mathematics

MDPI

Article

A Proposal to Fix the Number of Factors on Modeling the Dynamics of Futures Contracts on Commodity Prices †

Andrés García-Mirantes [1], Beatriz Larraz [2,*] and Javier Población [3]

[1] IES Juan del Enzina, 24001 Leon, Spain; andres.web.publicar@gmail.com
[2] Statistics Department/Faculty of Law and Social Sciences, Universidad de Castilla-La Mancha, 45071 Toledo, Spain
[3] Banco de España, 28014 Madrid, Spain; Javier.poblacion@bde.es
* Correspondence: beatriz.larraz@uclm.es; Tel.: +34-925-26-88-00

† This paper should not be reported as representing the views of the Banco de España (BdE) or European Central Bank (ECB). The views expressed herein are those of the authors and should not be attributed to the BdE or ECB.

Received: 8 May 2020; Accepted: 11 June 2020; Published: 14 June 2020

Abstract: In the literature on modeling commodity futures prices, we find that the stochastic behavior of the spot price is a response to between one and four factors, including both short- and long-term components. The more factors considered in modeling a spot price process, the better the fit to observed futures prices—but the more complex the procedure can be. With a view to contributing to the knowledge of how many factors should be considered, this study presents a new way of computing the best number of factors to be accounted for when modeling risk-management of energy derivatives. The new method identifies the number of factors one should consider in the model and the type of stochastic process to be followed. This study aims to add value to previous studies which consider principal components by assuming that the spot price can be modeled as a sum of several factors. When applied to four different commodities (weekly observations corresponding to futures prices traded at the NYMEX for WTI light sweet crude oil, heating oil, unleaded gasoline and Henry Hub natural gas) we find that, while crude oil and heating oil are satisfactorily well-modeled with two factors, unleaded gasoline and natural gas need a third factor to capture seasonality.

Keywords: commodity prices; futures prices; number of factors; eigenvalues

1. Introduction

Forecasting is not a highly regarded activity for economists and financiers. For some, it evokes images of speculators, chart analysts and questionable investor newsletters. For others, there are memories of the grandiose econometric forecasting failures of the 1970's. Nevertheless, there is a need for forecasting in risk management. A prudent corporate treasurer or fund manager must have some way of measuring the risk of earnings, cash flows or returns. Any measure of risk must incorporate some estimate of the probability distribution of the futures asset prices on which financial performance depends. Consequently, forecasting is an indispensable element of prudent financial management.

When a company is planning to develop a crude oil or natural gas field, the investment is significant, and production usually lasts many years. However, there must be an initial investment for there to be any return (see, for example, [1,2], among others). Assuming that futures values are not known after a certain date because there is no trade, it makes it difficult to measure the risk of these projects. Since commodities (crude oil, gas, gasoline, etc.) are physical assets, their price dynamic is much more complex than financial assets because their prices are affected by storage and transportation

cost (cost of carry). Due to such complexity, in order to model this price dynamic we need factor models such as in [3–9]. In addition, in the transport sector [10] and [11] use different factor models for modeling bulk shipping prices and freight prices.

In order to measure exposure to price risk due to a single underlying asset, it is necessary to know the dynamics of the term structure of asset prices. Specifically, the value-at-risk (VaR, [12]) of the underlying asset price, the most widely known measure of market risk [13], is characterized by knowing the stochastic dynamic of the price, the volatility of the price and the correlation of different prices at different times. For these reasons, to date, the behavior of commodity prices has been modeled under the assumption that the spot price and/or the convenience yield of the commodity follow a stochastic process.

In the literature we find that the spot price is considered as the sum of both short-term and long-term components (see, for example, [14,15]). Short-term factors account for the mean reverting components in commodity prices, while long-term factors account for the long-term dynamics of commodity prices, assuming they follow a random walk. Sometimes a deterministic seasonal component needs to be added [16].

Following this approach, some multifactor models have been proposed in the literature. Focusing on the number of factors initially considered, [17] developed a two-factor model to value oil-linked assets. Later, [14] planned a one-factor model, two-factor model and a three-factor model, adding stochastic interest rates to the previous factors. This was superseded by a new formulation which appeared in [15], enhancing the latter article and developing a short-term/long-term model. [18] added the long-term spot price return as a third risk factor. Finally, [19] offered researchers a general N-factor model.

At this point, it should be stressed that the decision regarding the number of factors to be used in the model needs to be made a priori. According to the above literature consulted, the models are usually planned with two, three or four factors. However, in this study, the need to assume a fixed number of factors in the model is discounted. We propose a new method that identifies the number of factors one should consider in the model and the type of stochastic process to be followed. This method avoids the necessity of inaccurately suggesting a concrete number of factors in the model. This is very useful for researchers and practitioners because the optimal number of factors could change, depending on the accuracy needed in each problem. Clearly, if we do not use the optimal number of factors in modeling the commodity price dynamics, the results will not be optimal.

To the best of our knowledge, there are three previous studies applying principal component analysis [20] to the modeling of commodity futures price dynamics [21–23]. However, they only model the futures prices dynamic and ignore the dynamic followed by the spot price and, consequently carrying the risk of being incoherent, since futures price are the spot price expected value under the Q measure.

This study aims to add value to previous contributions by assuming that is the spot price can be modeled as a sum of several factors (long term and short term, seasonality, etc.). Therefore, since it is widely accepted (see, for example, [24]) that the futures price is the spot price expected value under the Q measure ($F_{t,T} = E^*[S_{t+T}|I_t]$), where S_{t+T} is the spot price at time $t+T$, I_t is the information available at time t and $E^*[\,]$ is the expected value under the Q measure.), from the variance–covariance matrix of the futures prices we can deduce the best structure for modelling the spot prices dynamic.

