Application of Signal Processing to the Analysis of Financial Data

As the current recession grew to the point of being considered the worst economic downturn since the Great Depression of 1929, the news media has not only covered daily developments, but it has also discussed its causes and the lessons that can be learned. A few of the published articles, such as [1] and [2], shed some light on the fact that signal processing techniques play an important role in today’s finances. Indeed, today’s financial analysis and risk managers depend on mathematical tools that, at their core, are based on signal processing techniques. In this article, we highlight some of these techniques used to represent and predict the main features of price evolution and to classify stock so as to design diversified investment portfolios.

IS SIGNAL PROCESSING USEFUL IN ANALYZING FINANCIAL DATA?
Assume a manufacturer is asked by a client to sell some stock products to him at any moment in time that the client chooses within the next six months, but at today’s prices rather than that moment’s prices. Obviously, the client is prepared to pay today a premium to the manufacturer for the privilege of exercising this option. How much should the manufacturer ask so that the deal proves profitable for him? And how will the client choose the optimal point in time to exercise the option he paid for? Both the client and the manufacturer will have to monitor the price of the product and extrapolate its future evolution according to past available information. Viewing the price as a signal, this is a classical problem in signal processing.

Assume now an investor wishes to build a portfolio out of several available stocks. The principle of portfolio diversification stipulates that the stocks selected to be included in the portfolio must be as uncorrelated as possible to reduce the risk due to price fluctuations (this is an application of the law of large numbers). In effect, then, each stock is viewed as a mixture of “hidden market trends,” and the investor’s goal is not only to determine these trends but also which stocks depend on which trends to estimate their correlations. This is a direct analogue of the canonical example in sound source separation, known as the “cocktail party problem”: in a cocktail party, the signals reaching our ears are brouhahas of dozens of voices (corresponding to stock prices), and yet somehow we are able to isolate and focus on a specific speaker (corresponding to the hidden market trend).

TOOLS AND MODELS TO ANALYZE FINANCIAL DATA
A discussion of signal processing techniques in finance must necessarily begin with the Black-Scholes [7] model for price evolution. This is an excellent example of mathematically rigorous signal modeling that opened the door for the introduction of advanced quantitative techniques in finance and earned the Noble prize of economics to its inventors.

THE BLACK-SCHOLES MODEL
The most common signals in finance are prices (of goods, options, etc.), and a simple description of such a signal can be extracted from first principles and confirmed by fitting on financial data. Assume we know the price \( S(0) \) at \( t = 0 \); how does the price \( S(t) \) evolve within a short time interval thereafter? The relative price increase (or decrease) between times \( t \) and \( t + dt \), when \( dt \) is small, can be considered to consist of a linear term of slope \( r \) plus a random normal fluctuation of zero mean and variance proportional to \( dt \) (namely \( \sigma^2 dt \), where \( \sigma \) is known as the volatility)

\[
\frac{S(t + dt) - S(t)}{S(t)} = r dt + \sigma N(t) \sqrt{dt}.
\]

We assume here that any two normal random variables \( N(t) \) and \( N(t') \) corresponding to nonoverlapping intervals \([t, t + dt]\) and \([t', t' + dt]\) are independent. At the limit when \( dt \rightarrow 0 \), this can be written as

\[
\frac{dS(t)}{S(t)} = r dt + \sigma dW(t),
\]

where \( W \) denotes a random walk, namely a continuous function whose increments \( W(t + s) - W(t) \) are normal random variables of zero mean and variance equal to \( s \), and such that increments over nonoverlapping intervals are independent. The framework within which such stochastic differential equations are studied is known as Itô’s calculus [2], where this equation can be shown to admit the following closed form solution:

\[
S(t) = S(0)e^{r(t-\frac{\sigma^2}{2})+\sigma W(t)}.
\]

(continued on page 156)
This is the basic form of the Black-Scholes model. Comparison to real data reveals a certain degree of oversimplification; it is not the final answer, but it certainly is a very successful first step, which admits further extensions.

