Multivariate EMD-based modeling and forecasting of crude oil price

He, Kaijian; Zha, Rui; Wu, Jun; Lai, Kin Keung

Published in:
Sustainability (Switzerland)

Published: 01/01/2016

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.3390/su8040387

Publication details:
He, K., Zha, R., Wu, J., & Lai, K. K. (2016). Multivariate EMD-based modeling and forecasting of crude oil price. *Sustainability (Switzerland), 8*(4), [387]. https://doi.org/10.3390/su8040387

sustainability

MDPI

Article

Multivariate EMD-Based Modeling and Forecasting of Crude Oil Price

Kaijian He [1,2,*], Rui Zha [1], Jun Wu [1] and Kin Keung Lai [3,4]

1 School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China; 2015200852@mail.buct.edu.cn (R.Z.); wujun@mail.buct.edu.cn (J.W.)
2 Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong
3 International Business School, Shaanxi Normal University, Xi'an 710119, China
4 Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong; mskklai@cityu.edu.hk
* Correspondence: kaijian.he@my.cityu.edu.hk; Tel./Fax: +86-10-6443-8793

Academic Editor: Bing Wang
Received: 23 February 2016; Accepted: 8 April 2016; Published: 21 April 2016

Abstract: Recent empirical studies reveal evidence of the co-existence of heterogeneous data characteristics distinguishable by time scale in the movement crude oil prices. In this paper we propose a new multivariate Empirical Mode Decomposition (EMD)-based model to take advantage of these heterogeneous characteristics of the price movement and model them in the crude oil markets. Empirical studies in benchmark crude oil markets confirm that more diverse heterogeneous data characteristics can be revealed and modeled in the projected time delayed domain. The proposed model demonstrates the superior performance compared to the benchmark models.

Keywords: empirical mode decomposition (EMD); multivariate EMD analysis; crude oil price forecasting; time delay embedding; multiscale analysis; ARMA model

1. Introduction

The behavior of the price of crude oil has significant impacts on different parts of the economy, affecting government, enterprise, and investors, *etc.* It has become more volatile and exhibits nonlinear characteristics, since major markets in different countries worldwide became more deregulated and integrated from the mid-1980s. Therefore, there are increasing demands for better characterization and prediction of crude oil price [1,2]. The evolution of prices of crude oil over the past century, as reflected by movements in benchmark indices, such as the West Texas Intermediate (WTI) and Brent crude oil prices, show that the intensity of fluctuations in crude oil markets have become more severe in recent years. For example, if e, if we examine the crude oil price movement over the past century, we can see that the underlying mechanism of the crude oil market has experienced significant changes in the new millennium. Until 2002, the market was relatively smooth and stationary with a moderate level of fluctuations. Starting from the new millennium, and especially after 2002, the equilibrium level of the market has increased sharply compared to the previous 20 years [3]. Now it operates at a higher equilibrium level, subject to a diverse range of influencing factors.

Forecasting in the crude oil market is one of the utmost important issues and has received significant attention from practitioners and researchers alike over the years. As the financial system is subject to more volatility, uncertainty is constantly shifting, and is less understood, influencing factors more than those in the physical field, it becomes more important to identify and model the main driving factors. There are mainly two approaches to forecasting in the literature, *i.e.*, fundamentalist and reduced-form. The fundamentalist approach attempts to model, explicitly

and quantitatively, the relationship between the price movement and fundamental factors such as economic variables, *etc.*, which involves econometric models with an exponentially-increasing level of computational complexity. For example, they offer satisfactory performance over the medium-to-long term time horizon. However, this approach faces problems when it shifts to the short term time horizon, the influencing factors and their correlations become exponentially more complex. It is difficult to identify them appropriately and model their interrelationships in the modeling process. There are unknown nonlinear features in the case of time series models, as well as nonlinear interrelations with other macroeconomic factors in the case of multivariate models. A reduced-form approach resorts to reduced-form models, such as time series models, *etc.*, to extract information from the past data directly, without reliance on exogenous variables. It is less costly and more suitable when specific information related to price movement and the underlying fundamental factors are not available or are too costly to obtain. Thus, it is much cheaper, more robust, and more direct than the fundamentalist approach. Both fundamentalist and reduced-form approaches can only offer an insufficient level of explanatory and forecasting power for the price movement. They are usually very costly and infeasible in practice, except for rare circumstances. Recently, computational approaches emerge to take a more data driven approach and offer an important alternative. For example, artificial intelligence and machine learning techniques take a purely data-adaptive and data-driven approach. They have achieved some positive results, which show that the crude oil market contains more complex Data Generating Processes (DGPs). However, there are debates on their robustness and generalizability in the literature. Artificial intelligence and machine learning techniques, such as neural networks and support vector regression, rely on the data mining exercises to extract nonlinear data patterns [4,5]. They have shown some promising performance improvements. However, the performance improvement is not consistent for all test cases [6]. Meanwhile, arguments often arise with respect to their results, as they risk overfitting the data. Results solely relying on these approaches, powerful as they may be suffer from their "black box" nature, as limited insights into the underlying influencing factors with economic rationale can be inferred [4,5,7]. Therefore, better understanding of the underlying DGP and accurate forecasting in the crude oil market remain the most difficult problems in the field [1,2].

