PRIM Project
Application of PCA Analysis in Forecasting Gold Future Returns

LU Dongrui

February 10, 2020

Contents

1 Background

1.1 Motivations

Gold has always been a commodity that is traded actively on international markets. Investors typically increase their allocation to gold to position defensively as market conditions deteriorate, making it a well-known safe haven asset. Most recently, the gold price soared after the US assassinated Soleimani, one of the most prestigious Iranian general. Empirically indeed, gold has performed well during situations of stress such as war, high inflation, currency crisis or bear markets. Given the fact that the market is changing faster than ever, forecasting Gold futures prices is thus becoming increasingly important to investors for risk hedging as well as alpha generation.

1.2 Project Description

The aim is to forecast gold futures returns using multiple regression analysis. The regressors will include a large set of explanatory variables, for instance, US inflation swap, Europe inflation swap and precious metals data.

A principal component analysis (PCA) will be applied in order to avoid possible issues caused by the multicollinearity of the explanatory variables. Several numerical experiments will be conducted to assess the performance of the present method. In addition, the method will be compared with several other benchmarks in terms of 2 types of errors: mean absolute error (MAE) and mean square error (MAE).

1.3 Industrial Objectives and Contributions

From an industrial perspective, Alphien is seeking to enrich its current framework to allow its users to develop diverse, comprehensive quantitative strategies. Alphien is looking to endow its contract universe with powerful tools whether discretionary or quantitative to navigate through it. By probing into Gold future returns, the project will shed some light on the risk management as well as alpha generation for all Alphien users.

1.4 Problem Formulation and Notation

In the section, we mathematically formalize the problem studied in the project:

Given a time series $X = x_t, t = 1, 2, ..., T$ modelling Gold future returns, we would like to predict $x_{t+1}, x_{t+2}, ..., x_{t+k}$ (k-step ahead). Denote θ to be the parameter (set) for each model.

Remark: All the time series considered in the project is the continuous returns of financial price P_ts.

1.5 Prerequisites in Finance

- Asset
  An asset represents an economic resource for a company. Examples of assets are stocks, money on hands or deposit, bank loans, corporate bonds, mutual funds, etc. Financial contracts, for instance, are a subset of financial assets.

- Financial Contracts
  A financial contract is an arrangement that takes the form of an individually negotiated contract, agreement, or option to buy, sell, lend, swap, or repurchase, or other similar individually negotiated transaction commonly entered into by participants in the financial markets. A financial contract involves securities, commodities, currencies, interest or other rates, or any other financial or economic interest similar in function. Financial contracts are quoted in markets and have corresponding prices at each step of time.

- Prices
  Let p_t be the price of an asset defined on discrete time index t, then the sequence $p_1, p_2, ..., p_{t-1}, p_t$ forms a time series. The **price vector** P_t is defined as:

  $$P_t = [p_1, p_2, ..., p_t] \in \mathbb{R}^t_+$$

  Although easy to interpret, price series exhibit constraining statistical properties. First and foremost, price series are clearly positive definite from the definition. This largely restricts the model choice. Only those that are defined on positive definite series can be applied. Second, price series are usually non-stationary, which means that their distributions is time-varying rather than time invariant. In most cases this will greatly complicate the analysis, and in extreme cases, make it impossible to exploit the data.

- Returns
  The returns are a time series of the price changes between 2 consecutive discrete time stamps. The most frequently used forms of returns are:

  - **Discrete Returns**, denoted $r_{discrete}$

    $$r_{discrete}(t) = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

- **Continuous Returns**, denoted $r_{continuous}$

$$r_{continuous}(t) = \log(\frac{P_t}{P_{t-1}})$$

Usually, quantitative investment strategies are based on the returns over some period T. The returns exhibit more stable statistical properties as they assume negative values and are in some form stationary. As asset prices exhibit random behavior, so are the returns. Therefore, the returns are modelled as stochastic processes and useful information can be extracted from their statistical properties.

1.6 Prerequisites in Statistics

- Stationarity
  Stationarity is a property of time-varying process. Intuitively, stationarity means that the statistical properties of a process generating a time series do not change over time. In other words, the way a time series changes is time-invariant[1]. Consequently, if a time series is stationary, then its statistical properties, such as mean and variance, remain the same as time evolves. Mathematically, a stationary time series $\{x_t\}_t$ has the following properties:

  $$\mathbb{E}(x_t) = \mathbb{E}(x_t + \tau) \quad \mathrm{Var}(x_t) < \infty$$

  where τ is an arbitrary temporal delay.

- Multiple Linear Regression
  Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable[2]. The goal of MLR is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. Mathematically, a MLR with k explanatory variables $x_1, x_2, ..., x_k$ and one response variable y can be written as:

  $$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + \epsilon$$

  where
  α_0 = the intercept
  $\alpha_1, ..., \alpha_k$ = coefficients to be determined for each explanatory variable
  ϵ = the model error, also known as residual In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

- Mean Absolute Error
  Mean absolute error (MAE) measures the average difference between two continuous variables[3]. It is a common measure of forecast error in time series analysis. Given two time series $X_t = (x_1, .., x_N)$ and $Y_t = (y_1, ..., y_N)$, MAE is simply the average absolute vertical or horizontal distance between each point in a scatter plot and the $y = x$ line, therefore it is very easy to understand:

  $$MAE = \frac{1}{N}\sum_{i=1}^{N} |x_i - y_i|$$

  Compared with other widely used time series forecast performance metrics, MAE is less sensible to outliers when series values are bigger than 1. This property makes MAE appealing in many applications where heavy weighing outliers are undesirable. The disadvantage of MAE is that it is scale-dependent, thus it is inappropriate to compare MAEs between series with different scales.

- Mean Squared Error
  Mean squared error (MSE) measures the average of the squared difference between two variables[4]. It is a common measure of forecast error in time series analysis. Given two time series $X_t = (x_1, .., x_N)$ and $Y_t = (y_1, ..., y_N)$, MSE can be written as:

  $$MSE = \frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2$$

  Compared with other widely used time series forecast performance metrics, MSE is more sensible to outliers, especially when series values are bigger than 1. The mathematical benefits of MSE are particularly evident when it comes to analyze the performance of linear regression, as MSE allows one to partition the variation in a dataset into variation explained by the model and variation explained by randomness.

1.7 Model Hypothesis

We consider 4 forecast models for our problem: constant mean model, first-order auto-regressive model, second-order auto-regressive model and PCA-based multiple regression model. Each model has its requirements for the data.

- Constant Mean Model
  Constant mean model describes the data with 2 elements: a constant term and a stochastic term assumed to be a zero-mean white noise

series. This simple model applies for data that follow a normal distribution. Moreover, if we tend to forecast over a long period, the data are supposed to be stationary (its mean remains constant).

- Auto-regressive Model
  Auto-regressive model is a linear model that describes a time-varying process. The value of the variable at time t depends linearly on the values of its past values at $t-1$, $t-2$,... and a stochastic term. The order of an auto regressive model specifies how much history we take into account to describe the current state. For instance, first-order auto-regressive model (AR(1) model) explains a variable at time t by its history at time $t-1$ and a stochastic term. Second-order auto-regressive model (AR(2) model) depicts a variable at time t by its history at time $t-1$, $t-2$ and an error term.

  There are 4 principle hypotheses for auto-regressive model:

  - Normality
    This mean that the data are supposed to follow a normal distribution. In other words, the data comes from a random process.
  - Stationarity
    The data should be stationary, i.e., its mean, variance and other statistical properties don't change with time.
  - Exogeneity
    This means there is no correlation between process's history values and the error. Since the data are assumed to be the output of a random process, the error is centered: $\mathbb{E}(\epsilon) = 0$
  - Homoscedasticity
    The errors are supposed to have a fixed variance.

  If a time series behaves like a random process and is stationary, then auto-regressive model can be a simple and efficient representation of its evolution. Moreover, from the hypotheses we can deduce that the error series is in fact a white noise series with an i.i.d mean of zero and a constant variance. However, not all time-varying processes that fit an auto-regressive model are stationary. In cases where there is a unit root (non-stationary), the analysis becomes much more complicated. As a result, data stationarity is included as one of the main hypotheses.

- PCA-based Multiple Regression Model
  This is a multiple linear regression model preceded by a PCA transformation on data. There are 4 principle hypotheses for multiple linear regression model[5]:

- Non-collinearity between the explanatory variables
  This means none of the explanatory variable can be written as a linear combination of other explanatory variables.
- Independence in errors
  This means that for any of the two terms ϵ_i, ϵ_j $(i \neq j)$ in the error series $\epsilon_1, \epsilon_2, ..., \epsilon_n$, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$
- Exogeneity
  This means there is no correlation between the explanatory variables and the error. If the explanatory variable X is a random variable, we have $\mathbb{E}(\epsilon|X) = 0$, implying that the error is centered.
- Homoscedasticity
  The errors are supposed to have a fixed variance. For any explanatory variables X_i that are random variables, $\mathbb{E}(\epsilon^2|X_i) = \sigma^2$. Otherwise, $\mathbb{E}(\epsilon^2) = \sigma^2$.

The PCA addresses the worries raise by the potential multicollinearity of explanatory variables. Data are first transformed by PCA then applied to the multiple linear regression.

1.8 Project Summary

We reviewed a paper in which it uses 3 methods to forecast an index [19]. The paper finds that their method outperforms other benchmarks in terms of MAE and MSE. We applied similar analysis on Gold future forecast. First of all we established a data profile for Gold future to gain a comprehensive insight of the data property and nonetheless to check if our data meet the requirements of each model. Then we implemented 3 methods separately. In order to make the model more generalized, we incorporated the rolling prediction into the models. At last, we built 4 different strategies based on these models and examined the strategy performance over a year with common financial indicators.

2 Data Profiling

Before making forecasts, we have to first examine the data in order to choose the forecast methods. We establish Gold future profile from the following criteria and tests:

1. mean
2. standard deviation (sd)
3. skewness
4. kurtosis

5. Jarque-Bera (JB) Normality Test
6. Augmented Dickey-Fuller (ADF) Test
7. autocorrelation matrix truncated to lag=4
8. Barlett's Sphericity Test (together with other relevant series)
9. Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (together with other relevant series)

For the rest of this section, we regroup the above indicators according to their purposes and elaborate our observations. We analyse the Gold future returns during 3 time periods:

1. Two-year short-term data, from 2017-10-31 to 2019-10-31
2. Five-year mid-term data, from 2014-10-31 to 2019-10-31
3. Ten-year long-term data,from 2009-10-31 to 2019-10-31

First visualize the data.

