Non-stationary Stochastic Processes

(Revised April 10, 2012)

Robert Murray, Ph.D.
Omicron Research Institute

7.0 Contents

7.1 Stationary Stochastic Processes

A **stationary stochastic process** is one for which the statistical properties are constant in time. In the simplest case of a Gaussian process, the **mean** and **variance** of the process are time-independent. A good example of this is a periodic waveform in Electronics, with a Gaussian white noise background. This kind of periodic waveform is well adapted to Fourier techniques. The Fourier transform may be computed over a period of the signal, and from this a Periodogram constructed for the variance as a function of frequency. It has been shown [BD] that if a process is stationary, then in the **Fourier** basis the Fourier components of the stochastic process are orthogonal. This means that they are *independent* random variables, and hence *uncorrelated*. A stationary process is appropriate for processes that are very nearly regular and periodic in time. An example that comes to mind is the sunspot cycle. Evidently, the proper choice of stochastic process with which to model a given system depends on the choice of basis vectors, which *de-correlate* the time series. For a stationary time series the Fourier components de-correlate the time series and are an appropriate basis.

However, other systems may be better described by a **non-stationary stochastic process**, in which the covariance matrix and mean of the process can be functions of time. In this case, a more appropriate choice of basis in which the time series may be approximately de-correlated is the **Wavelet** basis. It is known [BD] that for many realistic time series, the Wavelet basis de-correlates the time series to a good approximation. Financial time series evidently constitute such a system. At least, it vastly simplifies the *estimation* problem to make this assumption. In the Wavelet basis, the basis functions are localized functions of time as well as frequency octave or level, as opposed to the Fourier basis in which the basis functions are infinite sine waves of precise frequency but completely unlocalized in time. So for most realistic time series, which are not exactly periodic, the Wavelet basis is a better description and a more accurate de-correlation of the data.

Ideally, or theoretically, a non-stationary time series has a well-defined, time-dependent, mean and covariance matrix, and an unlimited number of specific *realizations* of the time series may be generated with the same time-dependent statistics. Then the time-dependent mean and covariance matrix could be determined, in the limit of a large number of realizations, by suitable statistical averaging over all the realizations. This is the implicit assumption that is made when utilizing the concept of non-stationary time series (or any other stochastic time series). However, in the case of financial time series, there is only one realization of each time series, so in principle there is no way to determine precisely the statistical properties. Another way of saying this is that we cannot make any distinction, for financial time series, between the *marginal* distributions and the *conditional* distributions of the stochastic process. There are not enough data available to determine what the *marginal* distribution would be – we assume that all distributions are *conditional*.

The empirical mean and covariance can be measured over different time intervals, but these are subject to large statistical uncertainties that cannot be eliminated since there is only one time series to work with. So as a matter of practical necessity, at some point it is necessary to *assume* stationarity of the time series, of some sort. But the stationarity can exist at a higher level than that of the mean and covariance. In other words, we may define a time-dependent empirical mean and covariance matrix, use these to determine trading rules, then the validity of the trading rules tested over the whole time series using the assumption of stationarity of the *trading rules*. In other words, the correlation of the time-dependent trading rules with future returns is assumed to be *stationary*. But in these cases there is no "Law of Large Numbers" and it can never be proved that the trading rules work to within an arbitrarily small uncertainty, as would be the case if an infinite number of realizations of the stochastic process were available to test.

7.2 Non-stationary Stochastic Processes

For financial time series, it makes more sense to model the time series as a **non-stationary stochastic process**. This implies that the mean and covariance matrix of the time series of returns are taken to be functions of time. For a stationary time series, for example, the covariance matrix depends only on the *difference* between the two time indices, hence is a Toeplitz matrix, but for the non-stationary case the covariance matrix can be any positive-definite symmetric matrix. Similarly the mean of the returns process is assumed to be a constant, corresponding to a constant linear trend of the price process, although the *conditional mean* is a function of the past returns and is computed using a linear regression, assuming the constant covariance matrix of returns.

A problem arises, however, when it comes to actually estimating a time-dependent (conditional) covariance matrix. There are actually not enough data available in a typical financial time series to estimate such a time-dependent covariance matrix. So we cannot estimate the time-dependent covariance matrix in its full generality, and some simplifying assumptions must be made. Usually the assumption of *stationarity* is made to enable the time series itself to serve as a statistical sample. However, in fact the time series is not stationary, and when **Ordinary Linear Prediction (OLS)** is used to compute the future projection of the time series, the projection is very noisy. Most of the high-frequency fluctuations in the past data are random noise, but the OLS method fits the regression to this noise. So the question is how to reduce the problem of "fitting to the noise" to arrive at a meaningful future projection.

While Fourier methods are well adapted to stationary time series, for non-stationary time series Wavelet methods are more appropriate. In the **Discrete Wavelet**

Transform (DWT), the wavelet decomposition is in terms of wavelet coefficients, which depend on both the frequency octave and on time. For the DWT the *wavelet variance* provides an *Analysis of Variance* (ANOVA) [PW], just as the *Periodogram* does for the *Fourier Transform*. In realistic time series, it is a good approximation that the DWT wavelet coefficients are *independent* random variables. This assumption drastically reduces the degrees of freedom of the covariance matrix. So instead of estimating the correlation of the time series from the data, we make the assumption that the DWT coefficients are uncorrelated, and determine a fixed correlation matrix from this on each wavelet level. This is a good approximation for many time series, and for financial time series the data are so noisy that this simplifying assumption is necessary. This then leaves only the *wavelet variance* to be estimated, and this is estimated in terms of the square of the wavelet coefficients at each level, as functions of time. Also this implies that each wavelet coefficient has an influence only within a certain time interval, of the order of the time scale of the wavelet level. This influence is determined from a quadratic function of the DWT matrix called the **filter kernel**.

