

# Statistical Analysis of Correlated Fossil Fuel Securities

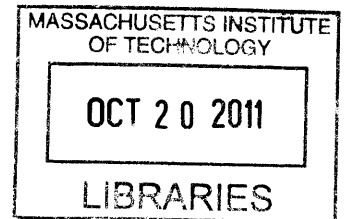
by

Derek Z. Li

SUBMITTED TO THE DEPARTMENT OF MECHANICAL ENGINEERING IN PARTIAL  
FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN MECHANICAL ENGINEERING  
AT THE  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2011



**ARCHIVES**

©2011 Massachusetts Institute of Technology. All Rights Reserved

Signature of Author: \_\_\_\_\_

\_\_\_\_\_ Department of Mechanical Engineering

May 6, 2011

Certified by: \_\_\_\_\_

\_\_\_\_\_ Paul D. Sclavounos  
Professor of Mechanical Engineering and Naval Architecture  
Thesis Supervisor

Accepted by: \_\_\_\_\_

\_\_\_\_\_ John H. Lienhard V  
Samuel C. Collins Professor of Mechanical Engineering  
Undergraduate Officer



# Statistical Analysis of Correlated Fossil Fuel Securities

by

Derek Z. Li

Submitted to the Department of Mechanical Engineering  
on May 6, 2011 in Partial Fulfillment  
of the Requirements for the Degree of Bachelor of Science in  
Mechanical Engineering.

## ABSTRACT

Forecasting the future prices or returns of a security is extraordinarily difficult if not impossible. However, statistical analysis of a basket of highly correlated securities offering a cross-sectional representation of a particular sector can yield information that is potentially tradable. Securities related to the fossil fuels industry are used as the basis of a practical application to two distinct forecasting techniques. The first method, forecasting using conditional multivariate Gaussian statistics, was shown to yield, in a relative sense, the best results for those securities which exhibited a high correlation with the rest of the basket. For the second method, principal component analysis was done on a basket of commodity futures to reveal a small number of dominant factors governing the movements of the portfolio. Autoregressive models were then applied to both the factors and futures, but results showed both to be essentially Markov processes.

Thesis Supervisor: Paul D. Sclavounos

Title: Professor of Mechanical Engineering and Naval Architecture

## Table of Contents

1.	INTRODUCTION.....	5
2.	PRELIMINARY STATISTICAL ANALYSIS.....	8
2.1	Processing Time Series Data	
2.2	Estimation of Global Covariance and Global Correlation Matrices	
3.	FORECASTING USING MULTIVARIATE GAUSSIAN STATISTICS.....	14
3.1	Extracting Information from the Fossil Fuel Sector	
3.2	Conditional Mean and Conditional Covariance	
3.3	Results	
4.	MODELING THE PORTFOLIO.....	21
4.1.	Principal Component Analysis	
4.2.	Modeling the Crude Oil Futures Curve	
4.3.	Extension of PCA onto a Portfolio of Correlated Commodities	
5.	FORECASTING USING AUTOREGRESSIVE METHODS.....	28
5.1.	Autoregressive Methods in Time Series Analysis	
5.2.	AR and ARMA Models	
5.3	Comparing Forecasts of Factors and Futures	
6.	CONCLUSIONS.....	33
6.1.	Summary of Results	
6.2.	Suggestions for Further Research	
7.	ACKNOWLEDGEMENTS.....	35
8.	REFERENCES.....	36

## 1. INTRODUCTION

### 1.1. The Fossil Fuels Market

The world's dependence on fossil fuels for energy has cemented the importance of fossil fuels in the world's financial markets. Various players including producers, refiners, airlines, banks, hedge funds, and retail investors all have an interest in the price movements in markets for fossil fuels.

It is, however, extraordinarily difficult if not impossible to predict these notoriously volatile price movements. Price movements of a security are described as a random walk, or a Gaussian distribution with zero mean return.

This thesis explores the idea that while a single security may be near impossible to forecast, it may be possible to extract additional information from a basket of related securities using various statistical methods. Such a basket should be formulated carefully to include highly correlated securities that collectively offer a complete, cross-sectional representation of the fossil fuel industry. For this study, the basket includes exposure to commodities, equities, and currency. An overview of the particular securities that are involved in this study follows here.

#### *West Texas Intermediate and Brent Crude Oil*

West Texas Intermediate (WTI) Crude oil is the benchmark crude oil in the US, traded on the New York Mercantile Exchange (NYMEX). Its delivery point is Cushing, Oklahoma. Brent Crude is the benchmark crude oil in Europe, generally extracted from oilfields in the North Sea. Contracts are traded electronically and cleared by the Intercontinental Exchange (ICE). Contracts are quoted in dollars / barrel and each contract is for one thousand barrels.

The two contracts usually trade closely together, with WTI at a small premium to Brent due to the slight premium in quality of the WTI grade. Crude oils around the world are classified by their API gravity, sulfur content, and acidity, among other things, and are refined into a diverse range of products such as gasoline, naphtha, heating oil, bunker fuel, and countless other products.

#### *RBOB Gasoline*

RBOB stands for reformulated gasoline blendstock for oxygenate blending. It is the lightest, most valuable refined product of crude oil. It is used mostly for transportation, which accounts for a significant percentage of energy consumption in the United States. RBOB contracts are

traded on the NYMEX and is the pricing benchmark for gasoline. Prices of the contracts are quoted in dollars / gallon, and the contract size is 42,000 gallons.

### *Gasoil*

Gasoil is another product refined from crude oil which includes diesel and heating oil. Gasoil contracts are traded electronically on ICE, and have delivery hubs around the world, notably in Singapore and in the Amsterdam-Rotterdam-Antwerp (ARA) area. Gasoil prices are quoted in dollars / metric ton, with a contract size of 100 metric tons.

### *Natural Gas*

Natural gas is used as both a source of energy for residential homes and for electricity generation at power plants. It is traded in New York on the NYMEX, and priced in dollars / MMBTU (million British Thermal Units), where each contract is for 10,000 MMBTU. Since natural gas is not a product of crude oil and used for different purposes than crude oil products, it is lowly correlated with those markets and generally trades as its own separate commodity. However, it is still an integral part of the world's energy demands and thus may still provide useful information in this study.

### *Trade Weighted Dollar Index*

The Trade Weighted Dollar Index (TWDI) may be any number of indices that value the US dollar based on a basket of currencies of the trading partners of the United States. The index used in this study is published by the Federal Reserve, and weighs the dollar against the currencies of seven of the largest trading partners of the US: the British Pound (GBP), the Euro (EUR), the Japanese Yen (JPY), the Swedish Krona (SEK), the Australian Dollar (AUD), the Swiss Franc (CHF), and the Canadian Dollar (CAD). The strength of the dollar heavily influences the prices of commodities as the two are often seen as substitute investments for one another. As the dollar weakens, investors generally pour into commodities as a higher yielding investment, and as the dollar gains, investors pour into the dollar for similar reasons. This relationship will result in a strong negative correlation between the prices of the TWDI and that of the fossil fuel commodities.

### *Exxon Mobil (XOM) and Chevron (CVX) Common Stock*

Both Exxon and Chevron are large, diversified oil and gas producers, each considered one of the six "supermajor" energy corporations in the world (the other four being Total, Royal Dutch Shell, BP, and ConocoPhillips). Their shares are listed on the New York Stock Exchange (NYSE) and are components of the thirty-company Dow Jones Industrial Average (DJIA). Both deal in the extraction, transportation, and refining operations of a broad range of fossil fuel products, and their share prices may reflect useful information about the commodity markets.

Using these securities, this thesis will explore the effectiveness of various statistical methods in forecasting the prices of these fossil fuel market securities, specifically the following two methods:

*Forecasting Using Multivariate Gaussian Statistics*

A basket of eight securities composed of the front month futures contracts of the five aforementioned commodities, the dollar index, and the two equities, will be analyzed to see if a stable covariance matrix can be produced. If so, a prediction model employing conditional multivariate Gaussian statistics will be used to forecast a particular security of interest using known prices of the rest of the basket.

*Forecasting Using Autoregressive Methods*

Factor analysis of a portfolio of securities composed of a number of futures contracts of WTI, Brent, and RBOB, will reveal a small number of dominant factors that explain movements of the portfolio. Five futures will be used for each commodity, with times to maturity of 1, 3, 6, 9, and 12 months, yielding a portfolio that contains 16 securities with the inclusion of the dollar index. The factors and the front month futures contracts will be fitted with autoregressive (AR) and autoregressive moving average (ARMA) models to determine the forecastability of the price series.

## 2. PRELIMINARY STATISTICAL ANALYSIS

### 2.1. Processing Time Series Data

The purpose of this section is to prepare the raw price series of the securities under study so that statistical analysis can be performed. First, it is assumed that the securities involved in this study follow diffusions of this type:

$$\frac{dS(t)}{S(t,T)} = \mu(t)dt + \sigma(t)dW(t)$$

where  $S(t)$  is the stochastic process,  $\mu(t)$  and  $\sigma(t)$  are the mean and standard deviations of that process, and  $W(t)$  is a Brownian motion. As it pertains to this study, the above equation governs the dynamics of spot and equity prices while futures prices will follow

$$\frac{dF(t,T)}{F(t,T)} = \mu(t,T)dt + \sigma(t,T)dW(t)$$

where  $F(t,T)$  is the futures contract with time to maturity  $T$ . Histograms in Fig. 1 show graphically that the processes follow a Gaussian distribution.

Exchange traded futures contracts have fixed expiration dates, which results in a floating time to maturity as time passes. As a contract nears expiration, its volatility increases, and as a result the security cannot be characterized as a stationary process, since it has non-constant volatility.

The process can be made stationary however by creating rolling futures contracts  $f(t,t+\tau)$  with constant relative tenor  $\tau$  by interpolating between two fixed-tenor futures contracts, thus

$$\ln f(t, t + \tau_j) = \frac{(t + \tau_j - T_j) * \ln F(t, T_{j+1}) + (T_{j+1} - t - \tau_j) * \ln F(t, T_j)}{T_{j+1} - T_j}, T_j < t + \tau_j < T_{j+1}$$

Since these pseudo-contracts have fixed times to maturity  $\tau_j$ , the volatilities will remain constant over time. This stationarity condition will be important in application to statistical forecasting models later on.

