A wavelet-based model for forecasting non-stationary processes

S. Van Bellegem, P. Fryźlewicz and R. von Sachs

FNRS Research Fellow, Institut de statistique, Université catholique de Louvain, Belgium
Graduate Student, Department of Mathematics, University of Bristol, UK
Professor, Institut de statistique, Université catholique de Louvain, Belgium

Abstract. In this article, we discuss recent results on modelling and forecasting covariance non-stationary stochastic processes using non-decimated wavelets.

1. Introduction

In this article, we are concerned with data generated by a univariate, discrete-time stochastic process. We focus on the analysis of its covariance structure, and therefore we assume that the process is zero-mean. Zero-mean processes arise, for example, when the global trend has been removed from the data. Trend estimation is a well studied problem, and some recent methods use wavelets.

The *autocovariance function* of a process X_t is denoted by $c_X(r,s) := \mathrm{Cov}(X_r, X_s)$, and, for *stationary processes*, it depends on the distance between r and s only, i.e. $c_X(r,s) = c_X(|r-s|)$. As it does not depend on any reference point in time, we say that the autocovariance function of a stationary process is homogeneous over time. All stationary processes X_t have the following Fourier representation:

$$X_t = \int_{(-\pi,\pi]} A(\omega)\exp(i\omega t)dZ(\omega), \quad t \in Z, \tag{1}$$

where $A(\omega)$ is the amplitude, and $Z(\omega)$ is a stochastic process with orthonormal increments [1]. Correspondingly, under mild conditions, the autocovariance function can be expressed as $c_X(\tau) = \int_{-\pi}^{\pi} f_X(\omega)\exp(i\omega\tau)d\omega$, where f_X is the *spectral density* of X_t [1].

The assumption of stationarity leads to an elegant theory from the point of view of both estimation and forecasting [1]. However, various studies based on statistical tests of stationarity have shown that many observed processes have a non-homogeneous autocovariance (spectral) structure [2, 3, 4]. Examples of such *non-stationary* processes abound e.g. in econometrics (returns on stock indices), biomedical statistics (electrocardiograms), meteorology (wind speed), and many other fields. The important question of how to model and forecast non-stationary processes arises, and one of the main motivations behind using wavelets here is that, being well-localised in both time and frequency, they have the potential to naturally handle phenomena whose spectral characteristics change over time.

2. The class of locally stationary wavelet processes and the wavelet spectrum

We now recall a definition of a class of zero-mean nonstationary processes built of non-decimated discrete wavelets [5], rather than harmonics $\exp(i\omega t)$ like in (1).

Definition 1 ([6]) *A triangular stochastic array* $X_{t,T}$ $(t = 0,\dots,T-1, T > 0)$ *is in the class of* locally stationary wavelet (LSW) processes *if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-[\log_2 T]}^{-1} \sum_{k=-\infty}^{\infty} w_{j,k;T}\ \psi_{jk}(t)\ \xi_{jk}, \tag{2}$$

where j *and* k *are scale and location parameters, respectively,* $w_{j,k;T}$ *are real constants,* $\{\psi_{jk}(t)\}_{jk}$ *is a non-decimated family of discrete compactly supported wavelets [5],* ξ_{jk} *is an orthonormal sequence of identically distributed zero mean random variables, and for each* $j \le -1$*, there exists a Lipschitz-continuous function* $W_j(z)$ *on* $[0,1)$ *such that*

- $\sum_{j=-\infty}^{-1} W_j(z)^2 < \infty$ *uniformly in* $z \in [0,1)$ *;*
- *the Lipschitz constants* L_j *satisfy* $\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty$*;*
- *there exists a sequence of constants* C_j *satisfying* $\sum_{j=-\infty}^{-1} C_j < \infty$*, such that, for all* T*,*

$$\sup_{k=0,\dots,T-1} |w_{j,k;T} - W_j(k/T)| \le C_j/T.$$

By analogy with Dahlhaus [7], the time-varying quantity $W_j(z)$ is defined in *rescaled time* $z = t/T \in [0,1)$. As the non-decimated wavelet system is overcomplete, the "amplitudes" $w^2_{j,k;T}$ are not uniquely defined and therefore not identifiable. However, due to the regularity of $W_j(z)$ in the rescaled time, the *wavelet spectrum* of $X_{t,T}$, defined by $S_j(z) := W_j(z)^2$, is unique. $S_j(z)$ measures the power of the process at a particular scale j and location z, and can be estimated by means of asymptotically unbiased multiscale estimators [6]. For stationary processes, $S_j(z)$ is independent of z for all j. Here, $j = -1$ denotes the finest scale.