The remainder of this study is organized as follows. Section 2 presents a general theoretical model and explains the methodology proposed to set an optimal set of factors. In Section 3, we describe the datasets used to show the methodology and these results are described. Finally, Section 4 sets out the conclusions.

2. Theoretical Model

2.1. Theoretical Model

In the main literature to date (for example, [19]), it is assumed that the commodity log spot price is the sum of several stochastic factors: $S_t = \exp(\mathbf{C}\mathbf{X}_t)$, $t = 0, \ldots, n$ where the vector of state variables $\mathbf{X}_t = (x_{1t}, \ldots, x_{Nt})$ follows the process: $d\mathbf{X}_t = \mathbf{M}dt + \mathbf{A}\mathbf{X}_t dt + \mathbf{R}d\mathbf{W}_t$, being $\mathbf{C}, \mathbf{M}, \mathbf{A}$ and $\mathbf{R}$ vectors' and matrices' parameters.

It is widely accepted that, for the model to be identifiable, some restrictions must be imposed. This means that if we assume that $\mathbf{A}$ is diagonalizable and all its eigenvalues are real (a different formula is available if some are complex), we can take $\mathbf{C} = (1, \ldots, 1)$, $\mathbf{M}' = (\mu, 0, \ldots, 0)$ and $\mathbf{A} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & k_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & k_N \end{pmatrix}$, with $k_i, i = 1, \ldots, N$ the eigenvalues and $k_1 = 0$, by simply changing the state space basis. Therefore, we already have $\mathbf{M}$, $\mathbf{A}$ and $\mathbf{C}$.

It is also easy to prove that as $d\mathbf{W}_t$ is a $N \times 1$ vector of correlated Brownian motion increments, $\mathbf{R}$ can be assumed as $\mathbf{R} = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{pmatrix}$. Note that $\mathbf{R}$ is not important, but the product $\mathbf{R}\mathbf{R}'$ is what appears in all formulae. In fact, it can be proved that any factorization of $\mathbf{R}\mathbf{R}'$ corresponds to a different definition of the noise, so we can safely take $\mathbf{R}$ as any Choleski factorization of $(\mathbf{R}\mathbf{R}')$. In the Black–Scholes world (risk-neutral world), knowing the real dynamics, the risk neutral one is $d\mathbf{X}_t = \mathbf{M}^* dt + \mathbf{A}\mathbf{X}_t dt + \mathbf{R}d\mathbf{W}_t^*$ where $\mathbf{M}^* = \mathbf{M} - \boldsymbol{\lambda}$ being $\boldsymbol{\lambda}' = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ the vector formed from each state variable's risk premium).

Following [25], the futures price is given by $F_{t,T} = \exp\left(g(T) + \mathbf{C}e^{\mathbf{A}T}\mathbf{X}_t\right)$, where we know explicitly $g(T) = \mathbf{C}\int_0^T e^{\mathbf{A}(T-s)}\mathbf{M}^* ds + C\left(\int_0^T e^{\mathbf{A}(T-s)}\mathbf{R}\mathbf{R}'\left(e^{\mathbf{A}(T-s)}\right)' ds\right)\mathbf{C}'$ and where both $g(T)$ and $C(T) = e^{\mathbf{A}T}$ are known deterministic functions independent of t and $\mathbf{X}_t$ is a stochastic process with known dynamics.

Defining in a more compact form, we have:

$$\begin{cases} d\mathbf{X}_t = (\mathbf{M} + \mathbf{A}\mathbf{X}_t)dt + \mathbf{R}d\mathbf{W}_t \\ F_{t,T} = \exp[\delta(T) + \phi(T)\mathbf{X}_t + \varphi(T)\mathbf{M}^* + \boldsymbol{\varepsilon}_{t,T}] \end{cases}$$

2.2. A General Procedure to Determine the Stochastic Factors

In the previous subsection, we have presented the general model for characterizing the commodity price dynamics based on the assumption that the log commodity spot price is the sum of several factors. However, to the best of the authors' knowledge, the optimal number of stochastic factors has not yet been studied, for these models.

This subsection presents a theoretical procedure to establish the optimal number of factors. It also presents a way to determine how those factors should be aligned (long-term, short-term, seasonal, etc.).

To address this problem, let us suppose that there are M futures maturities and n observations of the forward curve, that is, the matrix $\mathbf{U} = \log\left(F_{t,T_i}\right)$, $t = 0, \ldots, n$; $i = 1, \ldots, M$ has dimension $M \times (n+1)$. We further assume, as usual, that $n \gg M$. To determine the optimal number of stochastic factors needed to characterize the commodity price dynamic in the best way, first we must realize that the number of factors is equal to $rank(\mathbf{R})$ and, from the previous expression, $rank(\mathbf{R})$ has to be equal to the rank of the variance–covariance matrix of $\mathbf{U}$. If, as usual, the process $\mathbf{X}_t$ has a unit root, so it is non-stationary and the variance and covariances are infinity, we need another matrix to determine the rank of the variance–covariance matrix of $\mathbf{U}$.

