Application of PCA Analysis in Forecasting Gold Future Returns

LU Dongrui

Telecom-Paris

10/02/2020

Outline

Introduction - Background

- Motivations
  Gold is a well-known safe and stress-enduring asset. Forecasting Gold futures is of great significance to portfolio management in terms of both risk hedging and profit generation.

- Project Description
  1. Build data profile for Gold future returns
  2. Forecast gold futures returns by PCA-based multiple regression analysis
  3. Compare with other benchmarks in terms of MSE and MAE
  4. Build investment strategy with all forecast models and financially evaluate their performances

Introduction - Problem Formulation

Gold futures return forecast (k-step ahead)

Given a time series $X = x_t, t = 1, 2, ..., T$ modelling Gold future returns, we would like to predict $x_{t+1}, x_{t+2}, ..., x_{t+k}$.

Notations:

- X: time series
- x_t: X's price at time t
- T: window size, train data length
- k: forecast step
- $\hat{x}_t$: forecast price at time t

Forecast Modelling - Constant Mean

Hypotheses

Data normality and stationarity

Model

$$x_t = \mu + \epsilon_t$$

for $t = 1, 2, ..., T, T+1, ...T+k$
μ: constant, mean of X
ϵ_t: error, a white noise series with an i.i.d 0 mean

Theoretical Forecast

$$\hat{x}_t = \mu^* = \frac{1}{T}\sum_{t=1}^{T} x_t$$

for $t = T+1, ..., T+k$

Forecast Modelling - AR(1)

Hypotheses

Data normality, stationarity, exogeneity and homoscedasticity

Model

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

α, β: constant coefficient ($\beta = 0$ for centered data)
ϵ_t: error, a white noise series with an i.i.d 0 mean

Forecast Modelling - AR(1)

Theoretical Forecast

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

with

$$\begin{cases} \alpha^* = \frac{\mathrm{Cov}(X_t, X_{t-1})}{\mathrm{Var}(X_{t-1})} = \mathrm{acf}_1(X) \\ \beta^* = \frac{1}{T-1} \sum_{t=2}^{T} (x_t - \alpha^* x_{t-1}) \end{cases}$$

Equivalently,

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

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

acf_k: k^{th}-order auto-correlation coefficient

Forecast Modelling - AR(2)

Model

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

α, β, ω: constant coefficient ($\omega = 0$ for centered data)

ϵ_t: error, a white noise series with an i.i.d 0 mean

Theoretical Forecast

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

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

$$\begin{cases} \alpha^* = \frac{\mathrm{acf}_1(X)(1-\mathrm{acf}_2(X))}{1-\mathrm{acf}_1(X)^2} \\ \beta^* = \frac{\mathrm{acf}_2(X)-\mathrm{acf}_1(X)^2}{1-\mathrm{acf}_1(X)^2} = \mathrm{pacf}_2(X) \end{cases}$$

pacf_k: k^{th}-order partial auto-correlation coefficient

Forecast Modelling - PCA multiple regression

Model - Simple multiple regression

$$x_t = \mu_t + \epsilon_t = \mu + \sum_{i=1}^{n} \alpha_i y_{it} + \sum_{j=1}^{p} \beta_j x_{t-j} + \epsilon_t$$

x_{t-j}: j^{th}-order lagged series
y_{it}: explanatory variable
μ, α_i, β_j: constant coefficient ($\mu = 0$ for centered data)
ϵ_t: error, a white noise series with an i.i.d 0 mean and constant variance

