

Abstract: *Dario Villani, Raffaele Ghigliazza and René Carmona* extract the frequency content of a noisy signal using discrete Fourier transform. After calculating the deterministic component, we show the relevance of the method in removing spurious autocorrelations from the signal residuals. We present results for a temperature time series.

A discrete affair

★ The ability to calculate correlations for different assets or for the same asset over time is crucial for devising trading strategies or risk management decisions. The questions of how natural gas storage is correlated to temperature in a certain location or what is the persistence of gas prices over time are both extremely important for financial institutions. However, the calculation of correlations is plagued by the presence of deterministic periodic components, often called the trend or seasonal component. For example, natural gas storage and temperature decrease in winter and increase in summer. One should carefully remove this deterministic component before calculating the correlation. The estimate is ruined when part of the deterministic component is left in the noise or not properly removed.

In this article, we provide a solution for more precisely removing these effects. The calculation of the autocorrelation function of a noisy signal usually requires the removal of a deterministic component. We concentrate on the case of noisy periodic signals in order to tackle the important problem of the statistical analysis of temperature data. For this particular application, we need to carry out the following steps. First, we evaluate the frequencies of the embedded periodic components. Then we fully identify the deterministic component by a variational principle restricted to the class of functions consistent with the results of the first step. Finally, calculating the autocorrelation

function of the residuals completes the analysis of the signal.

We discuss the theoretical underpinnings of such a method in the special case of a signal with a single periodic component. We show that the three-step programme is often spoiled by the subtle consequences of the possible incommensurability of the sampling frequency and the intrinsic frequency of the signal in question. See, for example, Carmona (1998) for a discussion of the sampling theory of continuous signals. This paper quantifies one form of this undesirable effect and proposes a remedy for the ambiguity in the determination of the intrinsic frequency of a noisy periodic signal.

Fourier spectrum

To establish notation, we start by giving the explicit formula for the discrete Fourier transform (DFT) X_k of a given vector x_j of length N (see Cizek (1986)).

$$X_k = \frac{1}{\sqrt{N}} \sum_{j=1}^N x_j e^{2\pi i \frac{j-1}{N}(k-1)}, \quad k = 1, \dots, N. \quad (1)$$

With this definition of the Fourier transform, it follows that:

$$x_j = \frac{1}{\sqrt{N}} \sum_{k=1}^N X_k e^{-2\pi i \frac{k-1}{N}(j-1)}, \quad (2)$$

because of the orthogonality property

$$\frac{1}{N} \sum_{j=1}^N e^{-2\pi i \frac{n-k}{N} j} = \delta_{nk}. \quad (3)$$

Here and in the following sections, δ_{nk} is the Kronecker delta ($\delta_{nk}=1$ for $n=k$ and vanishes otherwise). Let us consider a monochromatic signal:

$$x_j = A e^{-2\pi i \frac{m-1}{N} j}, \quad j = 1, \dots, N$$

where m is a positive integer between 1 and N . For each integer $1 \leq k \leq N$, we have

$$X_k = A \sqrt{N} e^{-2\pi i \frac{k-1}{N} \delta_{km}}, \quad (5)$$

and hence:

$$|X_k| = |A| \sqrt{N} \delta_{km}. \quad (6)$$

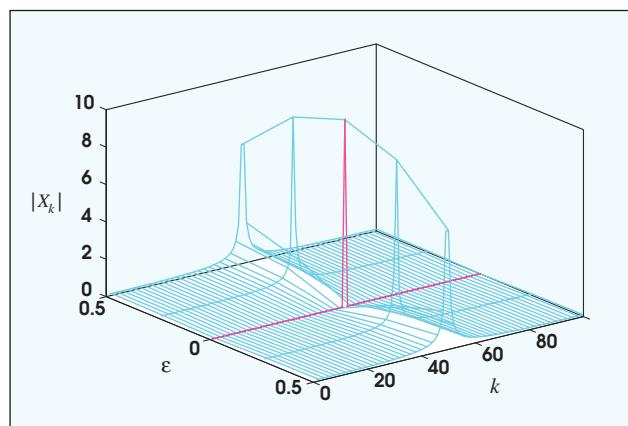


Fig 1. The spectrum $|X_k|$ as a function of k and ϵ . $|A|=1$, $N=100$ and $m=50$

In other words, the spectrum results in all coefficients being zero apart from a peak for the value of k equal to m . From the spectrum, we can calculate the frequency of the periodic signal given the length N (that is, $\omega = (m-1)/N$). A difficulty arises when the monochromatic signal is of the form

$$x_j = Ae^{-2\pi \frac{m-1+\epsilon}{N} j}, \quad j = 1, \dots, N, \quad -\frac{1}{2} \leq \epsilon < \frac{1}{2}. \quad (7)$$

In this case, the frequency cannot be expressed as a ratio of the form $(m-1)/N$, and we say we are facing an *incommensurate* lattice problem. Calculating the spectrum, we get

$$|X_k| = |A| \frac{1}{\sqrt{N}} \sqrt{\frac{1 - \cos 2\pi(m-k+\epsilon)}{1 - \cos 2\pi \frac{m-k+\epsilon}{N}}}. \quad (8)$$