If we define volatility (instantaneous variance) as $\sigma_{T_i}^2 = \lim_{h\to 0} \frac{Var\left(\log F_{t+h,T_i} - \log F_{t,T_i}\right)}{h}$, $i = 1, \ldots, M$ and cross-volatility (instantaneous covariance) as $\sigma_{T_i,T_j} = \lim_{h\to 0} \frac{Cov\left(\log F_{t+h,T_i} - \log F_{t,T_i}, \log F_{t+h,T_j} - \log F_{t,T_j}\right)}{h}$, $i, j = 1, \ldots, M$ (as expected, $\sigma_{T_i}^2 = \sigma_{T_i,T_j}$), we have the necessary matrix. Although we cannot compute the limit from the data, we can set h as the shortest time period available and estimate it directly as $\hat{\sigma}_{T_i,T_j} = \left[\frac{\hat{Cov}\left(\log F_{t+h,T_i} - \log F_{t,T_i}, \log F_{t+h,T_j} - \log F_{t,T_j}\right)}{h}\right]$, where $\hat{Cov}$ is the sample covariance.

We thus define the matrix $\mathbf{\Theta} = \left(\Theta_{ij}\right)$ (dim $M \times M$) as $\Theta_{ij} = \sigma_{T_i,T_j}$, $i, j = 1, \ldots, M$. We can estimate it directly from our database and we can also estimate its rank. Once we have this rank, as stated above $rank(\mathbf{\Theta}) = rank(\mathbf{R}) = N$, we know the number of stochastic factors (N) that define the commodity price dynamics.

From a practical point of view, however, if we follow this procedure as explained above, unless one futures maturity is a linear combination of the rest (which is not likely), we obtain $rank(\mathbf{\Theta}) = rank(\mathbf{R}) = N$. Nevertheless, the weights of these factors are going to be different and most of them will have an insignificant weight.

Fortunately, from this procedure, we can also estimate the eigenvalues $k_1, \ldots, k_N$ and, from there, determine the factor weight through the eigenvalues' relative weight. We can estimate the eigenvalues of $\mathbf{A}$ via a nonlinear search procedure by using the fact that $\sigma_{T_i,T_j} = \mathbf{C}e^{\mathbf{A}T_i}\mathbf{R}\mathbf{R}'\left(e^{\mathbf{A}T_j}\right)'\mathbf{C}'$ (see García et al. 2008) and therefore, $\mathbf{\Theta}$ can be expressed as $\mathbf{\Theta} = \mathbf{C}\begin{pmatrix} e^{\mathbf{A}T_1} & \cdots & e^{\mathbf{A}T_M} \end{pmatrix}\mathbf{R}\mathbf{R}'\begin{pmatrix} e^{\mathbf{A}T_1} \\ \vdots \\ e^{\mathbf{A}T_M} \end{pmatrix}'\mathbf{C}'$. σ_{T_i,T_j} is a linear combination of products of $e^{k_1T}, \ldots, e^{k_NT}$. In other words, if $k_1, \ldots, k_N$ are the eigenvalues of $\mathbf{A}$, $e^{k_1T}, \ldots, e^{k_NT}$ must be the eigenvalues of $\mathbf{\Theta}$.

Moreover, from the eigenvalues of matrix $\mathbf{A}$, it is also easy to determine the factors. Taking into account that factors´ Stochastic Differential Equation (SDE) is $d\mathbf{X}_t = \mathbf{M}dt + \mathbf{A}\mathbf{X}_t dt + \mathbf{R}d\mathbf{W}_t$, if, for example, the eigenvalue is $k = 0$, the factor is a long-term one because the SDE associated with this factor is a random walk (General Brownian Motion (GBM)): $dx_{it} = \mu_i dt + \sigma_i dW_{it}$. On the other hand, if the eigenvalue is $k \in (-1, 0)$, the factor is a short-term one because the SDE associated with this factor is an Ornstein–Uhlenbeck: $dx_{it} = \lambda x_{it} dt + \sigma_i dW_{it}$. If the eigenvalue is complex, the factor is a seasonal one.

From a practical point of view, when we carry out this procedure we get N eigenvalues and we need to decide how many of them to optimally choose. The way to decide this is through the relative weight of the eigenvalues. By normalizing the largest one to 1, the smallest eigenvalues represent negligible factors. This allows us to decide how many factors must be optimally chosen.

In order to clarify concepts, the following example could be useful, if we have $M = 9$ futures with maturities at times $T_1, \ldots, T_9$. The method is as follows.

1. Compute $\hat{\Theta}_{ij} = \left[\frac{\hat{Cov}\left(\log F_{t+h,T_i} - \log F_{t,T_i}, \log F_{t+h,T_j} - \log F_{t,T_j}\right)}{h}\right]$.
2. Compute the rank of $\hat{\Theta}$. Let us assume that this is 3.
3. As a result, we have three eigenvalues k_1, k_2 and k_3. It is usual to assume that $k_1 = 0$ as the futures process is not stationary, but k_1 can nevertheless be estimated. If we do assume it, however, we obtain that σ_{T_i,T_j} is a linear combination of the products of $e^{0T} = 1$, e^{k_2T} and e^{k_3T}. Therefore, we obtain the general equation $\Theta_{ij} = \alpha_{11} + \alpha_{12}e^{k_2T_j} + \alpha_{13}e^{k_3T_j} + \alpha_{21}e^{k_2T_i} + \alpha_{31}e^{k_3T_i} + \alpha_{22}e^{k_2(T_i+T_j)} + \alpha_{23}e^{k_2T_i+k_3T_j} + \alpha_{23}e^{k_2T_j+k_3T_i} + \alpha_{33}e^{k_3(T_i+T_j)}$ which can be estimated numerically as:
    a. Select an initial estimate of (k_2, k_3).
    b. Regress $\hat{\Theta}_{ij}$ and compute the error.

c. Iteratively select another estimate of (k_2, k_3) and get back to b.

To the best of the authors' knowledge, no method has combined the knowledge of this concrete specification $G = \Phi(T)$ with a nonlinear search procedure to identify factors, which is one of the contributions made by this article.