With other forecasting models, however, the stationarity condition is not as imperative as it was in Scлавounos and Ellefsen (2009) since it is not the purpose of this study to model the volatilities in commodity futures curves. The most important task for the time being is to find a



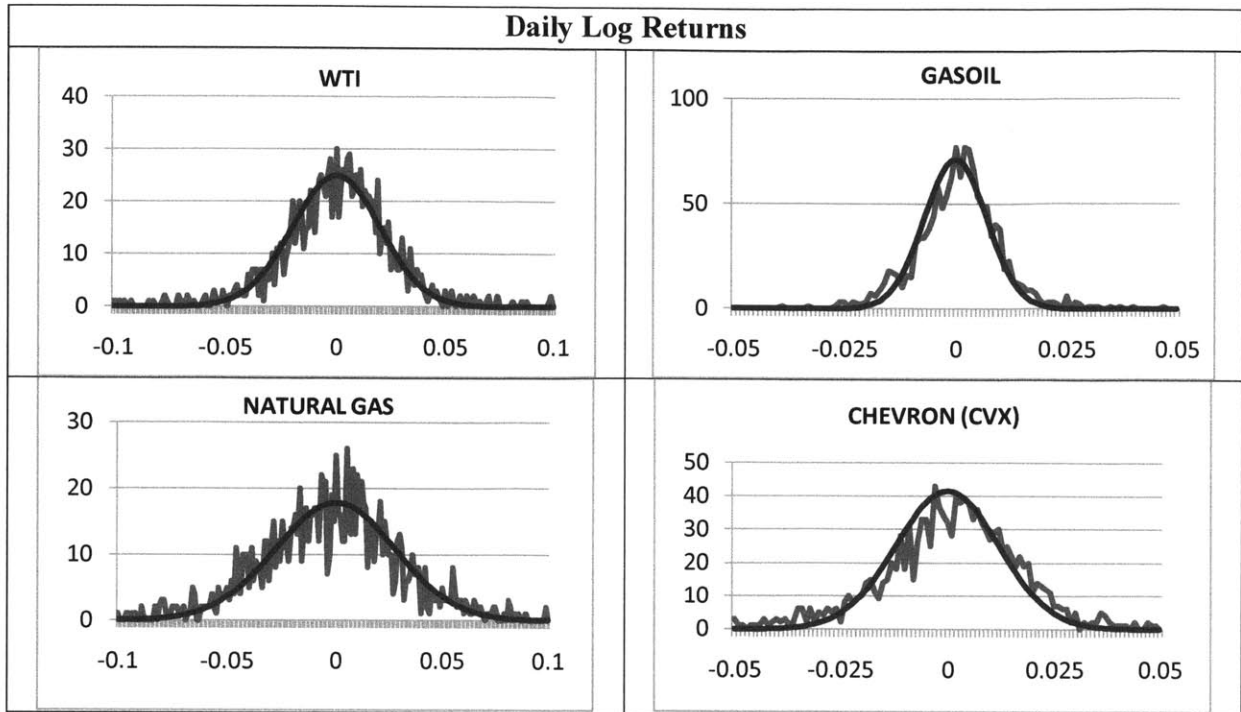


Figure 1: Histograms of daily log returns of selected securities, shown to be approximately Gaussian.

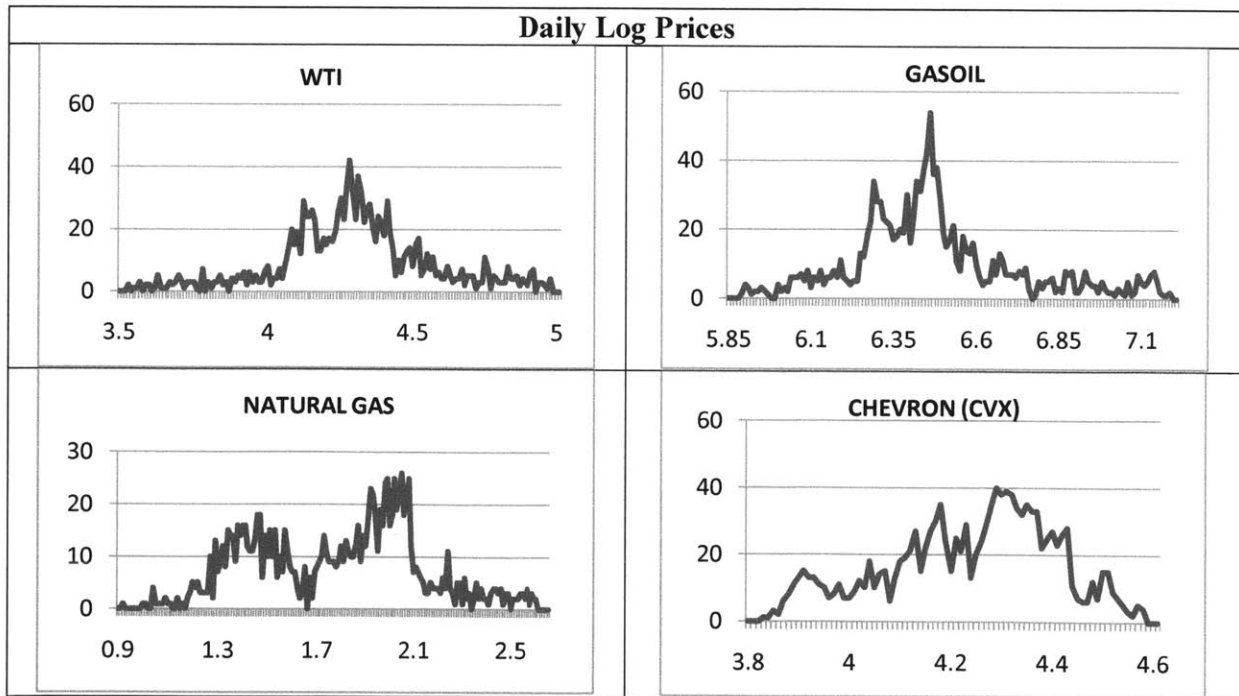


Figure 2: Histograms of front month future (spot) prices of various securities, shown to lack Gaussian characteristic.

basket of securities related to the fossil fuel industry that is both highly correlated and has a relatively stable covariance/correlation structure. To this end, price series were processed a few different ways to see if any method were preferable to the others, and from there it would be determined which data to proceed with. The target correlation range of 70-80% was considered, so the ideal method of processing would result in the largest number of element-wise correlations in that range.

The three types of price series analyzed were 1) daily prices on floating time to maturity (non-interpolated) futures, 2) daily prices on interpolated constant relative tenor futures, and 3) the de-trended daily log returns on the interpolated constant relative tenor futures. The floating time to maturity futures simply considered the price of the front month future contract to be the spot price, the price of the next contract to be the price one month forward, etc. Upon expiration, of the front month, the prior second month contract would be used for spot, and so on. Unlike the daily log returns, the daily log prices were shown in Fig. 2 to not follow a Gaussian distribution.

Thus for each security, daily log prices were obtained from January 3, 2006 to December 31, 2010, corresponding to  $N = 1259$  data points. The matrix of daily log prices has the form

$$\mathbf{X} = [X_1, X_2, \dots, X_i]$$

where the prices of each individual security  $X_i$  have the form

$$X_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

Similarly, the matrix of de-trended returns  $\mathbf{R}$  on the constant relative tenor futures  $f(t, t+\tau)$  has the form

$$\mathbf{R} = [R_1, R_2, \dots, R_i]$$

Where the de-trended returns  $R_i$  for a particular security  $X_i$  are calculated as

$$R_i = \begin{bmatrix} \ln\left(\frac{r_2}{r_1}\right) - \ln\left(\frac{r_{N+1}}{r_1}\right) \\ \ln\left(\frac{r_3}{r_2}\right) - \ln\left(\frac{r_{N+1}}{r_2}\right) \\ \vdots \\ \ln\left(\frac{r_{N+1}}{r_N}\right) - \ln\left(\frac{r_{N+1}}{r_1}\right) \end{bmatrix}$$

The sample means  $\mu_i$  of the prices and de-trended returns are calculated in the same way, where the sample means for the prices of the securities are

$$\mu_{x,i} = \frac{1}{N} \sum_{n=1}^N x_n$$

And the sample means for the de-trended log returns are

$$\mu_{r,i} = \frac{1}{N} \sum_{n=1}^N r_n$$

These sample means are also random variables with mean  $\bar{\mu}$  and variance  $\sigma_{\bar{\mu}}^2$ . This reflects the uncertainty of the sample means in determining the true mean, or global mean. For matrix  $\mathbf{R}$ , since the log returns have been de-trended, the sample means  $\mu_r$  are zero.

## 2.2. Estimation of Global Covariance and Global Correlation Matrices

The sample covariance of the time series  $\mathbf{X}$  and  $\mathbf{R}$  processed above can be calculated as an estimate of the global covariance  $\Sigma$  for the respective series. Element-wise, the covariance between security  $i$  and security  $j$  is defined as

$$\text{cov}(X_j, X_k) = \sigma_{j,k}^2 = \frac{1}{N} \sum_{n=1}^N (x_{j,n} - \mu_j)(x_{k,n} - \mu_k)$$

For the portfolio of fossil fuel securities, the sample covariance matrix is thus calculated as

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{X}_n - \boldsymbol{\mu})^T (\mathbf{X}_n - \boldsymbol{\mu})$$

The sample correlation matrix  $\mathbf{r}$  can be calculated as an estimate of the global correlation matrix  $\boldsymbol{\rho}$ . The element-wise correlation between security  $i$  and security  $j$  are calculated as

$$\rho_{j,k} = \frac{\sigma_{j,k}^2}{\sqrt{\sigma_{j,j}^2 \times \sigma_{k,k}^2}}$$

If the sample correlation matrix is sufficiently stable, it can be regarded as a good estimate of the global correlation matrix and used as such.

This preliminary statistical analysis showed that correlation structure exhibited similar stability as in Fig. 3 for each of the three different aforementioned time series. However, the daily prices on the floating time to maturity (non-interpolated) futures showed the highest correlation to the dollar index, in the 80% range shown in Fig. 4, while the de-trended daily log returns had a correlation near zero. The two equities were somewhat more correlated with the rest of the basket with the prices (in the 40-60% range) as opposed to the returns (25-50%), although the returns of the equities were correlated with each other in the 90% range. The correlation of gasoil prices with the basket of prices was also much higher, in the 95% range as opposed to the correlation of the gasoil returns with the basket of returns, which was in the 65% range. Therefore, the statistical analysis in Section 3 was done using the basket of prices of the non-interpolated futures due to the higher correlated price series, despite the absence of Gaussian behavior.

