Int. J. Production Economics ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Generalized optimal wavelet decomposing algorithm for big financial data

Edward W. Sun [a,*], Yi-Ting Chen [b], Min-Teh Yu [c,**]

[a] KEDGE Business School, 680 Cours de la Libèration, 33405 Talance Cedex, France
[b] School of Computer Science, National Chiao Tung University, Taiwan and CC-Tech LLC, Kaohsiung, Taiwan
[c] School of Management, National Chiao Tung University, Taiwan

ARTICLE INFO

Article history:
Received 31 December 2013
Accepted 27 December 2014

Keywords:
Big financial data
DWT
High-frequency data
MODWT
Wavelet

ABSTRACT

Using big financial data for the price dynamics of U.S. equities, we investigate the impact that market microstructure noise has on modeling volatility of the returns. Based on wavelet transforms (DWT and MODWT) for decomposing the systematic pattern and noise, we propose a new wavelet-based methodology (named GOWDA, i.e., the generalized optimal wavelet decomposition algorithm) that allows us to deconstruct price series into the true efficient price and microstructure noise, particularly for the noise that induces the phase transition behaviors. This approach optimally determines the wavelet function, level of decomposition, and threshold rule by using a multivariate score function that minimizes the overall approximation error in data reconstruction. The data decomposition method enables us to estimate and forecast the volatility in a more efficient way than the traditional methods proposed in the literature. Through the proposed method we illustrate our simulation and empirical results of improving the estimation and forecasting performance.

1. Introduction

Big financial data have become a torrent, flowing into every area of the financial service sector. Particularly when the business and regulation of high-frequency trading have increased tremendously, financial institutions are now churning out a burgeoning quantity of transactional data that capture jillion bytes of information about their sales, trades, and operations (see Sun et al., 2014, and references therein). Gazillion networks are being embedded to sense, create, and communicate data around the world with ultra low latency managed services delivering market data in nanosecond speeds.[1]

Big financial data refer to massive financial data (1) whose source is heterogeneous and autonomous,[2] (2) whose dimension is diverse,[3] (3) whose size and/or format is beyond the capacity of conventional processes or tools to effectively and affordably capture, store, manage, analyze, and exploit; and (4) whose relationship is complex, dynamic, and evolving. Institutions are increasingly facing more and more big data challenges, and a wide variety of techniques has been developed and adapted to aggregate, manipulate, organize, analyze, and visualize them. The techniques currently applied on big financial data usually draw from several fields, including statistics, applied mathematics and computer science and institutions that intend to derive value from the data should adopt a flexible, reliable, and multidisciplinary approach. Utilizing big financial data and its analytics will improve the performance of financial institutions. Turner et al. (2012) find that 71% of financial institutions they surveyed[4] report that the use

* Corresponding author.
** Corresponding author.
E-mail addresses: edward.sun@kedgebs.com (E.W. Sun), mtyu@nctu.edu.tw (M.-T. Yu).

[1] Fixnetix developed a microchip that can execute trades in nanoseconds and announced its iX-eCute chip is able to process trades in 740 ns in June 2011. Burstream announced the NanoSpeed Market Data Mesh service for NASDAQ OMX proprietary traders on September 19, 2011.

[2] We say that an information system is heterogeneous if the software that creates and manipulates data is different at all sites and such data follow different structure and format that do not adhere to all sites. Autonomy refers to databases being under separate and independent control. Early discussion can be found in Sheth and Larson (1990).

[3] For the same observation, we might have different perspectives in the change of view angle. Diverse dimensionality refers to the existence of different representations based on the feature of perspective, and the features involved to represent each single perspective are varied.

[4] They survey 1144 business and IT professionals in 95 countries, including 124 respondents from the banking and financial market industries.

http://dx.doi.org/10.1016/j.ijpe.2014.12.033
0925-5273/

of big data is creating a competitive advantage compared with 63% of cross-industry respondents.

Financial institutions have access to big data, but they have fewer tools to get value out of them as the data are sitting in their most coarsest form or in a semi-structured or unstructured format. Turner et al. (2012) point out that financial institutions lag behind their cross-industry counterparts in core capabilities of data technology, and the need for more advanced data technology increases with the introduction of big data. One broad way of using big financial data to create value is to collect the data at a tick-by-tick level, as Manyika et al. (2011) mention that big data can unlock significant value by making information transparent and usable at much higher frequency. When considering financial data at higher frequency, usually such data illustrate the complex structure of irregularities and roughness (i.e., multifractal phenomena) due to huge amounts of microstructure noise. The heterogeneity characterized by multifractal phenomena is caused by a large number of instantaneous changes in the markets and trading noises (see Sun et al., 2007, and references therein). Therefore, mining big financial market data needs to intelligently extract information conveyed at different frequencies.