It is plausible to assume that only the few most recent wavelet coefficients on each wavelet level are relevant for future prediction. In fact, the lowest frequency components of the time series, which are the most important for modeling, are precisely the ones with the fewest wavelet components. The high-frequency components in the distant past should have no effect on the future time series. For example, half the wavelet components lie in the highest wavelet level, and we regard these as stochastic noise and discard them. So by eliminating these high-frequency components, a significant improvement in the "signal to noise ratio" should be achieved.

Hence we will develop a **Wavelet Linear Prediction (WLP)** filter, based on the assumption that the wavelet components are independent random variables, which determines the fixed (stationary) correlation matrix at each wavelet level. The (conditional) non-stationarity arises from the time-dependent wavelet variance at each wavelet level, and this is incorporated into a time-dependent auto-covariance matrix for the process. Another aspect of non-stationarity is described by the class of models called **Auto-Regressive Conditional Heteroskedasticity (ARCH)** [G], which means that the volatility (square root of variance) is time-dependent and conditional on past variance or other data. So the ARCH models can also be treated effectively using wavelet techniques. In this way both the conditional mean and variance of the time series are projected forward in time using the Wavelet Linear Prediction filter.

Using the Wavelet LP filter, on each wavelet level the conditional mean of the returns is projected ahead only on the time scale of that level, thereby eliminating a large number of high-frequency components that are just stochastic noise. Similarly, the time-dependent variance or average absolute returns can be projected ahead with the same Wavelet LP filter and give an estimate of the future volatility. Thus the first and second moments of the conditional probability distribution of prices can be estimated. In fact, the volatility above and below the mean can be projected separately. This would give a measure of the skew, which is the third moment of the probability distribution of returns. A good predictor that can be used as a regressor for skew would be the price above and below some long-term average price. Finally, to complete the picture the kurtosis, or fourth moment of the probability distribution, can be determined by measuring the relative frequency of "outliers", or equivalently the Hurst exponent of the price deviations [M]. According to Mandelbrot [M], the volatility should typically increase like the 0.73 power of time rather than the 0.50 power as in the ordinary standard deviation, so perhaps instead of the average absolute deviation of returns or the squared deviation of returns as a measure of volatility, we should take the average absolute deviation to the power $1/0.73 = 1.37$ as the correct measure of volatility.

The first four moments of the future estimated conditional probability distribution can be displayed nicely on a price graph. The 1st moment is the expected conditional mean value of returns, integrated to give an expected price projection, which is displayed as a set of future projected price data points (say, by a series of blue squares). The 2nd moment is the standard deviation, which can be displayed as a series of blue error bars extending above and below the price projection. The 3rd moment is the skew, which is displayed as unequal error bars above and below the price projection. Finally, the 4th moment is kurtosis, which can be displayed by a second set of error bars (in dark blue, say), extending out to two standard deviations and denoting the probability of "outlier" events. The skew of these error bars can also be measured independently by measuring the outliers above and below the mean prices. In fact, the extension below the price projection of these two-sigma error bars denotes the likelihood of a rapid decline in prices or "crash" if the prices are over-extended on the upside, which is one of the most important estimation problems in Finance.

7.3 Time-dependent Mean and Covariance

By analogy with the stationary case, for the case of a (real) time series $X(t)$, we may define an **auto-covariance matrix**, which in this case will be time-dependent. In the general case, this covariance matrix will be a function of both time indexes, not just the difference of the two indexes as in the stationary case. For notational purposes, let us denote by t the *present* time, or the last *known* element of the time series, even though we suppose (for the moment) the stochastic process to extend from minus infinity to plus infinity in time. We then define the following notation:

$$X_n(t) \equiv X(t-n) \qquad n=0 \text{ (present)}, \quad n>0 \text{ (past)}, \quad n<0 \text{ (future)}$$

Then, we define a time-dependent auto-covariance matrix as follows:

$$\Gamma_{m,n}(t) \equiv \langle X_m(t), X_n(t) \rangle \equiv E\left[X_m(t) X_n(t) \right]$$

The expectation is over the hypothetical infinite set of *realizations* of the stochastic process. The other statistical quantity of interest in the case of a non-stationary process is the (time-dependent) **mean** of the process. This is defined by:

$$\mu_n(t) \equiv \langle X_n(t) \rangle$$

In what follows, we will assume that the average of the time-dependent mean over the whole (returns) series has been subtracted off from the time series of returns, so that the time series is **zero-mean**. Otherwise, it would be included in the wavelet "smooth" level of the wavelet decomposition. But this does not imply that the time-dependent mean is zero, only that its constant component is zero. In fact, this constant component of the mean of the returns is subtracted off before the Price Projection is computed, then added back on afterward, and in addition it can be plotted on the price graph as an overall linear trend line for the whole data series of prices.

We may also define a set of M time series which are thought to be correlated with future returns of the given time series; these are called **predictors** and are used in a multivariate linear regression which generalizes the univariate linear regression on the past values of the given time series alone. We then define a time-dependent **cross-covariance** matrix between any two of these time series. The set of all M predictors are labeled by an upper Greek index. We denote the original time series $X_n(t)$ as the predictor with index 0, or perhaps a surrogate for this time series, as described below. The auto-covariance matrix between all the predictors is then given by:

$$\Gamma_{a,b}^{(\alpha,\beta)}(t) \equiv \left\langle Y_a^{(\alpha)}(t), Y_b^{(\beta)}(t) \right\rangle \equiv E\left[Y_a^{(\alpha)}(t) Y_b^{(\beta)}(t) \right]$$

In general, we wish to regress the future returns on both the returns series and on various predictor series, so we will need to make use of both the auto-covariance matrix and the various cross-covariance matrices, all of which are contained in the covariance matrix of the predictors as given above.