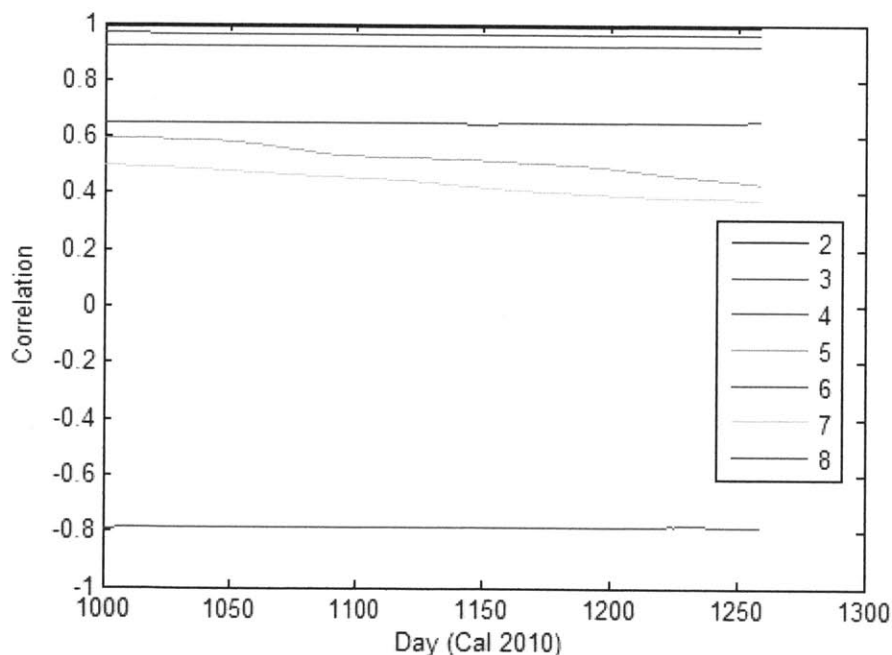


Figure 3: Correlation of daily log price of WTI against the rest of the basket, calendar year 2010. Correlations are recalculated daily over 2010 to show the stability of the correlations in the basket. Brent, RBOB, and Gasoil are highly correlated with WTI (securities 2,3,4), the trade weighted dollar index (TWDI) has correlation around -0.8, while natural gas and the equities Exxon (XOM) and Chevron (CVX) have lower/more moderate correlations (securities 5,7,8).

	WTI	Brent	RBOB	Gasoil	NatGas	TWDI	XOM	CVX
WTI	1	0.9916	0.9201	0.9629	0.432	-0.7838	0.3779	0.6531
Brent	0.9916	1	0.9345	0.9699	0.4241	-0.7995	0.3906	0.6789
RBOB	0.9201	0.9345	1	0.8769	0.4253	-0.6564	0.307	0.5484
Gasoil	0.9629	0.9699	0.8769	1	0.5397	-0.7519	0.4671	0.6988
NatGas	0.432	0.4241	0.4253	0.5397	1	-0.0944	0.5285	0.3096
TWDI	-0.7838	-0.7995	-0.6564	-0.7519	-0.0944	1	-0.472	-0.7731
XOM	0.3779	0.3906	0.307	0.4671	0.5285	-0.472	1	0.7748
CVX	0.6531	0.6789	0.5484	0.6988	0.3096	-0.7731	0.7748	1

**Figure 4: Correlations structure of the basket of fossil fuel securities. WTI, Brent, RBOB and Gasoil are highly correlated with each other, while the trade weighted dollar index shows a high negative correlation to those four fossil fuels. Natural gas and Exxon (XOM) exhibit low correlations with the rest of the basket, while Chevron (CVX) has a more moderate correlation.**

### 3. FORECASTING USING MULTIVARIATE GAUSSIAN STATISTICS

#### 3.1. Extracting Information from the Fossil Fuels Sector

The purpose of this section is to determine if a portfolio of securities representing the fossil fuels sector can be used to forecast the future prices of the securities in the portfolio using multivariate Gaussian statistics. This method requires two assumptions; first, that each of the price series follows a Gaussian distribution, so that the portfolio is multivariate Gaussian distributed; and second, that the portfolio has a stable covariance and correlation structure.

Careful selection of securities for the portfolio is important to ensure that a complete, representative picture of the fossil fuels market can be obtained. The portfolio is constructed to represent a cross-section of influencing factors in the fossil fuel sector, so that any information related to the sector that might affect one particular security first can be reflected properly in forecasts for the other securities. Securities chosen for the portfolio should ideally have correlations at least in the 70-80% range to ensure robust forecasts.

For the present analysis, eight securities were chosen, including various correlated commodity futures, equities of large diversified fossil fuel producers, and the US Dollar. Front month futures prices were used to represent their respective commodities (WTI Crude, Brent Crude, RBOB Gasoline, Gasoil, and natural gas). The Trade Weighted Dollar Index published by the Federal Reserve was used to represent the dollar, and the NYSE listed stock for Exxon Mobil and Chevron were used to represent the fossil fuel sector's exposure in the equity markets.

In Section 2.1, the daily log returns of the securities in the basket were shown to be Gaussian, although the daily log prices of those securities were not. In Section 2.2, the correlation structure of this basket was examined, appearing to be stable over the forecasting period of calendar year 2010. Thus the two conditions for the use of this analysis appear to be fulfilled for the daily log returns of the basket, but not for the daily log prices. However, this analysis attempts to apply the following statistics to the basket of log prices due to its preferable correlation structure.

#### 3.2. Conditional Mean and Conditional Covariance

According to Anderson (2003), the conditional mean and variance of a random variable may be different than the unconditional mean and variance. The conditional means and variances can be

calculated from the unconditional sample means and a stable sample covariance matrix, estimated from previous realizations of a collection of Gaussian random variables.

For a matrix of securities  $\mathbf{X}$ , the matrix can be split into two component block matrices

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}$$

Where,

$$\mathbf{X}^{(1)} = [X_1] = [x_{1,1}, x_{1,2}, \dots, x_{1,n-1}] = [x_1^{(1)}, x_2^{(1)}, \dots, x_{n-1}^{(1)}]$$

$$\mathbf{X}^{(2)} = \begin{bmatrix} X_2 \\ \vdots \\ X_i \end{bmatrix} = \begin{bmatrix} x_{2,1} & x_{2,2} & \dots & x_{2,n-1} \\ \vdots & \vdots & & \vdots \\ x_{i,1} & x_{i,2} & \dots & x_{i,n-1} \end{bmatrix} = [x_1^{(2)}, x_2^{(2)}, \dots, x_{n-1}^{(2)}]$$

The columns  $x_{i,1}^{(2)} \dots x_{i,n-1}^{(2)}$  are thus the known prices of security  $X_i$  in the portfolio for times  $t = 1 \dots n-1$ . The covariance matrix of  $\mathbf{X}$  can then be split into four component blocks

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Where,

$$\Sigma_{11} = [\sigma_{11}^2]$$

$$\Sigma_{12} = \Sigma_{21}^T = [\sigma_{12}^2 \quad \dots \quad \sigma_{1i}^2]$$

$$\Sigma_{22} = \begin{bmatrix} \sigma_{22}^2 & \dots & \sigma_{2i}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{i2}^2 & \dots & \sigma_{ii}^2 \end{bmatrix}$$

The unconditional mean of  $\mathbf{X}$  can also be split into two blocks

$$\boldsymbol{\mu} = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}, \text{ where } \boldsymbol{\mu}^{(1)} = [\mu_1] \text{ and } \boldsymbol{\mu}^{(2)} = \begin{bmatrix} \mu_2 \\ \vdots \\ \mu_i \end{bmatrix}$$

where  $\mu_i$  is the sample mean of security  $X_i$ , calculated above. The portfolio has been split into segments where  $\mathbf{X}^{(1)}$  is the security to be forecasted, while  $\mathbf{X}^{(2)}$  is the remaining securities in the

portfolio which will be used as the basis of the forecast. Unconditionally then,  $x_n^{(1)}$  will be the realization at  $t = n$  of random variable  $\mathbf{X}^{(1)}$ , having mean and variance  $\mu_1$  and  $\sigma_{11}^2$ . If values of highly correlated securities in  $x_n^{(2)}$  are known however, they could imply dramatically different probability distribution for  $x_n^{(1)}$ , characterized by the conditional mean and conditional variance/covariance. The conditional mean then, of  $\mathbf{X}^{(1)}$  at time  $n$ ,  $\mu_{X^{(1)}|X^{(2)}=x^{(2)}}^{(1)}$  is thus defined as

$$\mu_{X^{(1)}|X^{(2)}=x_n^{(2)}} = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x_n^{(2)} - \mu^{(2)})$$

While the conditional covariance matrix of  $\mathbf{X}^{(1)}$  at time  $n$  is defined as

$$\Sigma_{X^{(1)}|X^{(2)}=x^{(2)}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

The implication here is that, given a steady covariance matrix and price data of a portion of the portfolio at time  $t = n$ , the mean prices of the rest of the basket at  $t = n$  may be predicted with more certainty (lower variance) than before.

The predictions of this model essentially determine what the price of one particular security  $X_t$  should be in relation to the prices of the other securities. Potentially, this model could serve as a basis for some sort of statistical arbitrage trading strategy. This would rely heavily on reducing the error of the predictions, the optimization of the portfolio by including securities with ideal correlation levels that provide relevant information to the fossil fuel sector, and the persistence of a steady covariance/correlation structure. Given a sufficiently accurate model, price data for the portfolio on the order of milliseconds/microseconds would likely be necessary to exploit any arbitrage opportunities.

### 3.3. Results

For calendar year 2010, the prediction model was run for each of the eight securities in the portfolio. The price of the particular security to be predicted  $X_1$  was calculated every day. The covariance matrix, although determined to be steady in Section 2.2., was recalculated every day to reflect small fluctuations that might occur. In practice, daily recalculation of the correlation matrix would be useful in identifying a possible regime change, if it were to occur.