Figure 1: Short-term data

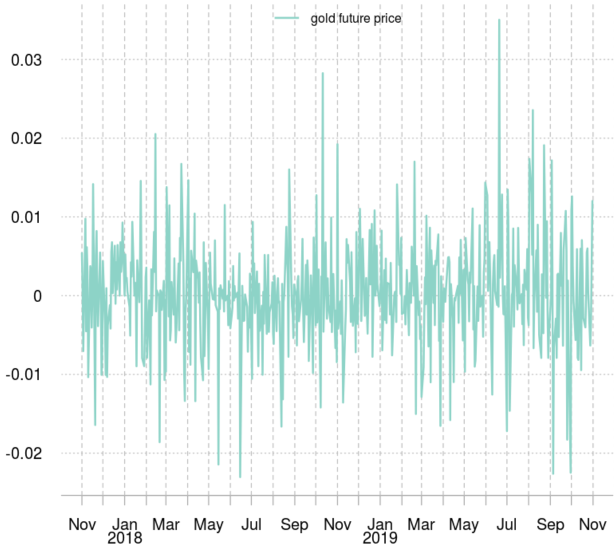


Figure 2: Mid-term data

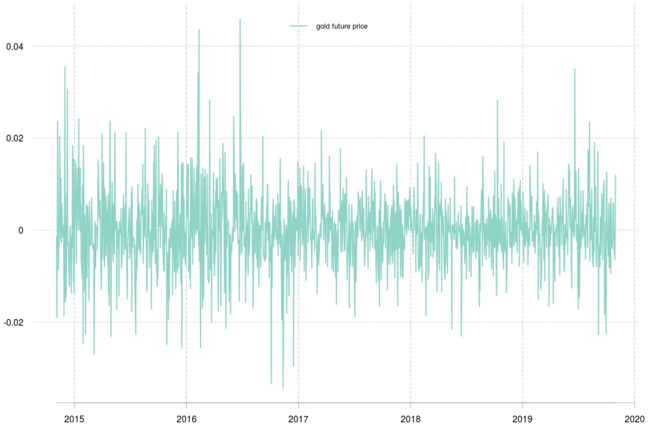
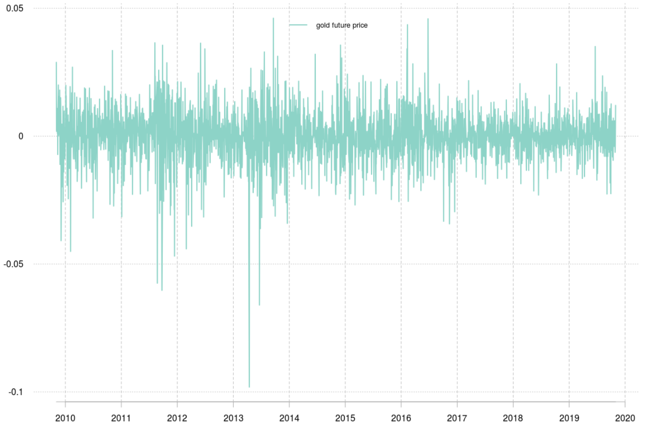


Figure 3: Long-term data



2.1 Stationarity

2.1.1 ADF test

The test tests the null hypothesis H_0 that a root unit is present in a time series sample, that is $r_t = r_{t-1} + ...$[6]. If H_0 is rejected the data is considered stationary.

The transformation $ln(P)$ equalize the variability over the length of a single series and suppresses larger fluctuations[7]. As a result, we expect the data to be stationary.

The table below shows the test values and the corresponding p-values:

	short-term	mid-term	long-term
ADF test value	-8.266	-11.426	-13.919
p-value	<0.01	<0.01	<0.01

For all our data, we get very negative test values with high reliability (p-value < 0.01). Moreover, the longer the time span, the more negative the test result. Thus H_0 can safely be rejected. It is possible that Gold future is stationary.

2.1.2 autocorrelation matrix truncated to lag=4

According to ADF test, the data seems to be stationary. Therefore, the autocorrelation coefficient (ACF) is time-invariant.

The table below shows the autocorrelation matrix for short-term data from 2017 to 2019[8]:

	x	$x_{lag=1}$	$x_{lag=2}$	$x_{lag=3}$	$x_{lag=4}$
x	1.0000	-0.0678	0.0433	-0.0335	-0.0628
$x_{lag=1}$	-0.0678	1.0000	-0.0725	0.0453	-0.0283
$x_{lag=2}$	0.0433	-0.0725	1.0000	-0.0720	0.0447
$x_{lag=3}$	-0.0335	0.0452	-0.0720	1.0000	-0.0730
$x_{lag=4}$	-0.0628	-0.0283	0.0447	-0.0730	1.0000

The null hypothesis H_0 corresponding to the p-values of correlation matrix is "the correlation coefficient is equal to 0". The alternative hypothesis is "the correlation coefficient is not 0", or in other words, "the data series are correlated."

The table below shows the p-values of each ACF:

	x	$x_{lag=1}$	$x_{lag=2}$	$x_{lag=3}$	$x_{lag=4}$
x	NA	0.1234	0.3257	0.4462	0.1536
$x_{lag=1}$	0.1234	NA	0.0993	0.3039	0.5199
$x_{lag=2}$	0.3257	0.0993	NA	0.1015	0.3098
$x_{lag=3}$	0.4462	0.3039	0.1015	NA	0.0972
$x_{lag=4}$	0.1536	0.5199	0.3098	0.0972	NA

The big p-values suggest we cannot reject H_0. It's possible that there is no correlation between gold future price and its lagged returns.

The correlation matrix for mid-term data from 2014 to 2019:

	x	$x_{lag=1}$	$x_{lag=2}$	$x_{lag=3}$	$x_{lag=4}$
x	1.0000	-0.0521	0.0048	0.0399	-0.0045
$x_{lag=1}$	-0.0521	1.0000	-0.0519	0.0053	0.0409
$x_{lag=2}$	0.0048	-0.0519	1.0000	-0.0512	0.0060
$x_{lag=3}$	0.0399	0.0053	-0.0512	1.0000	-0.0514
$x_{lag=4}$	-0.0045	0.0409	0.0060	-0.0514	1.0000

The corresponding p-values:

	x	$x_{lag=1}$	$x_{lag=2}$	$x_{lag=3}$	$x_{lag=4}$
x	NA	0.0606	0.8636	0.1502	0.8701
$x_{lag=1}$	0.0606	NA	0.0615	0.8479	0.1404
$x_{lag=2}$	0.8636	0.0615	NA	0.0651	0.8298
$x_{lag=3}$	0.1502	0.8479	0.0650	NA	0.0638
$x_{lag=4}$	0.8701	0.1404	0.8298	0.0638	NA

Similarly, despite the small correlation coefficients, the big p-values suggest that there may not be any correlation between gold future price and its lagged returns during the past 5 years.

The correlation matrix for long-term data from 2009 to 2019:

	x	$x_{lag=1}$	$x_{lag=2}$	$x_{lag=3}$	$x_{lag=4}$
x	1.0000	-0.0341	0.0063	0.0066	-0.0164
$x_{lag=1}$	-0.0341	1.0000	-0.0343	0.0065	0.0075
$x_{lag=2}$	0.0063	-0.0343	1.0000	-0.0342	0.0068
$x_{lag=3}$	0.0066	0.0065	-0.0342	1.0000	-0.0340
$x_{lag=4}$	-0.0164	0.0075	0.0068	-0.0340	1.0000

The corresponding p-values:

	x	$x_{lag=1}$	$x_{lag=2}$	$x_{lag=3}$	$x_{lag=4}$
x	NA	0.0819	0.7483	0.7379	0.4039
$x_{lag=1}$	0.0819	NA	0.0805	0.7401	0.7014
$x_{lag=2}$	0.7484	0.0805	NA	0.0811	0.7290
$x_{lag=3}$	0.7379	0.7401	0.0811	NA	0.0830
$x_{lag=4}$	0.4039	0.7014	0.7290	0.0830	NA

Similarly, the big p-values suggest that there may be no correlation between gold future price and its lagged returns during the past 10 years.

2.1.3 summary

ADF test suggests that Gold future returns are stationary. Therefore we proceed to compute the auto-correlation matrix and the corresponding p-values. There does not seem to be evident correlation between our data and its lagged series. However, the big p-values indicate that the correlation coefficients are not reliable, therefore it is possible that there exists some correlation.

2.2 Data Distribution

The values for the criteria and the tests concerning data distribution are summarized in the tables below.
For short-term data from 2017 to 2019:

Indicator	Value
mean	0.0002359912
sd	0.006838531
skewness	0.222637
kurtosis	5.54468
JB normality test (p-value $<$ 0.01)	147.58

For mid-term data from 2014 to 2019:

Indicator	Value
mean	0.0001336844
sd	0.008174698
skewness	0.3071127
kurtosis	6.009173
JB normality test (p-value $<$ 0.01)	515.56

For long-term data from 2009 to 2019:

Indicator	Value
mean	0.0000953887
sd	0.009892596
skewness	-0.7423951
kurtosis	10.19314
JB normality test (p-value $<$ 0.01)	5874.6

2.2.1 mean

The average gold future log-returns are always close to 0, which means $P_t \approx P_{t-1}$ for most of the days. Globally, the log-returns seem to have had minor fluctuations for the past 2,5 and 10 years. This conforms with the prior hypothesis that the data is stationary.

2.2.2 sd

The standard deviation is not trivial compared to the mean. Furthermore, the longer the time span, the bigger the standard deviation of gold future returns. Along with the graphs above, the results suggest that the data seem to behave like white noise.

2.2.3 skewness

Skewness is a measure of data symmetry. The skewness S for time series $X = (x_1, x_2, ...x_T)$ is

$$S = \frac{1}{T} \cdot \frac{\sum_{i=1}^{T}(x_i - \overline{X})^3}{sd(X)^3}$$

where $\overline{X}$ is the mean and $sd(X)$ is the standard deviation.

When X is a sample, we use the adjusted skewness $S^{'}$[9]:

$$S^{'} = \frac{\sqrt{T(T-1)}}{T-2} \cdot \frac{\sum_{i=1}^{T}(x_i - \overline{X})^3}{sd(X)^3}$$

The smaller S is, the more symmetric the data is. The skewness for normal distribution is 0.

The skewness for Gold future is around 0.25 for short-term and mid-term data, between $-\frac{1}{2}$ and $\frac{1}{2}$. The data is approximately symmetric. This corresponds with data's log-normality characteristics (close to 0). The skewness is positive. This means the data are positively skewed or skewed right. The right tail of the distribution is longer than the left. In other words, There are more outliers now than one or two years ago.

The skewness for Gold future is around -0.74 for long-term data. In the long run, the data is much less symmetric and the log-normality characteristics is less evident. The skewness is negative. This means the data are negatively skewed or skewed left. Contrary to the 2 cases above, the left tail of the distribution is longer than the right. In other words, There are more outliers around 10 years ago than recently.

2.2.4 kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution[10]. Data with high kurtosis tends to have many outliers (heavy-tailed). An extreme case is uniform distribution.

The kurtosis K for time series $X = (x_1, x_2, ...x_T)$ is [9]

$$K = \frac{1}{T} \cdot \frac{\sum_{i=1}^{T}(x_i - \overline{X})^4}{sd(X)^4}$$

where $\overline{X}$ is the mean and $sd(X)$ is the standard deviation.

The kurtosis of the normal distribution is 3. The excessive kurtosis compared to the normal distribution, also known as excess kurtosis, is not very big, which shows that data still have some log-normality. If data are perfectly log-normal, the excess kurtosis should be 0.

The data always have positive excess kurtosis and is therefore *leptokurtic*[11]. The data have fatter tails, i.e., more outliers than normal distributions. Moreover, the longer period we pick, the more outliers we observe. In particular, the kurtosis increases a lot when passing from 5-year observation to 10-year observation. As a result, we expect that there were more (bigger) fluctuations around 10 years ago.

2.2.5 JB normality test

This is a test of whether sample data have the skewness and kurtosis matching a normal distribution[12].

$$JB = \frac{T}{6}\left((S^2 + \frac{1}{4}(K-3)^2)\right)$$

The null hypothesis H_0 is "data follow normal distribution".

The JB statistic is far from 0, and the longer period we observe, the bigger the test statistics. This means that data don't follow normal distribution, especially in the long term. This conclusion conforms with the skewness and kurtosis which become less and less regular as the data period extends. But meanwhile, in all 3 cases, the test statistic asymptotically has a chi-squared distribution with two degrees of freedom, which is the same for normally-distributed data.

Overall, we can safely reject H_0. Although Gold returns bear some resemblance to normally distributed data, they don't come from normal distribution.

2.2.6 summary

From the above indicators, we can conclude that Gold future returns indeed behave somewhat like a normal distribution. The visualization as well as a high variance compared to the mean indicates a similarity with white noise. But meanwhile, with more outliers, Gold future returns are more irregular than a standard normal distribution.