Usually we will assume that the time series consists of N variables, with index 0 the most recent index and the indices increasing towards the past, so the time indices m,n,a,b, etc., will run over the interval $0,\ldots,N-1$. We may assume that there are a total of M predictors, which will include the past values of the time series itself as index 0, so the Greek indices will run over vales $0,\ldots,M-1$. We will frequently need the inverse of the covariance matrices defined above. However, due to the plethora of indices, we must use a different symbol for the inverse covariance matrix, denoted by the Greek letter Δ. Thus we for the inverse covariance matrices over the *past* data:

$$\sum_{b=0}^{N-1}\sum_{\beta=0}^{M-1} \Delta_{a,b}^{(\alpha,\beta)}(t)\Gamma_{b,c}^{(\beta,\gamma)}(t) = \sum_{b=0}^{N-1}\sum_{\beta=0}^{M-1} \Gamma_{a,b}^{(\alpha,\beta)}(t)\Delta_{b,c}^{(\beta,\gamma)}(t) \equiv \delta_{a,c}^{(\alpha,\gamma)}$$

In the case of the original time series $X_n(t)$, if there are other predictors present then it is counted among them, otherwise if it is an auto-regression then the Greek index is omitted. Notice also that this covariance matrix is inverted only over the *past* indices as defined above, corresponding to the choice of past returns (or other) data over which the future returns are regressed.

It is crucial for the solution of the Wavelet Linear Prediction problem to be able to estimate the past and future wavelet auto-variance and cross-covariances of the different predictors. This problem is similar to that of estimating the Periodogram in the stationary case. It can be seen that the above covariance matrix, which is a $MN \times MN$ matrix, would be extremely difficult to estimate. A major simplification comes from performing the Discrete Wavelet Transform (DWT) on all the predictor time series and working in the wavelet coefficient basis. Then we may make the plausible working assumption that all the DWT wavelet coefficients with different level index and time index are orthogonal, for all time series, both within and between the time series. Thus to compute the covariances, only products of wavelet coefficients with the same level and time index are taken – all the others are orthogonal by ansatz. Then, in the wavelet basis, we compute the time-dependent covariances by taking products of the like wavelet coefficients of the two series, then implement a suitable smoothing procedure. The stochastic uncertainty in the wavelet coefficients and variances is evidently, as in the case of the Periodogram, close to the value of the quantities we are trying to estimate, so some kind of smoothing procedure for the variance/covariance estimation is needed. The smoothing procedure is not specified a-priori, but generally the smoothing time scale on each level should be something of the order of the time scale of that level or greater. Some shrinkage procedure can also be implemented for further noise reduction, with the shrinkage adjustable from 0% to 100%, meaning that the time-dependent part is averaged on each level with the level average.

Furthermore, we need to also estimate the *future* auto-variances and cross-covariances, and for this it is essential that the average values of these quantities on each wavelet level be preserved in the estimate. Evidently the simplest way around this problem, as well as the most realistic, is to estimate the future wavelet auto-variance and cross-covariance as the level average of these quantities. This makes sense because we are approximating the DWT wavelet coefficients as independent random variables, so the future wavelet coefficients should not be correlated with the past ones. Also a reasonable estimate of the variances requires a lot of smoothing, so no precise future estimate is needed. But if some other method of projection such as OLP of the level wavelet variance were used, there would be no way to be sure that the average wavelet variance was preserved in the projection. Another alternative would be to subtract the level average from the wavelet variance, use OLP to project it ahead into the future, and then add the level average back to the future projection. However, since the wavelet coefficients are supposed to be uncorrelated and a high degree of smoothing is required anyway, this method seems unwarranted.

It might be tempting to try to define a Wavelet LP filter separately for each wavelet level, as in the case of the wavelet coefficients or the wavelet variances on each level, or the multi-resolution analysis (MRA). The filter kernels could be used to define an auto-covariance matrix for each level, and then this matrix could be inverted using a Toeplitz matrix inversion routine. However, this method cannot work. Confined to a single level, the covariance matrix is highly singular because it is equivalent to having a zero signal on all the other wavelet levels, so the matrix cannot be inverted. To compute a LP filter using the filter kernels, these must be weighted according to the wavelet variance on each level and then all the levels added together, and used to project the whole time series. The covariance matrix in this case will not be singular, and the correlation between past and future returns will depend on the difference in wavelet variance between levels – if the wavelet variance is the same on all levels, the correlation is zero. This means the LP projection must be done on the whole data set, not on each wavelet level individually. A related point is that only on the whole data set is the division between past and future sharp. On each individual wavelet level, the division between past and future is blurred, and it is not clear how an LP filter would be applied in this case to begin with.

7.4 Time-dependent Vector Auto-Regression (VAR)

Now let us note that normally a data set of N = 2048 past values will be used, starting with index 0, and to this another set of 2048 variables is attached for the future values plus zero padding. Hence the index of the "present" time t is always t = 2047, so we may henceforth drop the explicit reference to t. Then we may redefine the time index so that the index for the past data values will run from 0, denoting the present time, to +2047 in the distant past, with increasing index denoting the more distant past. Then the index for the future (estimated) data values will run from –1 to –2048 in the distant future. So the complete data set, consisting of past data values and future estimated values, will run from index +2047 in the past to –2048 in the future, making a data set of 4096 values in all.

To define the univariate auto-regression of financial time series, we need to make one more assumption. This treatment follows that in Numerical Recipes [NR, p.564].

The observed time series X_m, which is nearly random, is assumed to consist of the sum of a "true value" or signal, denoted $\hat{X}_m$, plus a random noise term ε_m. The observed covariance matrix of the past data is then denoted by $\Gamma_{m,n}$, while the estimated covariance matrix of the signal is denoted by $\hat{\Gamma}_{m,n}$. The estimated signal covariance matrix runs over both past and future time values of the indices. The covariance matrix of the noise, which is assumed to be uncorrelated with itself and with the (past and future) signal, is denoted by $\eta_{m,n} \equiv \left\langle \varepsilon^2 \right\rangle \delta_{m,n}$. The noise is presumed to be the same for both past and future time values of the indices. Thus we have:

$$
\begin{aligned}
&X_m = \hat{X}_m + \varepsilon_m \qquad \left\langle \hat{X}_m \varepsilon_n \right\rangle \equiv 0 \\
&\hat{\Gamma}_{m,n} \equiv \left\langle \hat{X}_m \hat{X}_n \right\rangle \qquad \eta_{m,n} \equiv \left\langle \varepsilon_m \varepsilon_n \right\rangle \\
&\Gamma_{m,n} \equiv \left\langle X_m X_n \right\rangle = \hat{\Gamma}_{m,n} + \left\langle \varepsilon^2 \right\rangle \delta_{m,n}
\end{aligned}
$$

We presume that the noise exists for all time values, past and future, but the signal consists of a past signal that exists only for the past time values, and a future estimated signal that exists for future time values. The past signal is also estimated, with the constraint that the sum of the past signal and the past noise equals the observed data series. Thus the covariance matrix for both past and future consists of the covariance of the noise together with the estimated covariance of the signal. Both of these components must be estimated when the covariance matrix itself is estimated in the wavelet basis.