Hypotheses

No mutlicollinearity of explanatory variables

Forecast Modelling - PCA multiple regression

Model and Theoretical Forecast - PCA multiple regression

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

z_{it}: new explanatory variable obtained by PCA
ω_i: constant coefficient

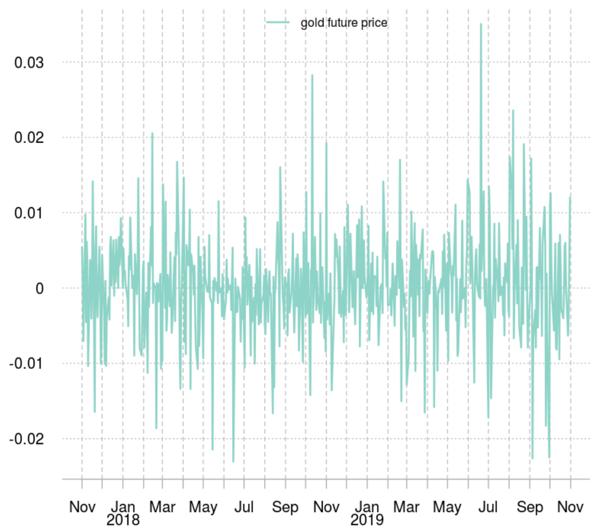
Hypotheses

PCA applicability

Forecast Experiments - Data

Take Gold futures log-returns, $X = \ln(\frac{P_t}{P_{t-1}})$. Consider 3 forecast tasks:

- Short term forecast (2 years): 2017-10-31 $\sim$ 2019-10-31
- Mid term forecast (5 years): 2014-10-31 $\sim$ 2019-10-31
- Long term forecast (10 years): 2009-10-31 $\sim$ 2019-10-31



Forecast Experiments - Data Profiling

- Stationarity
  Augmented Dickey-Fuller (ADF) Test, autocorrelation matrix
- Normality
  mean, standard deviation, skewness, kurtosis, Jarque-Bera (JB) Normality Test
- PCA Applicability
  Barlett's Sphericity Test, Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Conclusion

- Gold future returns are stationary. They have little (possibly no) correlation with their lagged series.
- Resemble white noise but have more outliers than a standard normal distribution
- Sufficient samples to apply PCA analysis (17 relevant series)

Forecast Experiments - General Settings and Metrics

Rolling forecast default settings:

- rolling window size $= 20$
- forecast horizon $= 1$

Metrics:

- MAE, mean absolute error
- MSE, mean square error

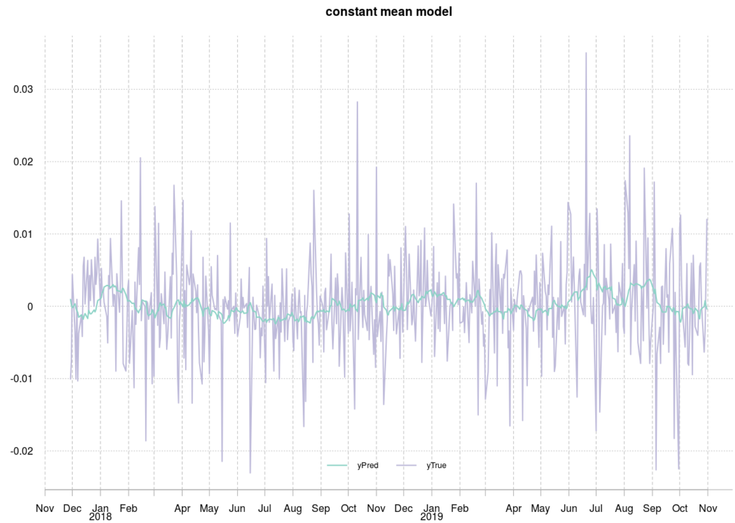
Forecast Experiments - R Function API

- getForecastCsteMean(pxs, windowSize, forecastHorizon, rollStep, showGraph = FALSE)
- getForecastAR(pxs, lags, windowSize, forecastHorizon, rollStep, showGraph = FALSE)
- getForecastPCA(pxs, maxLagOrder, windowSize, forecastHorizon, level = 0.8, showGraph = FALSE)

