Risk of Forecasting with Fitting Bloomfield Model to Independent Training Sample

Valeriy Voloshko
Research Institute for Applied Problems of Mathematics and Informatics, Independence Ave., 4, Room 701, 220030, valeravoloshko@yandex.ru

Abstract: *The risk of forecasting is analyzed for the scheme with independent training and forecasted samples of two identically distributed Gaussian stationary time series. The asymptotic expansion of risk is constructed when the forecasted sample is infinite and the spectrum is estimated by increasing training sample using fitting of finite order Bloomfield model.*

***Keywords*:** Risk, forecast, expansion, spectrum.

1. INTRODUCTION

Forecasting time series is a well-studied problem in the case of known model [1,2,3]. When the model is unknown, it should be identified by training data. Therefore, methods of identification are needed, and it is desirable them to be robust w.r.t. distortions of data [4,5] or w.r.t. small data volumes [6,7,8,9]. In this paper there is considered the problem of forecasting of stationary Gaussian time series with spectrum identification through fitting Bloomfield model [10,11] by independent training sequence.

2. MATHEMATICAL MODEL

Let $\{x_t : t \in \mathbb{Z}\}$ be a Gaussian stationary time series with zero mean $\mathrm{E}\{x_t\} = 0$ and some non-negative spectrum $S : \Pi \to \mathbb{R}_+$, $\Pi = [-\pi, \pi]$, which is assumed unknown.

Furthermore, let us assume that we have a sample $X = \left(x_j\right)_{j=1}^{T_f}$ of length $T_f \in \mathbb{N}$ to be forecasted and a training sample $\tilde{X} = \left(\tilde{x}_j\right)_{j=1}^{T}$ of length $T \in \mathbb{N}$ from time series $\{\tilde{x}_t\}$, which does not depend on $\{x_t\}$ and identically distributed with it.

Time inversion $t' = -t$ saves stationarity, Gaussianity and the spectrum of a time series [2], so let us without loss of generality consider the following problem of retrospective forecasting: by samples $\tilde{X}$, X to build a forecast for x_0 and to research its asymptotic properties. This problem is equivalent to the classical problem of forecasting [4] of the future value x_{T_f+1} by $\tilde{X}$, X.

3. PLUG-IN FORECAST

Let us use the mean-square risk [4]:

$$R = \mathrm{E}\left\{\left(\hat{x}_0 - x_0\right)^2\right\} \tag{1}$$

as a measure of the accuracy of forecasting statistics $\hat{x}_0 = \hat{x}_0\left(X, \tilde{X}\right)$. According to [1,2], if time series model is fully specified, then the need for training data is missing and the statistic $\hat{x}_0^* = \mathrm{E}\{x_0 \mid X\}$ has a minimal risk (1) among functions $\hat{x}_0(X)$ which depend only on X. Since $\{x_t\}$ is Gaussian, $\hat{x}_0^*$ is linear w.r.t. X [2]:

$$\hat{x}_0^* = \sum_{t=1}^{T_f} a_{T_f,t}^* x_t, \tag{2}$$

where $\left\{a_{T_f,t}^*(S)\right\}_{t=1}^{T_f}$ are uniquely expressed in terms of $S(\cdot)$ by the Durbin-Levinson formulas [12]. Therefore we consider only forecasting statistics linear w.r.t. X.

Since the spectrum $S(\cdot)$ is unspecified, according to the plug-in principle let us identify it by the training sample $\tilde{X}$, i.e. build the spectrum estimator $\hat{S}$. Then on the base of spectrum estimator it is built a plug-in forecasting statistics, which approximate the optimal forecast (2):

$$\hat{x}_0^* = \sum_{t=1}^{T_f} a_{T_f,t}^*\left(\hat{S}\right) x_t. \tag{3}$$

In this paper, following [6,7], it is proposed to identify the Bloomfield model [10] of order $p \in \Lambda = \{0, \ldots, [T/2]\}$, where $[z]$ is the integer part of $z \in \mathbb{R}$:

$$y(\lambda) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_t e^{it\lambda}, \quad \lambda \in \Pi,$$

$$c_\tau = \frac{1}{T} \sum_{t=1}^{T} \ln\left|y\left(\frac{2\pi t}{T}\right)\right|^2 \cos\left(\frac{2\pi t \tau}{T}\right), \quad \tau \in \mathbb{Z},$$

$$\hat{S}(\lambda) = \exp\left(\sum_{j=-p}^{p} c_j \cos(j\lambda)\right), \quad \lambda \in \Pi. \tag{4}$$

The order p of the Bloomfield model let us call the smoothing parameter.

