series theory one can use [4]. In this introductory chapter we briefly present some basic notions which are used in the rest of the book. The main methods to estimate trends from noisy time series are introduced in Sect. 1.2. In the last section we discuss the properties of the order one autoregressive stochastic process AR(1) which has the serial correlation described by a single parameter and which is a good first approximation for many noises encountered in real phenomena.

1.1 Discrete Stochastic Processes and Time Series

At the occurrence of an event ω the *random variable* X takes the value $X(\omega) = x$. We follow the practice of denoting by small letters the realizations of the random variable denoted by the corresponding capital letters. Throughout this book we consider only continuous random variables with real values. If the random variable is absolutely continuous, then it has a *probability density function* (pdf) denoted $p(x)$. The *cumulative distribution function* (cdf) $F(x) = P(X \leq x)$ is the probability that the random variable X takes on a value less than or equal to x. We denote the *mean* of the random variable by $\mu = \langle X \rangle$ and its *variance* by $\sigma^2 = \langle (X - \langle X \rangle)^2 \rangle$.

The evolution in time of a random phenomenon is modeled by a *stochastic process*, i.e., a family of random variables $\{X(t),\ t \in I \subset \mathbf{R}\}$ defined on the same probability space and indexed by a set of real numbers I. In this book we study only discrete stochastic processes for which I contains equidistant sampling moments. The observations are made at discrete time moments $t_n = t_0 + (n-1)\Delta t$, where $n = 1, 2, \ldots, N$, Δt is the sampling interval, and t_0 is the initial time. The observed values $x_n \equiv x(t_n)$ are realizations of the corresponding random variables $X_n \equiv X(t_n)$. Although the number of observations is always finite, we assume that there is an infinite stochastic process $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$ whose realizations for $n < 1$ and $n > N$ have not

C. Vamoş and M. Crăciun, *Automatic Trend Estimation*, SpringerBriefs in Physics,
DOI: 10.1007/978-94-007-4825-5_1, © The Author(s) 2012

been observed. To distinguish between the infinite stochastic process which models the time evolution of the natural phenomenon and its measurements, we call *time series* the finite sequence of real numbers $\{x_n, n = 1, 2, ..., N\}$.

The joint cdf of the random variables $X_{n_1}, X_{n_2}, \ldots, X_{n_m}$ is the probability that their values are smaller than some given values

$$F_{\mathbf{n}}(\mathbf{x}) = P(X_{n_1} \leq x_1, \ldots, X_{n_m} \leq x_m),$$

where $\mathbf{n} = (n_1, \ldots, n_m)$ and $\mathbf{x} = (x_1, \ldots, x_m)$. For absolutely continuous random variables there exists the joint pdf $p_{\mathbf{n}}(\mathbf{x})$. A stochastic process is (strictly) *stationary* if for every index vector $\mathbf{n}$ and integer d we have $F_{\mathbf{n}+d\mathbf{1}} = F_{\mathbf{n}}$ or $p_{\mathbf{n}+d\mathbf{1}} = p_{\mathbf{n}}$, where $\mathbf{1} = (1, 1, \ldots, 1)$, i.e., its joint probabilities do not change under temporal translations. From this definition, for $m = 1$ it follows that all the components of a stationary process have the same probability distribution $p_n(x) = p(x)$ for all integers n. Such a stochastic process is called *identically distributed*.

The *autocovariance function* of a stochastic process with finite variance for all its components ($\sigma_n^2 < \infty$) is defined as

$$\gamma(n, m) = \langle (X_n - \langle X_n \rangle)(X_m - \langle X_m \rangle) \rangle. \tag{1.1}$$

Obviously $\gamma(n, n) = \sigma_n^2$. If the stochastic process is stationary, then

$$\gamma(n + d, m + d) = \gamma(n, m), \tag{1.2}$$

for all n, m, d integers and the autocovariance function depends only on the lag $h = n - m$ so that $\gamma(h) \equiv \gamma(h, 0)$. It is easy to show that $\gamma(0) \geq 0$, $\gamma(h) = \gamma(-h)$, and $|\gamma(h)| \leq \gamma(0)$ for any h. The *autocorrelation function* of a stationary stochastic process is defined as $\rho(h) = \gamma(h)/\gamma(0)$ and then $\rho(0) = 1$.

Usually the observed time series do not satisfy the condition imposed to strictly stationary stochastic processes. Furthermore, the analysis of time series is often reduced only to the statistical moments of second order. Therefore one defines a subclass of the stationary process more suitable for modeling of real phenomena. A stochastic process is *weak-stationary* if $\langle |X_n^2| \rangle < \infty$, $\langle X_n \rangle = \mu$ for all integers n and satisfies Eq. (1.2). A special weak-stationary process is the *white noise*, for which the components are uncorrelated $\gamma(h) = \sigma^2 \delta_{h0}$, where δ_{nm} is the Kronecker delta. Such a stochastic process is denoted by $X_n \sim WN(\mu, \sigma^2)$.

Another subclass of stationary processes contains the *independent and identically distributed* (i.i.d.) stochastic processes. The components of an i.i.d. process are mutually independent $p_{\mathbf{n}}(\mathbf{x}) = p_{n_1}(x_1) p_{n_2}(x_2) \ldots p_{n_m}(x_m)$. They are also identically distributed $p_{n_i}(x_i) = p(x_i)$ and then $p_{\mathbf{n}+h\mathbf{1}}(\mathbf{x}) = p(x_1) p(x_2) \ldots p(x_m) = p_{\mathbf{n}}(\mathbf{x})$ so that, if the stochastic process is infinite, the stationarity condition (1.2) is satisfied.

