Least-Squares Spectral Analysis

Theory Summary

Reference: Mtamakaya, J. D. (2012). Assessment of Atmospheric Pressure Loading on the International GNSS REPRO1 Solutions Periodic Signatures. Ph.D. dissertation, Department of Geodesy and Geomatics Engineering, Technical Report No. 282, University of New Brunswick, Fredericton, New Brunswick, Canada, 208 pp.

1. LEAST-SQUARES SPECTRAL ANALYSIS

Least Squares Spectral Analysis is a powerful software developed at the University of New Brunswick, Fredericton. LSSA was first developed by Vaníček (1968, 1971) as an alternative to the classical Fourier methods (Pagiatakis, 1998). It has been revised by Wells, Vaníček and Pagiatakis in 1985 in order to eliminate certain "bugs" and to be more versatile (Wells, 1985). In this study version LSSA v. 5.02 is used developed by Spiros Pagiatakis.

LSSA has several advantages over the commonly known Fourier transform. Table 1 summarizes these and also the limitations of the Fourier transform based on Pagiatakis [1998, 2008].

Table 1. Advantages of LSSA in comparison with the classical Fourier transform

	LSSA	**FOURIER TRANSFORM**
Time series with unequally spaced values can be analyzed without pre-processing	True	False
Time series does not have to be continuous (allowance of data gaps)	True	False
Time series with an associated covariance matrix can be analyzed	True	False
Time series can have datum shifts and trends	True	False

No limitations for the length of the time series	True	False
Higher accuracy achieved with longer time series	True	False
Statistical testing on the significance of spectral peaks can be performed	True	False

Perhaps the most essential advantage of LSSA is the allowance of data gaps in the observed time series, what is common in GPS pseudorange and carrier-phase observations. The original time series can be directly used without interpolating or adding artificial data to fill the data gaps, hence not degrading the results of the spectral analysis. LSSA also permits to add bigger weights to statistically more significant observations and smaller weight to the other ones. This can be useful in geodetic applications when analyzing pseudorange and carrier-phase observations, assigning more weight to the more precise carrier-phase and less weight to pseudorange observations. Let us discuss the LSSA and the other great properties which this software offers in the next section.

4.1 Theory of signal and noise

The observed time series can be considered to be composed of *signal*, what we are interested to study, and *noise* which obscure the original signal. The noise can be *random* or *periodic*. Idealizing, we can think about the random noise as a *white noise* what is uncorrelated, has a constant spectral density, and it may or may not have a Gaussian distribution. Usually in practice we deal with a *non-white random noise*. Systematic noise is a noise what can be described by a certain functional form. It can be periodic (colored) or non-periodic. Non-periodic noise can be datum-shifts (offsets) and trends (linear, exponential, quadratic, etc.), and it causes the statistical properties of the series to be a function of time (Pagiatakis, 1998).

4.2 Mathematics of the least-squares spectrum

Let us consider a discrete time series $\mathbf{f}(t_i)$ in Hilbert space, (H) where t_i is a vector of observation times and $i = 1,2,\ldots,m$, m is the number of data points in the time series. We assume that a fully populated covariance matrix $\mathbf{C_f}$ is available for the time series. One of the main objectives of LSSA is to detect unknown periodic signals in **f,** especially when **f** contains systematic variations of unknown magnitude whose functional forms are known. After Vaníček (1971) we can call these known constituents *systematic noise* since their presence is nuisance from the point of view of detecting the unknown periodic signal (Vaníček, 1981).

The idea is to approximate **f** with another function **g** using the least squares principle, so that the differences between **f** and **g** (the vector of residuals **r**) are minimum in the least-squares sense. Thus the time series **f** can be modeled by **g** as follows (Pagiatakis, 1998):

$$\mathbf{g} = \mathbf{\Phi} \cdot \mathbf{x}, \qquad (9)$$

where **x** is the vector of unknowns and **Φ** is the design matrix (matrix of normal equations) which consist of several column vectors: $\mathbf{\Phi = [\Phi_1 , \Phi_2, \ldots,\Phi_m]}$ called based functions, each of which is a know function of the same dimension as **f** (Wells, 1985). Matrix **Φ** specifies the functional form of both signal and (systematic) noise. Using the standard least-squares notation we can write for the residual series (Pagiatakis, 1998).