2.3 PCA Applicability

We would like to perform PCA-based multiple regression model with Gold return series and other time series believed relevant to gold. These series are:

Name	Abbr. in R command
US Breakeven 10 Year	USGGBE10 Index
US Generic Govt 5 Year Yield	USGG5YR Index
CoffeeArabica	KC
EUR/JPY	RY
EuroBund	RX
Hang Seng Index	HI
INR/USD in SGX	XID
Korean Won	KU
Palladium	PA
Platinum	PL
Name	Abbr. in R command
SGX CNX Nifty	IH
Silver	SI
Swiss Market Index	SMix
US 2Yr Note	TU
Wheat	W
USD/JPY	JY

The table below shows the tests' results for short-term, mid-term and long-term data:

	short-term	mid-term	long-term
Barlett's sphericity test	3613.938	8382.648	16234.666
KMO sampling adequacy test	0.7662	0.7795	0.7870

2.3.1 Barlett's sphericity test

Let M be the matrix whose columns are Gold future and all other relevant series. This test tests the null hypothesis H_0 that the correlation matrix of M is an identity matrix[13, 14].

Correlation matrix is symmetrical, with diagonal elements being variances. All the rest are correlations between variable pairs.It's useless to apply PCA to the dataset if the correlation matrix is an identity matrix. Because if this is the case, every variable is already orthogonal to any other variable (correlation = 0). In other words, every variable is already a principal component itself, we thus cannot further reduce the dataset dimensionality.

Therefore, if we can reject H_0, then PCA is applicable.

We obtain big test values with a tiny p-value (always $< 2.2 \times 10^{-16}$), large enough to reject H_0. Consequently, We can perform PCA on dataset M.

2.3.2 KMO sampling adequacy test

This is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors[14, 15]. The test values varies from 0 to 1. The higher the test statistics, the more suited the data are for factor analysis. Therefore, a high value (close to 1.0) suggests that PCA may be useful with the data. A value close to 0 means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations in the data, which will cause a big problem for factor analysis[15, 16]. Generally speaking, if the statistics is less than 0.50, factor analysis such as PCA probably won't be very useful.

The KMO values stabilise around 0.78, which is between "middling" and "meritorious" levels. Our dataset M is sufficient for PCA.

2.3.3 summary

The 2 statistics examine the applicability of PCA from different perspectives. Barlett's sphericity test illustrates that our dataset has enough correlation to perform PCA analysis. KMO test verifies that our 3 datasets are all ample enough for PCA. Therefore, we can safely conclude that PCA analysis is appropriate for our problem.

3 Forecasting

This section aims to provide theoretical analysis for each forecast model. Based on Gold future profile, we select and compare the performances of the following forecast methods:

1. Constant Mean Model
2. First-order Auto-Regressive Model - AR(1)
3. Second-order Auto-Regressive Model - AR(2)
4. PCA-based Multiple Regression Model

3.1 Constant Mean Model

Constant mean model is a simple linear regression model that is suitable for data which have a quasi-normal distribution. It has only 1 parameter $\theta = \mu$:

$$x_t = \mu + \epsilon_t \quad t = 1, 2, ..., T, T+1, ...T+k$$

with μ a constant to be determined and ϵ_ts errors assumed to be a white noise series with an i.i.d mean of zero.

3.1.1 Model Resolution

μ is determined by ordinary least square (OLS) method. That is to say, we would like to find θ that minimizes squared error $L(\mu) = \sum_{t=1}^{T}(x_t - \mu)^2$:

$$\theta^* = \mu^* = \underset{\mu}{\operatorname{argmin}}\, L(\mu) = \underset{\mu}{\operatorname{argmin}} \sum_{t=1}^{T}(x_t - \mu)^2$$

Derive the error function $L(\mu)$ and set the derivative to be 0:

$$\frac{dL}{d\mu}(\mu^*) = -2\sum_{t=1}^{T} x_t + 2\mu^* \cdot T = 0$$

Thus, we have

$$\boxed{\mu^* = \frac{1}{T}\sum_{t=1}^{T} x_t}$$

The optimal parameter is the mean of training data series. We can easily verify that the constraint on ϵs is satisfied:

$$\mathbb{E}(\epsilon_t) = \mathbb{E}(x_t) - \mu^* = 0$$

3.1.2 Model Prediction

The predicted prices $\hat{x}_t, t = T+1, ..., T+k$ are:

$$\boxed{\hat{x}_t = \mu^*}$$

3.2 First-order Auto-Regressive Model – AR(1)

AR(1) model is a widely-used time series forecast model suitable for stationary series[17]. For a stationary time series X whose mean is not 0, AR(1) model has 2 parameters, $\theta = (\alpha, \beta)$:

$$x_t = \alpha\, x_{t-1} + \beta + \epsilon_t$$

with α,β 2 constant coefficients to be determined and ϵ_ts errors assumed to be a white noise series with an i.i.d mean of zero and a constant variance σ_ϵ^2.

If we preprocess X and subtract its mean, AR(1) model can be equivalently expressed as:

$$x_t = \alpha\, x_{t-1} + \epsilon_t$$

This is also the AR(1) model for zero-mean time series. In this case $\theta = \alpha$. The β in the previous representation is simply interpreted as the intercept of the model here.

3.2.1 Model Resolution

α and β can either be determined by ordinary least square (OLS) method or Yule-Walker (YW) method.

- linear regression method – OLS method

Under OLS method, we would like to find θ that minimizes squared error $L(\alpha, \beta) = \sum_{t=2}^{T} (x_t - \alpha\, x_{t-1} - \beta)^2$:

$$\theta^* = (\alpha^*, \beta^*) = \underset{\alpha,\beta}{\operatorname{argmin}}\, L(\alpha, \beta) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{t=2}^{T} (x_t - \alpha\, x_{t-1} - \beta)^2$$

Derive the error function $L(\alpha, \beta)$ respectively by α and β and set the derivative to be 0

$$\begin{cases} \frac{\partial L}{\partial \alpha}(\alpha^*, \beta^*) = \sum_{t=2}^{T} -2x_{t-1}(x_t - \beta^*) + 2\alpha^* x_{t-1}^2 = 0 \\ \frac{\partial L}{\partial \beta}(\alpha^*, \beta^*) = \sum_{t=2}^{T} -2(x_t - \alpha^* x_{t-1}) + 2\beta^*(T-1) = 0 \end{cases}$$

Since $\beta^* = \frac{1}{T-1} \sum_{t=2}^{T} (x_t - \alpha^* x_{t-1})$, we have:

$$-2x_{t-1}\left(x_t - \frac{1}{T-1} \sum_{t=2}^{T} (x_t - \alpha^* x_{t-1}) \right) + 2\alpha^* x_{t-1}^2 = 0$$

$$\alpha^* \cdot \sum_{t=2}^{T} x_{t-1}^2 + \frac{1}{T-1} \sum_{t=2}^{T} x_{t-1} \sum_{t=2}^{T} (x_t - \alpha^* x_{t-1}) = \sum_{t=2}^{T} x_{t-1} x_t$$

Divide $T-1$ on both sides, the left part of the last equation above can be rewritten as:

$$\alpha^* \cdot \frac{1}{T-1}\sum_{t=2}^{T} x_{t-1}^2 + \frac{1}{(T-1)^2}\sum_{t=2}^{T} x_{t-1}\sum_{t=2}^{T} x_t - \alpha^* \cdot \frac{1}{(T-1)^2}\left(\sum_{t=2}^{T} x_{t-1}\right)^2$$

Identify the definition of expectation, variance and covariance:

$$\alpha^* \left[\mathbb{E}(X_{t-1}^2) - \mathbb{E}(X_{t-1})^2\right] = \mathbb{E}(X_{t-1}X_t) - \mathbb{E}(X_{t-1})\mathbb{E}(X_t)$$

$$\alpha^* = \frac{\text{Cov}(X_t, X_{t-1})}{\text{Var}(X_{t-1})}$$

Finally, we have

$$\boxed{\begin{cases} \alpha^* = \frac{\text{Cov}(X_t, X_{t-1})}{\text{Var}(X_{t-1})} \\ \beta^* = \frac{1}{T-1}\sum_{t=2}^{T}(x_t - \alpha^* x_{t-1}) \end{cases}}$$

- time-series method – YW method

Under YW method, we would like to find θ that minimizes the expectation of the squared error $L(\alpha,\beta) = \mathbb{E}\left((x_t - \alpha\, x_{t-1} - \beta)^2\right)$:

$$\theta^* = (\alpha^*, \beta^*) = \underset{\alpha,\beta}{\text{argmin}}\, L(\alpha,\beta)$$

If we fit the model directly, our goal is to find (α, β) that minimizes

$$\frac{1}{T-1}\sum_{t=2}^{T}(x_t - \alpha\, x_{t-1} - \beta)^2$$

The objective function is almost the same as that of the OLS method, therefore we can apply the same analysis as before.

If we subtract the mean from X first and apply AR(1) model for zero-mean time series, our goal is to find α that minimizes:

$$\mathbb{E}\left((x_t - \alpha\, x_{t-1})^2\right) = \mathbb{E}\left(x_t^2 - 2\alpha\, x_t x_{t-1} + \alpha^2\, x_{t-1}^2\right)$$

Let $\gamma(k)$ be the autocovariance of X of order k. That is, $\gamma(k) = \text{Cov}(x_t x_{t-k})$. Because X is stationary, $\gamma(k)$ is only a function of k. Since we are dealing with zero-mean series, we have $\gamma(0) = \mathbb{E}(x_t^2)$ and $\gamma(1) = \mathbb{E}(x_t x_{t-1})$. Thus, the objective function $L(\alpha)$ becomes:

$$\gamma(0) - 2\alpha\gamma(1) + \alpha^2\gamma(0)$$

Derive L and set the derivative to be 0:

$$\frac{dL}{d\alpha}(\alpha^*) = -2\gamma(1) + 2\alpha^*\gamma(0) = 0$$

Recall the definition of autocorrelation $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$, finally we have

$$\boxed{\alpha^* = \frac{\gamma(1)}{\gamma(0)} = \rho(1)}$$

Remark: α^* is the first-order partial autocorrelation coefficient (PACF) of the stationary time series X. First-order PACF coincides with the autocorrelation coefficient (ACF)[17, 18].

3.2.2 Model Prediction

For a stationary time series X whose mean is not 0, the predicted prices $\hat{x}_t, t = T+1, ..., T+k$ are:

$$\boxed{\hat{x}_t = \alpha^*\, x_{t-1} + \beta^*}$$

Reorganize the above result, we have:

$$\boxed{\hat{x}_t - \mu_t = \alpha^*(x_{t-1} - \mu_{t-1})}$$

with $\mu_t = \frac{1}{T-1}\sum_{t=2}^{T} x_t$ and $\mu_{t-1} = \frac{1}{T}\sum_{t=2}^{T} x_{t-1}$. Notice that this is exactly the answer given by YW method with preprocessing.

For a zero-mean time series X, the predicted prices $\hat{x}_t, t = T+1, ..., T+k$ are:

$$\boxed{\hat{x}_t = \alpha^*\hat{x}_{t-1}}$$

3.3 Second-order Auto-Regressive Model – AR(2)

AR(2) model is a most popular time series forecast model suitable for stationary series[17]. In fact, as is indicated by their names, AR(1) and AR(2) models are from the same family. For a stationary time series X whose mean is not 0, AR(2) model has 3 parameters, $\theta = (\alpha, \beta, \omega)$:

$$x_t = \alpha\, x_{t-1} + \beta\, x_{t-2} + \omega + \epsilon_t$$

with α,β,ω 3 constant coefficients to be determined and ϵ_ts errors assumed to be a white noise series with an i.i.d mean of zero and a constant variance σ_ϵ^2

If we preprocess X and subtract its mean, AR(2) model can be equivalently expressed as:

$$x_t = \alpha\, x_{t-1} + \beta\, x_{t-2} + \epsilon_t$$

This is also the AR(2) model for zero-mean time series. In this case $\theta = (\alpha, \beta)$. The ω in the previous representation is simply interpreted as the intercept of the model here.