If the properties of the components of a stochastic process vary in time, then the stochastic process is *nonstationary*. As an example of nonstationary stochastic process we consider the *random walk* $\{X_n, n = 0, 1, 2, \ldots\}$ defined as

$$X_n = X_{n-1} + Z_n \quad \text{for} \quad n > 0, \tag{1.3}$$

where $\{Z_n\}$ is an i.i.d. stochastic process with zero mean, variance σ^2, and $X_0 = Z_0$. Obviously $\langle X_n \rangle = 0$ and for $n \leq m$ the autocovariance function given by Eq. (1.1) becomes

$$\gamma(n, m) = \langle X_n X_m \rangle = \langle X_n (X_n + Z_{n+1} + \cdots + Z_m) \rangle = \langle X_n^2 \rangle$$

because X_n depends only on $Z_0, Z_1, \ldots, Z_n$ which are independent of $Z_{n+1}, \ldots, Z_m$. Because

$$\langle X_n^2 \rangle = \sum_{k=0}^{n} \langle Z_k^2 \rangle + \sum_{k \neq l}^{n} \langle Z_k Z_l \rangle = (n+1)\sigma^2$$

we have

$$\gamma(n, m) = (1 + \min\{n, m\})\sigma^2. \tag{1.4}$$

Hence the autocovariance function of the random walk is not invariant to temporal translations and $\{X_n\}$ is a nonstationary process.

In practice we do not have access to random variables or stochastic processes but only to their realizations and we have to use the methods of the mathematical statistics in order to estimate the parameters of the observed phenomena. Let us consider a random variable X and one of its realizations $x^{(s)}$.[1] The set formed by S independent realizations $\{x^{(1)}, x^{(2)}, \ldots, x^{(S)}\}$ is called *sample of volume* S and it allows the estimation of the parameters of X. For instance the mean $\mu = \langle X \rangle$ is approximated by the *sample mean*

$$\mu^{\text{est}} \equiv \widehat{\mu} = \frac{1}{S} \sum_{s=1}^{S} x^{(s)}. \tag{1.5}$$

We make the convention that the quantities computed by means of a sample are denoted by a hat or with the superscript 'est'. By means of the law of large numbers one proves under rather general conditions that $\widehat{\mu}$ tends to μ when S tends to infinity. In the same way we define the *sample variance*

$$\widehat{\sigma}^2 = \frac{1}{S} \sum_{s=1}^{S} \left(x^{(s)} - \widehat{\mu} \right)^2. \tag{1.6}$$

Analogous relations can be used for a stationary time series $\{x_1, x_2, \ldots, x_N\}$. Instead of the sample mean (1.5) we define the *temporal mean*

[1] We have changed the usual notation x_s in order to avoid the confusion with the terms of a time series.

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{1.7}$$

Since the associated stochastic process $\{X_n\}$ is stationary, all the terms in the sum have identical pdfs and $\overline{x}$ tends to its theoretical mean μ when N tends to infinity. A similar analogy can be made for the sample variance (1.6). The serial correlation of a time series is characterized by the *sample autocovariance function*

$$\widehat{\gamma}(h) = \frac{1}{N}\sum_{n=1}^{N-h}(x_{n+h} - \overline{x})(x_n - \overline{x}), \quad 0 \le h < N. \tag{1.8}$$

For $-N < h \le 0$, we have $\widehat{\gamma}(h) = \widehat{\gamma}(-h)$. If X_n is a linear combination of the components of an i.i.d. stochastic process with finite fourth order moment and the sum of the absolute values of the linear combination coefficients is finite, then the estimator (1.8) is biased, but its asymptotic distribution has the mean equal to the theoretical autocovariance function ([4], Chap. 7). The *sample autocorrelation function* is given by

$$\widehat{\rho}(h) = \widehat{\gamma}(h)/\widehat{\gamma}(0), \quad |h| < N. \tag{1.9}$$

1.2 Trend Definition and Estimation

Stochastic processes model the random phenomena as opposed to the deterministic phenomena which are modeled by numerical functions of time. There are many situations when different random and deterministic phenomena overlap. In the simplest case, a deterministic and a random phenomenon, mutually independent, are superposed (for example the instrumental noise affecting a measured physical quantity). The most frequently used model is the stochastic process

$$X_n = f_n + Z_n, \tag{1.10}$$

where $\{Z_n\}$ is a stationary stochastic process with zero mean $\langle Z_n \rangle = 0$ named *additive noise* and $f_n = f(t_n)$ are the values at the sampling moments of the deterministic function named *trend*.

We denote by $p_Z(z)$ the pdf of Z_n and by $p_X(x, n)$ that of X_n. Because $\{Z_n\}$ is stationary, p_Z does not depend on n. According to Eq. (1.10), p_X is equal to p_Z translated by f_n

$$p_X(x, n) = p_Z(x - f_n). \tag{1.11}$$

The explicit dependence of p_X on the time index n indicates that $\{X_n\}$ is a nonstationary process with the mean varying in time