Recent empirical research has increasingly revealed and acknowledged the significance of these heterogeneous data behaviors. Some typical examples in the finance literature include autocorrelations, volatility clustering, *etc.* A more recent addition would be the multi-scale data features. The most popular approaches would be the wavelet analysis and the Empirical Mode Decomposition (EMD) model. They have been increasingly recognized as effective analysis models in the economics and finance fields. Until now these approaches in the literature have achieved significant progress in revisiting many economic and financial modeling issues in the time-scale framework. However, the current modeling attempts are scattered and provide much indirect evidence on the effectiveness of the multiscale-based approach in modeling, more accurately, the data features, in terms of the improved out-of-sample forecasting accuracy. For example, Plakandaras *et al.* [8] show that the combination of Ensemble Empirical Mode Decomposition (EEMD) and artificial intelligence techniques would result in the improved exchange rate forecasting accuracy. Lin *et al.* [9] combine the EMD and Support Vector Regression (SVR) and find the exchange rate prediction accuracy improved. Jammazi and Aloui [10] find the inclusion of Haar à trous wavelet analysis in the multilayer propagation neural network model leads to the improved forecasting accuracy. Zhang *et al.* [11] combine the EEMD, the Least Square Support Vector Machine mixed optimized with the Particle Swarm Optimization (LSSVM-PSO) method and the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model to achieve the improved forecasting performance using the WTI data at a different frequency.

Direct evidence of the use of multiscale-based approaches involves the modeling of some particular data features such as different regimes, cycles, and correlations in the multiscale domain. For example, Jammazi and Aloui [12] combine the Haar à trous wavelet model and Markov switching vector autoregressive (MS-VAR) model to analyze three regimes for the influence of the crude oil shocks on the stock market using data from the UK, France, and Japan [12]. Alvarez-Ramirez *et al.* [13]

propose an entropy time-asymmetry approach to analyze the dynamics of the crude oil price and macroeconomic data in the USA. They identify the cyclical behavior with the period of 4.5 years [13]. Martina *et al.* [14] construct a multiscale entropy model to discuss the dynamics and structure of the crude oil prices and the connection between oil price and macroeconomic activity in different time scales. They argue that the multiscale model provides a feasible approach to analyze the cyclic dynamics of crude oil prices and prove the cyclic period is 4.3 years [14]. Zheng and Lan [15] use a multifractal de-trend fluctuation analysis to analyze the dynamic characteristic of five different tanker markets. By comparing three different aspects including non-periodic cycles, the Hurst exponents, and origins of multifractality with crude oil markets, they find that the tanker markets are more fractal [15]. Benhmad [16] analyze the cyclical co-movement between crude oil prices and US Gross Domestic Product (GDP) using the wavelet analysis and the Granger causality test and found the existence of a cyclical relationship in the multiscale domain [16]. By decomposing the variance iteratively in different scales like low-frequency and high-frequency, Gençay [17] proposes multiscale tests of serial correlation with the improved performance [17]. Fernandez-Macho [18] examines the wavelet correlation and cross-correlation between different financial variables in the Euro stock market by using multiscale models [18]. Bouoiyour and Selmi [19] identify the regime switching phenomenon in the relationship between the exchange rate and export, implied by GARCH coefficients in the wavelet-transformed domain. Kristoufek [20] finds that the global financial crisis is dominated by short-term investment behaviors using wavelet analysis and Fractal Market Hypotheses (FMH) [20]. Benhmad [16] investigates the cyclical co-movement between crude oil prices and US GDP. The wavelet analysis is used to make possible the analysis of the lead-lag relationship between different time periods in the wavelet-transformed time scales [16]. Naccache [21] found 20–40 years as the specific cycle for the oil price-Morgan Stanley Capital International (MSCI) index relationship using the wavelet analysis [21]. However, as far as the EMD model is concerned, not much research focusing on the data feature analysis and on the interpretation of the economic meaning can be identified.

Despite the proliferation of the diverse range of data features identified in the literature, the performance of current approaches is rather limited when these data features are used alone. In practice these data features are often mixed together and reflected jointly in the observed price data. The price data is not dominated with single data features, such as the regime and cycles. This represents an interesting and important research problem. More specifically, when the boundary between the data features is blurred and may not be recognizable in the original price data, current multi-scale models, such as the wavelet analysis and a univariate EMD, have difficulty in extracting constituent components from the noisy data. In this paper we propose a new multivariate EMD-based methodology for combining distant data features during the modeling and the forecasting of crude oil price. The results of this study explore and unveil the complex market structure consisting of data components of different data characteristics modeled using a multivariate EMD analysis. Empirical studies have been conducted in the markers WTI and Brent to investigate the performance improvements of the proposed model against traditional benchmark models.