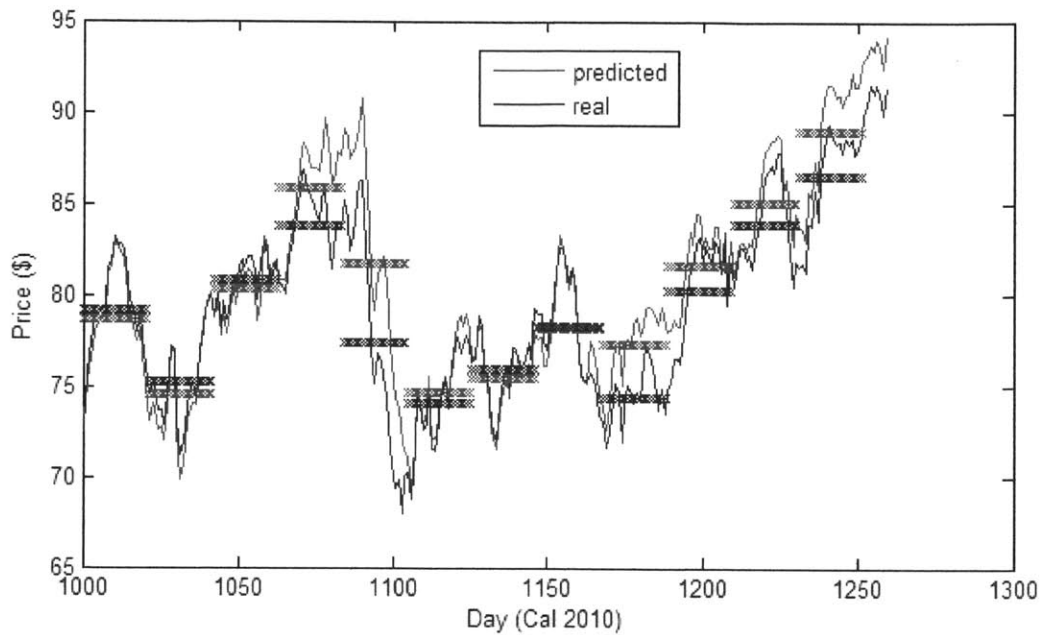


Figure 5: Daily price projections for WTI. Daily prices for WTI were projected based on the prices of the rest of the basket, using the conditional statistics outlined in Section 3.2. The thick red and blue lines represent the average of the projected prices and the actual prices, respectively. The closer together the thick lines are, the more accurate the projections were for a given month.

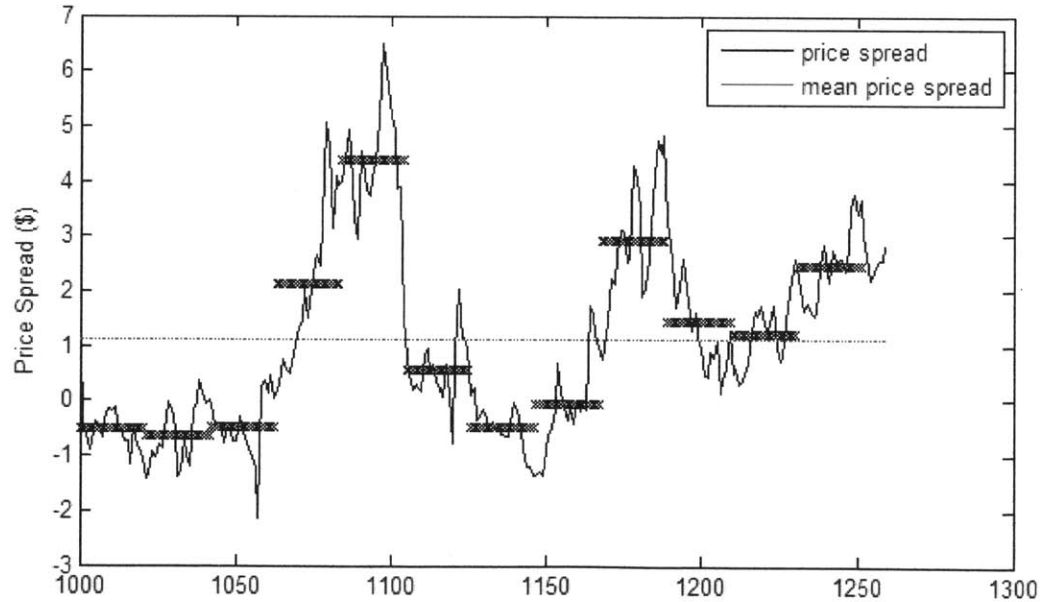
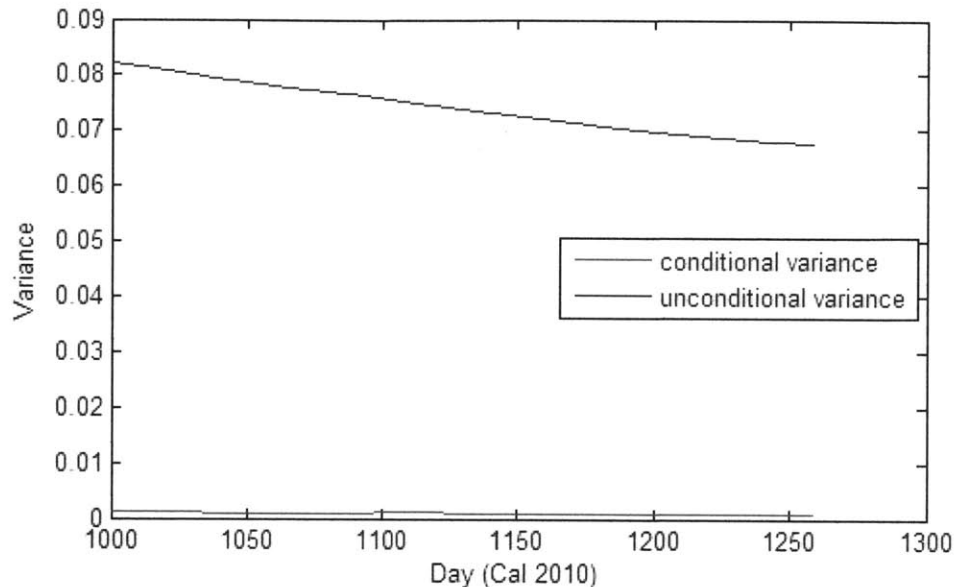


Figure 6: Price spread between daily projections and actual prices for WTI, calendar year 2010. Monthly mean spreads are shown by the short thick lines. Negative value indicates underestimate, while positive indicates overestimate. The prediction model overestimates the price by slightly more than one dollar over calendar year 2010.

The best results were obtained from predictions of WTI and Brent, likely due to the high correlation they have with each other. Results for WTI are shown in Figs. 5, 6, and 7. Good results were also obtained for Gasoil and RBOB, the latter shown in Fig.8. As the correlations of the particular security of interest decreased, so did the quality of the projections. As seen in Fig. 8, the daily projections for natural gas (correlations in the range of 40-50%) and Chevron (in the range of 50-65%) came with noticeable price spreads. While the general trend and slope



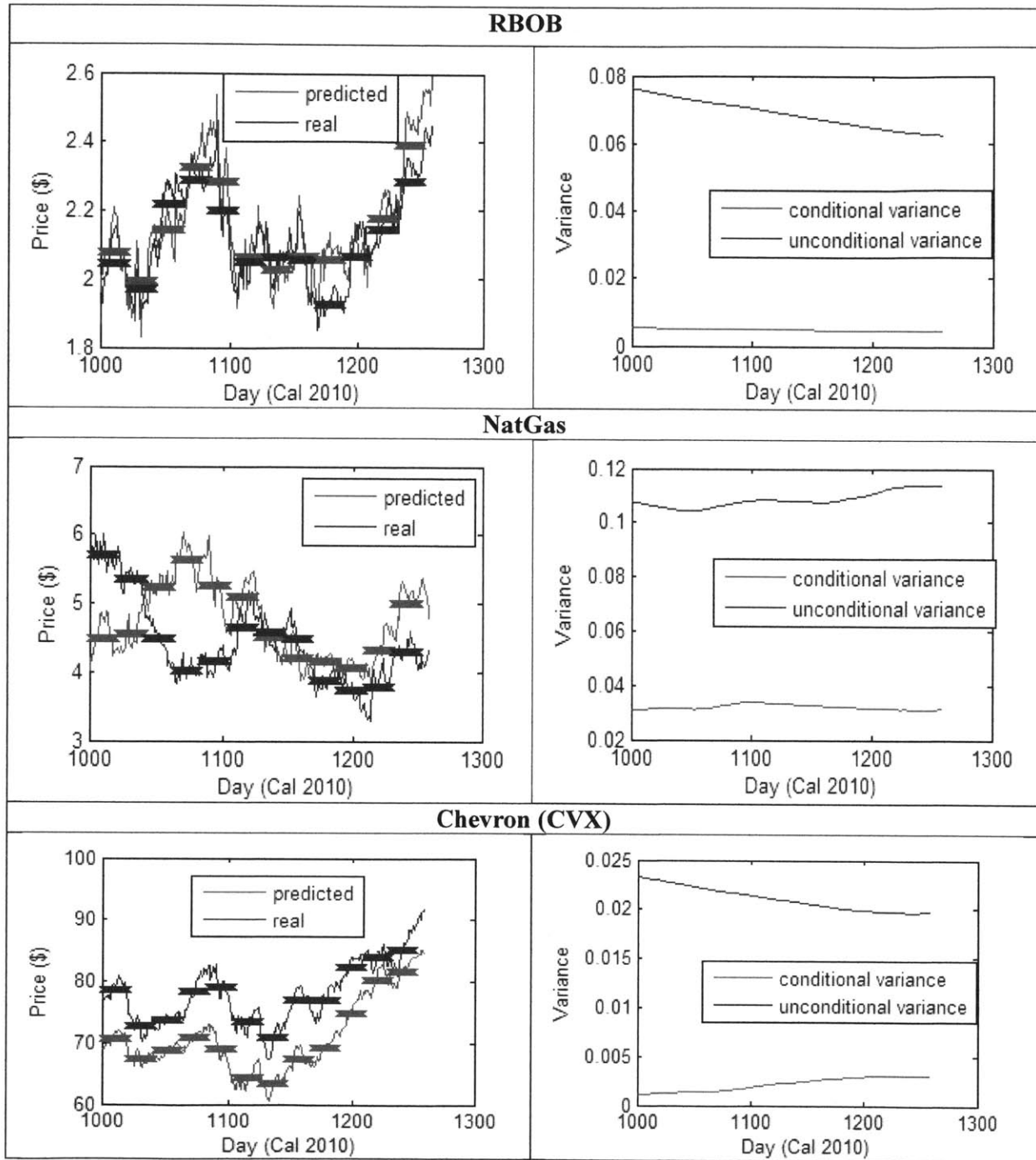
**Figure 7: Conditional versus unconditional variance for WTI daily prices, calendar year 2010. Using conditional statistics outlined in Section 3.2, the variance of the daily prices of WTI is decreased by two orders of magnitude.**

of the price series was still well represented in Chevron’s case, it appeared in the natural gas case that the projections predicted the exact opposite market response of what actually happened for the first 100 days of calendar year 2010.