With the classic assumption of data mining, data are generated by certain unknown function representing signals plus random noise (see Au et al., 2010, and references therein). Decomposing big financial data is equivalent to extracting the systematic patterns (i.e., approximate the unknown function) conveyed in the data from noise. A similar treatment can be found in the classic signal processing theory (see Oppenheim and Schafer, 2010). As Sun and Meinl (2012) point out, a specific problem arises when the trend component exhibits occasional jumps that are in contrast to the slow evolving longterm trend. In financial markets jumps are often caused by some unexpected large transactions or extreme prices. Traditional linear denoising methods (e.g., moving average) usually fail to capture these jumps accurately as these linear methods tend to blur them. On the other hand, nonlinear filters are not appropriate to smooth out these jumps sufficiently, because the patterns extracted by nonlinear filters are not stationary to present longrun dynamic information (see Sun and Meinl, 2012, and references therein).

The wavelet method has been shown to be one of a number of multifractal spectrum computing methods and proven to be a reliable tool not only in signal processing (e.g., Mallat, 2009), but also econometric analysis (e.g., Fan and Gençay, 2010; Fan and Wang, 2007; Hong and Kao, 2004). Particularly, it is suitable for time series analysis – for example, smoothing, denoising, and jump detection (e.g., Gençay et al., 2010; Donoho and Johnstone, 1998; Meinl and Sun, 2012; Chen et al., 2014). The advantage of the wavelet method is that it performs a multiresolution analysis – that is, it allows us to analyze the data at different scales (each one associated with a particular frequency passband) at the same time. In this way, wavelets enable us to identify single events truncated in one frequency range as well as coherent structures across different scales. Several studies have applied wavelet methods in mining financial data. For example, Ramsey and Lampart (1998) and Kim and In (2008) apply wavelets to analyze relationships and dependencies among key macroeconomic and financial variables. Laukaitis (2008) investigates credit card intraday cash flow and intensity of transactions with wavelet transforms for high frequency data denoising. Sun et al. (2011) propose a wavelet method for analyzing the currency market with high-frequency data. Ferbar et al. (2009) apply the wavelet denoising method for forecasting demand in supply chain management. Sun and Meinl (2012) investigate the local linear scaling approximation (LLSA) in denoising high frequency data and show its robustness in empirical application under statistical goodness-of-fit tests. Chen et al. (2014) provide an integrated wavelet denoising method to improve the performance of classical econometric models.

When applying wavelet decomposition, both DWT and MODWT need to determine the wavelet function, level of decomposition, and threshold rule (see Chen et al., 2014). A common approach in choosing the wavelet function is to use the shortest wavelet filter that can provide reasonable results (Percival and Walden, 2006). The level of decomposition leads to the choice that considers “the higher the better” in general. The threshold rule, which is a function identifying the wavelet coefficients to be deleted, has also been investigated in academic research (e.g., Gençay et al., 2002). The remaining challenge is how to determine the combination of wavelet function, level of decomposition, and threshold rule to reach an optimal smoothness that generally improves the performance of classic models after denoising the data.

In this paper we propose an algorithm, named the generalized optimal wavelet decomposing algorithm (GOWDA), to optimally determine the wavelet function, level of decomposition, and threshold rule by using a linear score function that considers six performance measures. The goal of our method is to denoise the big data and obtain the trend that (1) contains as much information as possible, (2) exhibits a certain degree of concentration that can utilize the classic model, and (3) preserves as few artifacts (i.e., undesired structures, like oscillating nature, generated through the denoising process) as possible. In our method, those performance measures we consider shall try to maximize the information preserved in wavelet coefficients. Intuitively, these measures describe the characteristics of the denoised data with wavelet transform that optimally provide output for further analysis, such as forecasting with the classical stationary model. We show that the resulting difference sequence between the denoised data and the original signal must converge in probability at a predetermined confidence level. This requires (1) the structural change (e.g., jumps) of the denoised data and the original signal should be synchronous, (2) there are no outliers in the denoised data, and (3) the extremum in the denoised data should be bounded. In this paper we exhibit that GOWDA can result in an optimal choice with respect to these requirements.

We investigate the performance of GOWDA with numerical simulations that consider two typical patterns often observed in big financial data: excessive volatility and regime switching. With a comparison, GOWDA results in a better performance than the alternative methods and the numerical results coincide with the analytical properties we have shown. From the results in this paper, we see that GOWDA not only maintains the computational complexity of the original wavelet transform but also minimizes its approximation errors. In order to confirm the computational reliability and consistency that we have shown in the simulations, we further perform an empirical investigation by applying GOWDA with DJIA 30 stock data that have been aggregated from the tick-by-tick level. The results we obtained from such a large sample investigation coincide with the previous simulation results. When using the denoised data generated by GOWDA for forecasting, we find that the performances (i.e., accuracy of forecasting) of classic models e.g., AR, ARMA, and ARMA-GARCH, have been significantly improved, confirming the robustness of GOWDA.

We organized the paper as follows. Section 2 introduces some background information focusing on big financial data and analytical methods. We attempt to describe the prevailing paradigm such as multiscale invariance and compressed sensing for data analytics. Section 3 presents the methodology in detail, summarizing it with an algorithm chart. In Section 4, we investigate the performance of our method by conducting a simulation study. The simulation results confirm the superior performance of our method. In Section 5, we conduct an empirical study by applying our method to analyze the big financial data of DJIA 30 stocks in

Please cite this article as: Sun, E.W., et al., Generalized optimal wavelet decomposing algorithm for big financial data. International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2014.12.033