As recalled in the Introduction, the autocovariance function of a stationary process is the Fourier transform of its spectral density. The next results shows an analogous link between the autocovariance function of an LSW process $X_{t,T}$, defined by $c_T(z,\tau) = \mathrm{Cov}(X_{[zT],T}, X_{[zT]+\tau,T})$, and its wavelet spectrum $S_j(z)$. We first define the *autocorrelation wavelets* $\Psi_j(\tau) = \sum_k \psi_{jk}(0)\psi_{jk}(\tau)$. They are symmetric with respect to τ and satisfy $\Psi_j(0) = 1$ for all scales. Also, like wavelets themselves, they enjoy good localisation properties.

Proposition 1 ([6]) *Under the assumptions of Definition 1,* $\|c_T - c\|_{L_\infty} = O\left(T^{-1}\right)$*, where*

$$c(z,\tau) = \sum_{j=-\infty}^{-1} S_j(z)\Psi_j(\tau). \tag{3}$$

Function $c(z,\tau)$ is called the *local autocovariance function* (LACV) of $X_{t,T}$. Formula (3) is a *multiscale representation of the nonstationary autocovariance function* $c(z,\tau)$. The representation is unique because the set $\{\Psi_j\}_j$ is linearly independent, [6]. By way of example, it can be proved [8] that $\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \delta_0(\tau)$ holds, which implies, in particular, that the wavelet spectrum of (stationary) white noise is proportional to $S_j(z) = 2^j$.

3. Forecasting

We now want to forecast an LSW process h steps ahead, basing on t observations $X_{0,T},\dots,X_{t-1,T}$. We set $T = t + h$ and consider a linear predictor $\hat{X}_{t+h-1,T} = \sum_{s=0}^{t-1} b^{(t,h)}_{t-1-s;T} X_{s,T}$, where the coefficients $b^{(t,h)}_{t-1-s}$ minimise the mean-square prediction error $\mathrm{E}(\hat{X}_{t+h-1,T} - X_{t+h-1,T})^2$.

Proposition 2 ([8]) *Assume $h = o(T)$ and let the length of ψ_j be ℓ_j. Assume also:*

$$\sum_{\tau=0}^{\infty} \sup_z |c(z,\tau)| < \infty \quad \sum_{j<0} (C_j + \ell_j L_j)\ell_j < \infty \quad \operatorname*{ess\,inf}_{z,\omega} \sum_{j<0} S_j(z)|\hat{\psi}_j(\omega)|^2 > 0,$$

where $\hat{\psi}_j(\omega) = \sum_{s=-\infty}^{\infty} \psi_{j0}(s)\exp(i\omega s)$. The mean-square prediction error can be written as

$$\tilde{b}^{(t,h)\prime} B^{(t+h;T)} \tilde{b}^{(t,h)} \, (1 + o_T(1)), \tag{4}$$

where $B^{(t+h;T)}_{m,n} = c(\frac{n+m}{2T}, n-m)$, $n,m = 0,\ldots,t-1+h$, and $\tilde{b}^{(t,h)}$ is the vector $(b^{(t,h)}_{t-1},\ldots,b^{(t,h)}_{0},b^{(t,h)}_{-1},\ldots,b^{(t,h)}_{-h})'$ with $b^{(t,h)}_{-1},\ldots,b^{(t,h)}_{-h+1} = 0$ and $b^{(t,h)}_{-h} = -1$.

In theory, the coefficients of $\tilde{b}^{(t,h)}$ (and therefore $b^{(t,h)}$) are now found by minimising (4). In practice, however, the elements of $B^{(t+h;T)}$ need to be estimated. It is easily observed [8] that Formula (3) suggests the following multiscale estimator of the LACV:

$$\hat{c}\left(\frac{k}{T},\tau\right) = \sum_{j,m=-\lceil \log_2 T \rceil}^{-1} A^{-1}_{mj} \left(\sum_{s=0}^{t-1} X_{s,T}\psi_{mk}(s) \right)^2 \Psi_j(\tau), \quad k = 0,\ldots,t-1, \tag{5}$$

where $A_{mj} = \sum_\tau \Psi_m(\tau)\Psi_j(\tau)$. The above estimator is asymptotically unbiased [8]. Also, it can be shown by simulation, and by exact calculation in some particular cases, that the above estimator enjoys better Mean-Square Error properties than other commonly used (asymptotically) unbiased estimators of local covariance. However, it is not consistent (its variance does not go to zero with T, [8]) and therefore has to be smoothed using e.g. a Gaussian kernel smoother with bandwidth g. Moreover, in (4), we also require the values of $c(k/T,\tau)$ for $k = t,\ldots,t+h-1$. Motivated by its slow evolution, we extrapolate the LACV by one-sided smoothing of the estimated autocovariances with the same bandwidth g. The next section discusses a data-driven method for choosing the smoothing parameter g.

4. The forecasting algorithm

As mentioned before, we estimate the entries of $B^{(t+h;T)}$ using (5) smoothed over k/T to achieve consistency. For simplicity, we choose the same bandwidth g for all τ. In this section and next, we only consider one-step-ahead forecasts.