This does not necessarily come as a surprise. The natural gas market is known to be lowly correlated with the markets of the other fossil fuels in the basket. Meanwhile, the Exxon and Chevron are exposed to the risks of the broader equities market as well as individual company risk that may have little to do with the fundamentals of the fossil fuels market.

The logical conclusion is that natural gas and the two equities, were outside of the ideal correlation range to include in a basket used for this type of statistical forecasting. Removing these three lower/moderately correlated securities may even yield more accurate forecasts for the remaining securities. Since the two equities used here were two of the largest, most stable diversified oil and gas companies, it seems that including individual equities in such a basket in general may be inappropriate, due to the likelihood that uncorrelated company risk would be introduced. However, exposure to equities in the fossil fuel industry could be obtained by a large,

liquid ETF that tracks a sizeable group of companies in the industry. This would reduce the company risk through diversification, possibly enough to obtain a stable correlation in the 80% range.



**Figure 8: Conditional price estimates for RBOB, Natural Gas, and Chevron (CVX), calendar year 2010. The conditional estimates for RBOB are on par with those of WTI, while large price spreads occur in the case of natural gas and Chevron (CVX). The lower correlations of the latter two with the rest of the basket are likely the source of the poorer estimates. The conditional variance of the latter two is also larger with respect to their unconditional variances.**

## 4. MODELING THE PORTFOLIO

### 4.1. Principal Component Analysis

Given a stable sample covariance matrix  $\mathbf{S}$  of a portfolio of highly correlated securities, a Principal Component Analysis (PCA) can be executed. The PCA begins with a Singular Value Decomposition of the sample covariance matrix  $\mathbf{S}$ , where  $\mathbf{U}$  is the matrix of eigenvectors, and  $\mathbf{\Lambda}$  is the diagonal matrix of eigenvalues. In the context of PCA, the eigenvectors are the factor weights, while the eigenvalues are the factor volatilities.

$$\Sigma = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

The factor volatilities show how many dominant factors  $d$  there are in the particular portfolio. Usually, the factor volatilities die down rather quickly, so that only a few  $d \ll i$  factors are required to explain most of the volatility of the portfolio. The more highly the securities are mutually correlated, the faster the factor volatilities will die down and there will be fewer dominant factors. A portfolio of lowly correlated securities will require a larger number of factors to explain the volatility of the portfolio, and the results of the PCA become less interesting.

Each factor is a linear combination of the securities in the portfolio, with the weights of each security  $k$  in factor  $j$  given by  $u_{jk}$  where

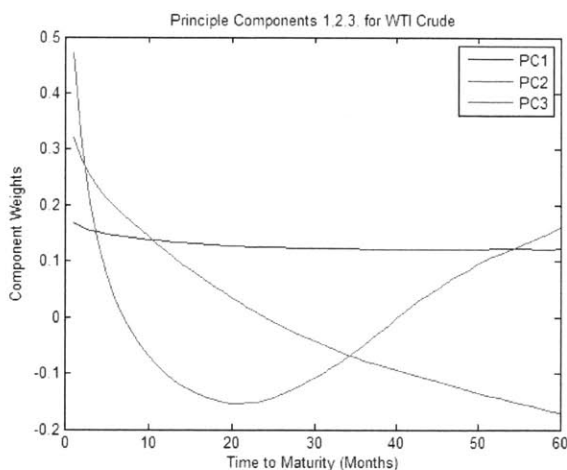
$$u_j = [u_{j1} u_{j2}, \dots, u_{ji}]$$

$$\mathbf{U} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_j \end{bmatrix}$$

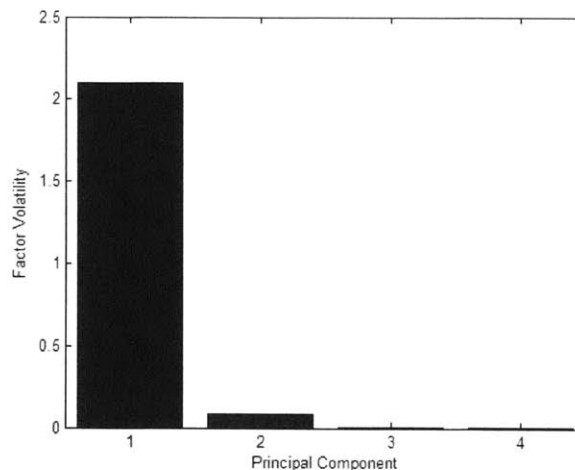
with  $i$  still as the  $i$ th security in the portfolio. Moreover, SVD decomposes the covariance matrix in such a way that the resulting eigenvectors are independent. This also means that their correlations will be zero. PCA is therefore useful in terms of determining the factors that govern the movement of the portfolio, but applying the analysis in Section 3.2 to the factors will not yield any useful information.

## 4.2. Modeling the Crude Oil Futures Curve

Since the best results of the statistical forecasting done in Section 3 were done with the log prices, factors determined from a basket of log prices of the portfolio of securities would be a useful mode of comparison. Previous PCA analyses such as that in Sclavounos and Ellefsen (2009) and Ellefsen (2010) were done on the de-trended log returns on constant relative tenor futures contracts interpolated from the traded fixed-tenor futures. PCA was done on just the crude oil market using the log prices of traded fixed-tenor crude oil futures with tenors of  $t = 1$  to 60 months. The results of the PCA are shown in Figs. 9 and 10 to demonstrate extremely similar behavior to the PCA done in Sclavounos and Ellefsen (2009) and Ellefsen (2010), giving credence to the possibility that PCA can analogously be done on a portfolio of securities using log prices of the securities, so long as accurately modeling the volatility of the portfolio is not the primary concern.



**Figure 9: First three factors for WTI Crude. Factor analysis done on WTI futures to tenors of 60 months. PCA done with log prices of those contracts seems to replicate of that in Sclavounos and Ellefsen (2009).**



**Figure 10: Factor Volatilities for WTI crude. Factor volatilities decrease exponentially, allowing the first few factors to explain the movements of the entire curve.**

## 4.3. Extension of PCA onto a Portfolio of Correlated Commodities

PCA was done on a portfolio of energy securities that consisted of the trade weighted dollar index (TWDI) and futures on three energy commodities: WTI Crude Oil (WTI), Brent Crude Oil (BR), RBOB Gasoline (RB). Five futures were included for each commodity, giving a total of 16 price series for the portfolio. The same maturities were used for each commodity future, the maturities being  $t = 1, 3, 6, 9,$  and 12 months, due to the limited availability of prices for the farther dated RBOB contracts from 2006 to mid 2007.

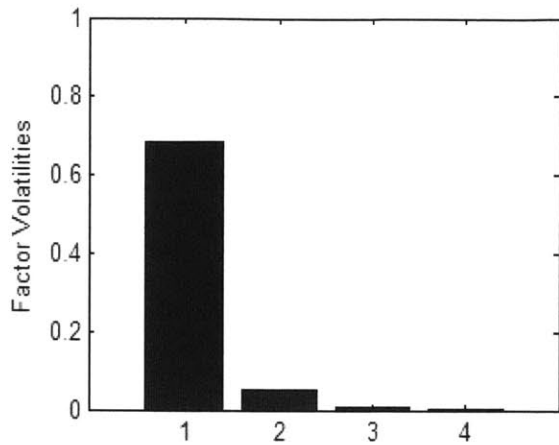


Figure 11: Volatility of first four factors of the commodity futures portfolio. The volatilities of the factors decrease exponentially as a larger percentage of the volatility of the portfolio is explained.

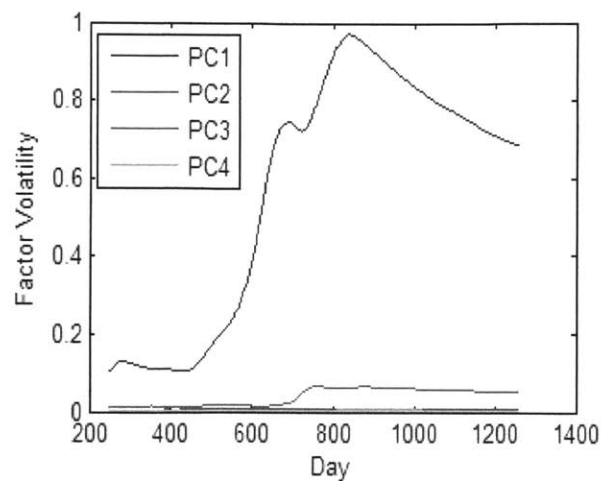


Figure 12: Stability of factor volatilities from 2007 to 2010. The large increase in the volatility of commodity markets in 2008 reflected in increased volatility of the first and second factors over the same time period. Day 250 corresponds to 1/3/2007, with day 1259 corresponding to 12/31/2010.

It appears that this particular portfolio was governed by four dominant factors, the volatilities of which can be seen in Figs. 11 and 12. Each factor also gave rise to a sensible interpretation, which could be deduced from the eigenfunctions in Fig. 13. The eigenfunctions are broken up into sections that represent the futures of one particular commodity, with the first curve from the left representing the five WTI futures, the second curve representing the five Brent futures, the third curve representing the five RBOB futures, and the last, disconnected point on the right, the price of the trade weighted dollar index.

The first, most dominant factor appeared to explain the up/down shift of the three commodity curves in tandem, negatively correlated with the dollar. From its eigenfunction (the first of the four graphs in Fig. 13) the weights of each of the three curves are positive while the weight on the dollar index is negative, supporting that conclusion. The higher loadings given to the front-month contract makes intuitive sense, since it is known that the volatility of futures contracts increase as they approach expiration. Fig. 14 plots the price of the first factor from 2006 to 2010, which appears to further support the interpretation of the first factor.

The second factor, as evidence from its eigenfunction in Fig. 13, appeared to explain the spread between the WTI futures curve and the Brent and RBOB curves, with the loadings of the former being positive and the latter two being negative. Fig. 15 supports this conclusion, with the price of the second factor looking starkly similar to the spot WTI-Brent spread.