It is now evident that an incommensurate frequency produces a spread of the peak – that is, $X_k \neq 0$ for $k \neq m$. Figure 1 shows this. It is worth noting that for $\epsilon \rightarrow 0$ and $N \rightarrow \infty$, we obtain

$$|X_k| \approx |A| \sqrt{N} \left| \frac{\sin \pi(m-k)}{\pi(m-k)} \right| \left\{ 1 + \left[\pi \cot \pi(m-k) - \frac{1}{m-k} \right] \epsilon \right\} + O(N^{-3/2}) \quad (9)$$

We recognise the zero-order term as the Fraunhofer diffraction pattern for the single slit (Born and Wolf (1965)). This should not be surprising, as the spread of the Fourier spectrum, X_k , is due to the interaction of two wavelengths – one for the lattice and the other for the signal.

The analysis outlined in equations 4 to 9 shows that the Fourier spectrum is widened even in the case of a noise-free signal. In the case of a noisy signal, the spectrum asymmetries introduced by the noise complicate the detection of the frequency.

The case of temperature data

We now analyse the average temperature data for the case of Seattle-Tacoma airport from January 1, 1960 to December 31, 1999 (14,610 entries). First, we remove the mean value of 52.20° Fahrenheit. Next, we determine the frequency of the embedded periodic signal by use of the Fourier paradigm. For any subset of the data, the Fourier spectrum gives a peak over a noisy background. However, each peak gives different estimates of the period, all inconsistent with the idea that the more points there are, the better is the estimate of the period. Figure 2 shows the value of the estimated period Λ as a function of N . We calculated the Λ values according to the following prescriptions. For each $N \leq 14,610$, we selected the first N points of the data set and calculated the period as

$$\Lambda = \frac{N}{k_{\max} - 1}, \quad (10)$$

where k_{\max} is the index for which the absolute value of the spectrum is maximum. We always chose the integer k_{\max} to be in the range below the Nyquist frequency (see Priestley (1981)). The uncertainty is substantial. Even with as many as 7,920 to 8,140 points (more than 20 times the number there would be in the ‘true’ period corresponding to the tropical year), the period takes values in the range 360–370

days. In experimental cases, where the value is not foreseeable from the very start, such uncertainty would be disastrous for estimating statistical properties such as variance and autocorrelation function.

Let us suppose we are given only the first $N_1 = 7,920$ data points. We would find a peak at $k_{\max} = 23$, which corresponds to $\Lambda_1 = 360$. If we had the first $N_2 = 8,140$ data, we would find a period of $\Lambda_2 = 370$, even with the same $k_{\max} = 23$. To fully identify the deterministic component d_j , we use the functional form:

$$d_j = \sum_{q=1}^Q \left[a_q \cos \frac{2\pi q}{\Lambda} j + b_q \sin \frac{2\pi q}{\Lambda} j \right], \quad (11)$$

where Q is the number of harmonics we use in the estimation. In each case, we minimise the least-squares distance between the signal diminished by its mean and the expression in equation 11. For $Q=3$, we obtain the following:

	a_1	b_1	a_2	b_2	a_3	b_3
$\Lambda_1 = 360$	-2.02	-10.58	-0.69	-0.51	-0.02	-0.27
$\Lambda_2 = 370$	-9.54	5.22	1.18	-0.18	0.07	0.17

The variances of the residuals are $\sigma_1^2 = 49.17$ and $\sigma_2^2 = 47.33$ for Λ_1 and Λ_2 . By using the tropical year estimate ($\Lambda_{tr} = 365.2422$), we get $\sigma_{tr}^2 = 26.65$ for $N = 7,920$, and $\sigma_{tr}^2 = 26.51$ for $N = 8,140$. It is evident that using a wrong value for the period leads us to overestimate the variance of the residuals. Figure 3 shows how strong the effect on the autocorrelation function is when a wrong period value is used. Not only does the variance of the residuals increase, but the residuals also show a false persistence. In the case of temperature time series, a strong persistence would imply the possibility of forecasting the weather beyond any reasonable range.

As already discussed, we must determine the ‘true’ frequency to make an effective analysis of the statistical properties of a signal. Here we provide a solution to this problem by analysing the functional dependence of the peaks appearing in figure 2. As well as providing a solution, we want to stress that any time the DFT is performed on a finite sample data set, a figure like figure 2 should be generated. One should not seek the largest possible data set, but rather study