Also, in practice, we only incorporate p past observations into the predictor, instead of the whole history of the process, which corresponds to only considering the bottom-right corner of $B^{(t+h;T)}$. The motivation behind this is that the "clipped" predictor often performs much better in practice, while also being less computationally expensive.

We select (g,p) by *adaptive forecasting*, i.e. we gradually update (g,p) according to the success of prediction. We first move backwards by s observations and choose the initial parameters (g_0,p_0) for forecasting $X_{t-s,T}$. Next, we forecast $X_{t-s,T}$ using not only (g_0,p_0) but also the 8 neighbouring pairs $(g_0+\delta\varepsilon_g, p_0+\varepsilon_p)$, for $\varepsilon_g,\varepsilon_p \in \{-1,0,1\}$ and δ fixed. As we already know the true value of $X_{t-s,T}$, we compare the 9 forecasts using a pre-selected criterion, and update (g,p) to be equal to the pair which gave the best forecast. We now use this updated pair, as well as its 8 neighbours, to forecast $X_{t-s+1,T}$, and continue in the same manner until we reach $X_{t-1,T}$. The updated pair (g_1,p_1) is used to perform the actual prediction, and can itself be updated later if we wish to forecast $X_{t,T}, X_{t+1,T},\ldots$

Various criteria can be used to compare the quality of the pairs of parameters at each step. Denote by $\hat{X}_{t-i,T}(g,p)$ the predictor of $X_{t-i,T}$ computed using pair (g,p), and by $\mathrm{P}_{t-i,T}(g,p)$

the length of the corresponding prediction interval. In [8], we propose the following criterion: choose the pair which minimises $|X_{t-i,T} - \hat{X}_{t-i,T}(g,p)|/\mathrm{P}_{t-i,T}(g,p)$. We suggest that the length s of the "training segment" be chosen in such a way that $X_{t-s-p},\ldots,X_{t-1}$ does not contain any apparent breakpoints observed after a visual inspection. To avoid dependence on the initial values (p_0,g_0), the algorithm can be iterated a few times along the training segment, e.g. until at least 95% of the observations fall within their 95% prediction intervals.

5. Example

We apply the algorithm to forecast the wind anomaly index (denoted here by X_t), plotted on `http://www.stats.bris.ac.uk/~mapzf/g24/wind.html`. This time series has been studied in meteorology in order to understand the El Niño effect in a specific region of the Pacific. The length of the series is $T = 910$. By trial and error, we have found that the pair $(g_0,p_0) = (70/910, 7)$ produces accurate one-step-ahead forecasts for many segments of the series. However, the results for the segment $X_{801},\ldots,X_{840}$ using (g_0,p_0) are extremely bad (only 5% of the observations fall within the 95% prediction intervals). Suppose that we want to forecast $X_{841}, X_{842},\ldots$. As we have no reason to believe that (g_0,p_0) has a chance of performing well here, we run the algorithm of the previous section with $s = 40$, $\delta = 1/910$, and the criterion described in the previous section. Three iterations along the training segment (with the "new" (g_0,p_0) always being set to the (g_1,p_1) obtained in the previous iteration) are sufficient to obtain 95% coverage of the 95% prediction intervals. The updated parameters are $(g_1,p_1) = (114/910, 14)$. Indeed, they do an excellent job in forecasting $X_{841},\ldots,X_{848}$ one-step-ahead: over this segment, all the true values of X_t lie within their prediction intervals, and the main gain from using our procedure here is that the prediction intervals themselves are more than 15% narrower than those obtained from optimally fitted stationary ARMA models. A similar effect has been observed at several other points of the series. For other examples and more details, the reader is referred to [8].

References

[1] Brillinger, D. Time Series. Data Analysis and Theory. Holt, Rinehart and Winston, Inc.; 1975.

[2] von Sachs R, Neumann MH. A wavelet-based test of stationarity. J Time Ser Anal 2000; 21:597-613.

[3] Fry´zlewicz P. Modelling and forecasting financial log-returns as locally stationary processes. Research Report. University of Bristol: Department of Mathematics; 2002.

[4] Van Bellegem S, von Sachs R. Forecasting economic time series using models of nonstationarity. Discussion paper No.0227. Universit´e catholique de Louvain: Institut de statistique; 2002.

[5] Nason GP, Silverman B. The stationary wavelet transform and some statistical applications. In: Antoniadis A, Oppenheim G, eds. Wavelets in Statistics. New-York: Springer; 1995:271-300.

[6] Nason GP, von Sachs R, Kroisandt G. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. J Roy Statist Soc Ser B 2000; 62:271-292.

[7] Dahlhaus R. Fitting time series models to nonstationary processes. Ann Statist 1997; 25:1-37.

[8] Fry´zlewicz P, Van Bellegem S, von Sachs R. Forecasting non-stationary time series by wavelet process modelling. Discussion paper No.0208. Universit´e catholique de Louvain: Institut de statistique; 2002.