Once we have determined the optimal number and form of the stochastic factors to characterize the commodity price dynamics, we can estimate model parameters using standard techniques. The Kalman filter (see, for example, [26]) uses a complex calibration technique. Other techniques include approximations such as [18] or [27]. Finally, the recently published option by [28] presents an optimal way of estimating model parameters by avoiding the use of the Kalman filter. Model parameters are estimated in the papers and so, for the sake of brevity we do not estimate the parameters in this study.

3. Data and Main Results

3.1. Data

In this subsection, we briefly describe the datasets used in this study. The datasets include weekly observations corresponding to futures prices for four commodities: WTI light sweet crude oil, heating oil, unleaded gasoline (RBOB) and Henry Hub natural gas. These futures were taken into consideration because they are the most representative and classic among the products. They are futures with many historical series and futures at many maturities. Therefore, they are considered as ideal for studying the optimal number of factors that should be chosen.

In this study, two data sets were considered for each commodity. Data set 1 contains less futures maturities, but more years of observations considered while data set 2 contains more futures maturities, but less years of observations. For dataset 1 (Table 1A), related to WTI crude oil, it comprised contracts from 4 September 1989 to 3 June 2013 (1240 weekly observations) for futures maturities from F1 to F17, F1 being the contract for the month closest to maturity, F2 the contract for the second-closest month to maturity, etc. In the case of heating oil, it contained contracts from 21 January 1991 to 3 June 2013 (1168 weekly observations) for futures maturities from F1 to F15. Meanwhile, RBOB gasoline first data set comprised contracts from 3 October 2005 to 3 June 2013 (401 weekly observations) for futures maturities from F1 to F12 and in the case of Henry Hub natural gas, it contained contracts from 27 January 1992 to 3 June 2013 (1115 weekly observations) for futures maturities from F1 to F16.

Looking at the dataset 2 (Table 1B), in the case of WTI crude oil, it comprised contracts from 18 September 1995 to 3 June 2013 (925 weekly observations) for futures maturities from F1 to F28 and in the case of heating oil it comprised contracts from 9 September 1996 to 3 June 2013 (874 weekly observations) for futures maturities from F1 to F18. In the meantime, RBOB gasoline comprised contracts from 2 February 2007 to 3 June 2013 (330 weekly observations) for futures maturities from F1 to F36 and, to end with, in the case of Henry Hub natural gas, dataset 2 (Table 1B) contained contracts from 24 March 1997 to 3 June 2013 (856 weekly observations) for futures maturities from F1 to F36.

Table 1 shows the main descriptive statistics of the futures, particularly the mean and volatility, for each dataset. It is interesting to note that the lack of low-cost transportation and the limited storability of natural gas made its supply unresponsive to seasonal variation in demand. Thus, natural gas prices were strongly seasonal [3]. The unleaded gasoline was also seasonal.

Table 1. Descriptive statistics.

(A) Dataset 1								
	WTI Crude Oil		**Gasoline**		**Natural Gas**		**Heating Oil**	
	Mean ($/bbl)	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**	**Mean ($/MMBtu)**	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**
F1	43.1	30%	96.7	32%	4.2	45%	63.6	28%
F2	43.3	27%	96.5	30%	4.3	40%	63.8	26%
F3	43.3	25%	96.4	29%	4.3	36%	64.0	24%
F4	43.4	24%	96.3	27%	4.4	32%	64.1	23%
F5	43.3	23%	96.1	27%	4.4	29%	64.1	22%
F6	43.3	22%	95.9	26%	4.4	27%	64.2	21%
F7	43.3	21%	95.7	25%	4.5	26%	64.2	20%
F8	43.2	20%	95.6	26%	4.5	24%	64.1	19%
F9	43.2	20%	95.6	25%	4.5	24%	64.1	19%
F10	43.1	19%	95.4	26%	4.5	22%	64.1	18%
F11	43.1	19%	95.5	25%	4.5	22%	64.1	18%
F12	43.0	18%	95.5	25%	4.5	21%	64.0	17%
F13	43.0	18%			4.5	20%	63.7	17%
F14	42.9	18%			4.5	20%	63.4	17%
F15	42.9	17%			4.5	20%	63.0	17%
F16	42.8	17%			4.5	20%		
F17	42.8	17%						
(B) Dataset 2								
	WTI Crude Oil		**Gasoline**		**Natural Gas**		**Heating oil**	
	Mean ($/bbl)	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**	**Mean ($/MMBtu)**	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**
F1	50.9	30%	101.1	32%	4.9	45%	53.3	30%
F2	51.2	28%	100.6	30%	5.0	40%	53.5	28%
F3	51.3	26%	100.3	30%	5.1	38%	53.6	26%
F4	51.3	25%	100.1	28%	5.1	33%	53.7	25%
F5	51.3	24%	99.8	28%	5.2	31%	53.7	24%
F6	51.3	23%	99.6	27%	5.2	28%	53.7	23%
F7	51.3	22%	99.3	26%	5.3	27%	53.7	22%
F8	51.3	21%	99.2	26%	5.3	26%	53.7	21%
F9	51.2	21%	99.1	25%	5.3	25%	53.7	21%
F10	51.2	20%	99.1	27%	5.3	24%	53.6	20%
F11	51.1	20%	99.1	26%	5.3	22%	53.6	19%
F12	51.1	19%	99.1	26%	5.3	21%	53.5	19%
F13	51.0	19%	99.1	26%	5.3	21%	53.3	19%
F14	50.9	19%	99.0	25%	5.3	21%	53.0	19%
F15	50.8	18%	98.9	25%	5.3	21%	53.0	18%
F16	50.8	18%	98.8	23%	5.3	20%	53.5	18%
F17	50.7	18%	98.5	24%	5.3	20%	55.4	18%
F18	50.7	18%	98.3	23%	5.3	19%	58.4	18%
F19	50.6	17%	98.1	23%	5.3	20%		
F20	50.5	17%	98.0	23%	5.3	19%		
F21	50.5	17%	98.0	23%	5.3	19%		
F22	50.4	17%	97.9	24%	5.3	20%		
F23	50.4	17%	97.9	23%	5.2	18%		
F24	50.3	17%	97.9	24%	5.2	18%		
F25	50.3	16%	97.9	24%	5.2	18%		
F26	50.2	16%	97.8	23%	5.2	18%		
F27	50.2	16%	97.8	24%	5.2	18%		
F28	50.1	16%	97.7	23%	5.2	18%		
F29			97.6	23%	5.2	18%		
F30			97.5	22%	5.2	18%		
F31			97.4	22%	5.2	18%		
F32			97.4	23%	5.2	19%		
F33			97.3	22%	5.2	18%		
F34			97.1	23%	5.2	19%		
F35			97.0	23%	5.2	18%		
F36			97.0	23%	5.2	17%		