Consider now the problem of finding the best *h*-step linear prediction of the signal $\hat{X}_m$, given *N* previous values of the time series X_m. If X_m is a zero-mean process, we may write the best linear prediction of $\hat{X}_{-h}$ in terms of the previous *N* values of the time series as:

$$\sum_{n=0}^{N-1} X_n \phi_{n(h)} \equiv X_{-h} \equiv \hat{X}_{-h} + \varepsilon_{-h}$$

The linear sum is written as the sum of the future estimated signal $\hat{X}_{-h}$ and a random noise term ε_{-h}, which is called the *discrepancy* or *residual*. The goal is to find the best linear prediction, which minimizes the mean square of the discrepancy. We cannot just multiply each side of the equation by X_m or $\hat{X}_m$ and take expectations, because we arrive at a different result depending on which we use. So, we instead write down the expression for the mean square discrepancy and then minimize this with respect to the LP coefficients [NR]:

$$\begin{aligned}\left\langle \varepsilon_{-h}^2 \right\rangle &= \left\langle \left(\sum_{m=0}^{N-1} X_m \phi_{m(h)} - \hat{X}_{-h} \right) \left(\sum_{n=0}^{N-1} X_n \phi_{n(h)} - \hat{X}_{-h} \right) \right\rangle \\ &= \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \left\langle X_m X_n \right\rangle \phi_{m(h)} \phi_{n(h)} - 2 \sum_{m=0}^{N-1} \left\langle X_m \hat{X}_{-h} \right\rangle \phi_{m(h)} + \left\langle \hat{X}_{-h}^2 \right\rangle\end{aligned}$$

Taking the derivative of this expression with respect to the LP coefficients and then setting it equal to zero yields:

$$\sum_{n=0}^{N-1} \left\langle X_m X_n \right\rangle \phi_{n(h)} - \sum_{m=0}^{N-1} \left\langle X_m \hat{X}_{-h} \right\rangle = 0$$

In terms of the covariance matrix, this becomes:

$$\sum_{n=0}^{N-1} \left[\hat{\Gamma}_{m,n} + \left\langle \varepsilon^2 \right\rangle \delta_{m,n} \right] \phi_{n(h)} = \hat{\Gamma}_{m,-h}$$

Thus, for each time *t*, considered to be the *present* time, the LP coefficients may be found by inverting the time-dependent covariance matrix, assuming that this is non-singular, where $\Delta_{n,m}$ denotes the inverse of the covariance matrix of signal plus noise in brackets:

$$\sum_{n=0}^{N-1} \left[\hat{\Gamma}_{m,n} + \left\langle \varepsilon^2 \right\rangle \delta_{m,n} \right] \Delta_{n,k} = \sum_{n=0}^{N-1} \Delta_{m,n} \left[\hat{\Gamma}_{n,k} + \left\langle \varepsilon^2 \right\rangle \delta_{n,k} \right] = \delta_{m,k}$$

Thus, inverting the covariance matrix yields the following expression for the LP coefficients:

$$\phi_{n(h)} = \sum_{m=0}^{N-1} \Delta_{n,m} \hat{\Gamma}_{m,-h}$$

Note again that the inverse covariance matrix is taken over past values of the time series only. This is a very important aspect of the linear regression, in that the past values are supposed to be the best linear predictors of the future values. Otherwise, if the future values were “known” and were included in the regression, then the future values would be the best linear predictors of themselves and we would have (leaving out the noise term, and taking the inverse over all past and future values of the indices):

$$\phi_{n(h)} \stackrel{?}{=} \sum_{\text{all } m} \Gamma^{-1}_{n,m} \Gamma_{m,-h} \equiv \delta_{n,-h}$$

Then the main problem in prediction is how to estimate the time-dependent covariance matrix of past values of the time series, along with the estimated covariance $\hat{\Gamma}_{m,-h}$ between the future signal $\hat{X}_{-h}$ and the past signal $\hat{X}_m$. Taking the inverse of the covariance matrix is the major problem of computing the LP, and this problem will be treated in more detail shortly.

As an alternative, the best linear predictor of $\hat{X}_{-h}$ may be computed by computing the best linear predictor of $\hat{X}_{-1}$ and then iterating this forward in time, using the previously computed predicted values as data in the next iteration. This method is better suited to stationary time series, however. In the non-stationary case the LP coefficients should be re-computed for each iteration, and this would involve using the previously computed predictors as data in the computation. Using this iterative method does, however, introduce an additional element of "feedback" into the regression.

7.5 Inverting the Covariance Matrix

The inversion of the covariance matrix is the major problem of computing the Linear Prediction. Fortunately the Wavelet formalism yields an elegant solution of this problem and a vast simplification of it. This is fortunate since financial data are sparse, and to try to estimate the full multivariate covariance matrix would be hopeless. But first it is necessary to clarify a point concerning the difference between past and future states of the time series.

