Regression and Hidden Markov Models for Gold Price Prediction

Li, Shen
China Asset Management Co., Ltd.
Beijing, 100033, China
shenli19941220@163.com

Kun, Shen
China Asset Management Co., Ltd.
Beijing, 100033, China
kongk@chinaamc.com

Chao, Yi
China Asset Management Co., Ltd.
Beijing, 100033, China
yic@chinaamc.com

Yixin, Chen
China Asset Management Co., Ltd.
Beijing, 100033, China
chenyix@chinaamc.com

Abstract—In the long run, gold price is positively related to inflation rates because gold is a perfect asset to hedge against inflation. In the short run, gold price fluctuates a lot. Many factors can cause gold price volatility, such as economic and political uncertainties, exchange rates, interest rates and so on. Here we try several models to predict monthly gold prices, including linear regression model and ARIMA model. We also try to predict monthly gold returns with hidden Markov model. It turns out that HMM is much better.

Keywords—gold, prediction, HMM, OLS, ARIMA

I. INTRODUCTION

It is well known that gold can be a perfect asset to hedge against inflation. Therefore, gold price has been rising for decades as the inflation rate has been going up at the same time. In March 2020, since US Federal Reserve pledged unlimited quantitative easing to keep markets functioning, there has been a significant rise in the gold price. At the same time, gold price also fluctuates a lot in the short run. For example, gold price plunged suddenly on August 11, 2020, which was affected by US real interest rates because it hit the bottom and started to rise again. Therefore, several factors need taking into account when we predict gold price or returns.

In [2], a framework is developed based on the simple view of “supply and demand” where gold can be perfectly hedged against inflations in the long run and gold price is volatile in the short run. Therefore, gold price is positively linked with CPI, M2 and negatively linked with interest rates, including fed rates and US dollar 3 month LIBOR rates. In the short run, gold supply is affected by its historical price and leasing rates, both of which have a negative relationship with gold supply. And short-run gold demand includes mainly investment demand, production demand and jewelry demand. So we take into account PMI, GDP, dollar index, credit risk, S&P 500 index and economic political uncertainties as factors affecting gold demand.

The prediction of Gold price is a typical time-series application. And several models are considered here. The first model is Autoregressive Integrated Moving Average model-ARIMA. In this model, gold price is predicted based on its own historical price and irrelevant to other variables. Three key coefficients p, d, q determining ARIMA model are the order of autoregression, the order of differencing and the order of moving average, respectively. It starts with ADF test to check if the time-series is stationary or conduct d-order differencing if otherwise.

Except for univariate time-series model, multi-variate time-series models are appropriate to fit the supply-and-demand system here, including Vector Autoregression and Vector Error Correction model, both of which are perfect for gold price prediction. In such multi-variate models, all variables are put into models and then are predicted together more than one step forward in the future.

Also, we can try linear model to predict gold price [2]. A linear model is different from time-series models in which we can only specify one dependent variable and predict one step forward in the future.

Although we have been talking about predicting exact price or returns so far, there is another perspective, which is to predict the probability of rising and falling of prices. So we try Hidden Markov Model as one of these methods. In HMM, we split economic conditions into several combinations and then predict possible gold performance in each possible combination.

So, let’s compare all these models and see how they perform in the prediction.

II. DATA DESCRIPTION

This paper is mainly based on [2]. In the long run, gold price is mainly determined by inflation rates. In the short run, gold price is influenced by supply and demand.

So, there are several variables: CPI index, 3-month dollar LIBOR, dollar index, M2, GDP, PMI index, credit risk, S&P 500 index and economic political uncertainties ranging from Jan, 1985 to Jul, 2020.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mark</th>
<th>Description</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPI index</td>
<td>cpi</td>
<td>Monthly data, positively linked with gold price.</td>
<td>As inflation rate.</td>
</tr>
<tr>
<td>3-month dollar LIBOR</td>
<td>libor3</td>
<td>Daily data, negatively linked with gold price.</td>
<td>As gold supply cost.</td>
</tr>
</tbody>
</table>
Dollar index dollar_index | Daily data, negatively linked with gold price. | As gold nominal price.
---|---|---
M2 m2 | Monthly data, positively linked with gold price. | As inflation rate.
GDP q_gdp | Quarterly data, positively linked with gold price. | As gold demand.
PMI index pmi | Monthly data. | As gold demand.
Credit risk credit_risk | Monthly data, t-bill rate minus AAA bond rate, negatively linked with gold price. | As gold demand.
S&P500 index sp500 | Daily data, positively linked with gold price. | As gold demand.
Economic political uncertainties epu | Monthly data, positively linked with gold price. | As gold demand.