3.2. Main Results

We now present the results after applying the method proposed to the 4 commodities (2 datasets per commodity) described above in order to select the number of factors to model the behavior of commodity prices. The results correspond to the eigenvalues in decreasing order, the percentage of the overall variability that they explain and the cumulative proportion of explained variance. These are reported in Tables 2–5.

Table 2. Eigenvalues for both datasets of the WTI light sweet crude oil.

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	99.6713	99.6713	100	99.5448	99.5448
0.3202	0.3191	99.9904	0.4428	0.4408	99.9855
0.0084	0.0084	99.9988	0.0126	0.0125	99.9980
0.0010	0.0010	99.9998	0.0017	0.0017	99.9997
0.0001	0.0001	99.9999	0.0002	0.0002	99.9998
2.9905×10^{-5}	2.9806×10^{-5}	100	0.0001	0.0001	99.9999
1.2209×10^{-5}	1.2169×10^{-5}	100	3.7318×10^{-5}	3.7148×10^{-5}	100
5.4907×10^{-6}	5.4727×10^{-6}	100	1.6898×10^{-5}	1.6821×10^{-5}	100
2.7838×10^{-6}	2.7746×10^{-6}	100	8.7477×10^{-6}	8.7079×10^{-6}	100
1.5250×10^{-6}	1.5200×10^{-6}	100	4.4713×10^{-6}	4.4509×10^{-6}	100
7.5290×10^{-7}	7.5043×10^{-7}	100	3.0355×10^{-6}	3.0217×10^{-6}	100
4.3460×10^{-7}	4.3317×10^{-7}	100	2.3004×10^{-6}	2.2899×10^{-6}	100
3.3010×10^{-7}	3.2901×10^{-7}	100	1.3628×10^{-6}	1.3566×10^{-6}	100
1.8100×10^{-7}	1.8041×10^{-7}	100	8.7940×10^{-7}	8.7540×10^{-7}	100
1.1490×10^{-7}	1.1452×10^{-7}	100	4.7100×10^{-7}	4.6886×10^{-7}	100
1.0350×10^{-7}	1.0316×10^{-7}	100	3.6450×10^{-7}	3.6284×10^{-7}	100
3.9300×10^{-8}	3.9171×10^{-8}	100	2.3880×10^{-7}	2.3771×10^{-7}	100
			1.8840×10^{-7}	1.8754×10^{-7}	100
			1.4180×10^{-7}	1.4115×10^{-7}	100
			1.1480×10^{-7}	1.1428×10^{-7}	100
			1.0070×10^{-7}	1.0024×10^{-7}	100
			7.7800×10^{-8}	7.7446×10^{-8}	100
			5.4200×10^{-8}	5.3953×10^{-8}	100
			4.7800×10^{-8}	4.7582×10^{-8}	100
			3.8600×10^{-8}	3.8424×10^{-8}	100
			2.3000×10^{-8}	2.2895×10^{-8}	100
			2.1600×10^{-8}	2.1502×10^{-8}	100
			8.9000×10^{-9}	8.8595×10^{-9}	100

As a general rule, we can consider that the first factor, which corresponds to the first eigenvalue, was clearly dominant in the sense that it can explain a percentage of the total variance ranging between 95.2% and 99.7%, depending on the commodity. It captures qualitative long-run effects. However, it is always necessary to consider a second factor capable of taking up short-term effects. Both the first and second factors explain a cumulative proportion of overall variance between 97.5% and 99.9%, depending on the case under study. In WTI light sweet crude oil, these two factors explain more than a 99.99% of the total variance is explained, while in heating oil case studies, these percentages were approximately 99.88% and in unleaded gasoline and Henry Hub natural gas, they were approximately 97–98%.

Consequently, in the first commodity (crude oil) it is recommended that just the first two factors are considered. The reason is that a third factor will impose a larger estimating effort and a minimum reduction in terms of error measures. The first factor will capture long-term effects, such as world economic events, which significantly impact on commodity prices. The second factor will capture the nature of short-term components such as temporary issues and unforeseen situations. The third and

following stochastic factors can be considered as seasonal factors [28] and, as we know, crude oil is a non-seasonal commodity. This matter reinforces the idea that it is suitable to consider a model with only the first two factors.

The next commodity, heating oil, presents some seasonal behavior, which could be captured by a third factor. The fact that the gain in the percentage of cumulative proportion of overall variance goes from 99.88 to 99.94 and from 99.90 to 99.94 in its respective datasets suggest the inclusion of a third factor was not necessary.