Forecast Experiments - Results

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

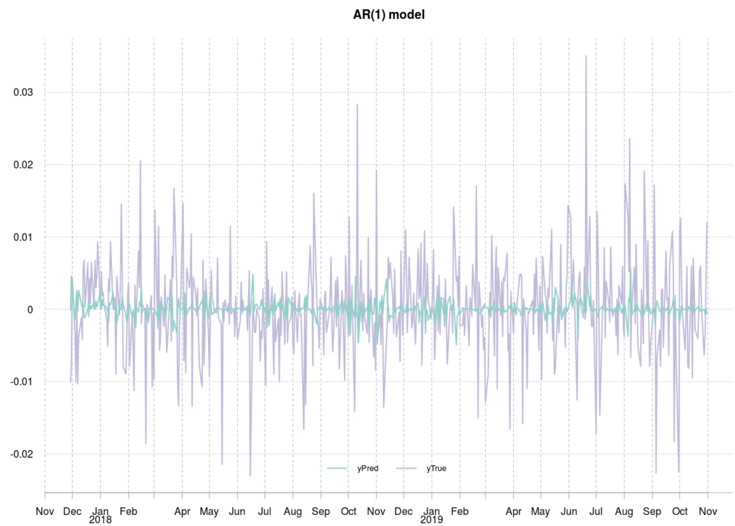
Table 1: Constant Mean Model



Forecast Experiments - Results

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

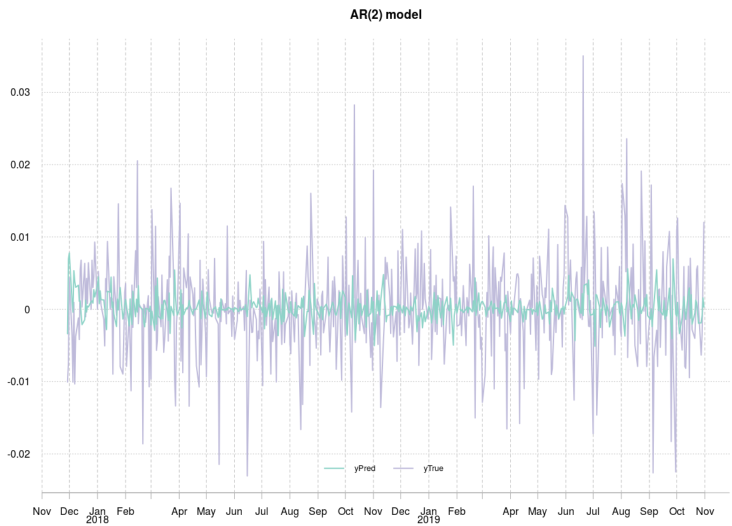
Table 2: AR(1) Model



Forecast Experiments - Results

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

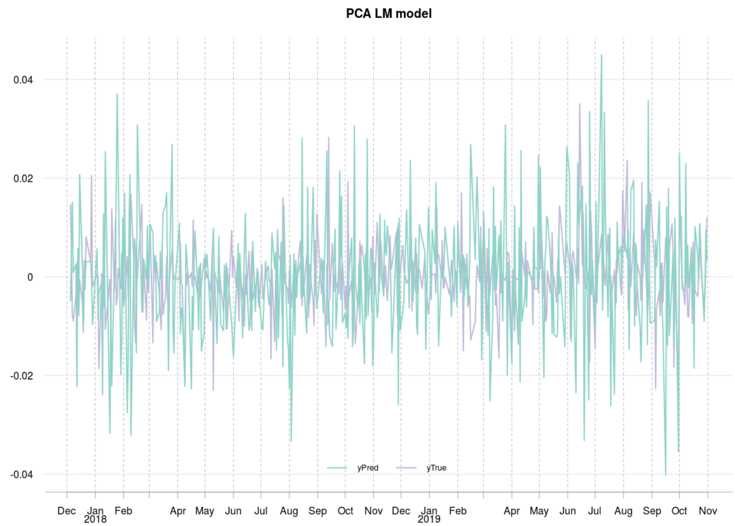
Table 3: AR(2) Model



Forecast Experiments - Results

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

Table 4: PCA multiple regression Model



Forecast Experiments - Summary

1. For every model, MAE and MSE error increase as the forecast period gets longer
2. PCA multiple regression model performs the best in short-term forecast, but its power relies on dataset quality and could significantly suffer as forecast period expands
3. AR(1) model performs the best for mid-term and long-term forecast
4. AR(2) model performs worse than AR(1) model in all experiments
5. Constant Mean model has satisfactory performances in all experiments

Investment Strategy Building - Strategy Description

Strategy

A signal time series S_t that takes values in $\{0, 1\}$. 1 means entering the position. 0 means exiting the position. On day t, the possible 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.

Strategy Return Calculation Logic

- Generate the allocation (0/1) 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)

Investment Strategy Building - Performance Metrics

Generate strategy based on every forecast model and compare to the underlying strategy in terms of:

- Profitability:
  Annualized Return
- Risk:
  Annualized Volatility, Maximum Drawdown
- Risk-adjusted profitability:
  Sharpe Ratio, Sortino Ratio

Build strategy for the period 2018-12-01 to 2020-02-05.