All of the data are transformed into monthly data. Daily data are averaged on monthly basis and quarterly data are duplicated on monthly basis.

In [2], Levin and Wright (2006) mainly studied the relationship between gold prices and other macroeconomic variables. They used same-time or one-period lagged macroeconomic variables to predict gold prices in the next period. But when we trade gold, we have to think about the availability of macroeconomic data. Therefore, our paper is distinctively differentiated from [2] in data usage. When we think about how to use data appropriately, this issue is tricky to handle.

III. MODELS

Gold price belongs to time-series data, so we mainly focus on time-series models. Usually, we split data into training and testing data once, then train models and estimate results just once. However, in our model, we try to estimate gold price in a rolling way. For example, we take 1st–nth month data to train the model, then predict n+1th–n+3th results, and then take 2nd–n+1th month data to train the new model and predict n+2th–n+4th results, and so on. You have noticed that we estimate n+2th twice. Here, we take the mean value as the final estimate. The whole process is shown in Fig. 1.

<table>
<thead>
<tr>
<th>All data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train 1st–nth</td>
</tr>
<tr>
<td>Forecast n+1th–n+3th</td>
</tr>
<tr>
<td>Train 2nd–n+1th</td>
</tr>
<tr>
<td>Forecast n+2th–n+4th</td>
</tr>
<tr>
<td>Train 3rd–n+2th</td>
</tr>
<tr>
<td>Forecast n+3th–n+5th</td>
</tr>
<tr>
<td>……</td>
</tr>
<tr>
<td>Take the mean value of duplicated estimates</td>
</tr>
</tbody>
</table>

So we get ARIMA(1,1,1) and predict results from 2010 to 2020. As we can see in Fig. 4, the result is really terrible, where the grey confidence area can’t cover actual results.

A. ARIMA

ARIMA model is a univariate model short for “Autoregressive Integrated Moving Average model”, which means that we use historical gold price to forecast future price. It is defined with three parameters p, q, d where p is the autoregressive order, q is moving average order and d is differencing order, respectively. The formula is (1).

\[
ARIMA(d,p,q) \Delta^d Y_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta^d Y_{t-i} + \sum_{j=1}^{q} \phi_j \epsilon_{t-j} + \epsilon_t \tag{1}
\]

If we train and predict once, the whole process is as follows. We split data into training data from 1970 to 2010 and testing data from 2010 to 2020. First of all, make sure that data are stationary - there doesn’t exit a unit root in time-series data. We can difference data until the pre-processed data is stationary based on ADF test. Secondly, we determine the AR order based on the stationary data, where the partial autocorrelation function (PACF) is used to determine AR order in Fig. 2. Thirdly, we determine MA order and autocorrelation function (ACF) is used to get MA order in Fig. 3. All of these can be done together in Python pmdarima package.
We also try rolling prediction. We take 120 months as training length starting from Jan 1980, predict next 3 monthly results, move forward with one-month step and take the mean value of duplicated estimates. Although forecasted results look like perfectly being covered with actual gold price, it is not good actually.

Let’s take a very close look at the predicted result in Fig. 6 and the predicted results are lagged behind the actual results. And if we use predicted signal to trade, we can’t get a good result in Fig. 7.

Each time, we predict the next n-month gold-price. Again, the result is not good although it looks good. The prediction always follows the real price in Fig. 8 and the predicted accumulated returns are terrible in Fig. 9.

C. VAR

VAR is short for Vector Auto Regression. And we can take it as multivariate version of ARIMA. Gold price is not only determined by its trend, but also determined by other time-series.

In k-variate VAR($p$) model

\[ r_t = \phi_0 + \phi_1 r_{t-1} + \Phi_2 r_{t-2} + \cdots + \Phi_p r_{t-p} + \alpha_t \]  \hspace{1cm} (3)

where \( r_t = (r_{1,t}, \ldots, r_{k,t})^T \).
First of all, we use Granger Causality test to see whether it is appropriate to use A variable to predict B variable. As is showed in Fig. 10, most of the variables are causes of others.

Also, we use rolling VAR model to predict the result continuously. Here, we take 90 months as training length and 5 months as prediction length each time. In each step, we determine the order p based on AIC coefficient.