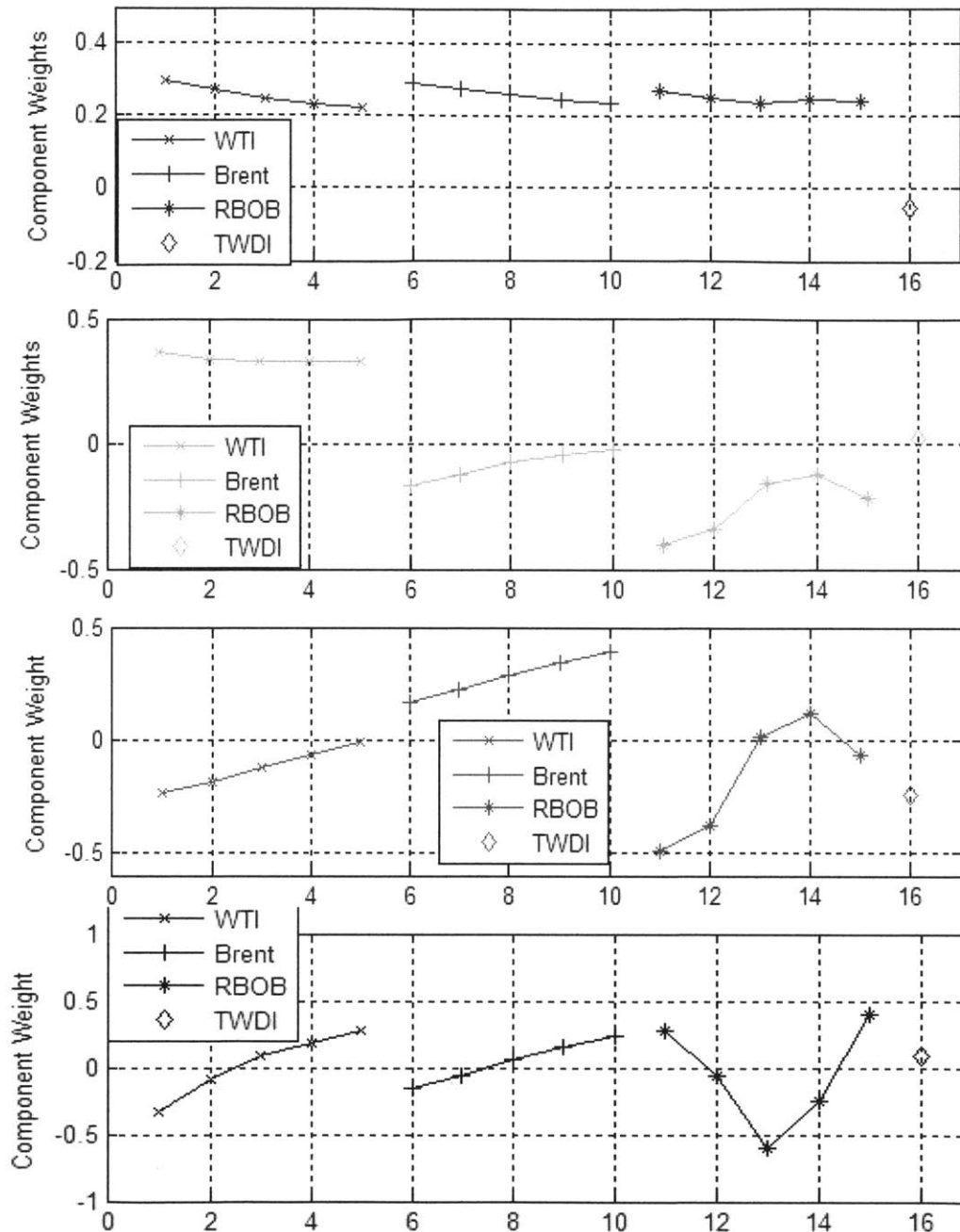


Figure 13: First four eigenfunctions of the commodity portfolio. The eigenfunctions, from the top, show the shape of factors 1 to 4. The first factor reflects a uniform up/down shift of the three commodity curves in unison. The second factor shows the spread between the WTI curve and the Brent and RBOB curves. The third factor shows the spread between the Brent curve and the WTI and RBOB curves. The fourth factor shows the tilt of the WTI and Brent curves towards either contango or backwardation, while affecting the seasonality of the RBOB curve.

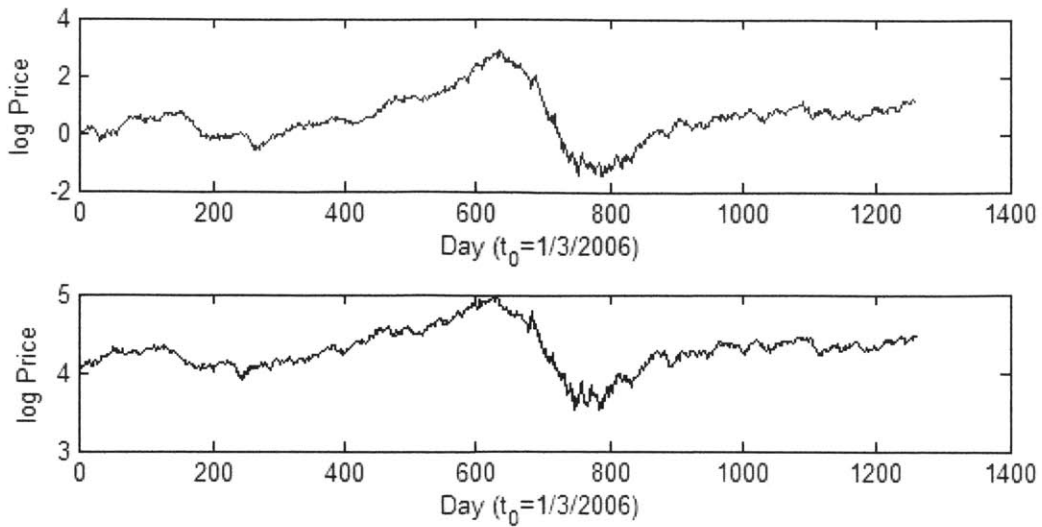


The third factor appears to explain the spread between the Brent futures curve and the RBOB and WTI curves, with the loadings of the former being positive and the latter two being negative. However, since much of the WTI-Brent relationship should have been explained in the second factor, this factor should largely be dominated by the Brent-RBOB relationship. This assertion is supported by Fig. 16, which shows the similarities of the price of the third factor and the spot Brent-RBOB spread.

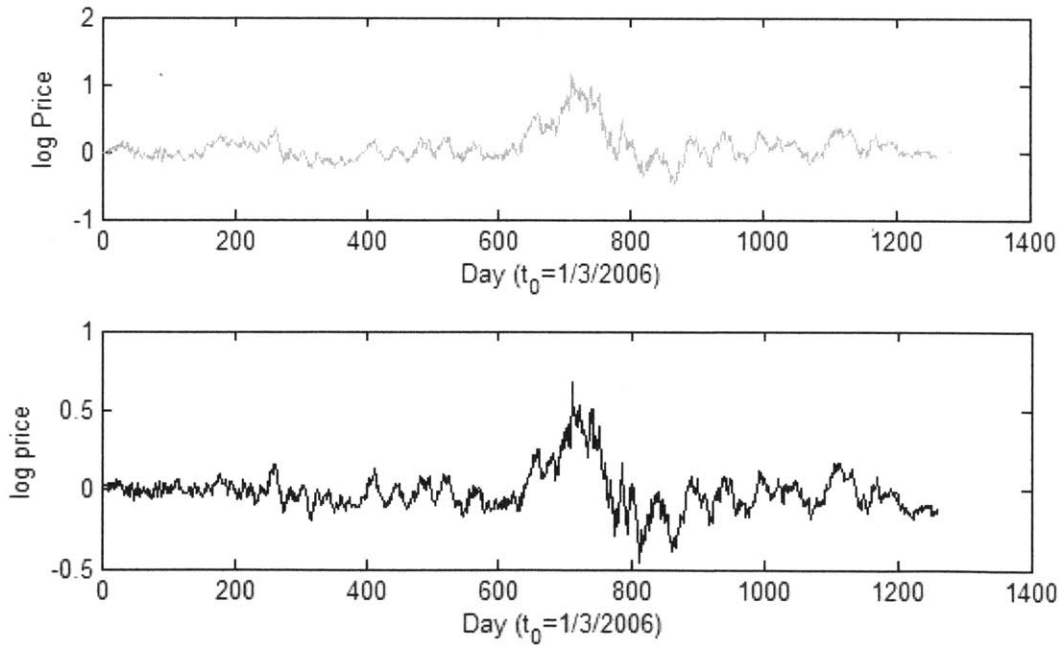
The fourth factor, from interpreting the eigenfunction, appears to represent the shift of the WTI and Brent futures curves into contango and backwardation, and some sort of seasonality with the RBOB curve. From Fig. 17, the seasonality of the RBOB curve in the price of the fourth factor is evident. Since the portfolio only includes futures with tenors up to 12 months, it seems justified that the effect of the tilt in the WTI and Brent curves on the price of the fourth factor may be small in comparison to the seasonality of the RBOB. However, if the portfolio were to include futures with much larger tenors, out to 24 or 36 months or more, the effect of the tilt in the price of the fourth factor may be more apparent.

Looking at the PCA done on this particular portfolio as a whole, the apparent dependence of the first three factors on spot price spreads could have been foreseen, due to the exclusive use of near dated futures in the portfolio with relative tenors of 12 months or less. These futures have a high correlation to the front month contract (95-99%) and thus would fluctuate similarly to spot. The tilt of the Brent and WTI futures curves, supposedly reflected in the fourth factor, is decidedly muted due to the exclusive use of near dated contracts as well. As seen in Fig. 9, referring to the analysis done in Section 4.2, the pivot point of the second factor of the WTI futures curve is around 20 months. In the very least, to fully capture the tilt effect explained by the second factor in Section 4.2, futures with tenors greater than 24 months would need to be included in the portfolio.