3.3.1 Model Resolution

Due to the dependency between the explanatory variables, we can no longer apply multiple linear regression. The parameters can be determined by YW method. Similarly, we preprocess X to ensure it has zero-mean. The goal is to find θ that minimizes the expectation of the squared error $L(\alpha, \beta) = \mathbb{E}\left((x_t - \alpha\, x_{t-1} - \beta\, x_{t-2})^2\right)$:

$$\theta^* = (\alpha^*, \beta^*) = \underset{\alpha,\beta}{\operatorname{argmin}}\, \mathbb{E}\left((x_t - \alpha\, x_{t-1} - \beta\, x_{t-2})^2\right)$$

Rewrite the objective function with autocovariance $\gamma(k) = \operatorname{Cov}(x_t x_{t-k})$:

$$\begin{aligned}\mathbb{E}\left((x_t - \alpha\, x_{t-1} - \beta\, x_{t-2})^2\right) &= \mathbb{E}\left(x_t^2 + \alpha^2 x_{t-1}^2 + \beta^2 x_{t-2}^2 - 2\alpha x_t x_{t-1} - 2\beta x_t x_{t-2} + 2\alpha\beta x_{t-1} x_{t-2}\right) \\ &= \gamma(0) + \alpha^2\gamma(0) + \beta^2\gamma(0) - 2\alpha\gamma(1) - 2\beta\gamma(2) + 2\alpha\beta\gamma(1) \\ &= (1 + \alpha^2 + \beta^2)\,\gamma(0) - 2\alpha\gamma(1) - 2\beta\gamma(2) + 2\alpha\beta\gamma(1)\end{aligned}$$

Derive objective function L respectively by α and β, and set the derivatives to be 0:

$$\begin{cases}\frac{\partial L}{\partial \alpha}(\alpha^*, \beta^*) = 2\alpha^*\gamma(0) - 2\gamma(1) + 2\beta^*\gamma(1) = 0 \\ \frac{\partial L}{\partial \beta}(\alpha^*, \beta^*) = 2\beta^*\gamma(0) - 2\gamma(2) + 2\alpha^*\gamma(1) = 0\end{cases}$$

Recall autocorrelation $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$:

$$\begin{cases}\alpha^* = \rho(1)(1 - \beta^*) \\ \beta^* = \rho(2) - \alpha^*\rho(1)\end{cases}$$

The linear system above yields:

$$\boxed{\begin{cases}\alpha^* = \frac{\rho(1)(1-\rho(2))}{1-\rho(1)^2} \\ \beta^* = \frac{\rho(2)-\rho(1)^2}{1-\rho(1)^2}\end{cases}}$$

Remark: β^* is the second-order PACF of the stationary time series X[18].

3.3.2 Model Prediction

For a stationary time series X whose mean is not 0, the predicted prices $\hat{x}_t, t = T+1, ..., T+k$ are:

$$\boxed{\hat{x}_t - \mu_t = \alpha^*(x_{t-1} - \mu_{t-1}) + \beta^*(x_{t-2} - \mu_{t-2})}$$

with $\mu_t = \frac{1}{T-2}\sum_{t=3}^{T} x_t$, $\mu_{t-2} = \frac{1}{T-2}\sum_{t=3}^{T} x_{t-1}$ and $\mu_{t-1} = \frac{1}{T-2}\sum_{t=3}^{T} x_{t-2}$.

For a zero-mean time series X, the predicted prices $\hat{x}_t, t = T+1, ..., T+k$ are:

$$\boxed{\hat{x}_t = \alpha^* x_{t-1} + \beta^* x_{t-2}}$$

3.4 PCA-based Multiple Regression Model

Multiple regression model is in fact a combination of multiple linear regression model and the PCA analysis [19]. Suppose a stationary time series X. Multiple regression model explains the mean of X by some explanatory variables (i.e. other relevant time series) and its own lag series:

$$\begin{aligned} x_t &= \mu_t + \epsilon_t & (1) \\ &= \mu + \sum_{i=1}^{n} \alpha_i y_{it} + \sum_{j=1}^{p} \beta_j x_{t-j} + \epsilon_t & (2) \end{aligned}$$

with μ,α_i,β_j constant coefficients to be determined and ϵ_ts errors assumed to be a white noise series with an i.i.d mean of zero and a constant variance σ_ϵ^2. n,p are positive integers.

It's possible that there exists multicollinearity between the explanatory variables y_{it}s and x_{t-j}s. To avoid the complication brought by the potential multicollinearity, PCA is applied to transform the explanatory variables into the principal components z_{it}s, which are used as the new explanatory variables:

$$x_t = \mu + \sum_{i=1}^{m} \omega_i z_{it} + \epsilon_t \tag{3}$$

Suppose the original explanatory variables form a m-dimensional random variable R_t with corvariance matrix Σ_R. Each z_{it} is a linear combination of y_{it}s and x_{t-j}s. Any two principal components z_{it} and z_{jt} $(i \neq j)$ are uncorrelated. Furthermore, the variance of the z_{it}s are as large as possible so that z_{it} can best explain the structure of Σ_R.

As Σ_R is non-negative definite, it has a spectral decomposition. Let $(\lambda_1, e_1), ...,$ (λ_m, e_m) be the eigenvalue-eigenvector pairs of Σ_R $(\lambda_{1,...,m} > 0)$. Thus, the principal components and their variances are, $\forall i \in [\![1, 2, ..., m]\!]$:

$$z_{it} = e_i' R_t = \sum_{j=1}^{m} e_{ij} R_{jt}$$

$$\text{Var}(z_{it}) = e_i' \Sigma_R e_i$$

In addition, $\text{Var}(z_{1t}) > \text{Var}(z_{2t}) > ... > \text{Var}(z_{mt})$.

3.4.1 Model Resolution

The explanatory variables being all independent from one another, the coefficients can be obtained via normal multiple linear regression methods, for example OLS method. Under OLS, it's sufficient to solve the linear system made up of the partial derivatives of the objective function $L(\mu, \omega_1, \omega_2, ...\omega_m) = \sum_{t=p}^{T} (x_t - \mu - \sum_{i=1}^{m} \omega_i z_{it})^2$.

3.4.2 Model Prediction

The predicted prices $\hat{x}_t, t = T+1, ..., T+k$ are:

$$\boxed{\hat{x}_t = \mu + \sum_{i=1}^{m} \omega_i z_{it}}$$

4 Forecast Implementation and Experiments

In this section, we first show some toy examples and compare the model output with the analytical one. Then we conduct experiments on Gold future returns. We have conducted 3 experiments with short-term (2-year), mid-term (5-year) and long-term (10-year) data respectively.
All the functions are written in R language. Below is a list of all the packages that are used in the forecast method implementation:

Package Name	Purpose
base	basic arithmetic and operations
stats	basic statistic indicator, model building and prediction
timeDate	data profile, data preprocessing
tseries	data profile
REdaS	data profile
Hmisc	series correlation coefficient
zoo	time series operations
xts	time series operations
MLmetrics	model evaluation
grDevices	plotting
graphics	plotting
AlphienData	plotting

In addition to the forecast model, we've incorporated the notion "rolling forecast" into the implementation. A rolling forecast predicts the future over a set period of time[20]. Its first in/first out (FIFO) mechanism ensures that the forecast always covers the same amount of time. Because a rolling forecast window requires routine revisions, it is also referred to as a continuous forecast or an iterative forecast.

In our experiment, we set the default rolling window size to be 20 days and the forecast horizon to be 1 day, that is, one-day ahead forecast with history data of 20 days.

4.1 Constant Mean Model

Function API:
getForecastCsteMean(pxs, trainDataLen = 20, forecastStep = 1, rollStep = 1, showGraph = FALSE)

pxs: data
trainDataLen: rolling window size
forecastStep: forecast horizon
rollStep: rolling step, move forward rollStep days after each forecast
showGraph: boolean parameter, whether to plot the forecast values and the true values

4.1.1 Toy Example

$X = (-0.0001, -0.0183, 0.0064, -0.0182, -0.0024)$.
Q1. What is the predicted value for x_6?

Answer: First deduce by hand.

$$T = 5 \qquad \hat{x}_6 = \mu^* = \frac{1}{T}\sum_{t=1}^{T} x_t = -0.0065$$

Now predict x_6 with our function *getForecastCsteMean*. We get $\hat{x}_6 = \mu^* == -0.00652$
Q2. What is the predicted value for $x_{6:8} = (x_6, x_7, x_8)$?

Answer:

$$T = 5 \qquad \forall j \in [\![6, 8]\!], \hat{x}_j = \mu^* = \frac{1}{T}\sum_{t=1}^{T} x_t = -0.0065$$

With function *getForecastCsteMean*. We get $\hat{x}_6 = \mu^* == -0.00652$

4.1.2 Experiment on Gold Future Returns

The table below shows the mean absolute error (MAE) and mean square error (MSE) for each forecast:

	short-term	mid-term	long-term
MAE	5.0702e-03	6.0636e-03	0.00709086
MSE	4.8249e-05	6.9511e-05	0.00010291