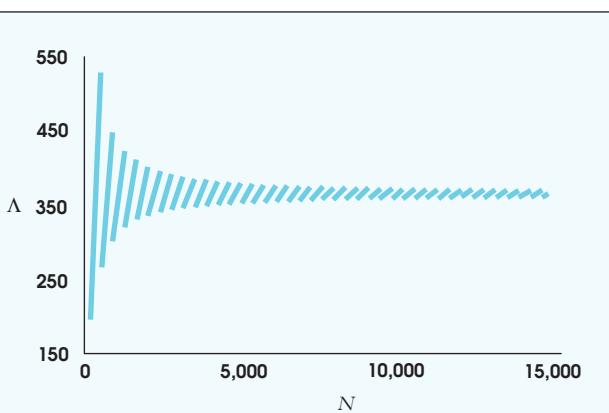
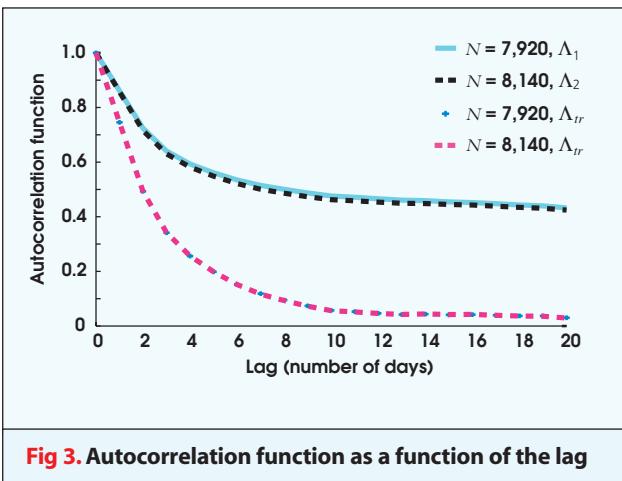


Fig 2. The period Λ as a function of the number of points, N

**Fig 3. Autocorrelation function as a function of the lag**

the estimated period as a function of the number of points. Our ansatz is that both the maximum and the minimum ‘peaks’ of each segment in figure 2 lie on a curve of the type:

$$\Lambda_{\min, \max}(x) = \frac{Ax}{B+x} \quad (12)$$

By using least-squares minimisation, we find:

	<i>A</i>	<i>B</i>
Λ_{\min}	365.70	194.26
Λ_{\max}	366.19	-165.67

In both cases, the R^2 is 1 apart from round-off errors. This means the function in equation 12 is not just a good guess, but in fact the ‘true’ decay of the estimated period to its large N limit – that is, A . The results above for Λ_{\min} and Λ_{\max} show that the estimates are fairly close to the ‘true’ period. Furthermore, the variance and the autocorrelation function are very close to those calculated for the tropical year value, 365.2422. More specifically, the variance for $N=8,140$ is 26.79 and 27.54 for Λ_{\min} and Λ_{\max} , respectively. In both

cases, we cannot distinguish the autocorrelation function from the one obtained after using Λ_{tr} .

Conclusion

We have shown how to calculate the autocorrelation function of noisy periodic signals in the case of a single-frequency mode. Our scheme aims to improve on a naive application of the DFT. The main point is that more is not necessarily better when it comes to the DFT. The dependence on the number of points for the estimated frequency is more important than the position of the peak for a specific N . Our solution stems from building envelopes of minima and maxima of piecewise linear functions. Other researchers could certainly develop an improved version of this method. Yet there is no question that we must find another solution by looking at the results in figure 2. ER

Dario Villani is a commodity trader at Hess Energy Trading Company in New York.

email: dvillani@alumni.princeton.edu

Raffaele Ghiglazza is a PhD student in the department of mechanical and aerospace engineering at Princeton University.

René Carmona is the Paul M Wythes ’55 professor of engineering and finance at Princeton University.

References

- Carmona, R, Hwang, WL and Torrésani, B,** *Practical Time-Frequency Analysis: Gabor and Wavelet Transforms with an Implementation in S*, Academic Press, New York, 1998.
- Cizek, V,** *Discrete Fourier Transforms and Their Applications*, Adam Hilger, Boston, 1986.
- Born, M and Wolf, E,** *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, Pergamon Press, Oxford, 1965.
- Priestley, MB,** *Spectral Analysis and Time Series*, Academic Press, New York, 1981.

Call for papers

Energy Risk welcomes technical article submissions on topics relevant to our readership. Core areas include market and credit risk measurement & management, the pricing and hedging of derivatives and/or structured securities, and the theoretical modelling and empirical observation of markets and portfolios with particular emphasis on the energy industry. This list is not exhaustive.

The most important publication criteria are originality, exclusivity and relevance – we try to strike a balance between them. Given that *Energy Risk* technical articles are shorter than those in dedicated academic journals, clarity of exposition is another yardstick for publication.

Once received by the editor, we log and check submissions against the criteria above. We reject articles that obviously fail to meet one or more criteria at this stage.

We then send articles zto one or more anonymous referees for peer review. Our referees are drawn from the research groups, risk

management departments and trading desks of major financial and energy institutions, as well as from academia.

Depending on the feedback from referees, the editor makes a decision to reject or accept the submitted article. His decision is final. Submissions should be sent, preferably by email, to the editor, James Ockenden (jockenden@riskwaters.com).

The preferred format is Microsoft Word, with equations in Mathstype format, although Adobe PDFs are acceptable. The maximum recommended length for articles is 3,500 words, with some allowance for charts and/or formulas – that is, this wordcount should be reduced proportionately, depending on the number of charts/tables/formulas included.

We expect all articles to contain references to previous literature. We reserve the right to cut articles to satisfy production considerations. Authors should allow four to eight weeks for the refereeing process.