Table 3. Eigenvalues for both datasets of the heating oil.

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	99.6133	99.6133	100	99.5365	99.5365
0.2698	0.2687	99.8820	0.3666	0.3649	99.9014
0.0658	0.0655	99.9475	0.0475	0.0472	99.9486
0.0474	0.0472	99.9947	0.0448	0.0446	99.9932
0.0028	0.0028	99.9975	0.0037	0.0037	99.9969
0.0013	0.0013	99.9988	0.0012	0.0012	99.9981
0.0009	0.0008	99.9997	0.0011	0.0011	99.9992
0.0001	0.0001	99.9998	0.0005	0.0005	99.9997
0.0001	0.0001	99.9999	0.0001	0.0001	99.9998
4.7937×10^{-5}	4.7752×10^{-5}	99.9999	0.0001	0.0001	99.9999
1.9734×10^{-5}	1.9658×10^{-5}	100	4.1450×10^{-5}	4.1257×10^{-5}	99.9999
1.1626×10^{-5}	1.1581×10^{-5}	100	2.8767×10^{-5}	2.8633×10^{-5}	99.9999
1.0482×10^{-5}	1.0441×10^{-5}	100	1.5784×10^{-5}	1.5711×10^{-5}	100
9.6273×10^{-6}	9.5901×10^{-6}	100	1.3478×10^{-5}	1.3416×10^{-5}	100
6.3346×10^{-6}	6.3101×10^{-6}	100	9.3425×10^{-6}	9.2992×10^{-6}	100
			7.9877×10^{-6}	7.9507×10^{-6}	100
			6.1859×10^{-6}	6.1572×10^{-6}	100
			5.7507×10^{-6}	5.7240×10^{-6}	100

Conversely, for the unleaded gasoline and Henry hub natural gas, at least a third factor seemed to be necessary. Both were seasonal commodities (see, for example, [3]). They were characterized by very limited storability and their prices were highly dependent on the commodity demand. Third and fourth factors will acknowledge this behavior. It seems necessary to capture more than long-term and short-term dynamics. Depending on the cumulative variance, if we would like to explain (98–99%), we need to consider at least a third factor or two more. In the unleaded gasoline case, the inclusion of a third factor would increase the cumulative proportion of overall variance from 98.48% to 99.73% and from 97.49% to 98.73%. However, with a fourth factor, we would reach 99.86% and 99.76%, respectively. When we apply the methodology proposed to Henry Hub natural gas datasets, we also verify the need to consider a third and even a fourth factor to explain 99.80% and 99.65% of the total variance, respectively.

Table 4. Eigenvalues for both datasets of the unleaded gasoline (RBOB).

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	96.8591	96.8591	100	95.2473	95.2473
1.6748	1.6222	98.4813	2.3570	2.2450	97.4924
1.2901	1.2496	99.7308	1.3050	1.2429	98.7353
0.1334	0.1292	99.8600	1.0762	1.0250	99.7603
0.0558	0.0540	99.9140	0.0639	0.0608	99.8212
0.0386	0.0374	99.9515	0.0599	0.0570	99.8782
0.0217	0.0210	99.9724	0.0437	0.0416	99.9198
0.0156	0.0151	99.9876	0.0217	0.0206	99.9405
0.0093	0.0090	99.9966	0.0208	0.0198	99.9602
0.0022	0.0022	99.9988	0.0171	0.0163	99.9765
0.0009	0.0009	99.9997	0.0074	0.0070	99.9835
0.0003	0.0003	100	0.0049	0.0047	99.9882
			0.0030	0.0029	99.9910
			0.0025	0.0024	99.9935
			0.0019	0.0018	99.9953
			0.0012	0.0011	99.9964
			0.0008	0.0008	99.9972
			0.0006	0.0006	99.9978
			0.0004	0.0004	99.9982
			0.0004	0.0003	99.9986
			0.0003	0.0003	99.9989
			0.0003	0.0003	99.9991
			0.0002	0.0002	99.9993
			0.0001	0.0001	99.9995
			0.0001	0.0001	99.9996
			0.0001	0.0001	99.9997
			0.0001	0.0001	99.9997
			0.0001	0.0001	99.9998
			0.0001	0.0000	99.9999
			3.8439×10^{-5}	3.6612×10^{-5}	99.9999
			2.8300×10^{-5}	2.6955×10^{-5}	99.9999
			2.4205×10^{-5}	2.3055×10^{-5}	99.9999
			1.9530×10^{-5}	1.8602×10^{-5}	100
			1.5016×10^{-5}	1.4303×10^{-5}	100
			1.2694×10^{-5}	1.2091×10^{-5}	100
			9.9475×10^{-6}	9.4747×10^{-6}	100

These results are coherent with the patterns shown in the futures contracts of each commodity. By considering seasonality as a stochastic factor instead of a deterministic one, we can choose from two- to four-factor models to better model the behavior of commodity prices. It should be noted that the long-term and short-term effects, captured by the first two factors, are clearly dominant in terms of their eigenvalues' relative weight. However, the seasonality should be considered if necessary.

It is important to bear in mind that the distinction between long term and short term is not always direct. It is related to the eigenvalue of the factor, which, as we have stated, is always in the form e^k with $k \leq 0$ (a positive k would mean an explosive process, which is clearly not observed in the data).

If $k = 0$, we have a long-term effect (a unit root). The more negative k is, the shorter the effect. Therefore, $k = -1$ means a much shorter effect than $k = -0.01$, for example.

Explanation capacities of each factor are measured according to their (relative) contribution to the global variance. For example, if there is a unique factor related to eigenvalue $k = 0$ that gives 90% of variance, we would conclude that long term dynamics explain 90% of the variance.