Figure 4: short-term forecast – CsteMean

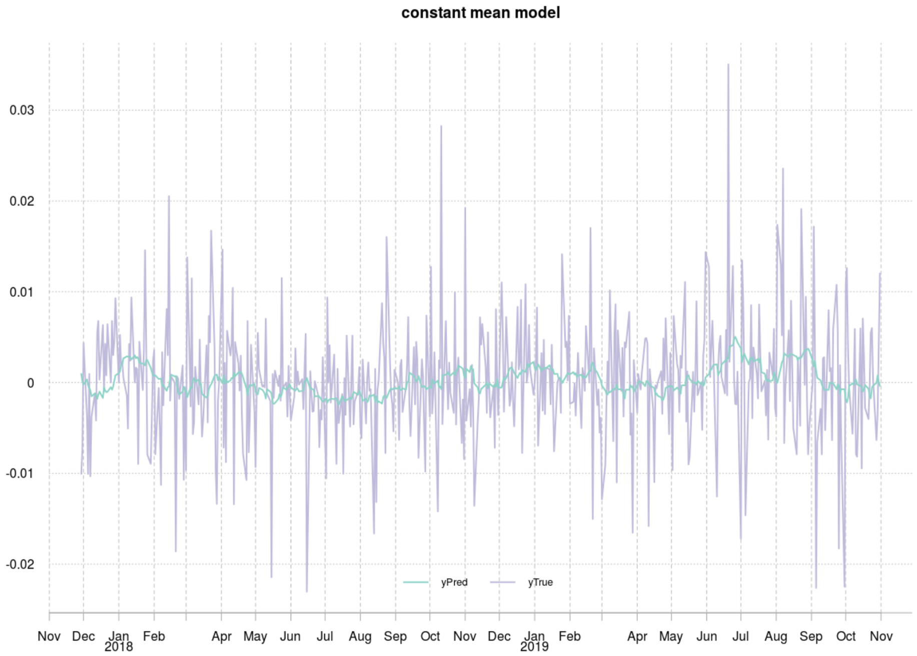


Figure 5: mid-term forecast – CsteMean

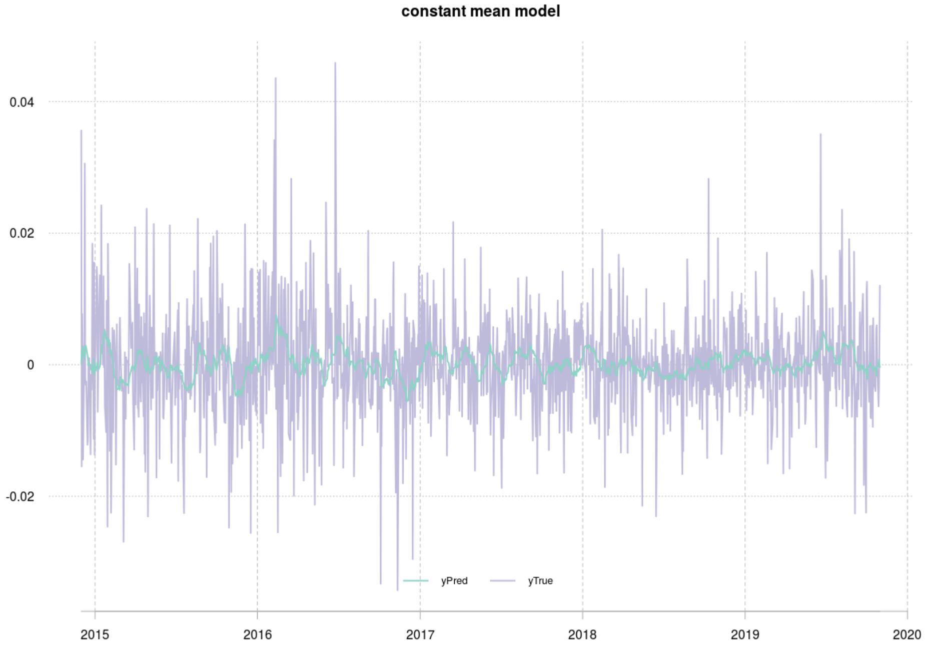
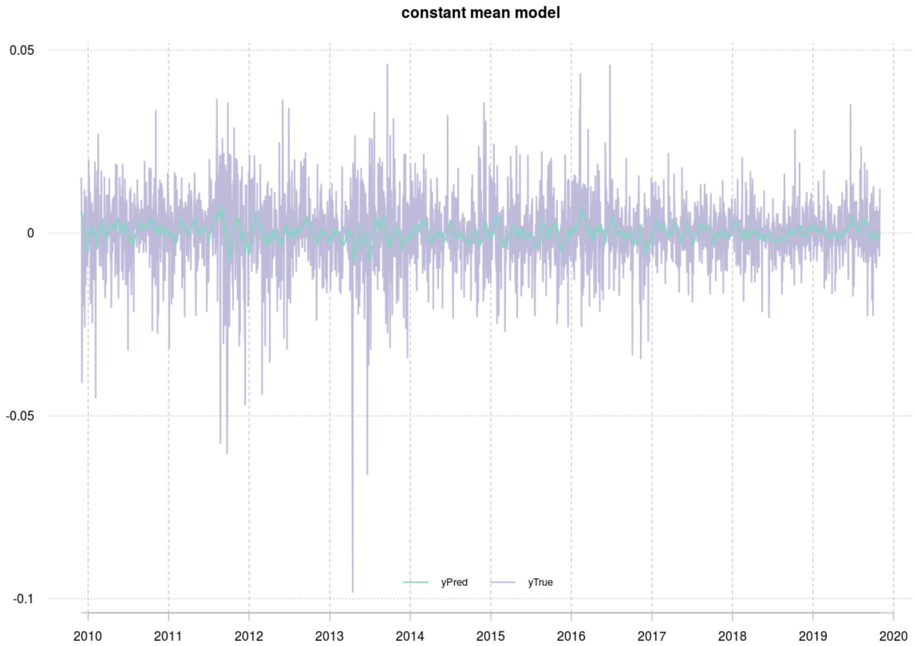


Figure 6: long-term forecast – CsteMean



4.2 First-order Auto-regressive Model

Function API:
getForecastAR(pxs, lags = 1, trainDataLen = 20, forecastStep = 1, rollStep = 1, showGraph = FALSE)

pxs: data
lags: the maximal lag order
trainDataLen: rolling window size
forecastStep: forecast horizon
rollStep: rolling step, move forward rollStep days after each forecast
showGraph: boolean parameter, whether to plot the forecast values and the true values

4.2.1 Toy Example

$X = (-0.0001, -0.0183, 0.0064, -0.0182, -0.0024)$.
Q1. What is the predicted value for x_6?

Answer:

First deduce by hand. $T = 5$. First preprocess X by subtracting its mean $\mu_X = -0.0065$

$$X' = X - \mu_X = (0.0064, -0.0118, 0.0129, -0.0117, 0.0041)$$

The optimal parameter is the ACF between $Y_{T-1} = X'_{1:T-1}$ and $Y_T = X'_{2:T}$.

$$\alpha^* = \mathrm{acf}(Y_{T-1}, Y_T) = -0.9348$$

$$\mu_6 = \frac{1}{T-1}\sum_{t=2}^{T} x_t = -0.0081 \quad \mu_5 = \frac{1}{T-1}\sum_{t=2}^{T} x_{t-1} = -0.0076$$

$$\hat{x}_6 = \mu_6 + \alpha^*(x_5 - \mu_5) = -0.0129$$

Now predict x_6 with our function *getForecastAR*. We get $\alpha^* = -0.905$ and $\hat{x}_6 = -0.0128$.
Q2. What is the predicted value for $x_{6:8} = (x_6, x_7, x_8)$?
Answer:

$$\forall j \in [\![6, 8]\!], \hat{x}_j = \mu_j + \alpha^*(x_{j-1} - \mu_{j-1})$$

The optimal parameter α^* is the same as in Q1. We need to know the true value of x_6 to compute $\hat{x}_7$ and the true values of x_6, x_7 to compute $\hat{x}_8$.

4.2.2 Experiment on Gold Future Returns

The table below shows the mean absolute error (MAE) and mean square error (MSE) for each forecast:

	short-term	mid-term	long-term
MAE	4.9791e-03	5.9162e-03	0.00700516
MSE	4.8094e-05	6.8220e-05	0.00010231

Figure 7: short-term forecast – AR(1)

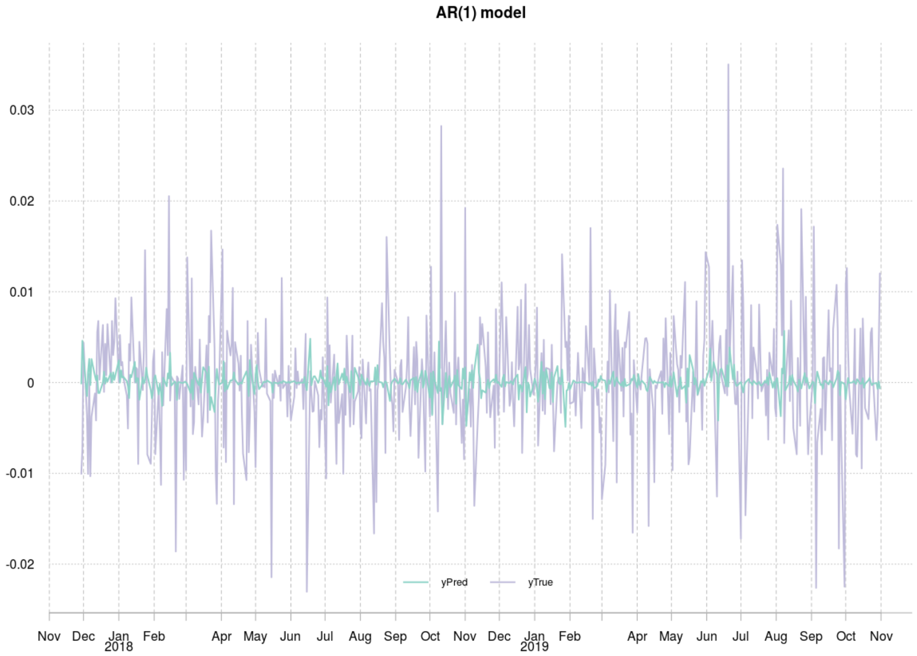


Figure 8: mid-term forecast – AR(1)

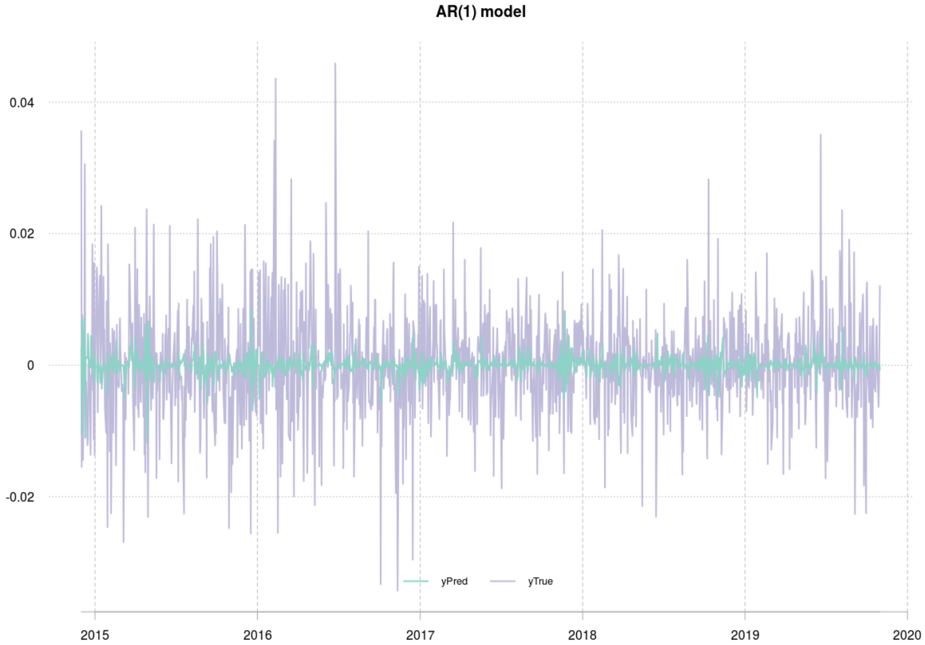
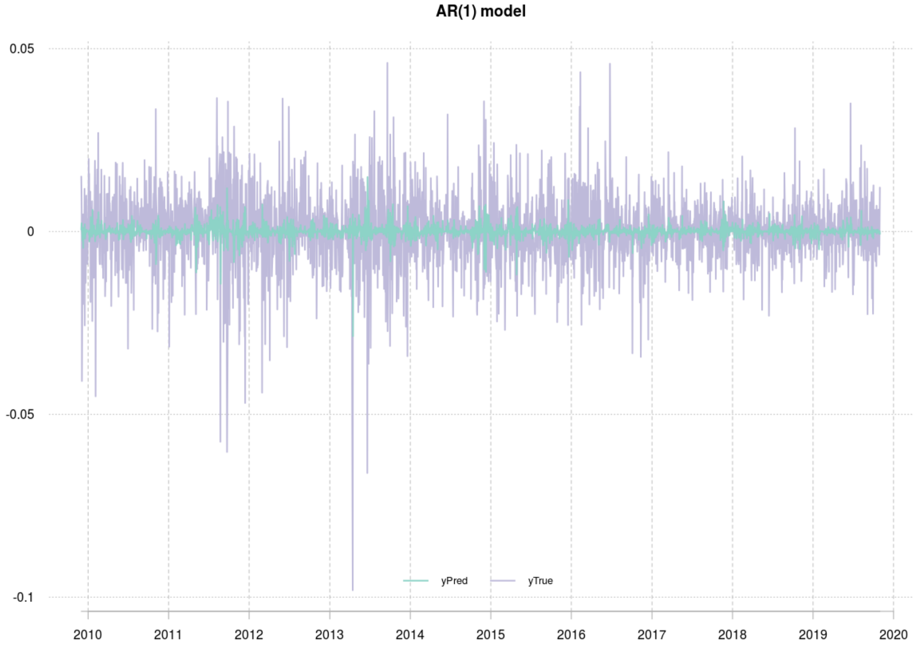


Figure 9: long-term forecast – AR(1)



4.3 Second-order Auto-regressive Model

The function API is the same as AR(1) model.
getForecastAR(pxs, lags = 2, trainDataLen = 20, forecastStep = 1, rollStep = 1, showGraph = FALSE)

4.3.1 Toy Example

$X = (-0.0001, -0.0183, 0.0064, -0.0182, -0.0024)$.
Q1. What is the predicted value for x_6?