This paper contributes to the relevant literature in two aspects. Firstly, we introduce the time delay embedding method to extend the analysis from a one dimensional domain to a higher dimensional domain. This makes it possible to analyze the distinct data features of different underlying DGPs not visible in the original time domain. Secondly, we introduce the multivariate EMD model to analyze and model the geometric multiscale data features in the higher dimensional domain. The data components are distinguished by their geometric differences in the time delay embedding domain. Thus, we can identify the DGPs of the most crucial factor.

The rest of the paper is organized as follows: In Section 2, we provide a brief account of the multivariate EMD theory. The multivariate EMD-based forecasting algorithm is proposed in Section 3. Results from the empirical studies are reported and analyzed in Section 4. Section 5 concludes with the summarizing remarks.

2. Multivariate EMD Theory

Recently, EMD emerges as a new multiscale approach by relaxing the basic assumptions and taking an empirical, intuitive, direct, and self-adaptive data approach [4,22]. EMD has been widely applied in different physical disciplines, such as signal processing, structured health monitoring, and biomedical engineering, to name just a few [23]. In recent years we have also witnessed the increasing number of research utilizing EMD as an important and useful tool to analyze the multiscale data structure in the economics and finance fields. Compared with Fourier and wavelet analyses, EMD offers much better temporal and frequency resolutions when employed for data analysis [22]. EMD can adaptively decompose a time series into several independent Intrinsic Mode Function (IMF) components and one residual component. The band of fluctuations for different underlying components are automatically and adaptively selected from the time series. EMD is a data-driven method with very few assumptions; thus, it can be used for data series of nonlinear and nonstationary nature [24].

When applied to multivariate data, the univariate EMD is only useful when the multivariate data is loosely coupled with the least level of correlations and dependence among different channels. In practice, there are further complications that invalidate its applications to multivariate data. This includes non-uniformity, scale alignment, *etc.* To model the multivariate price data, as well as the correlations and dependence among different individual price data accurately, the univariate EMD is extended to the multivariate case with a significantly different underlying conceptualization. In the multivariate EMD, the rotation concept is proposed as a generalization of the original oscillation concept in the univariate EMD to accommodate the correlations and dependence among different individual channels [25,26]. The intrinsic mode is characterized and separated with the rotation concept adopted in the multivariate EMD. The multivariate data is viewed as fast rotations superimposed on a slow rotation, where intrinsic models are jointly determined and scale-aligned. Thus, the multivariate EMD would enjoy the advantage of synchronous and coherent treatment of multivariate data. It is capable of exploring the underlying causality.

To estimate the high dimensional mean, the original high dimensional data are projected into a lower dimension, by taking different directions. By interpolating the extrema of projections at different directions via cubic splines, multiple envelopes at different directions are obtained. Then they are averaged to obtain the mean. The computational complexity increases exponentially with the increasing number of directions used in the projections. Thus, the number of directions used are usually kept to the minimal level that would offer a sufficient level of accuracy. One popular approach is to use the uniform distribution for the direction vectors and assume it may follow the uniform quasi-Monte Carlo sampling series, such as the low-discrepancy Hammersley sequences [25].

The multivariate EMD algorithm is illustrated as follows [26]:

(1) Sample over an (n-1) dimensional sphere to obtain a point set.
(2) Along the direction vector $x^{\theta k}$, calculate the projection $p^{\theta k}(t)_{t-1}^{T}$ of the input signal $v(t)_{t=1}^{T}$.
(3) Find the time instants $t_j^{\theta k}$ corresponding to the maxima of the set of projected signals $p^{\theta k}(t)_{k=1}^{k}$ for the whole set of direction vector *k*.
(4) Calculate the multivariate envelope curve $e^{\theta k}(t)_{k=1}^{k}$ using the interpolation method over the interval $[t_j^{\theta k}, v(t_j^{\theta k})]$
(5) Calculate the mean *m(t)* of the envelope curves as in $m(t) = \frac{1}{K}\sum_{k=1}^{K} e^{\theta_k}(t)$.
(6) Calculate the $c_i(t) = v(t) - m(t)$ for i-th order of IMF. Evaluate the $c_i(t)$ using the stoppage criterion. If the stoppage criterion is satisfied, apply the above procedure to $v(t) - c_i(t)$, otherwise apply it to $c_i(t)$.