$$\hat{\mathbf{r}} = \mathbf{f} - \hat{\mathbf{g}} = \mathbf{f} - \mathbf{\Phi}\left(\mathbf{\Phi}^{\mathbf{T}}\mathbf{C}_{\mathbf{f}}^{-\mathbf{1}}\mathbf{\Phi}\right)^{-\mathbf{1}}\mathbf{\Phi}^{\mathbf{T}}\mathbf{C}_{\mathbf{f}}^{-\mathbf{1}}\mathbf{f}\,. \tag{10}$$

In equation (10), $\hat{\mathbf{g}}$ is the orthogonal projection of **f** onto the subspace **S** (from Hilbert space) generated by the column vector of **Φ** (see Figure 1 for illustration). It follows from the projection theorem that $\hat{\mathbf{r}} \perp \hat{\mathbf{g}}$. This means that **f** has been decomposed into a *signal* $\hat{\mathbf{g}}$ and *noise* $\hat{\mathbf{r}}$.

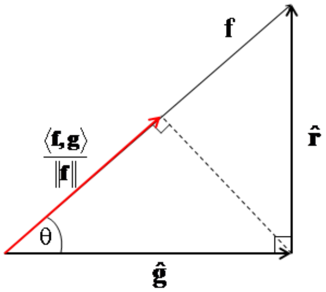
$\mathbf{f}$
$\hat{\mathbf{r}}$
$\frac{\langle \mathbf{f},\mathbf{g}\rangle}{\|\mathbf{f}\|}$
θ
$\hat{\mathbf{g}}$

Figure 1. Orthogonal projection of vectors $\hat{\mathbf{g}}$ and $\hat{\mathbf{r}}$

In Figure 1 the red arrow represents the orthogonal projection of $\hat{\mathbf{g}}$ into **f**. Then the inner product (dot product) of these two vectors is:

$$\langle \mathbf{f}, \hat{\mathbf{g}} \rangle = |\mathbf{f}| \cdot |\hat{\mathbf{g}}| \, cos\,\theta \tag{11}$$

and the orthogonal projection is then:

$$\frac{\langle \mathbf{f}, \hat{\mathbf{g}} \rangle}{|\mathbf{f}|} = |\hat{\mathbf{g}}| \, cos\,\theta\,. \tag{12}$$

From equation (11) one can conclude that if the angle between $\hat{\mathbf{g}}$ and **f** is 0 degrees, then $\hat{\mathbf{g}}$ fully represents **f**. Thus the ratio of the length of the orthogonal projection multiplied by 100% and the length of **f** tells how much the least-squares approximation of **g** represents **f** in terms of percent. The ratio is smaller than unity and can be expressed as follows:

$$\mathbf{s} = \frac{\langle \mathbf{f}, \hat{\mathbf{g}} \rangle / |\mathbf{f}|}{|\mathbf{f}|} = \frac{\langle \mathbf{f}, \hat{\mathbf{g}} \rangle}{|\mathbf{f}|^2} = \frac{\mathbf{f}^{\mathbf{T}} \mathbf{C}_{\mathbf{f}}^{-\mathbf{1}} \hat{\mathbf{g}}}{\mathbf{f}^{\mathbf{T}} \mathbf{C}_{\mathbf{f}}^{-\mathbf{1}} \mathbf{f}} \tag{13}$$

where **s** is the least-squares spectrum in percentage ($0 < \mathbf{s} < 1$). The larger the **s** (closer to 1) the better is the least-squares fit to the data (Pagiatakis, 2008).