4. ASYMPTOTIC EXPANSION OF RISK

Introduce some notations: $1\{m \mid n\} = 1$, if $m \in \mathbb{N} \setminus \{0\}$ divides $n \in \mathbb{N}$, otherwise $1\{m \mid n\} = 0$,

$$\kappa_n = 1\{2 \mid n\}, \quad \varepsilon_n = 1 - \kappa_n, \quad n \in \mathbb{N},$$

$$\Theta_n(\lambda) = \sum_{j \in \wp^n} \sum_{z = j_1 \pm \ldots \pm j_n} 1\{T \mid z\} \prod_{r=1}^{n} \cos(j_r \lambda),$$

$$\Upsilon_n(\lambda) = \sum_{j \in \wp^n} \left(1 + \kappa_T (-1)^{j_1 + \ldots + j_n}\right) \prod_{r=1}^{n} \cos(j_r \lambda),$$

where $\lambda \in \Pi$, $\wp = \{1, \ldots, p\}$. Also for $K : \Pi^n \to \mathbb{R}$:

$$[K]_j = (2\pi)^{-n} \int_{\Pi^n} K(z) \prod_{r=1}^{n} \cos(j_r z_r) dz, \quad j \in \mathbb{Z}^n,$$

$$\tilde{K}(\lambda) = \sum_{j \in \wp^n} [K]_j \prod_{r=1}^{n} \cos(j_r \lambda), \quad \lambda \in \mathbb{R},$$

$$\chi(K) = \overline{\lim_{\tau \to +\infty}} \sqrt[\tau]{|[K]_\tau|}, \quad \pi(K) = e^{[\ln K]_0}, \quad K : \Pi \to \mathbb{R}.$$

To formulate the main result we need the following functions [8,9] depending on spectrum $S(\cdot)$ ($\lambda, \nu, \mu \in \mathbb{R}$):

$$H(\lambda)=2\sum_{j>0}[S]_j\sin(j\lambda),\quad f(\lambda)=\frac{H'(\lambda)}{S(\lambda)},$$

$$g(\lambda)=\frac{H(\lambda)}{S(\lambda)\sin\lambda},\quad h(\lambda)=f^2(\lambda)+\frac{g^2(\lambda)}{2},$$

$$\Psi(\nu,\mu)=\frac{H(\nu)+H(\mu)}{2\sin\frac{\nu+\mu}{2}},\quad \Phi(\nu,\mu)=\frac{2\Psi^2(\nu,\mu)}{S(\nu)S(\mu)}.$$

Let us build on the base of it the functions:

$$Q(\lambda)=\sum_{j>p}[\ln S]_j\cos(j\lambda),$$

$$Q_0\equiv 1,\quad Q_1=\frac{\pi^2}{3}\Theta_2+2\left(\tilde f+\Upsilon_1\ln 2\right),$$

$$Q_2=\frac{Q_1^2}{2}+\frac{\pi^2}{3}\Upsilon_2+\frac{8A}{3}\Theta_3+2\tilde\Phi+\tilde h,$$

where $A=\sum_{n>0}n^{-3}\approx 1{,}202$ is known as the Apery's constant [13].

Theorem 2. *If* $\pi(S)>0$ *and* $\chi(\ln S)<1$, *then at* $T_f=+\infty$ *the risk (1) of forecasting statistics (3) on the base of spectrum identification (4) uniformly w.r.t.* $p\in\Lambda$ *satisfies the asymptotic expansion as* $T\to+\infty$*:*

$$R=\pi(S)\left(\sum_{j=0}^{2}\left[Q_j(\lambda)e^{2Q(\lambda)}\right]_0T^{-j}+O\left(\frac{p^6}{T^3}\right)\right).\qquad(5)$$

Note that if the smoothing parameter p is fixed and $T\to+\infty$, then the risk (5) tends to the value $\pi(S)\left[e^{2Q}\right]_0$, which according to Theorem 2 of [6] coincides with the risk of plug-in forecasting statistics on the base of Bloomfield model of order p in the case of known spectrum.