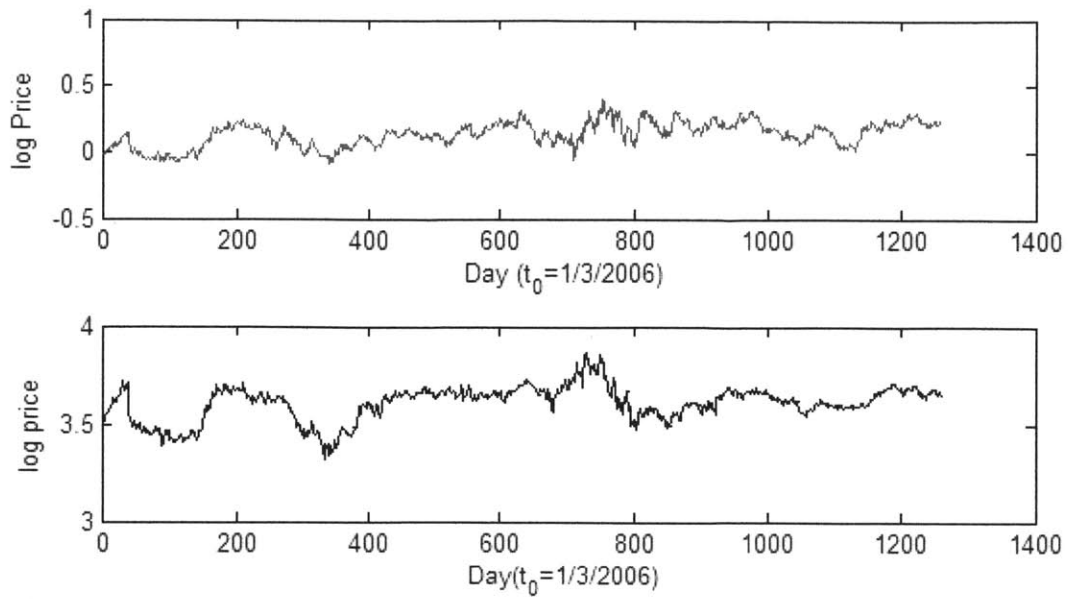
By including farther-dated futures in the portfolio, the first three factors may change in character as well. The prices of the three factors would likely be less characteristic of the spot price spreads in Figs. 15 and 16, as the farther-dated contract would be less correlated with spot. New factors could also emerge to explain the movements of the expanded portfolio.



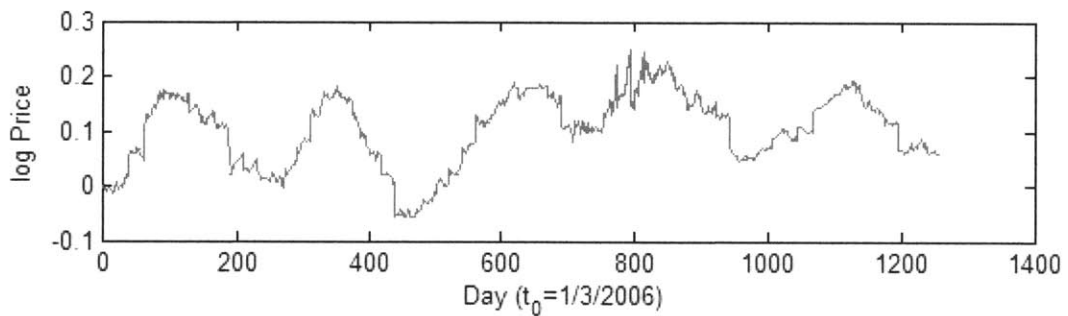
**Figure 14: Price of Factor 1 above, with spot price of WTI below. The price of the first factor appears to represent the unified upward/downward shift of the commodity portfolio.**



**Figure 15: Price of Factor 2 above, with spot WTI-Brent spread below. The second factor appears to represent the spread between the WTI and Brent futures curves.**



**Figure 16: Price of Factor 3 above, with spot Brent-RBOB spread below. The third factor appears to represent the spread between the Brent and RBOB futures curves.**



**Figure 17: Price of Factor 4. The fourth factor appears to characterize the seasonality of the RBOB curve, with peaks in the summers and lows in the winters.**

## 5. FORECASTING USING AUTOREGRESSIVE METHODS

### 5.1 Autoregressive Methods in Time Series Analysis

In this section, autoregressive methods are used in an attempt to determine the forecastability of both the factors and the commodity futures. It is hypothesized that the factors, as linear combinations of the futures in the portfolio, may contain more forecastable information than the factors themselves. Both the factors and the commodity futures will be fitted with two types of models to determine if either price series can be forecasted with any level of confidence.

Autoregression is a widely used technique to forecasting time series data sourced from a variety of disciplines including engineering, signal processing, and finance. While regression in essence minimizes the mean squared error by providing a formula for the values of a dependent variable  $z$  as a function of an independent variable  $x$ , autoregression minimizes the mean squared error by providing a formula of  $z$  at time  $t$  as a function of previous values of  $z$ , from  $z_{t-1}$  to  $z_{t-k}$  where  $k$  is the indicated lag, or order of the model.

The purpose of using this technique is to determine the degree to which the principal components, or factors, found in the previous section can be forecasted, and whether they can be forecasted more accurately than the actual securities themselves.

Given a random variable  $Z$ , discrete, successive observations of  $Z$  at time  $t$ , in constant intervals, take the form

$$\mathbf{Z} = z_1, z_2, \dots, z_N$$

And are regarded as a time series. Observations of  $\mathbf{Z}$  at times  $t, t-1, t-2$ , are denoted  $z_t, z_{t-1}, z_{t-2}$ , etc. The deviation of the observations from the mean  $\mu$  are denoted

$$\tilde{z}_t = z_t - \mu$$

Here, it is assumed that  $\mathbf{Z}$  is a stationary process that fluctuates about a constant mean  $\mu$  with constant variance  $\sigma_z^2$ .

The autocorrelation function is a useful tool in describing a stationary stochastic time series. The autocorrelation function determines the degree of correlation between an observation  $z_t$  and another observation at lag  $k, z_{t-k}$ . Under the strict stationarity condition, the joint probability

distribution between  $z_t$  and  $z_{t-k}$ , no matter the value of  $t$ , are identical. Thus given a stationary time series, the autocovariance at lag  $k$ ,  $\gamma_k$ , and the autocorrelation at lag  $k$ ,  $\rho_k$ , are calculated as

$$\gamma_k = E[(z_t - \mu)(z_{t-k} - \mu)]$$

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{E[(z_t - \mu)(z_{t-k} - \mu)]}{\sigma_z^2}$$

## 5.2. AR and ARMA Models

The two models used to fit the price series are the Autoregressive (AR) Model and the Autoregressive Moving Average (ARMA) Model. This section outlines the forms of both these models.

### *Autoregressive (AR) Model*

The autoregressive model of order  $p$  has the form

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \dots + \phi_p \tilde{z}_{t-p} + a_t$$

Where, as before, the terms  $\tilde{z}_t, \tilde{z}_{t-1}$ , are deviations from  $\mu$  at time  $t$ , and the terms  $\phi_1, \phi_2 \dots \phi_p$  are regression constants. The term  $a_t$  is a Gaussian white noise process that accounts for the error between the forecasted  $\tilde{z}_t$  and the actual  $\tilde{z}_t$ .

The constants  $\phi_1 \dots \phi_p$  were found by using the AR() function in MATLAB, which computes the matrix of constants  $\phi$  as the solution to the Yule-Walker equations.

### *Autoregressive Moving Average (ARMA) Model*

The autoregressive moving average (ARMA) model is a combination of the autoregressive model of order  $p$  and the moving average model of order  $q$ . The ARMA model of order  $p, q$ , has the form

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \dots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}$$

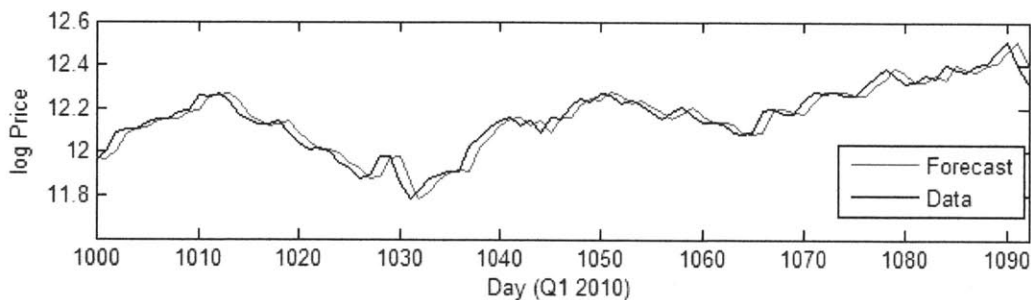
where the phi terms and theta terms are regression constants. The matrix of autoregression constants  $\phi$  and the matrix of moving average constants  $\theta$  were found by using the ARMAX()

function in MATLAB. The  $a_t \dots a_{t-q}$  terms each represent realizations of a Gaussian white noise process at times  $t \dots t-q$ .

### 5.3. Comparing Forecasts of Factors and Futures

The price series for the first four factors, as well as that for the front month futures for WTI, Brent, and RBOB, were each fitted with autoregressive and autoregressive moving-average models. However, neither the factors nor the commodity futures themselves could be forecasted to any degree of confidence. This could be explained by examining the matrix of autoregressive constants  $\phi$ . For both the AR and ARMA models,  $\phi_1$  was consistently in the range of 0.97-0.99, while constants for lags  $k$  greater than 1 were consistently on the order of 0.03 or less. Prices at time  $t$  were essentially unaffected by prices at lags  $k$  greater than 1, while were heavily dependent on  $z_{t-1}$ . The resulting fit therefore looked identical to the input prices with the values of the fit shifted over +1 day.

Thus, both models seemed to imply that the time series were all first order autoregressive processes, or Markov processes, supporting the theory of the random walk of securities prices. The conclusion that these price series were first order autoregressive processes was supported by the partial autocorrelation functions, shown in Fig. 19, which indicated values of  $\phi_{kk}$  to be essentially zero at lags  $k$  greater than 1.