In spectral analysis it is usual to search for periodic signals which can be expressed in terms of sine and cosine base functions. Specifically for LSSA we know **f(*t*)** and we use:

$$\mathbf{\Phi} = \lfloor cos\, \omega_j t_i , sin\, \omega_j t_i \rfloor \tag{14}$$

where ω_j is a spectral frequency for which spectral values $\left(s\left(\omega_j \right) \right)$ are desired. The vector of spectral frequencies can be written as: $\Omega = \{\omega_j\}$ where $j = 1,2,\ldots,k$ (Wells, 1985). The orthogonal projection of **f** onto **S** will be different for each $\omega_j \in \Omega$ and each $\omega_j \in \Omega$ is tried independently from the rest (Pagiatakis, 1998). For each ω_j for which we want $\left(s(\omega_j)\right)$ we compute:

$$g(\omega_j) = \hat{x}_1 \cos\omega_j t + \hat{x}_2 \sin\omega_j t \,, \tag{15}$$

where $\hat{\mathbf{x}} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}$ is determined from (Wells, 1985):

$$\hat{\mathbf{x}} = \left(\mathbf{\Phi}^{\mathbf{T}} \mathbf{C}_{\mathbf{f}}^{-\mathbf{1}} \mathbf{\Phi}\right)^{-\mathbf{1}} \mathbf{\Phi}^{\mathbf{T}} \mathbf{C}_{\mathbf{f}}^{-\mathbf{1}} \mathbf{f} \,. \tag{16}$$

Then the LSS is defined by

$$\mathbf{s}\left(\omega_j\right) = \frac{\mathbf{f}^{\mathbf{T}} \mathbf{C}_{\mathbf{f}}^{-\mathbf{1}} \hat{\mathbf{g}}\left(\omega_j\right)}{\mathbf{f}^{\mathbf{T}} \mathbf{C}_{\mathbf{f}}^{-\mathbf{1}} \mathbf{f}} \,, \ j = 1,2,\ldots,k \,. \tag{17}$$

When calculating the LSS there is a simultaneous least-squares solution for the parameters of the process. This is a rigorous approach to the hidden periodicities: the parameters of the assumed linear system driven by noise are determined simultaneously with the amplitudes and phases of the periodic components, and with other parameters that describe systematic noise (Pagiatakis, 1998).

For example the approximating function can be a sine wave with known frequency:

$$\mathbf{g} = a \cos(2\pi\omega_j + \phi) \tag{18}$$

where a is the amplitude and ϕ is the phase for a specific frequency ω_j which gives the best approximation (least-squares fit) to **f**. For different frequencies the LSS **s** varies (**s** is a function of frequency) and so it becomes the representation of how well the sine wave fits the data **f**. The time series is best approximated when **s** equals to 1 for a certain predefined frequency ω_j (Pagiatakis, 2008).

4.3 Least-squares spectrum evaluation

Table 1 also lists that LSSA allows performing statistical testing on the significance of spectral peaks. This is very important since we need to know which peak is statistically significant and which one can be suppressed. Vaníček [1971] derived the expected (mean) spectral value of white noise in the LSS for series consisting of statistically independent random values. He pointed out the possibility of deriving magnitudes (threshold values) above which spectral peaks are statistically significant (Pagiatakis, 1998).

Postulating that series **f** has been derived from a population of random variables following the multidimensional normal distribution $n(\mathbf{0}, \mathbf{C_f})$. This null hypothesis ($H_0$) implies that if:

$$Q_n / Q_s \geq (v/2) F_{v,2,\alpha} \tag{19}$$

then the series will contain statistically insignificant noise. In equation (18) Q is a random variable with subscripts n and s referring to signal and noise, respectively; v is the degree of freedom and F stands for *F-distribution* and α is the significance level. The alternative hypothesis H_1 is:

$$Q_n / Q_s < (v/2) F_{v,2,\alpha} \ . \tag{20}$$

It makes sense to use only the lower tail end of *F*, since large values of the ratio Q_n/Q_s (upper tail of F) imply the signal is significantly larger than the noise and hence can not be detected. Then the statistically significant spectral peaks will satisfy the following inequality (Pagiatakis, 1998):

$$\mathbf{s}(\omega_j) \geq \left[1 + \frac{v}{2} F_{v,2,\alpha}\right]^{-1}. \qquad (21)$$

Once a statistically significant signal is detected it should be related to the physics of the observed system. Once it is understood the detected signal becomes systematic noise and then further hidden periodicities can be searched (Vaníček, 1981).