Each element of a time series is a random variable, which may be interpreted as a linearly independent basis vector of a Hilbert space. This includes both past values and future values. But, of course, the future values are not yet known. The job of (linearly) regressing the future values on the past values means trying to find the best estimate of the future value as a linear function of the past values in some set. So, for example, we may regress the future values on a set of past values consisting of N = 2048 values, which form a Hilbert space of dimension 2048. Then the linear regression means estimating the projection of the future values on this 2048-dimensional Hilbert space of past values. If the future values are orthogonal to the past values, then this means they are uncorrelated with the past values, and thus cannot be estimated at all. So the task is to estimate the projection of the future values on the Hilbert space of the past values, which is what the estimated future covariance matrix $\hat{\Gamma}_{m,-h}$ measures.

When the covariance matrix is expressed in terms of wavelet filter kernels and a constant wavelet variance on each level in the DWT approach, it is Toeplitz since the filter kernels depend only on the difference of their indices. In general, taking a time-dependent wavelet variance with different values for different values of the indices, the covariance matrix is then not Toeplitz. But in any case the past covariance matrix is truncated to a finite range of index values and does not range over the whole index set of the filter kernels. Otherwise the properties of the filter kernels, as described later, could be used to easily invert the covariance matrix. The conventional Toeplitz matrix inversion routine [NR, p.90] could be used to invert the covariance matrix in the case where the level wavelet variance is taken to be constant, but this is somewhat computation intensive and awkward. In fact, this is the same method as used in the OLP filter in the stationary approximation, since the covariance matrix is in fact stationary in this case. More generally, the Cholesky decomposition [NR, p.96] for inverting a positive definite symmetric matrix should be used to invert the WLP covariance matrix over the space of past values. This method is somewhat time-consuming, but to be safe this is probably the preferred method.

Projected Covariance Matrix

The major advantage of the wavelet approach is that, when the simplifying assumption is made that all DWT coefficients except for those with the same level index and time index are orthogonal; the covariance matrix in the wavelet basis is diagonal. But this wavelet basis is over the whole $2N$-dimensional space of the time series and includes both past and future values. This means that the covariance matrix $\tilde{\Gamma}_{m,n}$ in the time basis over the whole space has the Singular Value Decomposition (SVD) form [NR] and hence is easy to invert using the DWT wavelet transform and its inverse. But this cannot be applied to the space of the past values of the time series alone. To demonstrate this, let us try to use this method to solve the LP equation, by first extending it to the whole 2N-dimensional space of the time series. As noted previously, if we simply apply the inverse covariance matrix to both sides of the extended equation, we arrive at zero for the (past) LP coefficients. To avoid this, we must define a projection operator onto the space of past values and define sums in the LP equation to be zero for future values of the indexes m,n:

$$\sum_{n=-N}^{N-1} P_m \tilde{\Gamma}_{m,n} P_n \tilde{\phi}_n^{(h)} = P_m \hat{\Gamma}_{m,-h} \qquad P_{m,n} \equiv \begin{cases} 1 & (m,n \geq 0) \\ 0 & (m,n < 0) \end{cases}$$

The extension of the covariance matrix and its inverse to the whole 2N-dimensional space is denoted by a tilde, as opposed to the usual definition of the inverse covariance matrix Δ, which is the inverse over the past values of the time series only. Now we apply the inverse covariance matrix $\tilde{\Delta}_{k,m}$, defined over the whole 2N-dimensional space in terms of the DWT transform and wavelet variance over this space:

$$\sum_{n=-N}^{N-1} \left[\sum_{m=-N}^{N-1} \left(P_k \tilde{\Delta}_{k,m} P_m \right) \tilde{\Gamma}_{m,n} \right] P_n \tilde{\phi}_n^{(h)} = \sum_{m=-N}^{N-1} \left(P_k \tilde{\Delta}_{k,m} P_m \right) \hat{\Gamma}_{m,-h}$$

Once again we have projected both sides onto the Hilbert space of past values of the time series. The quantities in parentheses are the projection of the inverse covariance matrix over the whole 2N-dimensional space onto the space of past values. The quantity in brackets on the left side would be the identity matrix if the projection operators were not present, but due to the projection operators the projected inverse covariance matrix is no longer the inverse of the full covariance matrix over the whole space. So we see that we cannot use this method to invert exactly the covariance matrix when it is expressed in terms of the DWT transform and wavelet variance over the whole 2N-dimensional space. However, this method will serve as an *approximate* method of inverting the covariance matrix defined over past values only, when the matrix $\tilde{\Gamma}_{m,n}$ is restricted to past values of the indices. If that is the case, then the LP coefficients are given by the expression on the right side of the above equation. This means that we are using the projection $P_k \tilde{\Delta}_{k,m} P_m$ of the inverse covariance matrix $\tilde{\Delta}_{k,m}$ over the whole space as an approximation for the inverse covariance matrix $\Delta_{k,m}$ over the past values only.

We may recast the formalism of the projection operators in a slightly different form if we define a new covariance matrix over the whole data range, both past and future. Let us define on the left hand side of the regression equation a new covariance matrix:

$$\begin{aligned} \widehat{\Gamma}_{m,n} &\equiv \hat{\Gamma}_{m,n} && \text{for } m,n \geq 0 \\ &\equiv 0 && \text{otherwise} \end{aligned}$$

For the right hand side of the regression equation we define a new covariance matrix as follows:

$$\begin{aligned} \widehat{\Gamma}_{m,-h} &\equiv \hat{\Gamma}_{m,-h} && \text{for } m \geq 0 \\ &\equiv 0 && \text{for } m < 0 \end{aligned}$$

In terms of these new matrices, the linear regression equation may be written, for all *m*:

$$\sum_{n=N-1}^{-N} \left[\widehat{\Gamma}_{m,n} + \left\langle \varepsilon^2 \right\rangle \delta_{m,n} \right] \phi_{n(h)} = \widehat{\Gamma}_{m,-h}$$

This breaks down into two separate equations:

$$\begin{aligned} \sum_{n=N-1}^{0} \left[\hat{\Gamma}_{m,n} + \left\langle \varepsilon^2 \right\rangle \delta_{m,n} \right] \phi_{n(h)} &= \hat{\Gamma}_{m,-h} && \text{for } m \geq 0 \\ \sum_{n=-1}^{-N} \left\langle \varepsilon^2 \right\rangle \delta_{m,n} \phi_{n(h)} = \left\langle \varepsilon^2 \right\rangle \phi_{m(h)} &= 0 && \text{for } m < 0 \end{aligned}$$