Investment Strategy Building - Experiment

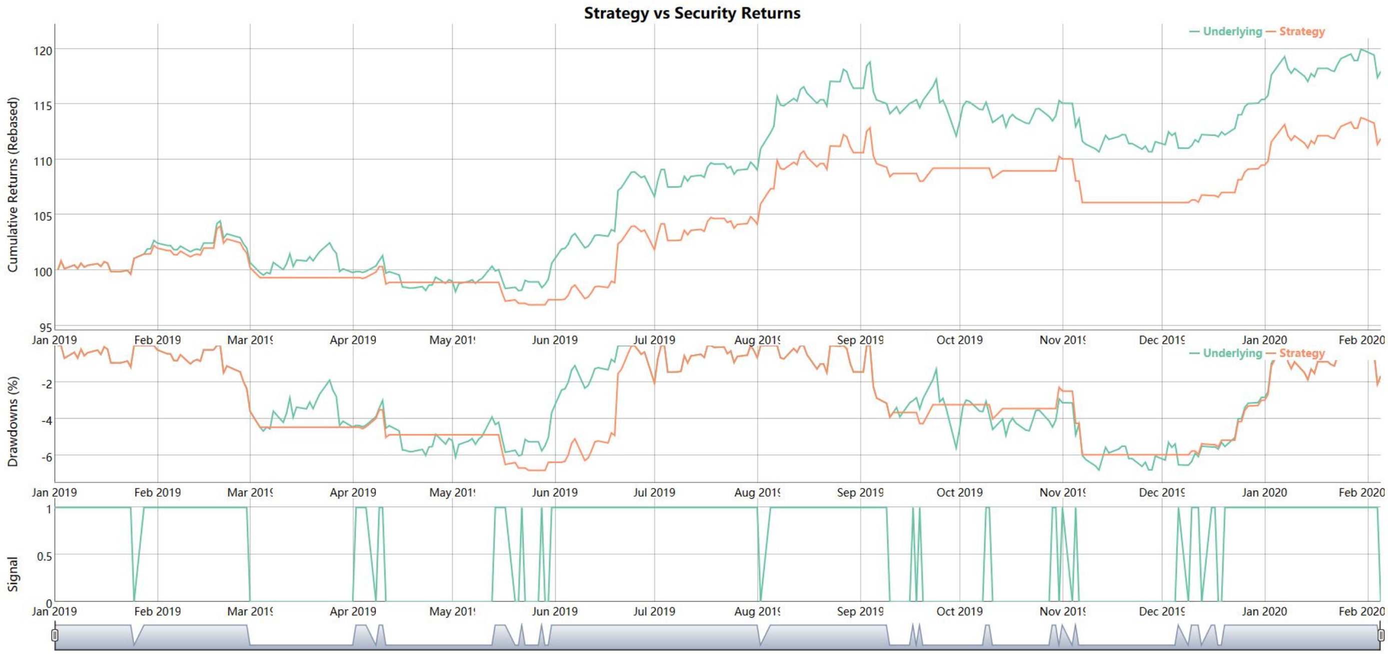


Figure 1: Constant Mean Strategy

Investment Strategy Building - Experiment

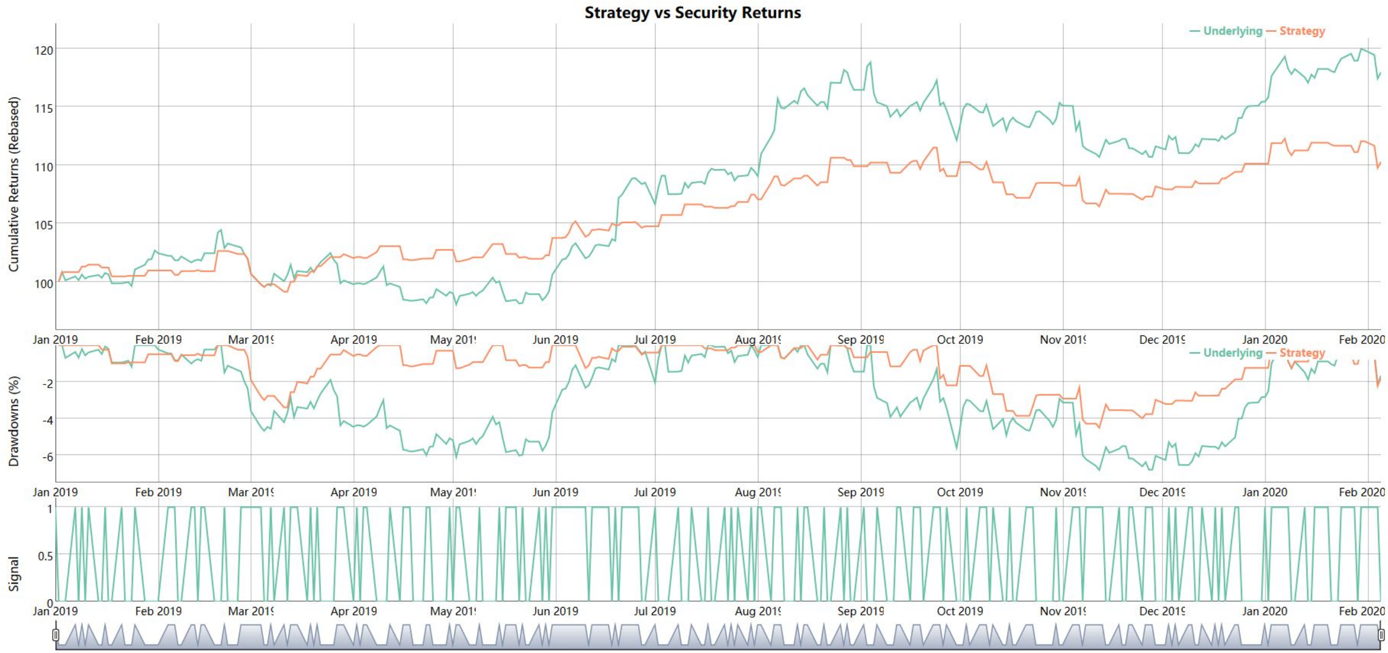


Figure 2: AR(1) Strategy

Investment Strategy Building - Experiment

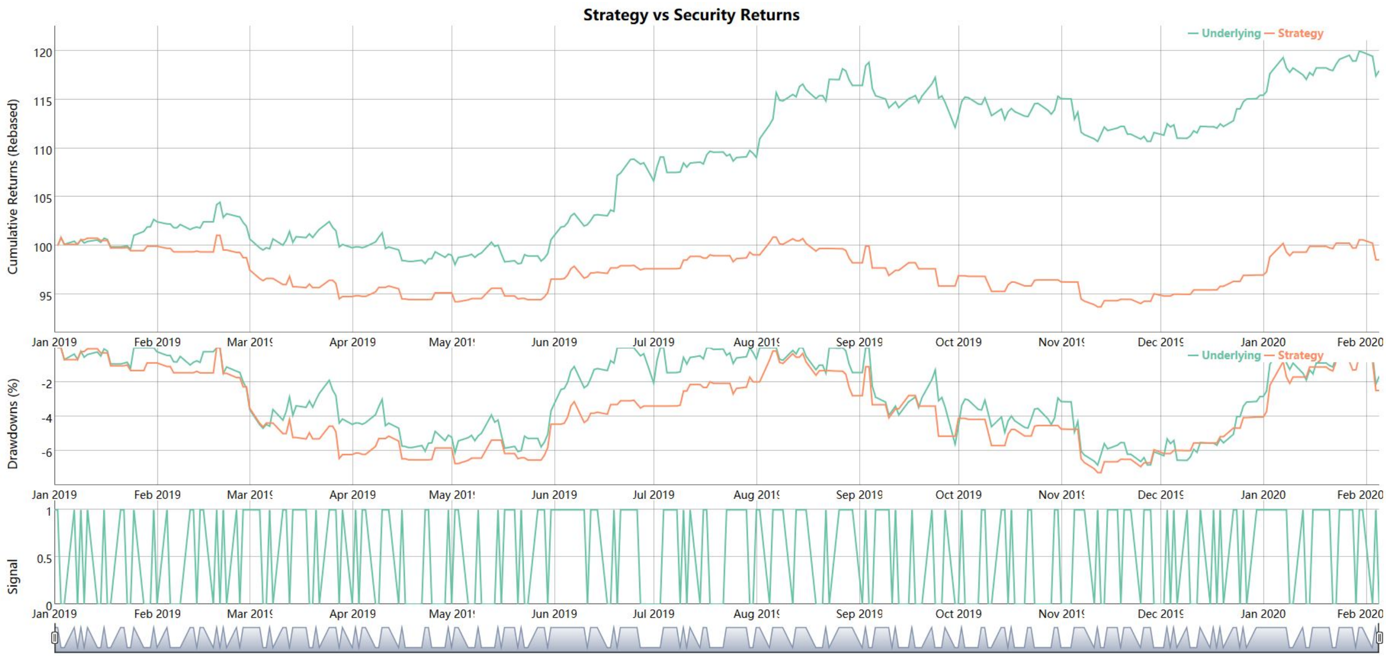


Figure 3: AR(2) Strategy

Investment Strategy Building - Experiment

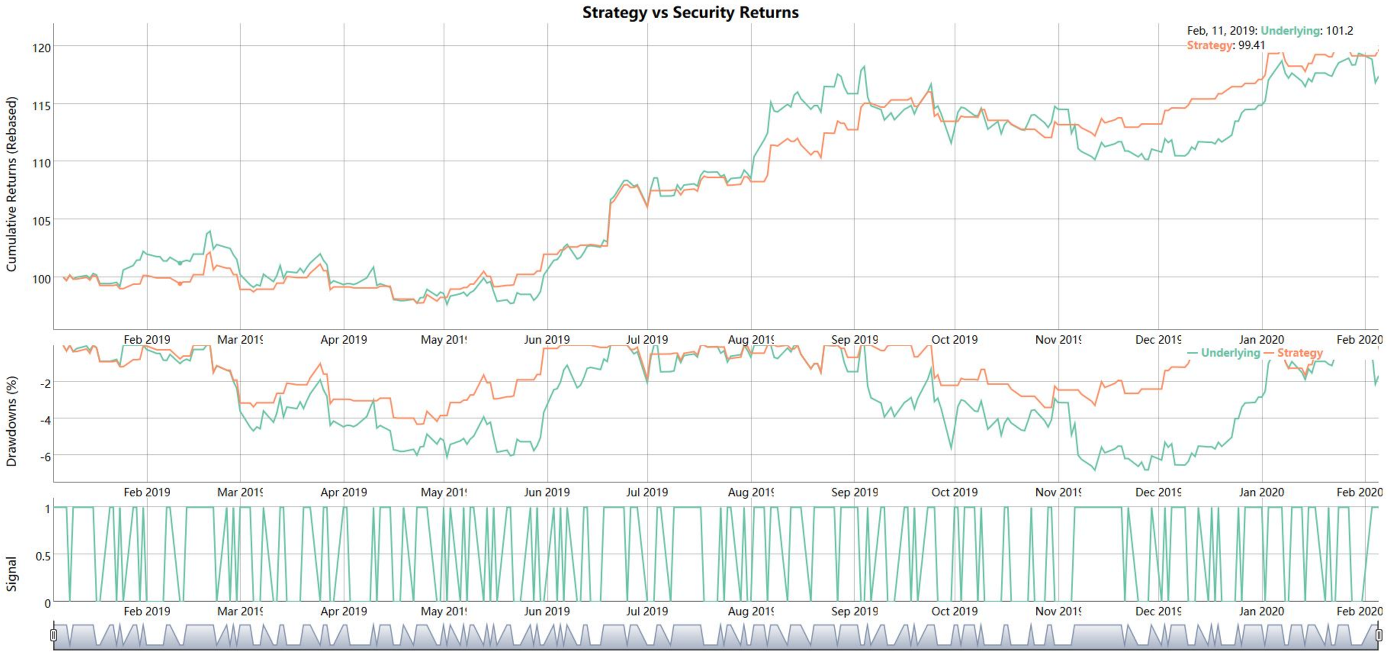