4.4 Input and output parameters for LSSA

The *input parameters* to compute the spectrum consist of the time series, the limits and density of the spectral band to be produced and the known constituents' base functions (representing the systematic noise functional forms) (Vaníček, 1985). The input file, which contains the user-defined parameters, is called *lssa.in*. There are eight blocks of parameters that need to be specified for the analysis of the series before each run. Let us shortly describe the parameters to be defined in *lssa.in* (Vaníček, 1985).

First, the user must enter the name of the file to be analyzed, for example **data_ser.dat**, the number of data points, the units of time of the series and the units of the values of the series. The first line in the **data_ser.dat** is a text line containing the name of the project, followed by three columns. The first column stands for the time of the series (years, days, seconds, etc.), the second column is the time series itself and the third column represents the standard deviation of the time series if known. If the standard deviations are unknown the user can replace them with values he/she may think are reasonable. The units of the standard deviations are the same as the units of the time series.

Next, the user can identify the known constituents in the time series according to his/her best knowledge. There are four optional types of known constituents build in the algorithm. The software makes it possible for the user to specify his/her own functions as well. The known constituents build in LSSA are the follows:

- **Random constant (Datum Shifts)**.

The user can specify the number of datum biases in the time series and the epochs when the shifts are suspected. The minimum number of shifts is one which represents the beginning of the time series.

- **Linear trend**.

The user can decide whether to calculate linear trend or not.

- **Periodic signals.**

Certain periodic constituents can be forced (fitted) to the series. The result is a sine wave given by its amplitude and phase, along with a statistical test on its significance. The user can enter the number of sine waves to be forced, the periods and names of each sine wave.

- **User specified**

If the user believes that some nonlinear trend (exponential, polynomial) exists he/she can specify it here. In general, a polynomial fit of maximum order 5 can be achieved.

The other parameters to be defined in *lssa.in* are:

- **Processes**

Certain random processes can be tried on the series. If an autoregressive (AR) process up to the order 5 is suspected then the user should switch to 1. In this case the coefficient of the AR process along with a statistical test is calculated. The latter indicates on output whether it is significant or not. In addition a statistical test is performed on the calculated coefficient to specify whether it is different from unity (random walk). If the coefficient is statistically equal to unity then it is a random walk.

- **Characteristics of series**

This block gives information about the covariance of the series. Only a diagonal covariance matrix is accepted. If the a-priori variance factor of the input series is known the user should switch to 1. If so, the values of the series are considered equally weighted and the weight is set to unity. Otherwise the weight of each value of the series is calculated as the inverse of the variance. As mentioned above the standard deviations of the time series are given in the 3rd column of the input series.

- **Spectrum characteristics**

The user can define the output spectrum, as well as the number of spectral values in each band. The user should keep in mind that the spectrum is band limited between the fundamental frequency and the Nyquist frequency.

- **Statistics**

The user can define the critical level for statistical testing. Usually values α =0.05□ or α=0.01 are used for critical levels to detect significant peaks in spectrum.

After executing LSSA using the input file *lssa.in* with properly defined parameters, several output files are created. These are: *lssa.out, residual.dat, spectrum.dat* and *hist.dat*.

Lssa.out contains the summary of results together with the statistical tests. The spectrum is described in three different forms: percentage variance (least-squares spectrum), power spectral density in dB (what is identical with the Fourier transform when the series is equally spaced), and power spectral density in unit^2/frequency (where unit is the unit of the time series e.g. cycles/day etc.).

The resulting spectrum is given in six columns which list the period of spectrum in units of the time series, frequency in cycles/unit time, fidelity (in units of time). If a significant peak is present in the spectrum it shows up in this column. The 4th column stands for the least-squares spectrum in percentage variance, next follows with the power spectral density in dB and last, the power spectral density in unit^2/frequency is given. The spectrum is also printed with asterisks for quick identification of peaks

Residual.dat lists the input and residual series. There are four columns which contain the time, the input time series, normalized residual series (after removing trends, periods, processes, etc.) and the standard deviation of the residuals.

Spectrum.dat contains the aforementioned 6 columns of the output spectrum as described in *lssa.out*. This file can be easily exported into Matlab or other mathematical package used for plotting or any other further analysis.

Hist.dat contains the histogram of the normalized residual series.