Thus these regression equations are equivalent to the previous one. Inverting the matrix in brackets over the whole index space solves the regression equation. But since this matrix is block-diagonal, its inverse is given in terms of the previous inverse as follows:

$$\begin{aligned} \left[\widehat{\Gamma}_{m,n} + \left\langle \varepsilon^2 \right\rangle \delta_{m,n} \right]^{-1} &\equiv \Delta_{m,n} && \text{for } m,n \geq 0 \\ &\equiv \left\langle \varepsilon^2 \right\rangle^{-1} \delta_{m,n} && \text{otherwise} \end{aligned}$$

Thus when we invert the covariance matrix over the *whole* index space with these definitions, we arrive back at the same solution to the regression problem. But defining these covariance matrices over the whole space in such a way prevents their estimation in the wavelet basis. This is because in the wavelet basis there is no sharp cut-off as defined above.

Covariance Matrix in Wavelet Basis

The matrix $\hat{\Gamma}_{m,n}$, for all values of indices both past and future, is what we estimate in the diagonal wavelet basis by estimating the past and future signal variance. It is very convenient therefore to be able to invert the covariance matrix in this diagonal basis. Let us therefore write the regression equation again as follows, extending the indices on the signal covariance matrix $\hat{\Gamma}_{m,n}$ to the whole data set and projecting onto the past values of the indices:

$$\sum_{k=N-1}^{-N} P_m \left[\hat{\Gamma}_{m,k} + \left\langle \varepsilon^2 \right\rangle \delta_{m,k} \right] P_k \phi_{k(h)} = P_m \hat{\Gamma}_{m,-h}$$

This means that the covariance matrix has been extended to future values of the index *m*. We may then define a new inverse covariance matrix over the whole data set, both past and future values of the indices, as follows:

$$\sum_{m=N-1}^{-N} \left[\hat{\Gamma}_{n,m} + \left\langle \varepsilon^2 \right\rangle \delta_{n,m} \right] \tilde{\Delta}_{m,k} = \sum_{m=N-1}^{-N} \tilde{\Delta}_{n,m} \left[\hat{\Gamma}_{m,k} + \left\langle \varepsilon^2 \right\rangle \delta_{m,k} \right] = \delta_{n,k}$$

Now we make the *approximation* that the matrix $\hat{\Gamma}_{m,k}$ is zero for future values of its indices (compared to past values and the noise term), apart from the noise term:

$$\hat{\Gamma}_{m,k} \approx 0 \qquad (m < 0, k < 0)$$

Thus we have *approximately*:

$$\sum_{m=N-1}^{-N} \tilde{\Delta}_{n,m} P_m \left[\hat{\Gamma}_{m,k} + \left\langle \varepsilon^2 \right\rangle \delta_{m,k} \right] P_k \approx \delta_{n,k} P_k$$

The regression equation is then solved *approximately* in terms of this inverse as follows:

$$\phi_{n(h)} \approx \sum_{m=N-1}^{-N} \tilde{\Delta}_{n,m} P_m \hat{\Gamma}_{m,-h} \qquad (n \geq 0)$$

In this equation we simply truncate the sum and sum only over values of the index *m* corresponding to past times. These covariance matrices and their inverses are then easily calculated in the diagonal wavelet basis. The approximation consists of the assumption that the covariance matrix goes to that of random noise in the future, corresponding to equal wavelet variances on all wavelet levels for all future times, and that the signal covariance matrix for future times is small (compared to the past signal covariance matrix and noise term). The right hand side is first-order in the small matrix $\hat{\Gamma}_{m,-h}$, but the approximation applies to the matrix $\hat{\Gamma}_{m,k}$ only.

Small-Correlation Approximation

We note that the above is a reasonable approximation because the signal covariance matrix $\hat{\Gamma}_{m,n}$ is small compared to the noise covariance matrix $\left\langle \varepsilon^2 \right\rangle \delta_{m,n}$. In fact for the covariance matrix on the left hand side it might even be reasonable to drop the signal covariance matrix entirely and approximate it as the noise covariance alone. Then we arrive at the following approximation:

$$\phi_{n(h)} \approx \left\langle \varepsilon^2 \right\rangle^{-1} \hat{\Gamma}_{n,-h} \qquad (n \geq 0)$$

This may be considered a "0$^{\text{th}}$ order approximation". For a 1$^{\text{st}}$ order approximation, we may write:

$$\begin{aligned} \tilde{\Gamma}_{m,n} &\equiv \left\langle \varepsilon^2 \right\rangle^{+1} \left[\delta_{m,n} + \hat{\Gamma}_{m,n} \big/ \left\langle \varepsilon^2 \right\rangle \right] \\ \tilde{\Delta}_{m,n} &\approx \left\langle \varepsilon^2 \right\rangle^{-1} \left[\delta_{m,n} - \hat{\Gamma}_{m,n} \big/ \left\langle \varepsilon^2 \right\rangle \right] \end{aligned}$$

This inverse can also be calculated easily in the wavelet basis, since it is also diagonal. Due to the large uncertainty in the covariance matrix and the necessity for smoothing, these approximations should be just as good as trying to compute the inverse over past values only of the smoothed covariance matrix in the time basis.

Stationary Approximation

Instead of the decomposition of the full covariance matrix $\tilde{\Gamma}_{m,n}$ into a "signal" term $\hat{\Gamma}_{m,n}$ and a noise term, which is a somewhat arbitrary decomposition, it might make more sense to decompose the covariance matrix into a stationary positive-definite covariance matrix $\bar{\Gamma}_{m,n}$ and a non-positive definite matrix $\Upsilon_{m,n}$, which represents the non-stationary part or deviation from stationarity. All these matrices are defined for all index values, both past and future. In the wavelet basis the stationary covariance matrix corresponds to the level average wavelet variances, and the non-stationary part corresponds to the deviations of each level wavelet variance from the level average value (hence it is non-positive definite).