3. A Multivariate EMD Based Forecasting Model for Crude Oil Price Movement

In the empirical studies, a price formation process in the social system, such as the crude oil market as one typical financial system, is subject to more complicated factors than in the case of a

physical system in nature. The data movement is less stationary and is influenced by more complex driving DGPs, which are difficult to be identified in current approaches. In this paper, we make three assumptions:

(1) We assume that the data are of heterogeneous nature, *i.e.*, the underlying DGPs show distinct behavioral patterns. We follow the Heterogeneous Market Hypothesis (HMH)-based approach to analyze and model the heterogeneous characteristics of the market microstructure. In contrast to the traditional efficient market hypothesis-based approaches that assume that markets consist of homogeneous agents with rational expectations, in an over-simplified manner, ignoring the existence of heterogeneous data features, HMH proposes that the market consists of heterogeneous agents with heterogeneous investment strategies, and investment time horizons [27]. In HMH, we assume that crude oil price evolves in a complicated and dynamic manner, subject to the influences of different factors. The market investors or agents react to news shocks differently based on their own characteristics, resulting in DGPs of distinctively different characteristics [28,29]. Following HMH, we further assume that there are diverse data characteristics, which reflect investment time horizon, frequency, scale, and individual characteristics for investors.

(2) We assume that the system is governed by some subject of existing data generating processes, which exhibit the multi-scale data structure. We further assume that among different DGPs, one DGP serves as the main driving factor.

(3) We assume that the data are deterministically chaotic, so that the phase space can be reconstructed from the univariate crude oil price data using methods such as the time delay embedding method where both systems have a topologically conjugate dynamic [30]. The time delay reconstruction holds for finite-dimensional subsets of infinite-dimensional spaces, thereby generalizing previous results which were valid only for subsets of finite-dimensional spaces [31].

With these assumptions, we view the seemingly nonstationary nonlinear data as the result of mixture and joint influence of different stationary data of both linear and nonlinear characteristics over different investment time horizons. The boundary between data features of different interests and relevance can be set using some multiscale models, or any other models recognizing the distinctions between different data features. The governing DGP is more stationary and has the most significant impact on the joint price movement, compared to others.

The multivariate EMD-based forecasting model consists of several phases including feature transformation, feature extraction, individual forecasts, dimension reduction, and ensemble forecasts.

In the feature transformation phase, we project the original time series into the reconstructed phase space in the higher dimensions using the delayed time method [32]. For data assumed to be deterministically chaotic, the system dynamics in the reconstructed data series in the phase space is the same as those of the original data. For a crude oil price time series r_i, $i = (1, 2, \ldots, N)$, the m-dimensional phase space x is constructed as in Equation (1):

$$x = [r_i, r_{i+\tau}, \cdots, r_{i+(m-1)\tau}], i = 1, 2, \ldots, N_m \quad (1)$$

where N is the length of the time series. τ is the time lag. m is the embedding dimension of the phase space. $N_m = N - (m-1)\tau$ is the size of the vector point. x is the two-dimensional matrix that is constructed using the delayed embedding method.

In the feature extraction phase, the multivariate EMD algorithm is used to extract the distinct data components across different scales. The accuracy of the feature extraction is sensitive to a different parameter set, including different decomposition scale, direction, *etc.*

$$x = \sum_{j=1}^{J} c_j + \epsilon, c_j, \epsilon \in R^n \quad (2)$$

where c_j is the p-variate IMF aligned with scale at time t. ϵ is the residual.

In the individual forecasting phase, for each scale j we follow Equation (1) to transform the data component matrix c_j in the higher dimension back to the univariate data y_k, $k = (1, 2, \ldots, N_m)$ at a lower dimension. Different statistical tests are performed on the univariate data to help determine the appropriate model specification, *i.e.*, the model equations and the lag orders. For example, the lag orders for time series equations are determined following the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) minimization principles. Then, the optimal parameters are determined for the model specification chosen, using appropriate econometric, machine learning, and optimization techniques. The parameters of Autoregressive Moving Average (ARMA) model are determined using the maximum likelihood estimation (MLE) technique as in Equation (3):

$$\widehat{y_k} = \delta + \sum_{i=1}^{m} \phi y_{k-i} + \sum_{j=1}^{n} \theta \varepsilon_{k-j} + \epsilon \tag{3}$$

where y_k is the conditional mean of the data, y_{k-i} is the lag m returns with parameter ϕ, and ε_{k-j} is the lag n residuals in the previous period with parameter θ. δ is the constant coefficient. ϵ is the error term. This transforms the original feature extract matrix into the individual forecasting matrix.

In the model tuning phase, the model specification and parameters of the multiscale models used are determined using the model tuning data. Particular performance measures, such as Mean Square Errors (MSE) and statistical test results, are used to guide the parameters searching process. The minimization of MSE corresponds to the minimization of error variance. Then we use the robust regression model to determine the adjustment ratio α_1. The model parameters, such as the weighting functions, are determined using the model tuning dataset.

$$r_i = \alpha_0 + \alpha_1 \widehat{y_i} + \epsilon_i, i = 1, \ldots, k \tag{4}$$

where r_i is the return observation, $\widehat{y_i}$ is the forecasted conditional mean, ϵ_i is the error term, α_0 is the intercept coefficient, α_1 is the slope coefficient.