For example, to model price evolution over longer time intervals, the obvious solution is to extend the model to include time-varying (or even stochastic, as a further step) \( r \) and \( \sigma \), so that

\[
\frac{dS(t)}{S(t)} = r(t)dt + \sigma(t)dW(t).
\]

Even this extension, however, cannot account for the large price jumps occasionally observed over very small time intervals and that are not captured by normal fluctuations. We can include such jumps in the model by adding a jump process \( J(t) \), namely a piecewise constant function whose interjump times are independent and whose jumps are typically given by a distribution that lacks a variance and often even a mean (e.g., a Cauchy distribution)

\[
\frac{dS(t)}{S(t)} = r(t)dt + \sigma(t)dW(t) + dJ(t).
\]

Establishing a rigorous model for price evolution is important from a theoretical point of view, but, in practice, we seek signal processing techniques to analyze data sets (e.g., of actual price evolution). For example, a technique found to perform very well in the cocktail party problem/portfolio creation mentioned in the introduction is nonnegative matrix factorization (NMF).

### NONNEGATIVE MATRIX FACTORIZATION

NMF is very simple: we collect the \( m \) observed signals of length \( n \) into an \( m \times n \) array \( Y \) (adding perhaps a constant to each signal to form nonnegative signals), and we determine an \( m \times k \) array \( A \) (the mixing array) and a \( k \times n \) array \( X \) (the sources array), such that the factorization error \( e = Y - AX \) is minimized. The number of sources \( k \) is usually determined through a “threshold effect;” increasing \( k \) beyond a certain value does not reduce the error significantly. The arrays \( A \) and \( X \) can be determined through an iterative process.

Based on \( A \), a clustering technique (e.g., \( k \)-means clustering) can be used to group the observed signals according to the sources they depend on. In practical terms, this means that we can now choose one stock out of each group and build our portfolio.

In the traditional approach to investing, stock groups were practically a nonissue, because they were taken (by mere convention, apparently) to reflect different sectors: in other words, IT stocks were grouped together in a group, pharmaceuticals stocks were grouped together in a separate group, etc. Applications of NMF on actual stock data (e.g., the Dow-Jones index [4]), however, led to the perhaps shocking result that, contrary to common belief, grouping does not reflect sector partition but rather spans across sectors. This result stresses the importance of analyzing and monitoring the market instead of relying on intuition.

### EMPIRICAL MODE DECOMPOSITION

The Black-Scholes model expresses the price variation of a stock in terms of a trend plus a fluctuation. As mentioned above, though, this is a theoretical model that oversimplifies reality. In actual data sets, fluctuation components over different time scales are observed: coarse time scales reflect periodicities meaningful to analysis and prediction, while fine time scales reflect noise. Both the trend and the fluctuations are important for the prediction of future values of the stock. How can all these components (trend and fluctuations) be meaningfully and reliably determined?

Periodicities are usually detected through application of the Fourier transform, which, in our case, however, suffers from two obvious drawbacks: a) it cannot cope with the nonstationarity introduced by the presence of the trend, and b) it chooses its expansion functions (sines and cosines) in advance, hence it tends to over-decompose: for example, a periodic function that is close to (but not exactly) a sine will be expanded into an infinite series. Localized expansions, such as the windowed Fourier transform or the wavelet transform, address the former issue but still suffer from the latter.

The empirical mode decomposition (EMD) [5] is specifically designed to expand a function into a trend plus a number of intrinsic mode functions (IMFs), whose defining features are a) that all local minima and maxima are negative and positive, respectively, and b) that their local mean is zero. Clearly, sines and cosines fall in this class, so IMFs can be viewed as “sines with slowly varying amplitude and frequency.”

To apply EMD, fit a (cubic) spline interpolant through all local maxima of the function (upper envelope), and do the same for local minima (lower envelope); then, compute the pointwise average of the two interpolants (the local mean) and subtract it from the function. Repeat this process until the local mean is zero everywhere (within some preset tolerance). The resulting function is the first IMF; this process is depicted in Figure 1 for an example signal. Subtract this IMF from the original function and repeat on the remainder function to obtain the second IMF. Repeat until all IMFs have been recovered and the remainder does not have enough local extrema to continue: this is the trend.

### PHASE-RECTIFYING SIGNAL AVERAGING

An alternative approach to determining hidden periodicities in price variations, which can also be applied in conjunction with EMD, is phase-rectifying signal averaging (PRSA) [6]. Here, we adopt the point of view that the signal at hand is essentially the sum of periodic signals plus some impurities (such as white noise and spikes). PRSA constructs a new signal, where periodicities appear stronger and impurities weaker, so that
Periodicities can be reliably detected by applying ordinary Fourier transform. For this, we assume the signal is stationary and, in particular, detrended: trend is removed either by linear (or, in general, polynomial) fitting, or through EMD.