Figure 4: PCA multiple regression Strategy

Investment Strategy Building - Experiment

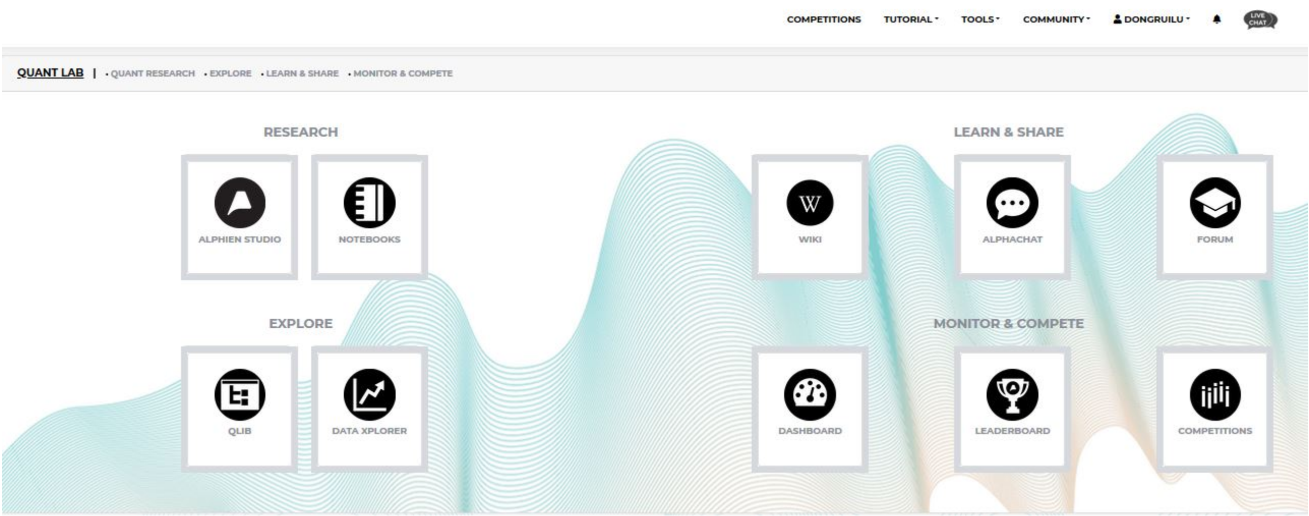
	CsteMean	AR(1)	AR(2)	PCA-MR
Annualized Return (%)	10.82	9.36	-1.33	18.17
Annualized Volatility (%)	9.57	7.24	7.81	8.45
Sharpe Ratio	1.13	1.29	-0.16	2.15
Sortino Ratio	1.79	1.94	-0.23	3.85
Maximum Drawdown (%)	-6.84	-4.52	-7.26	-4.31

Table 5: Strategy Performance

Conclusion