In the forecasting phase, we use the model specification determined in the model tuning phase to advance crude oil return forecasts r_{k+1} using the rolling window method:

$$\widehat{r}_{k+1} = \alpha_1 \widehat{y}_{k+1} \tag{5}$$

4. Empirical Results

The US WTI and the European Brent crude oil markets are the two largest and representative markets in the world and are considered as the benchmarks by most government and private players in the market, including the Energy Information Administration (EIA) in the US. These benchmark prices have, therefore, been used for empirical studies in this paper. Oil prices in these markets are recognized as indicative worldwide, prices in other markets are invariably affected by, and related to, these prices. The data source is the EIA of the Department of Energy (DoE) in the US. The dataset used in the experiment consists of daily observations, denominated in USD/barrel unit. We conducted the empirical evaluation of the proposed algorithm against the benchmark ones, using daily observations in these two markets. The time period covered for both markets extend from 2 January 2002 to 28 December 2015, when the latest events and data are covered. The exact number of observations is not equal due to different holiday breaks in two different countries. This includes 3520 daily observations for the WTI dataset and 3551 daily observations for the Brent dataset. Following the convention in the machine learning literature, we use 70% as the dividing criteria to ensure sufficient samples for statistical significance of results [33,34]. The first 49% of the dataset serves as the training set for estimating model specifications and parameters initially. The next 21% dataset is used to determine the optimal multivariate EMD model parameters. The remaining 30% of dataset is reserved as the test set for the performance evaluation of different models. The original data is log-differenced at the first order as $r_t = \ln \frac{p_t}{p_{t-1}}$. Then the dataset is constructed. The returns are transformed to be

scale-free, which corresponds to percentage changes in financial positions and have more attractive statistical properties, such as stationarity, *etc.* The holding period is assumed to be one day. For each experiment, a portfolio of one asset position worth 1 USD is assumed. The statistical predictive accuracy of different models is evaluated using the MSE and the Clark West statistical predictive accuracy test for nested models, since the models evaluated are nested with one another [35,36].

Popular tests for the nonlinearity effect in the data include the Brock-Dechert-Scheinkman (BDS) test [37,38], bispectrum test [39], and bicorrelation test [40], *etc.* Among them, the BDS test, since its introduction, has become the standard for testing independence or nonlinear dependencies in the data. The null hypothesis for the test is that elements of the time series are independently and identically distributed (IID).The null hypothesis is rejected when tested statistics exceed a given critical value (1.95 at 95% confidence level). The rejection of null hypothesis implies that there are some kind of nonlinear dependencies, or even a chaotic hidden structure in the data. The null hypothesis for tests of departures from normality is that the data distribution is symmetric and mesokurtic (of normality). By performing statistical hypothesis test, it can be found out whether the data distribution statistically conforms to normal assumptions. There is a variety of tests available, including the Jarque-Bera (JB) test for normality and Pearson chi-square tests, *etc.* [41]. Among them, the JB test for normality is the most commonly applied test in practice. The JB test for normality tests whether the third and fourth moments (*i.e.*, skewness and kurtosis) of data series are jointly zero. When the data size is not very large, rejection of null hypothesis implies that the sample distribution deviates from normal distribution significantly. The descriptive statistics of the crude oil daily prices in both WTI and Brent markets are illustrated in Table 1.

Table 1. Descriptive statistics and statistical tests.

Statistics	Mean	Standard Deviation	Skewness	Kurtosis	p_{JB}	p_{BDS}
r_{WTI}	0.0001	0.0229	−0.4367	6.4688	0.001	0.0013
r_{Brent}	0.0001	0.0169	0.0664	6.3167	0.001	0

Descriptive statistics in Table 1 show that the distribution of crude oil price in both markets deviates significantly from the standard normal distribution and may contain nonlinear data characteristics. The standard deviation is significant for both markets. As indicated by the skewness value, the average return in the WTI market is more negative-oriented, while it is more positive-oriented in the Brent market. The kurtosis appears to deviate from the normal level, which indicates that the market exhibits significant abnormal return change events. Additionally, since the null hypothesis of both the JB test of normality and BDS test of independence are rejected, this further indicates that the market return contains unknown nonlinear dynamics, not easily captured by traditional linear models [42,43].

Random Walk (RW) and ARMA models are two widely used models, which represents the most stable and robust models in the literature. They are used as the benchmark models in the model evaluation process [44]. The lag order for benchmark ARMA(*r*,*m*) during the forecasting process is set to ARMA(1,1). The lag order for ARMA(*r*,*m*) in the forecasting process is determined based on the AIC and BIC minimization principles. The embedding dimension is set to 16. The direction is set to 32. We also tried different weight functions using the model tuning data. These weight functions include Andrews, Bisquare, Cauchy, fair, Huber, logistic, Talwar, and Welsch.