Table 5. Eigenvalues for both datasets of the henry hub natural gas.

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	97.8179	97.8179	100	95.8957	95.8957
1.1564	1.1311	98.9491	2.7972	2.6824	98.5782
0.4785	0.4681	99.4172	0.5993	0.5747	99.1529
0.3960	0.3874	99.8046	0.5221	0.5007	99.6535
0.0839	0.0821	99.8867	0.1178	0.1130	99.7665
0.0730	0.0714	99.9580	0.0993	0.0952	99.8617
0.0223	0.0218	99.9798	0.0782	0.0750	99.9367
0.0052	0.0050	99.9849	0.0166	0.0159	99.9527
0.0039	0.0038	99.9887	0.0074	0.0071	99.9598
0.0031	0.0030	99.9917	0.0070	0.0067	99.9665
0.0027	0.0027	99.9944	0.0060	0.0057	99.9722
0.0023	0.0023	99.9967	0.0051	0.0049	99.9772
0.0012	0.0012	99.9979	0.0048	0.0046	99.9818
0.0010	0.0010	99.9988	0.0038	0.0037	99.9854
0.0007	0.0007	99.9995	0.0032	0.0031	99.9886
0.0005	0.0005	100	0.0024	0.0023	99.9908
			0.0020	0.0019	99.9927
			0.0018	0.0017	99.9944
			0.0017	0.0016	99.9961
			0.0013	0.0012	99.9973
			0.0010	0.0009	99.9983
			0.0006	0.0006	99.9988
			0.0003	0.0003	99.9991
			0.0002	0.0002	99.9993
			0.0002	0.0002	99.9995
			0.0001	0.0001	99.9996
			0.0001	0.0001	99.9997
			0.0001	0.0001	99.9998
			0.0001	0.0001	99.9998
			0.0001	0.0001	99.9999
			3.2345×10^{-5}	3.1018×10^{-5}	99.9999
			2.8128×10^{-5}	2.6974×10^{-5}	100
			1.5565×10^{-5}	1.4926×10^{-5}	100
			1.2489×10^{-5}	1.1977×10^{-5}	100
			9.3473×10^{-6}	8.9637×10^{-6}	100
			7.6972×10^{-6}	7.3813×10^{-6}	100

It should be noted that this article focuses on the econometric theory and identifies the optimal number of factors to characterize the dynamics of commodity prices. Apart from this econometric approach, where each factor represents a component—long term, short term, seasonal, etc.—these factors may also capture economic forces [29–31]. In other words, there are economic forces that are being captured by these factors, such as technology effects (long term) or the functioning of the market (short term). Following [15], we argue that the long-term factor reflects expectations of the exhaustion of the existing supply, improvements in technology for the production and discovery of the commodity, inflation, as well as political and regulatory effects. The short-term factor reflects short-term changes in demand or intermittent supply disruptions. An interpretation of seasonal factors can be found in [3].

This method provides a new selection criterion for obtaining the optimal number of factors. It is always important to keep in mind the purpose of modeling such commodity prices. If we need more accuracy because, for example, we are designing investment strategies, the consideration of more factors is understandable. We could also use fewer factors in a different case.

This is important because, on one hand, if we use too many factors the model will be too complex and parameter estimation may not be accurate. On the other hand, if we use too few factors the model will not be acceptable because it will not capture all the characteristic of the price dynamics that we need to consider in order to solve our problem.

We believe our findings to be very useful for researchers and practitioners. Based on our findings, a researcher who needs to model a commodity price dynamic can use our method to identify the number and the characteristics of the factors to be included in the model. Moreover, a practitioner who is investing or measuring risk can also use our methodology in order to identify the optimal number of factors needed and their characteristics.

Finally, as stated above, we have chosen to order the factors according to their relative (joint) contribution to variance because it is a direct and simple way to interpret the results. We are aware that collinearity and, in general, correlation structures can modify the results. However, since the first eigenvalue explains around 95% of variance, it seems unlikely that results are going to change substantially by a more refined analysis.

4. Summary and Conclusions

In this article, we propose a novel methodology for choosing the optimal number of stochastic factors to be included in a model of the term structure of futures commodity prices. With this method, we add to the research related to the way we characterize commodity price dynamics.

The procedure is based on the eigenvalues of the variance–covariance matrix. Moreover, in deciding how many of them to choose, we propose using the relative weight of the eigenvalues and the percentage of the total variance explained by them and balancing this with the effort of estimating more parameters.

In this article, we applied our method to eight datasets, corresponding with four different commodities: crude oil, heating oil, unleaded gasoline and natural gas. Results indicate that to model the first two commodity prices two factors are suitable, which corresponds with the two biggest eigenvalues, since they are sufficient to account for both long-term and short-term structures. Nevertheless, in the case of unleaded gasoline and natural gas, a third or even fourth factor is needed. We think that, in accordance with the literature, this is related to their seasonal behavior.

Our results support the notion that including too many or too few factors or factors with characteristics which are not optimal in a model for commodity prices could lead to results which may not be as accurate as they should be.

Author Contributions: Conceptualization, B.L. and J.P.; methodology, A.G.-M.; software, A.G.-M.; validation, Población and A.G.-M.; formal analysis, A.G.-M.; investigation, J.P. and A.G.-M.; resources, A.G.-M., B.L. and J.P.; data curation, A.G.-M., B.L. and J.P.; writing—original draft preparation, B.L.; writing—review and editing, B.L. y Población; visualization, B.L.; supervision, Población; project administration, B.L.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the financial support of the Spanish Ministerio de Economía, Industria y Competitividad Grant Number ECO2017-89,715-P (Javier Población).