Corollary 1. *Let* $\pi(S)>0$ *and* $\chi(\ln S)<\chi_+<1$. *Then at* $T_f=+\infty$ *the risk (1) of forecasting statistics (3) on the base of spectrum identification (4) uniformly w.r.t.* $0<p<T/3$ *satisfies the asymptotic expansion as* $T\to+\infty$*:*

$$R=\pi(S)\left(1+\frac{K_1p}{T}+\frac{K_2p^2+K_3p+K_4}{T^2}+O\left(\frac{p^6}{T^3}+\chi_+^p\right)\right),\qquad(6)$$

where

$$K_1=\frac{\pi^2}{6},\ K_2=\frac{\pi^4}{72}+A,$$

$$K_3=(1+\kappa_T)\left(\frac{\pi^2}{12}\ln 2+\ln^2 2+\frac{\pi^2}{6}\right)+\frac{\pi^4}{144}-A,$$

$$K_4=-\varepsilon_p\left(\kappa_T\left(\frac{\pi^2}{12}+2\ln 2\right)+\frac{\pi^2}{12}\right)\ln 2+[h]_0-\frac{[\Phi]_{0,0}+[f]_0^2}{2}$$

$$+\left(\frac{\pi^2}{24}+\ln 2\right)(f(0)-[f]_0)+\left(\frac{\pi^2}{24}+\kappa_T\ln 2\right)(f(\pi)-[f]_0).$$

Note that the only one coefficient K_4 at the smallest term T^{-2} of expansion (6) depends on spectrum $S(\cdot)$. This demonstrates the robustness of forecasting statistics (3) w.r.t. distortions of the time series model, which keep the regularity condition $\chi(\ln S)<1$ holding. The expansion (6) is applicable when the value χ_+^p is commensurate with p^6/T^3, i.e. when $p>C\ln T$, where $C>0$ is some model dependent constant. Since coefficients K_1, K_2 and K_3 are positive, the first term of the expansion (6) increases with increasing of p. Therefore the optimal choice of the smoothing parameter p on the base of this expansion is the nearest odd integer to $C\ln T$. The choice of an odd number minimizes the term with the factor $-\varepsilon_p$ of the coefficient K_4. The asymptotic analysis of the expansion (5) in the interval $0<p<C\ln T$ as $T\to+\infty$ seems to be difficult problem and requires a special research.

6. REFERENCES

[1] А.Н. Колмогоров. Интерполяция и экстраполяция стационарных случайных последовательностей, *Известия АН СССР. Сер. мат.* 5 (1) (1941). с. 3-14.

[2] Т.В. Андерсон. *Статистический анализ временных рядов.* Москва, 1976.

[3] Д.Р. Бриллинджер. *Временные ряды. Обработка данных и теория.* Москва, 1980.

[4] Ю.С. Харин. *Оптимальность и робастность в статистическом прогнозировании.* Минск, 2008.

[5] Yu.S. Kharin. V.A. Voloshko. Robust estimation of AR coefficients under simultaneously influencing outliers and missing values, *J. Statist. Plann. Inference* 141 (9) (2011). p. 3276-3288.

[6] Ю.С. Харин. В.А. Волошко. Прогнозирование стационарных временных рядов на основе малопараметрической модели Блумфилда, *Доклады НАН Беларуси* 54 (6) (2010). с. 27-32.

[7] V.A. Voloshko. Yu.S. Kharin. Statistical forecasting based on Bloomfield exponential model. *Proceedings of the conference "Computer Data Analysis and Modeling (CDAM'2010)",* Minsk, Belarus 7-11 September 2010, pp. 268-271.

[8] Yu.S. Kharin. V.A. Voloshko. On asymptotic properties of the plug-in cepstrum estimator for Gaussian time series, *Math. Methods of Stat.* 21 (1) (2012). p. 43-60.

[9] В.А. Волошко. Ю.С. Харин. Об асимптотических свойствах оценок кепстральных коэффициентов для гауссовского временного ряда, *Доклады НАН Беларуси* 55 (6) (2011). с. 23-28.

[10] P. Bloomfield. An exponential model for the spectrum of a scalar time series, *Biometrika* 60 (2) (1973). p. 217-226.

[11] N. Merhav. On the asymptotic statistical behavior of empirical cepstral coefficients, *IEEE Trans. on Sig. Proc.* 41 (5) (1993). p. 1990-1993.

[12] R.A. Maronna. R.D. Martin. V.J. Yohai. *Robust Statistics: Theory and Methods.* NY, 2006.

[13] R. Apéry. Irrationalité de $\zeta(2)$ et $\zeta(3)$, *Astérisque* 61 (1979). p. 11-13.