Assuming a different scale *i* as the main underlying driving factor, predictive accuracy of the proposed multivariate EMD forecast (MEMDF) model against alternative benchmark models is listed in Table 2.

Table 2. In-sample performance comparison of different models.

Models	RW	ARMA	MEMDF							
			1	2	3	4	5	6	7	8
$\mathbf{WTI}_{MSE,\times 10^{-4}}$	6.7194	6.7133	9.9698	7.6461	8.2590	7.5943	7.0964	7.0676	6.8689	6.9802
$\mathbf{Brent}_{MSE,\times 10^{-4}}$	5.1652	5.1837	7.6798	6.2551	6.3701	5.9294	5.5196	5.3251	5.2520	5.6616

Where MEMDF$_i$, $i = 1, 2, \ldots, 8$ refer to the proposed MEMDF with scale i assumed as the main underlying factor for forecasting. Experiment results in Table 2 show that as we assume different scales as the main underlying driving factor for the forecasting model, different performance of the proposed model would result. The performance variation suggests that the multivariate EMD-based forecasting model provides different assumptions on the market microstructure; some may be twisted and biased. One common criterion to choose the optimal model specification and parameters is to use the in-sample MSE as the optimization objective. This implies that we assume that the main driving at the particular scale chosen with the optimal model specification is the true one that would remain in the future out-of-sample data. Using optimal in-sample MSE as the optimization criterion using the model tuning data, we choose scale 7 with a Huber weight function for the WTI market with an in-sample MSE 6.7061×10^{-4} and choose scale 7 with a Talwar weight function for the Brent market with an in-sample MSE 5.1590×10^{-4}.

With the chosen model specifications, we further evaluate the out-of-sample prediction performance of the proposed algorithm. The level predictive accuracy of the proposed algorithm against the benchmark models is listed in Table 3. We further evaluate the generalizability of the proposed algorithm, against the benchmark RW and ARMA models. Results are listed in Table 3.

Table 3. Out-of-sample performance comparison of different models.

Models	RW	ARMA	MEMDF
$\mathbf{MSE}_{WTI,10^{-4}}$	3.7891	3.8271	3.7856
$\mathbf{CW}_{WTI,ARMA}$	0.0061	N/A	0.0036
$\mathbf{CW}_{WTI,RW}$	N/A	0.6897	0.0828
$\mathbf{MSE}_{Brent,10^{-4}}$	4.7533	4.7565	4.7529
$\mathbf{CW}_{Brent,ARMA}$	0.1177	N/A	0.1748
$\mathbf{CW}_{Brent,RW}$	N/A	0.7460	0.3006

Where MSE$_i$ refers to the calculated MSE for different models in the market i, CW$_{i,j}$ refers to the calculated Clark West test of predictive accuracy for different models against the benchmark model j in the market i. Experiment results in Table 3 show that the performance of the proposed MEMD-based algorithm is significantly better than the benchmark RW and ARMA models for WTI, in terms of the level of predictive accuracy. Overall, the proposed model achieves lower MSE than the benchmark RW and ARMA models. The forecasting performance gap is statistically significant at 95% confidence level for the WTI market. For the Brent market, it is statistically significant at 80% against the ARMA model and 70% confidence level against the RW model. For both markets, the ARMA model is inferior to the RW model in terms of forecasting accuracy. This suggests that the market contains nonlinear dynamics instead of the linear features captured by simple linear model, such as the ARMA model. The slight inferior performance of the proposed model in the Brent market may indicate that the Brent market may be less efficient than the WTI market., the dominating factor in the Brent market, unlike the WTI market, is less linear, which results in a larger estimation bias for the ARMA model. More accurate nonlinear modeling of the underlying DGPs than the currently-employed ARMA model are needed. As our out-of-sample test results confirm that the model specification chosen in the model tuning phase leads to the improved forecasting results, scale 7 is supposed to be the optimal scale where the underlying data components are extracted. Since different weighting functions are chosen for the

robust regression model in different markets, this suggests that the market has different levels of noise, which would result in different levels of disruption on the main driving factors in different markets.

Experimental results in this paper have some important implications. They show that the data contain complicated data characteristics of a diverse nature. Only very few of the extracted underlying data components can be classified as the main driving factors. Most of other underlying factors are of no relevance to the main data trends, which can be classified as disruptive noise and contribute very little to the improvement of the forecasting accuracy. More importantly, this finding suggests that during the forecasting exercise it is of critical importance to analyze and model the diverse data characteristics in the data. This explains why current universal function approximation techniques, such as neural networks, can achieve superior performance compared to the benchmark models in some circumstances, but their performance is far from stable and inconsistent in different circumstances. Thus, the proposed model provides important insights into the market microstructure.