Acknowledgments: This study should not be reported as representing the views of the Banco de España (BdE) or European Central Bank (ECB). The views in this study are those of the author and do not necessarily reflect those of the Banco de España (BdE) or European Central Bank (ECB). We thank the anonymous referees. Any errors are caused by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jahn, F.; Cook, M.; Graham, M. *Hydrocarbon Exploration and Production*; Elsevier: Aberdeen, UK, 2008.
2. Smit, H.T.J. Investment analysis of offshore concessions in the Netherlands. *Financ. Manag.* **1997**, *26*, 5–17. [CrossRef]
3. García, A.; Población, J.; Serna, G. The Stochastic Seasonal Behaviour of Natural Gas Prices. *Eur. Financ. Manag.* **2012**, *18*, 410–443.
4. García, A.; Población, J.; Serna, G. The stochastic seasonal behavior of energy commodity convenience yields. *Energy Econ.* **2013**, *40*, 155–166.
5. García, A.; Población, J.; Serna, G. Analyzing the dynamics of the refining margin: Implications for valuation and hedging. *Quant. Financ.* **2013**, *12*, 1839–1855.

6. Alquist, R.; Bhattarai, S.; Coibion, O. Commodity-price comovement and global economic activity. *J. Monet. Econ.* **2019**. [CrossRef]
7. Jacks, D.S. From boom to bust: A typology of real commodity prices in the long run. *Cliometrica* **2019**, *13*, 201–220. [CrossRef]
8. Nazlioglu, S. Oil and Agricultural Commodity Prices. In *Routledge Handbook of Energy Economics*; Soytas, U., San, R., Eds.; Routledge: London, UK, 2020; pp. 385–405.
9. Ayres, J.; Hevia, C.; Nicolini, J.P. Real exchange rates and primary commodity prices. *J. Intern. Econ.* **2020**, *122*. [CrossRef]
10. Población, J.; Serna, G. A common long-term trend for bulk shipping prices. *Marit. Econ. Logist.* **2018**, *20*, 421–432. [CrossRef]
11. García, A.; Población, J.; Serna, G. Hedging voyage charter rates on illiquid routes. *Intern. J. Shipp. Transp. Logist.* **2020**, *12*, 197–211.
12. Morgan, J.P. *Risk Metrics–Technical Document*; Reuters: New York, NY, USA, 1996.
13. Echaust, K.; Just, M. Value at Risk Estimation Using the GARCH-EVT Approach with Optimal Tail Selection. *Mathematics* **2020**, *8*, 114. [CrossRef]
14. Schwartz, E.S. The stochastic behavior of commodity prices: Implication for valuation and hedging. *J. Financ.* **1997**, *52*, 923–973. [CrossRef]
15. Schwartz, E.S.; Smith, J.E. Short-term variations and long-term dynamics in commodity prices. *Manag. Sci.* **2000**, *46*, 893–911. [CrossRef]
16. Sorensen, C. Modeling seasonality in agricultural commodity futures. *J. Futures Mark.* **2002**, *22*, 393–426. [CrossRef]
17. Gibson, R.; Schwartz, E.S. Stochastic convenience yield and the pricing of oil contingent claims. *J. Financ.* **1990**, *45*, 959–976. [CrossRef]
18. Cortazar, G.; Schwartz, E.S. Implementing a stochastic model for oil futures prices. *Energy Econ.* **2003**, *25*, 215–218. [CrossRef]
19. Cortazar, G.; Naranjo, L. An N-Factor gaussian model of oil futures prices. *J. Futures Mark.* **2006**, *26*, 209–313. [CrossRef]
20. Camiz, S.; Pillar, V.D. Identifying the Informational/Signal Dimension in Principal Component Analysis. *Mathematics* **2018**, *6*, 269. [CrossRef]
21. Cortazar, G.; Schwartz, E.S. The valuation of commodity-contingent claims. *J. Deriv.* **1994**, *1*, 27–39. [CrossRef]
22. Clewlow, L.; Strickland, C. *Energy Derivatives, Pricing and Risk Management*; Lamina Publication: London, UK, 2000.
23. Tolmasky, C.; Hindanov, D. Principal components analysis for correlated curves and seasonal commodities: The case of the petroleum market. *J. Futures Mark.* **2002**, *22*, 1019–1035. [CrossRef]
24. Hull, J. *Options, Futures and Other Derivatives*, 5th ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2003.
25. García, A.; Población, J.; Serna, G. A Note on Commodity Contingent Valuation. *J. Deriv. Hedge Fund.* **2008**, *13*, 311–320. [CrossRef]
26. Harvey, A.C. *Forecasting Structural Time Series Models and the Kalman Filter*; Cambridge University Press: Cambridge, UK, 1989.
27. Kolos, S.P.; Rohn, E.I. Estimating the commodity market price of risk for energy prices. *Energy Econ.* **2008**, *30*, 621–641. [CrossRef]
28. García, A.; Larraz, B.; Población, J. An alternative method to estimate parameters in modeling the behavior of commodity prices. *Quant. Financ.* **2016**, *16*, 1111–1127. [CrossRef]
29. Coles, J.L.; Li, Z.F. An Empirical Assessment of Empirical Corporate Finance. *SSRN* **2019**. [CrossRef]
30. Coles, J.L.; Li, Z.F. Managerial Attributes, Incentives, and Performance. *Rev. Corp. Financ. Stud.* **2019**. [CrossRef]
31. Dang, C.; Foerster, S.R.; Li, Z.F.; Tang, Z. Analyst Talent, Information, and Insider Trading. *SSRN* **2020**. [CrossRef]