Answer:
First deduce by hand. $T = 5$. First prepocess X by subtracting its mean $\mu_X = -0.0065$

$$X' = X - \mu_X = (0.0064, -0.0118, 0.0129, -0.0117, 0.0041)$$

Now extract the series and its lagged series from X': $Y_T = X'_{3:T}$, $Y_{T-2} = X'_{1:T-2}$ and $Y_{T-1} = X'_{2:T-1}$.
$\rho(1) = \text{acf}(Y_T, Y_{T-1}) = -0.9369$, $\rho(2) = \frac{\gamma(2)}{\gamma(0)} = \frac{\mathbb{E}(x_t x_{t-2})}{\mathbb{E}(x_t^2)} = \frac{0.000317}{0.000378} = 0.8389$

Therefore, the optimal parameters are:

$$\begin{cases} \alpha^* = \frac{\rho(1)(1-\rho(2))}{1-\rho(1)^2} = -1.235 \\ \beta^* = \frac{\rho(2)-\rho(1)^2}{1-\rho(1)^2} = -0.3181 \end{cases}$$

$$\mu_6 = \frac{1}{3}\sum_{t=3}^{5} x_t = -0.0047 \quad \mu_5 = \frac{1}{3}\sum_{t=3}^{5} x_{t-1} = -0.01 \quad \mu_4 = \frac{1}{3}\sum_{t=3}^{5} x_{t-2} = -0.004$$

Finally, the predicted value for x_6 is:

$$\hat{x}_6 = \mu_6 + \alpha^*(x_5 - \mu_5) + \beta^*(x_4 - \mu_4) = 0.0048$$

Now predict x_6 with our function *getForecastAR*. We get $\alpha^* = -1.97, \beta^* = -1.32$ and $\hat{x}_6 = -0.0009$. The value is different because the $\rho(2) = -0.664$ and $\rho(1) = -0.853$ in r.

4.3.2 Experiment on Gold Future Returns

The table below shows the mean absolute error (MAE) and mean square error (MSE) for each forecast:

	short-term	mid-term	long-term
MAE	5.1276e-03	6.0940e-03	0.00717417
MSE	4.9773e-05	7.0166e-05	0.00010502

Figure 10: short-term forecast – AR(2)

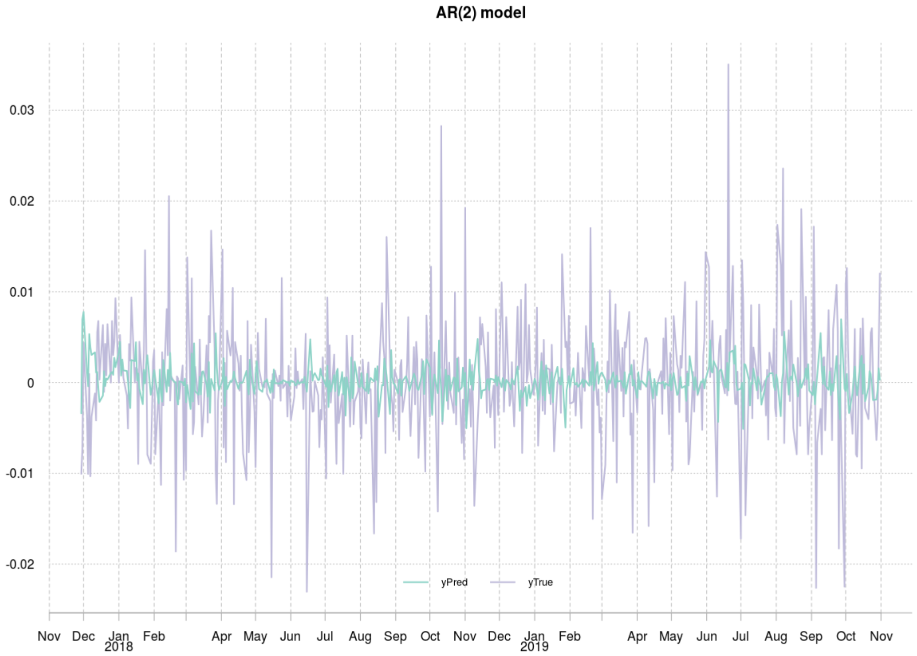


Figure 11: mid-term forecast – AR(2)

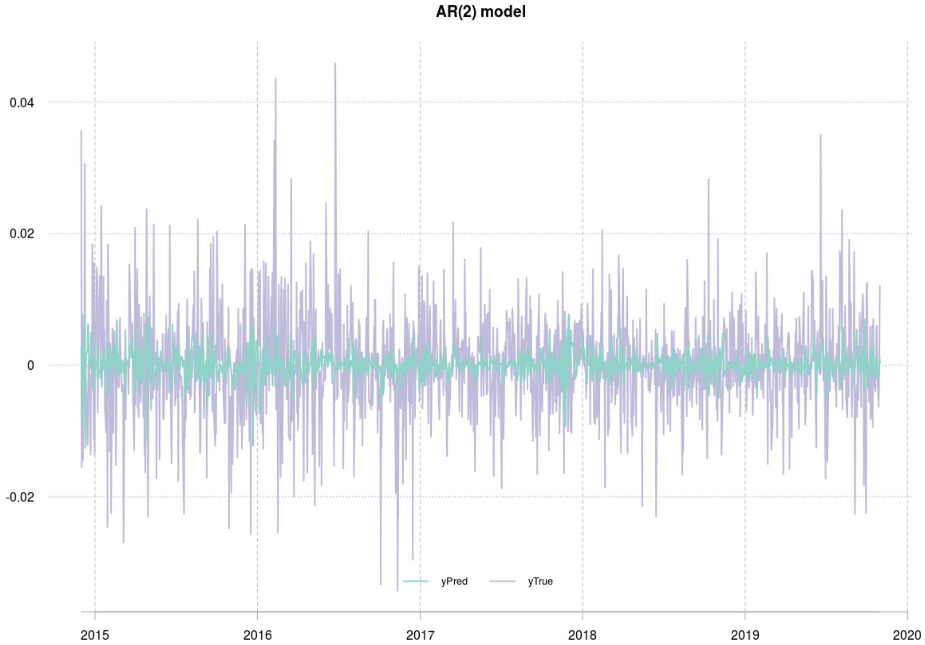
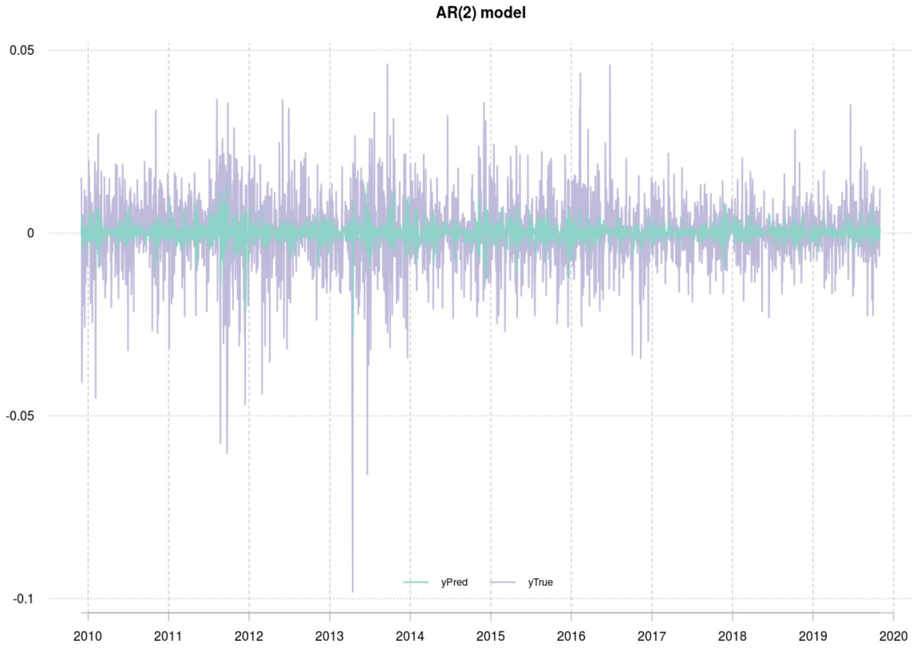


Figure 12: long-term forecast – AR(2)



4.4 PCA-based Multiple Regression Model

Function API:
getForecastPCA(pxs, maxLagOrder = 2, windowSize = 20, forecastHorizon = 1, level = 0.8)

pxs: data, including train data and test data
maxLagOrder: the maximal lag order
windowSize: rolling window size
forecastHorizon: forecast step for one rolling forecast
level: PCA precision control

4.4.1 Experiment on Gold Future Returns

The table below shows the mean absolute error (MAE) and mean square error (MSE) for each forecast:

	short-term	mid-term	long-term
MAE	4.5524e-03	0.00717584	0.00785713
MSE	3.8419e-05	0.00011482	0.00013621