- Constant Mean model generates much profit but has more risks.
- AR(1) model has good risk-adjusted performance.
- AR(2) model is sub-optimal.
- PCA multiple regression model performs the best

Alphien platform



Alphien platform

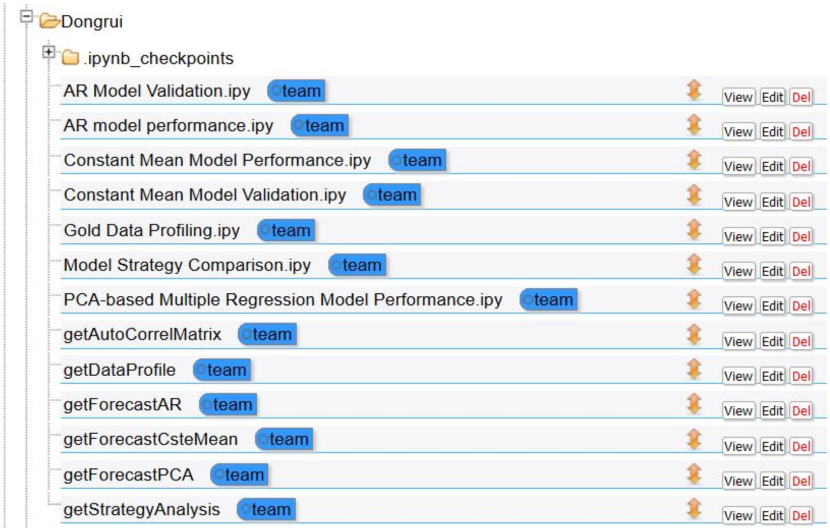


Figure 5: project folder

Summary

- Reviewed N.Sopipan's paper on PCA based multiple regression
- Built data profile for Gold future returns and checked PCA applicability with 17 relevant series
- Implemented forecast methods in R
- Developed theoretical analysis for every forecast method and tested forecast functions' validity with the help of closed-form predictions
- Conducted short term, mid term and long term forecast experiments on the dataset
- Built investment strategies based on forecast methods and financially analysed their performance