5. Conclusions

In this paper we propose a novel multivariate EMD crude oil price forecasting model to analyze and incorporate the multiscale geometric data characteristics in the crude oil price movement model. The proposed multivariate EMD model has demonstrated impressive capability to extract the distinct main driving factor from the transformed crude oil price in the higher domain. Empirical studies on two major benchmark crude oil markets of the world suggest its effectiveness in analyzing the heterogeneous market structure and demonstrate significant positive performance improvement as a result.

Compared to the previous approaches addressing the multiscale characteristics in the multiscale domain, work in this paper provides an alternative approach to model the market microstructure of the crude oil market and gain better insights into the underlying driving factors. Moreover, we also show that multiscale-based modeling could provide a redundant and twisted view of the market microstructure. Thus the model specification and parameters need to be carefully selected and statistically valid.

Acknowledgments: This work is supported the National Natural Science Foundation of China (NSFC nos. 71201054, 91224001,71433001), the Strategic Research Grant of City University of Hong Kong (No. 7004574), and the Fundamental Research Funds for the Central Universities in BUCT (buctrc201618).

Author Contributions: Kaijian He conceived, designed and performed the experiments. Kaijian He and Rui Zha analyzed the data. Jun Wu and Kin Keung Lai contributed reagents, materials and analysis tools. Kaijian He wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Yang, C.; Hwang, M.; Huang, B. An Analysis of Factors Affecting Price Volatility of the Us Oil Market. *Energy Econ.* **2002**, *24*, 107–119. [CrossRef]
2. Plourde, A.; Watkins, G. Crude Oil Prices between 1985 and 1994: How Volatile in Relation to Other Commodities? *Resour. Energy Econ.* **1998**, *20*, 245–262. [CrossRef]
3. Askari, H.; Krichene, N. Oil price dynamics (2002–2006). *Energy Econ.* **2008**, *30*, 2134–2153. [CrossRef]
4. Yu, L.; Wang, S.; Lai, K.K. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Econ.* **2008**, *30*, 2623–2635. [CrossRef]
5. Zhang, X.; Lai, K.; Wang, S.Y. A new approach for crude oil price analysis based on Empirical Mode Decomposition. *Energy Econ.* **2008**, *30*, 905–918. [CrossRef]
6. Clements, M.P.; Franses, P.H.; Swanson, N.R. Forecasting Economic and Financial Time-Series with Non-Linear Models. *Int. J. Forecast.* **2004**, *20*, 169–183. [CrossRef]