To apply PRSA, we first determine the $m$ anchor points of the signal $S$, namely the points of increase $i$ where $S(i) > S(i-1)$, labeled as $i_1, \ldots, i_m$, and place a (rectangular) window of length $2L$ around each with the anchor point in the middle. We then “fold” the signal, averaging over all windows. More explicitly, we construct the auxiliary signal

$$\tilde{S}(n) = \frac{1}{m} \sum_{j=1}^{m} S(n + i_j),$$

for $n = -L, \ldots, -1, 0, \ldots, L - 1$ (edge effects can be dealt with in several obvious ways). Finally, we take the Fourier transform of $\tilde{S}$. A windowed version of PRSA exists where PRSA is applied on sections of $S$, just like the windowed Fourier transform.

**THE FUTURE OF SIGNAL PROCESSING IN FINANCE**

Today’s recession is widely recognized to be a manifestation of a change in the rules of the financial game. Markets expand due to globalization; investors’ numbers increase; investors are offered an increasing number of increasingly complex alternatives to invest in; abundant computational power gives investors the potential to be (or the illusion of being) smarter. Indeed, today’s recession is partly (mostly or exclusively, according to some) attributed to mispricing “exotic” investment options, a result of glorifying brute force (i.e., computational power) at the expense of the models and techniques used.

Finance and economy developed for centuries without reference to advanced mathematics (and signal processing). Economists were trained by a combination of tradition and psychology, being taught for each situation what the traditionally appropriate action is, and above all to develop and trust their instinct. Indeed, most established economy and finance textbooks (such as [8]...
genes, clustering, and classification. Many existing computing technologies such as statistic inference, neural network, data mining, and pattern recognition work well on these tasks. Some of these tools are open source software so that you are free to copy and redistribute them. MeV is one of them and is a good starting point if you are going to see what you can do about gene expression analysis. It integrates several popular analysis tools [e.g., hierarchical clustering, K-means, significance analysis of microarrays (SAM), and clustering affinity search technique (CAST)] and supports various data formats. It can give you an idea of what you need for analysis of gene expression data.

**SYSTEMS BIOLOGY**

http://cytoscape.org/ (Cytoscape)
http://www.genome.jp/kegg/ (Kyoto Encyclopedia of Genes and Genomes, KEGG)
http://www.bioconductor.org/ (Bioconductor) [tutorials]

Instead of just analyzing individual components, systems biology, which focuses on several components simultaneously, is the systematic study of complex or integrated interactions of gene, proteins, and biological reactions. The goal of systems biology is to allow people to delve deeper into the foundation of biological activities and have a more general view on the function or mechanism of gene and protein. Because of the complex system, it needs interdisciplinary studies that involve statistics, computation, chemistry, and other fields to discover the biological silhouette. These systems biology works usually include a network of interaction or biological signal pathway. KEGG is one of the most famous pathway databases and, by some knowledge-based methods, it provides information of cell and organism behaviors at the genomic and molecular level. As for the network of interaction, Cytoscape gives the convenient visualization of molecular interaction networks and has many open source plug-in modules to help users integrate the network with the data of the gene profiles or analyze the properties of the network and subnetwork. Bioconductor is a powerful open source and open development software project for the analysis and comprehension of genomic data. Because it is based on the R project (http://www.r-project.org/), users have many resources that they can do for data analysis, such as gene expression analysis, sequence and genome analysis, structural analysis, statistical analysis, and computational modeling.

**AUTHOR**

Jung-Hsien Chiang (jchiang@mail.ncu.edu.tw) is a professor in the Department of Computer Science and Information Engineering at National Cheng Kung University, Taiwan. He is currently a visiting professor with the Institute for Systems Biology, Seattle, Washington.

**REFERENCES**


Careful quantitative risk analysis seems to be the antidote to today’s recession. The “mathematization” of the financial system places policy making in the hands of the experts and increases investors’ trust, as decisions are directly based on realistic, observable, and reproducible quantitative analysis instead of the “good hunch” of a “market-seasoned veteran.” Signal processing is the reliable way to our financial future.