Figure 13: short-term forecast – PCA

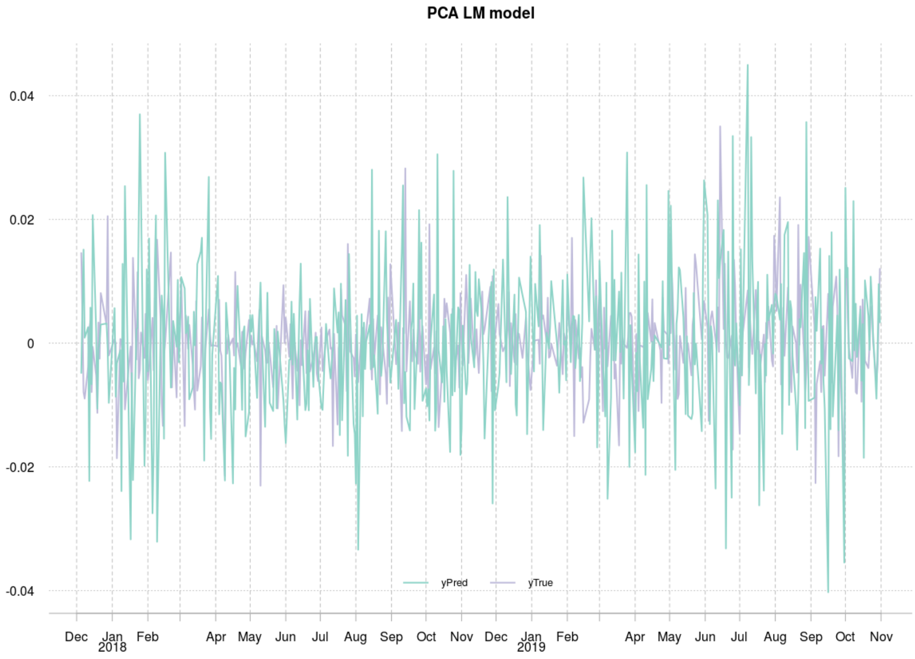


Figure 14: mid-term forecast – PCA

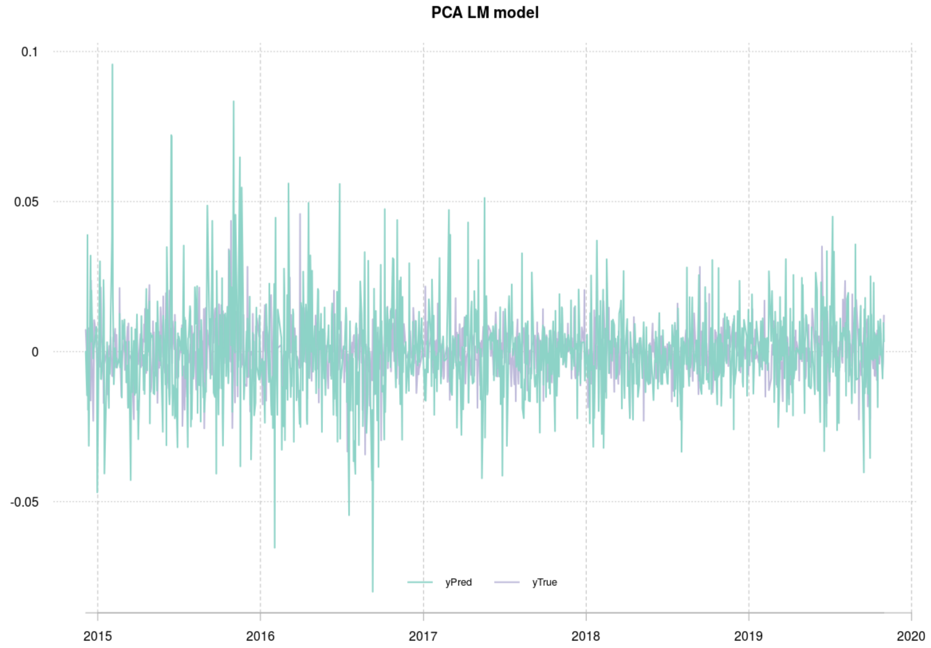
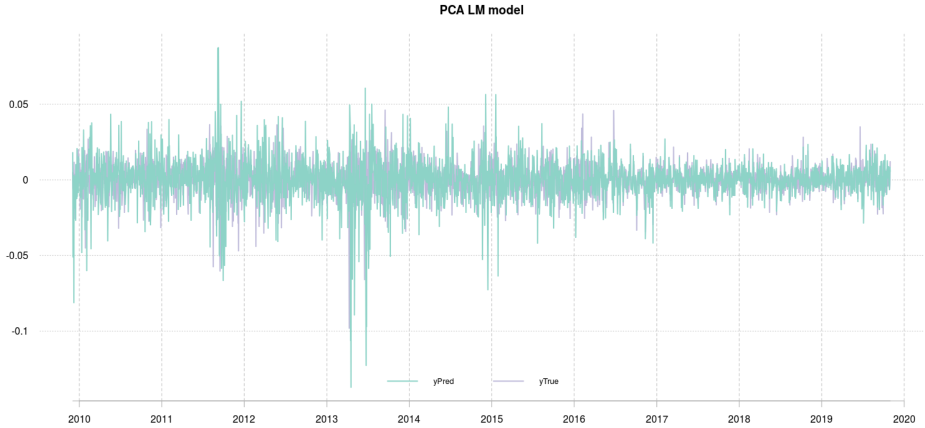


Figure 15: long-term forecast – PCA



4.4.2 Dataset Profiling

Let's have a closer look at the dataset used by PCA multiple regression method. We calculate the first-order auto-correlation coefficients and the second-order partial auto-correlation of all relevant series for the last 2 years, 5 years and 10 years and compare them to those of Gold futures.

The table below summarises the autocorrelation coefficients for lagged series (lag order = 1, round to 0.0001):

Series Name	2 yr	5 yr	10 yr
Gold	-0.0690	-0.0504	-0.0075
USGGBE10 Index	0.1005	0.1150	0.1078
USGG5YR Index	-0.0318	-0.0864	-0.0705
KC	-0.0907	-0.0209	-0.0352
RY	-0.0307	0.0032	0.0187
RX	-0.0161	0.0044	0.0137
HI	-0.0006	-0.0304	-0.0037
XID	-0.0390	-0.0200	-0.0445
KU	-0.0411	-0.0467	-0.0423
PA	-0.1062	-0.0205	0.0335
PL	-0.0927	-0.0490	0.0011
IH	-0.0041	-0.0063	-0.0206
SI	-0.1079	-0.0607	-0.0317
SMix	-0.0226	0.0226	0.0375
TU	-0.0757	-0.0838	-0.0488
W	-0.0425	0.0132	0.0049
JY	-0.0815	-0.0419	-0.0075

Note that for the last 2 years, the auto-correlation coefficients of relevant series are all negative except one and are mostly very close to that of Gold future. This suggests that the relevant series we selected have had very similar behavior for the past 2 years. The similarity decreases as the observation period gets longer.

The table below summarises the partial autocorrelation coefficients for lagged series (lag order = 2, round to 0.0001):

Series Name	2 yr	5 yr	10 yr
Gold	0.0368	0.0026	-0.0050
USGGBE10 Index	0.0596	-0.0337	-0.0476
USGG5YR Index	0.0535	-0.0081	-0.0063
KC	-0.0003	0.0188	0.0128
RY	0.0067	-0.0218	-0.0260
RX	-0.0091	-0.0374	-0.0298
HI	0.0135	0.0195	0.0119
XID	-0.0092	-0.0270	-0.0364
KU	0.0137	-0.0227	-0.0151
PA	0.0319	0.0141	-0.0156
PL	0.0699	0.0391	0.0314
IH	0.0843	0.0229	0.0300
SI	-0.0053	-0.0009	-0.0047
SMix	0.0271	-0.0542	-0.0408
TU	0.0500	0.0147	-0.0262
W	-0.0447	-0.0060	-0.0147
JY	0.0391	0.0437	-0.0050

The partial correlation coefficient of second order measures the correlation of the residuals which remains after removing the correlation between series and first order lagged series (i.e. lag order=1). A closer examination allows us to draw the conclusion that the relevant series have been very similar to Gold futures for the last 2 years. Same as first-order auto-correlation coefficients, the similarity suffers as we expand our observation period.

To sum up, we have a high quality dataset for short-term forecast. As for mid-term and long-term forecast, we don't have an ideal dataset.

4.5 Model Comparison and Conclusion

We regroup our experiments according to forecast duration and summarise the metrics in the following 3 tables.

For short-term forecast, the average MAE error and the average MSE error of each model are:

Model	avg MAE error	avg MSE error
CsteMean	5.0702e-03	4.8249e-05
AR(1)	4.9791e-03	4.8094e-05
AR(2)	5.1276e-03	4.9773e-05
PCA ML	4.5524e-03	3.8419e-05

We can see that PCA-based multiple regression model beats all other methods in the short-term forecast task, whereas AR(2) model performs the worst. Though very simple, constant mean model has a fairly good performance. This is because our data bear resemblance to white noise. The poor performance of AR(2) model is mainly due to the fact that Gold future returns don't have a strong correlation with its lagged data. This fact also explains why the AR(1) model, despite being more sophisticated, has similar performance to constant mean model. The satisfactory results of PCA-ML model, on one hand, showcase the power of PCA analysis. On the other hand, its excellent performance can be attributed to fact that the relevant series behaved likewise during the experiment period (from November 2017 to October 2019). The dataset quality helps ensure the forecast accuracy.

For mid-term forecast, the average MAE error and the average MSE error of each model are:

Model	avg MAE error	avg MSE error
CsteMean	6.0636e-03	6.9511e-05
AR(1)	5.9162e-03	6.8220e-05
AR(2)	6.0940e-03	7.0166e-05
PCA ML	7.1758e-03	11.482e-05

For mid-term forecast task, the errors of every model increase. Contrary to the short-term forecast, PCA-based multiple regression model performs the worst. Indeed, from the graphs we can easily observe that the forecast series (green line) in general has bigger fluctuations than the ground-truth series (purple line). We think the drastic error increase in PCA forecast, compared to the other methods, is due to a lower dataset quality. The relevant series haven't had similar behaviors than Gold future returns over the past 5 years, therefore the PCA is unable to capture Gold future's dynamics. The AR models have better performances than the previous case. This confirms our observation that Gold future returns, despite being weakly correlated to the its lagged series, are stationary. Moreover, the AR(1) model has the best performance among the 4 methods.

For long-term forecast, the average MAE error and the average MSE error of each model are:

Model	avg MAE error	avg MSE error
CsteMean	0.00709086	0.00010291
AR(1)	0.00700516	0.00010231
AR(2)	0.00717417	0.00010502
PCA ML	0.01076292	0.00027106

Every model has a bigger MSE error and MAE error. Similarly to mid-term forecast task, AR(1) method outperforms all other models and PCA multiple regression method performs the worst of all. From our experiment, we can draw similar conclusions for long-term forecast as for mid-term forecast.

Overall, we can see that PCA multiple regression is the best forecast strategy for short-term forecast and AR(1) model is more suitable for mid-term and long-term forecast. However, the accuracy of PCA-based prediction depends on the dataset quality, therefore we have to choose our relevant series carefully before applying the forecast. Nonetheless, the log-normality of Gold future returns ensures that the simpliest model – Constant Mean model always has an acceptable result.

5 Investment Strategy Building

Based on our 4 forecast methods, we built and analysed 4 different Gold future trading strategies for the period 2018-12-01 to 2020-02-05. We also compared each strategy with the underlying strategy, i.e., the true Gold future price evolution. Same as before, we set the rolling window 20 and the rolling forecast step 1.

We considered the simplest strategy: either all-in or all-out. Define a signal time series that takes values in $\{0, 1\}$. 1 means entering the position (we buy with all the money we have). 0 means exiting the position (we sell all the assets we hold) [21]. On day t, there are altogether 4 possible actions. The actions and the corresponding previous signals are:

- buy: 0 at day $t-2$ and 1 at day $t-1$. Enter the position on day t
- hold: 1 at day $t-2$ and 1 at day $t-1$. Hold the position on day t
- sell: 1 at day $t-2$ and 0 at day $t-1$. Exit the position on day t
- keep clear: 0 at day $t-2$ and 0 at day $t-1$. Keep empty position on day t

If at day t, the sign flips from 0 to 1 or from 1 to 0, we have to calculate and accumulate the return. Nothing happens when the sign remains constant $((0,0)$ or $(1,1))$. Let's show the logic of strategy return calculation with an example [22]:

- Generate the allocation (0 or 1) on day t (10/02/2020) after the close
- Enter/exit the position before the close on day $t+1$ (11/02/2020)
- Calculate the returns after the close on day $t+2$ (12/02/2020)

We will rely on the following performance indicators to evaluate each strategy:

1. Annualized Return
   This is the most direct indicateur of strategy profitability. For example, if a stock has a price of 100 at the beginning of a year and 110 at the end of the year, the annualized return is 10%.

2. Annualized Volatility
   This metrics measures the uncertainty of the strategy return on a yearly basis. It is a general risk indicator.

3. Sharpe Ratio
   This indicator measures the return of an investment compared to its risk [23]. The value is the average return earned in excess of the risk-free rate per unit of volatility or total risk. If an investment has a sharpe ratio of 1, it means that for every unit of risk the investment will yield an equal amount of return. Generally speaking [24],

   - A sharpe ratio under 1 is sub-optimal
   - A sharpe ratio between 1 and 2 is good (acceptable)
   - A sharpe ratio between 2 and 3 is very good.
   - A sharpe ratio greater than 3 is excellent.

4. Sortino Ratio
   This is a variation of the Sharpe ratio. The Sortino ratio is the average return earned in excess of the risk-free rate per unit of down volatility or downside risk (an asset's standard deviation of negative portfolio returns)[25]. As the ratio focuses only on the negative deviation of a portfolio's returns from the mean, it is thought to give a better view of a portfolio's risk-adjusted performance since positive volatility is a benefit. Similar to sharpe ratio, the greater the sortino ratio, the better. Usually an investment with a sortino ratio greater than 2 is considered good.