**Figure 18: Autoregressive Model (order  $p = 6$ ) of price of the first factor, first quarter (Q1) 2010. Here, extra terms  $\phi_2 \dots \phi_6$  have been included to demonstrate that the model is still essentially a first order process. The forecasted curve appears very similar to the actual data, just lagging by one day. Identical results were obtained for the other factors as well as the commodity futures.**

A possible explanation for this is that the price series used in the autoregression, referring to both the factors and the commodity futures, were not stationary processes. As seen in Fig. 12, the volatilities of the factors and the commodity futures vary significantly from 2006 to 2010, while Fig. 5 shows that WTI certainly does not fluctuate about a constant mean, along with the rest of the commodity futures. Thus it could be said that the AR and ARMA models were

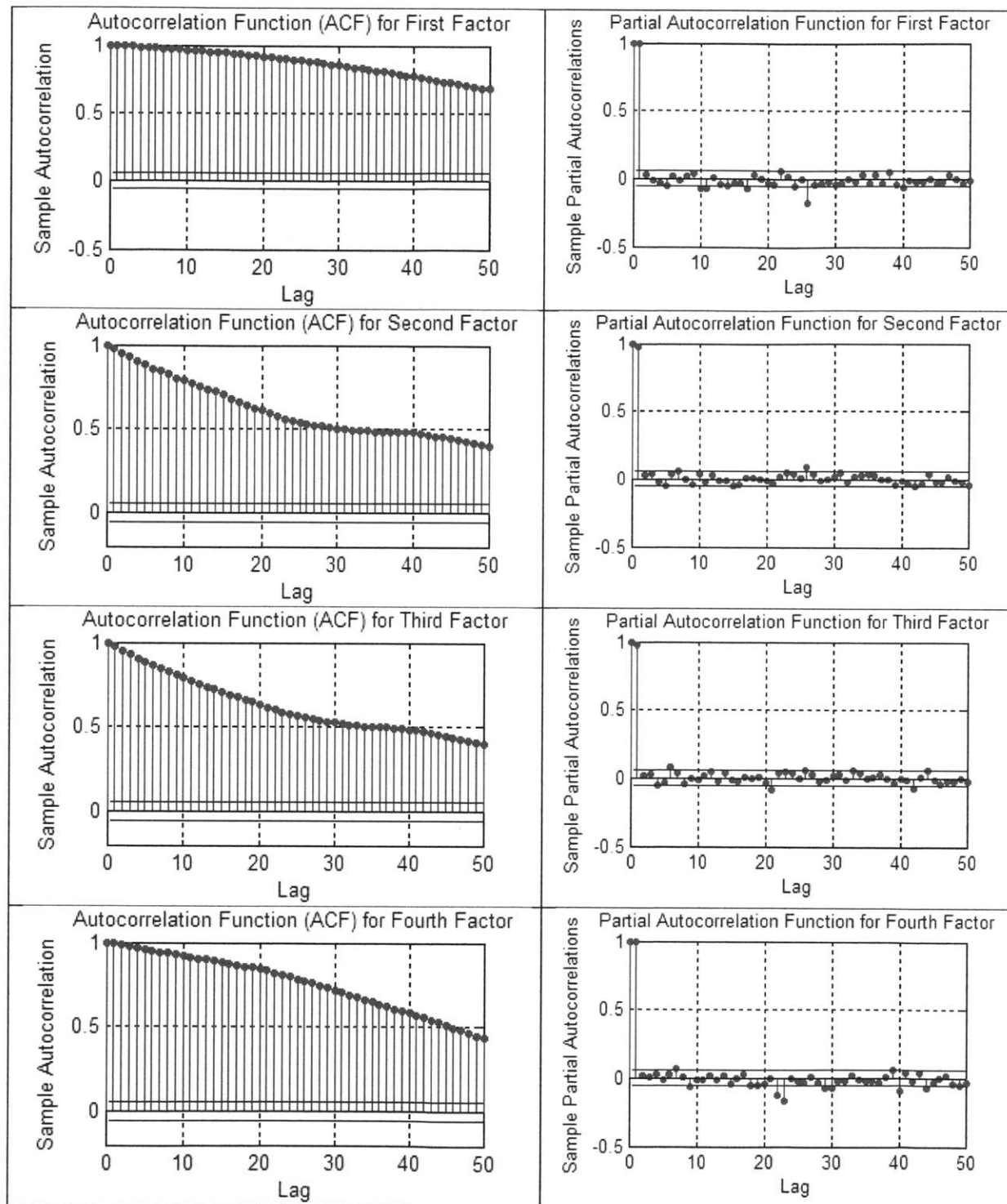
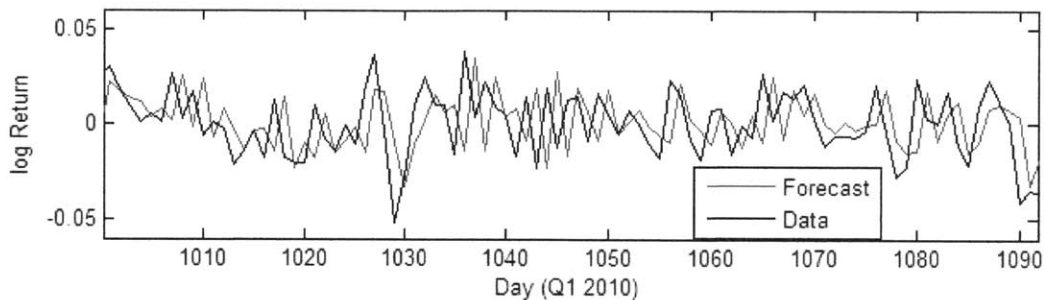


Figure 19: Autocorrelation and Partial Autocorrelation functions of the four factors. Autocorrelation functions show high autocorrelation of  $z_t$  to  $z_{t-k}$  even through lag  $k = 50$  days. Partial autocorrelation functions show that the factors are first order autoregressive processes, or Markov processes.

wrongly applied, and that the characterization of the factors and the futures as a first order autoregressive process was inaccurate.

However, applying the autoregression models to the de-trended log returns on the constant relative tenor futures from Section 2.1. yielded nearly identical results, shown in Fig. 20. These de-trended log returns have constant mean of 0 and much more stable variance, so that they are effectively stationary. The autoregression constant  $\phi_1$ , like before, was consistently in the range of 0.97 to 0.99. As a result, the forecast curve was largely identical to the actual data, but shifted over one time interval, again like before. The results of this application support the idea that although the price series used at first may not have been stationary, the characterization of the factors and commodity futures as Markov processes was still correct.



**Figure 20: ARMA Model of de-trended daily log returns on WTI,  $p = 6$ ,  $q=1$ , first quarter (Q1) 2010. The results of the regression on a stationary process are about the same as for the non-stationary price series. This supports the conclusion that the factors and the futures are first order autoregressive processes.**

In furtherance of this research, it would be interesting to see how effective a non-stationary model such as the Autoregressive Integrated Moving Average (ARIMA) model is in forecasting the factors and the commodity futures, whether this model still supports the conclusion that both the factors and the futures are first order autoregressive processes, and whether or not one can be forecasted more than the other.



## 6. CONCLUSIONS

### 6.1. Summary of Results

In Section 2, a basket of eight securities were compiled to offer a complete, cross-sectional representation of the fossil fuels industry. The securities used were the front month future prices of five fossil fuels commodities (WTI crude, Brent Crude, RBOB Gasoline, gasoil, and natural gas), the trade weighted dollar index, and the equities of two diversified oil and gas companies Exxon Mobil and Chevron. Forecasts of the daily prices of each security over calendar year 2010 based on the prices of the remaining securities in the basket were conducted using the conditional statistics presented in Section 3.2. The accuracy of the forecasts was discussed in Section 3.3. Projected prices for WTI averaged within a dollar over the course of calendar year 2010, while projections for RBOB averaged within 3 cents per gallon. While the projections for the four crude oil/crude oil product contracts were quite reasonable, the projections for natural gas and the two equities diverged significantly from actual data, likely due to insufficient correlation with the rest of the basket. The projections for the first four contracts could potentially be optimized by eliminating the lesser correlated securities and adding other, higher correlated securities related to the fossil fuels market.

The factor analysis done on the crude oil futures curve in Scлавounos and Ellefsen (2009) using de-trended daily log returns on constant relative tenor futures was repeated in Section 4.2. with simple log prices on non-interpolated futures, yielding eigenfunctions for the first three factors that shared the same characteristics. This gave credence to the possibility that factor analysis in general could be done using simple log prices of non-interpolated futures, so long as accurately modeling volatility was not a priority. This factor analysis was extended to a portfolio of three commodities (WTI, Brent, and RBOB), each with five futures of maturities 12 months or less, in addition to the trade weighted dollar index, yielding a total of sixteen securities. Four primary factors were found to govern the price process of that particular portfolio. Since the portfolio only contained near dated futures contracts correlated 95% or greater with spot, the four factors seemed to be dominated by movements in spot prices and spot price spreads. These factors may or may not have similar characteristics if father dated futures were included.

Autoregressive forecasting methods were applied to the price series of both the factors found in Section 4.3 as well as the front month futures prices of WTI, Brent, and RBOB, with each series being fitted with autoregressive (AR) and autoregressive moving average (ARMA) models. Due to high autocorrelations at lag  $k = 1$ , the autoregressive constant  $\phi_1$  was consistently in the range of 0.97 to 0.99, resulting in a forecast curve that was identical to the actual realized data, but shifted over one time interval. AR and ARMA models were fitted to de-trended daily log returns

of both the factors and the futures and yielded the same results, implying that the price series are all Markov processes.

## **6.2. Suggestions for Future Research**

The optimization of the basket in Section 3 for the purpose of forecasting using the conditional Gaussian statistics could be an interesting problem to pursue. Studies could be done into the benefits of removing lowly correlated securities and inserting more highly correlated securities. The ideal number of securities to include into such a basket poses another question, along with the possibility that there is some correlation horizon above/below which adding a particular security to the basket will improve/degrade the projections. A comparison of the accuracy and dependability of forecasts obtained from the non-Gaussian log prices of a basket and the Gaussian log returns of a basket would be of interest as well.

Section 4.3. found four significant factors that governed the movement of a commodity futures portfolio. However, that portfolio did not include any futures with tenors larger than 12 months. Inclusion of back month contracts into the portfolio could result in factors with decidedly different characteristics, which could be cause for further study.

Although the conclusions of Section 5 was that the price series of both the factors and the futures were Markov processes, fitting the price series with a non-stationary autoregressive integrated moving average (ARIMA) model could yield different results. If that is the case, and either the factors or futures (or both) are seen to not be Markov processes, this type of forecasting method could potentially yield tradable information.

## 7. ACKNOWLEDGEMENTS

The author would like to thank Professor Paul Sclavounos for his guidance throughout the research process. Per Ellefsen and Nicolas Hadjiyiannis were also very willing to provide access to personal copies of their theses when they had yet to be published by MIT Libraries.

## 8. REFERENCES

- Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 3<sup>rd</sup> ed.
- Box, G.E.P, Jenkins, G.M., and Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall, 3<sup>rd</sup> ed.
- Downey, Morgan (2009). *Oil 101*. Prentice-Hall.
- Efron, B., and Morris, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, Vol. 70, No. 350, (Feb. 1975).
- Ellefsen, P.E. (2010). Commodity market modeling and physical trading strategies. S.M. Thesis, *Massachusetts Institute of Technology*.
- Hadjiyiannis, Nicolas (2010). Canonical Correlation of Shipping Forward Curves. S.M. Thesis, *Massachusetts Institute of Technology*.
- Sclavounos, P.D., and Ellefsen, P.E. (2009). Multi-factor model of correlated commodity curves for crude oil and shipping markets. *Massachusetts Institute of Technology. Working Paper. Center for Energy and Environmental Policy Research (CEEPR)*