Hence, we have for the full covariance matrix $\tilde{\Gamma}_{m,n}$:

$$\tilde{\Gamma}_{m,n} \equiv \left\langle \varepsilon^2 \right\rangle \delta_{m,n} + \hat{\Gamma}_{m,n} \equiv \bar{\Gamma}_{m,n} + \Upsilon_{m,n}$$

We now define the inverse $\bar{\Delta}_{m,n}$ of the stationary part of the covariance matrix:

$$\sum_{m=N-1}^{-N} \bar{\Gamma}_{n,m} \bar{\Delta}_{m,k} = \sum_{m=N-1}^{-N} \bar{\Delta}_{n,m} \bar{\Gamma}_{m,k} = \delta_{n,k}$$

We may now write a 1st order approximation to the full covariance matrix and its inverse in terms of these new matrices as:

$$\begin{aligned} \tilde{\Gamma} &\equiv \bar{\Gamma}\left[I + \bar{\Delta}\Upsilon \right] \equiv \left[I + \Upsilon\bar{\Delta} \right]\bar{\Gamma} \\ \tilde{\Delta} &\approx \bar{\Delta}\left[I - \Upsilon\bar{\Delta} \right] \approx \left[I - \bar{\Delta}\Upsilon \right]\bar{\Delta} \end{aligned}$$

Depending on the smoothing, we presume that the non-stationary part $\Upsilon_{m,n}$ of the covariance matrix is small compared to the stationary part, and that most of the variation in the covariance matrix is random noise. Furthermore, it falls off rapidly to zero for future time indices due to the smoothing. In fact the stationary covariance matrix is the correct asymptotic limit for times far in the future. Hence in a 1st order approximation we may approximate the $\Upsilon_{m,n}$ term as negligible in the inverse matrix, and approximate the regression equation as:

$$\phi_{n(h)} \approx \sum_{m=N-1}^{-N} \bar{\Delta}_{n,m} P_m \hat{\Gamma}_{m,-h} \qquad (n \geq 0)$$

Using this approximation can perhaps increase the stability and reduce the stochastic noise of the filter. Also the stationary part of the covariance matrix can be inverted by conventional means such as a Toeplitz routine if desired, thereby further increasing the accuracy of the regression equation. The non-stationary part of the inverse covariance matrix is a 2nd order correction, which can be ignored.

Estimation of Wavelet Variance

To compute the matrix $\hat{\Gamma}_{m,n}$ in the wavelet basis, it is necessary to estimate the time-dependent wavelet variance on each level, for both past and future index values.

These need to be smoothed, and the future wavelet variance should approach a constant average wavelet variance value on each level corresponding to a constant stationary covariance matrix $\overline{\Gamma}_{m,n}$. To compute the level wavelet variance, the most practical method is to compute the wavelet variance using reflection boundary conditions. (This makes the most recent values of wavelet variance count a little more in the average in lieu of future values.) Then the time-dependent wavelet variance is cut off at the wavelet index corresponding to time zero, and the level averages substituted for the future values. Then the entire level wavelet variance is smoothed, preferably with Savitzky-Golay acausal smoothing to preserve the long-term variations. At the present point in time, the smoothed value is roughly the average of the most recent variance values and the average level variance, leading to a form of "shrinkage" [PW]. The future values of the covariance matrix tend rapidly to the stationary average values. This then leads to the optimum method of capturing non-stationary variations in the wavelet variance while simultaneously ensuring that the stochastic noise is adequately suppressed.

This covariance matrix is presumed to be diagonal in the wavelet basis, so it may be immediately inverted over the whole index set using the filter kernels, or equivalently, DWT or MODWT wavelet transforms. The covariance matrix may be separated into a stationary and non-stationary part by simply choosing the average wavelet variance or deviations from average on each level. The inverse wavelet variance is obtained by simply taking the inverse of the level wavelet variances, whether smoothed or in the stationary approximation (averaged). The matrix $\hat{\Gamma}_{m,-h}$ may be similarly estimated in terms of the same wavelet variance and the filter kernels, over the whole data set, with an arbitrary constant noise term subtracted off of each level to give only the estimated signal covariance, since this noise term is diagonal and does not make a contribution. But this covariance matrix is expected to fall off to zero rapidly as the time lag increases, so its main contribution will be in the region of "shrinkage" close to time zero. Hence the shrinkage will help stabilize the non-stationary part of this covariance matrix. This then solves the regression problem in the wavelet basis.

7.6 Predictors: Relative Price, Velocity, Acceleration

For a more effective linear prediction filter, we may extend the univariate regression described above to a multivariate regression. To do this we use a set of arbitrary functions of the past data set as a set of basis functions or **predictors**, and measure the correlation of these functions with future returns to determine their predictive power. We may then generalize the univariate regression to a multivariate regression, on a set of *M* predictors. The regression equation may be written as a sum over the predictors, labeled by a Greek index running from 0 to *M*, with the index 0 denoting the original time series itself (or some surrogate) as a predictor:

$$\sum_{\beta=0}^{M-1}\sum_{n=0}^{N-1} Y_n^{(\beta)} \phi_{n(h)}^{(\beta)} \equiv \hat{X}_{-h}$$

The covariance matrix between the predictors is given as before by:

$$\Gamma_{m,n}^{(\alpha,\beta)} \equiv \left\langle Y_m^{(\alpha)}, Y_n^{(\beta)} \right\rangle \qquad \Gamma_{m,n}^{(\alpha)} \equiv \left\langle Y_m^{(\alpha)}, X_n \right\rangle$$

Thus, taking covariances on both sides, we find the following multivariate LP equation:

$$\sum_{\beta=0}^{M-1}\sum_{n=0}^{N-1} \Gamma_{m,n}^{(\alpha,\beta)} \phi_{n(h)}^{(\beta)} = \hat{\Gamma}_{m,-h}^{(\alpha)}$$

Finally, computing the inverse of this covariance matrix over the space of the past data, we find the following equation for the generalized LP coefficients:

$$\phi_{n(h)}^{(\beta)} = \sum_{\alpha=0}^{M-1}\sum_{m=0}^{N-1} \Delta_{n,m}^{(\beta,\alpha)} \hat{\Gamma}_{m,-h}^{(\alpha)}$$

This generalized covariance matrix and its inverse are extremely complicated and virtually impossible to "estimate", as noted above. But the wavelet approach affords a vast simplification. We *assume* that the DWT wavelet coefficients of each predictor are orthogonal except for those with the same level and time index. Then we compute the time-dependent covariance matrix between the different predictors just using these wavelet coefficients with the same index. The result may be smoothed and used as an element of the multivariate covariance matrix, by direct analogy with the time-dependent variance in the univariate covariance matrix.