5. Maximum Drawdown
   Drawdown is a measure of downside volatility. It helps determine the historical financial risk of an investment [26]. Maximum drawdown is the ratio between the peak-to-bottom difference and the peak price.

The table below summarizes the strategy performance (round to 0.01):

	CsteMean	AR(1)	AR(2)	PCA-MLR
Annualized Return (%)	10.82	9.36	-1.33	18.17
Annualized Volatility (%)	9.57	7.24	7.81	8.45
Sharpe Ratio	1.13	1.29	-0.16	2.15
Sortino Ratio	1.79	1.94	-0.23	3.85
Maximum Drawdown (%)	-6.84	-4.52	-7.26	-4.31

The strategy generated by constant mean model has a good (acceptable) performance in terms of sharpe ratio and sortino ratio. Among the 4, this strategy generates a fairly good annualized return but at the same time has bigger volatility.

The strategy generated by AR(1) model performs better than constant mean model in terms of sharpe ratio and sortino ratio. A lower volatility ratio and a smaller maximum drawdown suggest that it has a better risk-adjusted return performance than constant mean strategy. Meanwhile, its annualized return is still satisfying (nearly 10%).

The strategy generated by AR(2) model has a rather poor performance in terms of sharpe ratio and sortino ratio. The investment is sub-optimal. In our experiment, this strategy doesn't generates any profit at all.

The strategy generated by PCA-based multiple regression model performs the best. The investment is considered very good in terms of sharpe ratio and sortino ratio. Not only does it generate the highest annualized return, but it also boasts the lowest maximum drawdown and a reasonable volatility ratio. Overall, we conclude that it has the best risk-adjusted return performance. Recall that in the forecast experiment PCA-based multiple regression model has excellent performance in short-term (one-year) forecast. Here this method again proves its effectiveness in short-term strategy building.

Now visualise all the strategies. The green line represents Gold future and the orange line the strategy.

Figure 16: Constant Mean Strategy

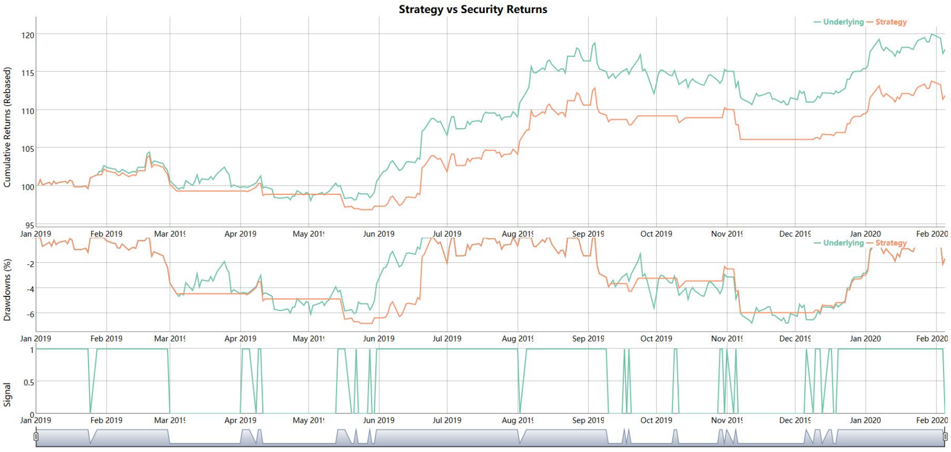


Figure 17: AR(1) Strategy

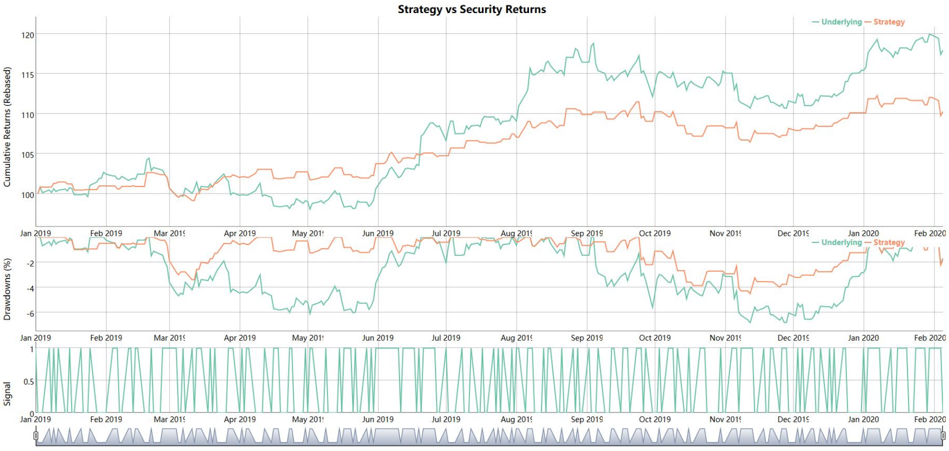


Figure 18: AR(2) Strategy

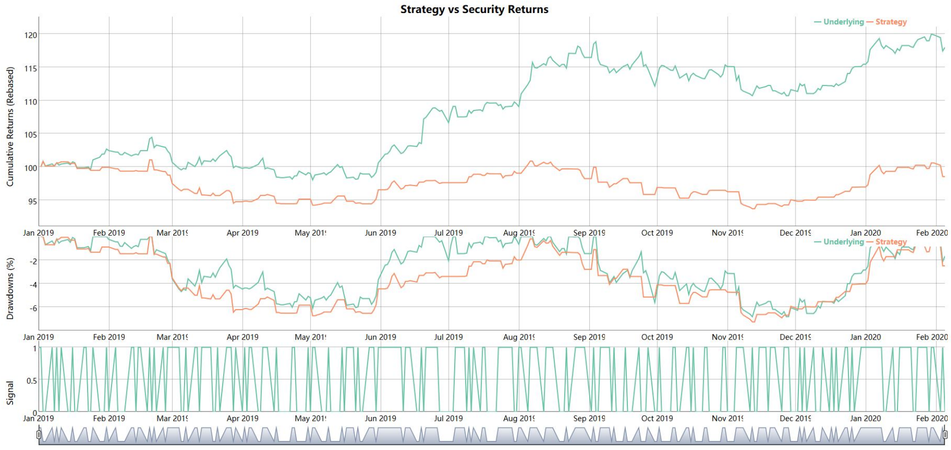
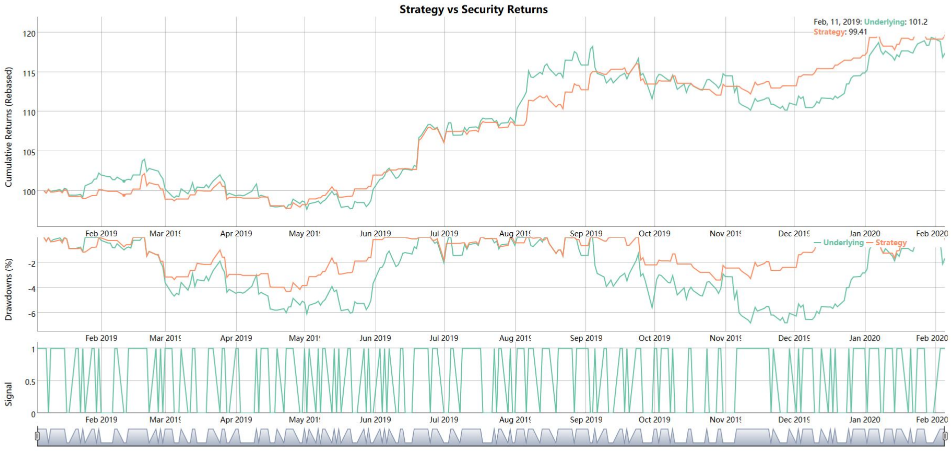


Figure 19: PCA-multiple regression Strategy



For each strategy we illustrate the (rebased) cumulative return, drawdowns (in %) and the transaction signals.

From the cumulative return plots we observe that all strategies except AR(2) are able to follow the Gold future's alpha generation trend.

The drawdown plots qualitatively shows the relative downside volatility of each strategy when compared to Gold future prices. It's quite clear that AR(1) and PCA-based multiple regression strategy both have lower drawdowns than Gold futures most of the time.

Conclusion

In this PRIM project, we first reviewed N.Sopipan's paper on PCA forecast. Then, with a similar approach, we address the Gold future returns forecast by applying 4 different models: constant mean model, AR(1) model, AR(2) model and PCA-based multiple regression model. In particular, we incorporate rolling prediction to all models to make the forecast more generic. After that, we go further and analyse our methods from a financial point of view by applying them to strategy building. The forecast experiments show that the PCA-based multiple regression model yields the best results in short-term forecast and that AR(1) model is the most promising method for mid-term and long-term prediction. In our strategy building experiments, the PCA-based multiple regression model outperforms all the other strategy both in terms of alpha generation and risk management. This somehow again shows the excellent short-term forecast ability for PCA-based multiple regression model.

References

[1] Stationarity in time series analysis
https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322

[2] multiple linear regression - MLR Definition
https://www.investopedia.com/terms/m/mlr.asp

[3] mean absolute error
https://en.wikipedia.org/wiki/Mean_absolute_error

[4] mean squared error
https://en.wikipedia.org/wiki/Mean_squared_error

[5] multiple regression model
https://fr.m.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire#Mod%C3%A8le_lin%C3%A9aire_multiple

[6] Augmented Dickey–Fuller test
https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

[7] Robert H. Shumway, David S. Stoffer. *Time Series Analysis and Its Applications - 2017*, Springer, Cham, https://doi.org/10.1007/978-3-319-52452-8

[8] Rdocumentation-acf
https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/acf

[9] Measures of Skewness and Kurtosis
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

[10] Measures of Shape: Skewness and Kurtosis
https://brownmath.com/stat/shape.htm

[11] kurtosis
https://en.wikipedia.org/wiki/Kurtosis

[12] Jarque–Bera test
https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

[13] Rdocumentation-Bartlett's sphericity test
https://www.rdocumentation.org/packages/REdaS/versions/0.9.0/topics/Bartlett-Sphericity

[14] KMO and Bartlett's Test
https://www.ibm.com/support/knowledgecenter/SSLVMB_26.0.0/statistics_casestudies_project_ddita/spss/tutorials/fac_telco_kmo_01.html

[15] Kaiser-Meyer-Olkin-Statistics
https://www.statisticshowto.datasciencecentral.com/kaiser-meyer-olkin/

[16] R practice: Factor analysis
http://minato.sip21c.org/swtips/factor-in-R.pdf

[17] AR model
https://mcs.utm.utoronto.ca/~nosedal/sta457/ar1-and-ar2.pdf

[18] PACF
https://mcs.utm.utoronto.ca/~nosedal/sta457/pacf.pdf

[19] N.Sopipan.*FORECASTING THE FINANCIAL RETURNS FOR USING MULTIPLE REGRESSION BASED ON PRINCIPAL COMPONENT ANALYSIS*, Journal of Mathematics and Statistics, 9 (1): 65-71, 2013. Available at https://pdfs.semanticscholar.org/9815/fe6cb1d3009d1d722d96d41f4ad9beed8113.pdf

[20] rolling forecast
https://whatis.techtarget.com/definition/rolling-forecast

[21] payout strategy
https://wiki.alphien.com/ALwiki/Guide_to_lag_in_Alphien

[22] strategy return calculation logic
https://wiki.alphien.com/ALwiki/Backtest_Process#Daily_Returns_Computation

[23] Sharpe ratio
https://www.investopedia.com/terms/s/sharperatio.asp

[24] Sharpe ratio
https://www.investopedia.com/ask/answers/010815/what-good-sharpe-ratio.asp

[25] Sortino ratio
https://www.investopedia.com/terms/s/sortinoratio.asp

[26] Drawdown
https://www.investopedia.com/terms/d/drawdown.asp