7. He, K.; Xie, C.; Chen, S.; Lai, K.K. Estimating VaR in crude oil market: A novel multi-scale non-linear ensemble approach incorporating wavelet analysis and neural network. *Neurocomputing* **2009**, *72*, 3428–3438. [CrossRef]
8. Plakandaras, V.; Gupta, R.; Gogas, P.; Papadimitriou, T. Forecasting the US real house price index. *Econ. Model.* **2015**, *45*, 259–267. [CrossRef]
9. Lin, C.S.; Chiu, S.H.; Lin, T.Y. Empirical mode decomposition-based least squares support vector regression for foreign exchange rate forecasting. *Econ. Model.* **2012**, *29*, 2583–2590. [CrossRef]
10. Jammazi, R.; Aloui, C. Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling. *Energy Econ.* **2012**, *34*, 828–841. [CrossRef]
11. Zhang, J.L.; Zhang, Y.J.; Zhang, L. A novel hybrid method for crude oil price forecasting. *Energy Econ.* **2015**, *49*, 649–659. [CrossRef]
12. Jammazi, R.; Aloui, C. Wavelet decomposition and regime shifts: Assessing the effects of crude oil shocks on stock market returns. *Energy Policy* **2010**, *38*, 1415–1435. [CrossRef]
13. Alvarez-Ramirez, J.; Rodriguez, E.; Martina, E.; Ibarra-Valdez, C. Cyclical behavior of crude oil markets and economic recessions in the period 1986–2010. *Technol. Forecast. Soc. Change* **2012**, *79*, 47–58. [CrossRef]
14. Martina, E.; Rodriguez, E.; Escarela-Perez, R.; Alvarez-Ramirez, J. Multiscale entropy analysis of crude oil price dynamics. *Energy Econ.* **2011**, *33*, 936–947. [CrossRef]
15. Zheng, S.Y.; Lan, X.G. Multifractal analysis of spot rates in tanker markets and their comparisons with crude oil markets. *Phys. A Stat. Mech. Its Appl.* **2016**, *444*, 547–559. [CrossRef]
16. Benhmad, F. Dynamic cyclical comovements between oil prices and US GDP: A wavelet perspective. *Energy Policy* **2013**, *57*, 141–151. [CrossRef]
17. Gençay, R. Multi-scale tests for serial correlation. *J. Econ.* **2015**, *184*, 62–80. [CrossRef]
18. Fernandez-Macho, J. Wavelet multiple correlation and cross-correlation: A multiscale analysis of euro zone stock markets. *Phys. A Stat. Mech. Its Appl.* **2012**, *391*, 1097–1104. [CrossRef]
19. Bouoiyour, J.; Selmi, R. Exchange volatility and export performance in Egypt: New insights from wavelet decomposition and optimal GARCH model. *J. Int. Trade Econ. Dev.* **2015**, *24*, 201–227. [CrossRef]
20. Kristoufek, L. Fractal Markets Hypothesis and the Global Financial Crisis: Wavelet Power Evidence. *Sci. Rep.* **2013**, *3*, 2857. [CrossRef] [PubMed]
21. Naccache, T. Oil price cycles and wavelets. *Energy Econ.* **2011**, *33*, 338–352. [CrossRef]
22. Huang, N.E.; Wu, M.L.; Qu, W.; Long, S.R.; Shen, S.S. Applications of Hilbert–Huang transform to non-stationary financial time series analysis. *Appl. Stoch. Models Bus. Ind.* **2003**, *19*, 245–268. [CrossRef]
23. Premanode, B.; Toumazou, C. Improving prediction of exchange rates using Differential EMD. *Expert Syst. Appl.* **2013**, *40*, 377–384. [CrossRef]
24. Yu, L.; Wang, S.; Lai, K.K. Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Syst. Appl.* **2008**, *34*, 1434–1444. [CrossRef]
25. Mandic, D.P.; Rehman, N.U.; Wu, Z.H.; Huang, N.E. Empirical Mode Decomposition-Based Time-Frequency Analysis of Multivariate Signals. *IEEE Signal Process. Mag.* **2013**, *30*, 74–86. [CrossRef]
26. Park, C.; Looney, D.; Rehman, N.U.; Ahrabian, A.; Mandic, D.P. Classification of Motor Imagery BCI Using Multivariate Empirical Mode Decomposition. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2013**, *21*, 10–22. [CrossRef] [PubMed]
27. Dacorogna, M.M.; Gençay, R.; Müller, U.A.; Olsen, R.B.; Pictet, O.V. *An Introduction to High-Frequency Finance*; Academic Press: San Diego, CA, USA, 2001.
28. Müller, U.A.; Dacorogna, M.M.; Davé, R.D.; Olsen, R.B.; Pictet, O.V.; von Weizsäcker, J.E. Volatilities of different time resolutions. Analyzing the dynamics of market components. *J. Empir. Finance* **1997**, *4*, 213–239. [CrossRef]
29. Lux, T.; Marchesi, M. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* **1999**, *397*, 498–500. [CrossRef]
30. Takens, F. Detecting Strange Attractors in Turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*; Volume 898 of the series Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1981; pp. 366–381.
31. Robinson, J.C. A topological delay embedding theorem for infinite-dimensional dynamical systems. *Nonlinearity* **2005**, *18*, 2135. [CrossRef]

32. Wang, T.; Liu, X.; Zhang, Z. Characterization of chaotic multiscale features on the time series of melt index in industrial propylene polymerization system. *J. Franklin Inst.* **2014**, *351*, 878–906. [CrossRef]
33. Walczak, S. An empirical analysis of data requirements for financial forecasting with neural networks. *J. Manag. Inf. Syst.* **2001**, *17*, 203–222.
34. Zou, H.; Xia, G.; Yang, F.; Wang, H. An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. *Neurocomputing* **2007**, *70*, 2913–2923. [CrossRef]
35. Clark, T.E.; West, K.D. Using Out-Of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis. *J. Econ.* **2006**, *135*, 155–186. [CrossRef]
36. Clark, T.E.; West, K.D. Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *J. Econ.* **2007**, *138*, 291–311. [CrossRef]
37. Brock, W.A.; Hsieh, D.A.; LeBaron, B.D. *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*; MIT Press: Cambridge, MA, USA, 1991.
38. Panagiotidis, T. Testing the Assumption of Linearity. *Econ. Bull.* **2002**, *3*, 1–9.
39. Hinich, M. Testing for Gaussianity and Linearity of a Stationary Time Series. *J. Time Ser. Anal.* **1982**, *3*, 169–176. [CrossRef]
40. Hsieh, D. Implications of Nonlinear Dynamics for Financial Risk Management. *J. Financ. Quant. Anal.* **1993**, *28*, 41–64. [CrossRef]
41. Brooks, C. *Introductory Econometrics for Finance*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2008.
42. Jarque, C.M.; Bera, A.K. A test for normality of observations and regression residuals. *Int. Stat. Rev.* **1987**, *55*, 163–172. [CrossRef]
43. Broock, W.; Scheinkman, J.A.; Dechert, W.D.; LeBaron, B. A test for independence based on the correlation dimension. *Econ. Rev.* **1996**, *15*, 197–235. [CrossRef]
44. Meese, R.A.; Rogoff, K. Empirical Exchange-Rate Models of the Seventies—Do They Fit out of Sample. *J. Int. Econ.* **1983**, *14*, 3–24. [CrossRef]