Generally, in a multivariate regression, such as for example taking the returns from all the securities in a portfolio as predictors and doing a vector auto-regression of all these securities on each other, to compute a time-dependent covariance between all the securities would be too difficult and time-consuming. So the variances and cross-covariances between the securities can be computed on each wavelet level, by multiplying the wavelet coefficients of the same level and time index, and level averages taken to form the M×M covariance matrix of the portfolio (level-by-level). This level-average covariance matrix may be diagonalized, and in the diagonal basis the components of the diagonal vector of securities can be projected by univariate regression. After this is completed, the inverse diagonalization may be performed to yield the separate securities again together with the projection yielded by the vector auto-regression.

We may arrive at a variant of the usual auto-regression in the form of a univariate regression that uses a single predictor, by adapting the above multivariate regression formalism. In the case of a single predictor the regression equation takes the form:

$$\sum_{n=0}^{N-1} Y_n^{(\alpha)} \phi_{n(h)}^{(\alpha)} \equiv \hat{X}_{-h}$$

Taking covariances on both sides with the single predictor, we find the following univariate LP equation:

$$\sum_{n=0}^{N-1} \Gamma_{m,n}^{(\alpha,\alpha)} \phi_{n(h)}^{(\alpha)} = \hat{\Gamma}_{m,-h}^{(\alpha)}$$

Finally, computing the inverse of the covariance matrix for the predictor over the space of the past data, we find the following equation for the LP coefficients:

$$\phi_{n(h)}^{(\alpha)} = \sum_{m=0}^{N-1} \Delta_{n,m}^{(\alpha,\alpha)} \hat{\Gamma}_{m,-h}^{(\alpha)}$$

In this way the univariate regression for a single predictor is very similar in form to that for an auto-regression over the returns data itself. In the future estimated covariance, the covariance between the predictor and the returns in the wavelet basis takes the place of the wavelet variance in the auto-regression case. For example the single predictor might be some linear combination of the smoothed price/returns data consisting of a sum of suitably chosen predictors. This eliminates the need to compute and diagonalized a time-dependent covariance matrix between several predictors.

In particular, we may define a set of three predictors from the original time series (returns), consisting of **Relative Price**, **Velocity**, and **Acceleration**. The Relative Price is the difference of two smoothings of the price data, the Velocity is just the returns themselves (1^{st} difference of price), and the Acceleration is the 1^{st} difference of returns (2^{nd} difference of price). All three of these predictors are linear combinations of the original returns data. However, we may take all of these predictors to have additional smoothing of different types such as **Savitzky-Golay** smoothing, and each predictor should be projected using the OLP filter or some other method. This is to facilitate measuring the correlation with future returns. Also the time lag or phase of each predictor should be adjustable for maximum correlation with future returns. Then the time-dependent wavelet covariance on each level between each predictor and past returns, with the phase of the predictor suitably adjusted, is used to estimate the future covariance matrix of that predictor in the LP equation, by replacing the future values by the level averages and then smoothing the covariances.

The covariances between these three predictors could be computed individually as described previously. However, after adjusting the time lag of each predictor, they will probably be very highly correlated and the covariance matrix between them close to singular. So a better procedure would be to simply add these three predictors together, after adjusting their phases for maximum correlation with future returns, with adjustable weights. This avoids the problem of having to compute a nearly singular covariance matrix, and at the same time should result in some noise reduction and increased predictive power. This predictor can then be taken as the index-0 predictor in the univariate regression, replacing the returns series itself.

We then compute the time-dependent wavelet variance of this predictor over the past data set, and smooth it using a smoothing window or SG smoothing, to get the past covariance matrix of the predictor. We must also compute the time-dependent wavelet covariance between the predictor and future returns. The future covariance between the predictor and future returns may be estimated by computing the level average of this covariance, using this level average for the future estimated values of the covariance, then applying the smoothing. From these time-dependent variances of the predictors and covariances between predictors and returns, the corresponding covariance matrices are constructed by use of the filter kernel. The inverse over past values of the predictor covariance matrix is then taken and multiplied by the estimated future covariance matrix to obtain the LP coefficients and from these the future projection of the returns series.

We could also regress over the Relative Price, Velocity, and Acceleration as three independent predictors. By regressing over the Relative Price, Velocity, and Acceleration with Wavelet smoothing, we are actually extending the correlation structure of the returns series to a 3-parameter (with 1 constraint) set of fixed correlations. Then the individual wavelet coefficients of the returns are no longer regarded as statistically independent. Evidently the Relative Price (integral of returns) corresponds to a positive correlation between neighboring wavelet coefficients (and correlation with lower levels), while the Acceleration (derivative of returns) corresponds to a negative correlation between neighboring wavelet coefficients (and correlation with higher levels). Also the time lag of these three predictors must be taken into account. So these correspond to definite assumptions about the correlation structure of the wavelet coefficients of returns, so they do not need to be "estimated" as in OLS. By measuring the correlation of these three predictors with future returns and adjusting the time lags and weights for optimum correlation, we may then simply add three predictors to form a single predictor, thereby optimizing the correlation structure of the wavelet coefficients of returns.