When it comes to VAR model, we have to pay attention to the model stationarity. $VAR(p)$ model can be transformed $VAR(1)$.

$$\begin{bmatrix} X_{1,t} \\ \vdots \\ X_{k,t} \end{bmatrix} = \phi_0 + \phi_1 \begin{bmatrix} X_{1,t-1} \\ \vdots \\ X_{k,t-1} \end{bmatrix} + \phi_p \begin{bmatrix} X_{1,t-p} \\ \vdots \\ X_{k,t-p} \end{bmatrix} + \epsilon_{1,t}$$

(4)

When we get matrix $A$ from (4), we can use it to calculate eigenvalues to judge if the overall system is stationary according to (5).

$$VAR(p)$$ is stationary if $|A - \lambda I| = 0, \max(\lambda) < 1$$

(5)

If not, then there is unit root in this system which is non-stationary. In our example, most models are $VAR(1)$ and has unit root in this system.

The rolling regression get the prediction result as follows. We take part of the predicted results out and the predicted results still follow the real price. However, if we use the predicted signals to trade, accumulate returns are terrible.

$$\begin{array}{c}
\text{Predicted prices} \\
\text{Actual prices}
\end{array}$$

(6)

Fig. 10. Granger Causality test

When we get matrix $A$ from (4), we can use it to calculate eigenvalues to judge if the overall system is stationary.

$$\begin{array}{c}
\text{Part of the results of VAR} \\
\text{VAR accumulated returns}
\end{array}$$

D. HMM

After terrible trials above, we need to think about another method. First and foremost, data were used in a terribly inefficient way because we lagged all macro data in 2 months to predict future price. Things change a lot every day, let alone 2 months later. Secondly, we always use numerical values of all variables. Maybe the calculation of probability of up and down is more appropriate.

So, here we try another method called Hidden Markov Model. In this model, we split economic conditions into several states and then guess what would happen in each situation at next step.

In HMM, we have to estimate transition matrix $A$ in (6), emission matrix $B$ in (7) and start probability $\pi$ in (8). Transition matrix is the probability matrix where it demonstrates how much probability the present situation changes to the next situation, the sum of each row being 1. Emission matrix is also a probability matrix showing the probability of the price change direction in one specific situation, the sum of each row being 1. Start probability shows the prior probability of each state, the sum of $\pi$ being 1. These three matrices in (9) are estimated based on the historical data.

$$A = \begin{pmatrix} P(q_t = 1|q_{t-1} = 1), \ldots, P(q_t = k|q_{t-1} = 1) \\ P(q_t = 1|q_{t-1} = k), \ldots, P(q_t = k|q_{t-1} = k) \end{pmatrix}_{kk}$$

(6)
We have two assumptions in this model. The LIMITED HORIZION ASSUMPTION is that the probability of being in a state at time $t$ depends only on the state at time $t - 1$ in (10). And observations are CONDITIONALLY INDEPENDENT of other variables given the current state in (11). And we have to evaluate the probability of next observation $P(y|A)$ according to all coefficients.

$$P(q_{t}|y_{1}, ..., y_{t-1}, q_{1}, ..., q_{t-1}) = P(q_{t}|q_{t-1})$$ (10)

$$P(y_{t}|y_{1}, ..., y_{t-1}, q_{1}, ..., q_{t}) = P(y_{t}|q_{t})$$ (11)

We have two assumptions in this model. The LIMITED HORIZION ASSUMPTION is that the probability of being in a state at time $t$ depends only on the state at time $t - 1$ in (10). And observations are CONDITIONALLY INDEPENDENT of other variables given the current state in (11). And we have to evaluate the probability of next observation $P(y|A)$ according to all coefficients.

$$P(y_{t}|y_{1}, ..., y_{t-1}, q_{1}, ..., q_{t-1}) = P(y_{t}|q_{t-1})$$ (10)

$$P(y_{t}|y_{1}, ..., y_{t-1}, q_{1}, ..., q_{t}) = P(y_{t}|q_{t})$$ (11)

However, there are 3 distinctive states actually. We should label (0,0,1), (1,0,1) and (1,1,1) as state 1, (0,0,0), (0,1,0) and (1,1,0) as state 2, (0,1,1) and (1,0,0) as state 3. So the transition matrix is 3 * 3.

Then, we roll the HMM estimation. We split gold returns into 3 states, rising, unchanged and falling based on different groups of percentiles. Again, we take fixed-length training data each time, and predict gold state probabilities at the next step. And we take the greatest probability of prediction as the final result. And based on our predicted signals, we trade gold accordingly. If the signal is ‘rising’, then we buy gold; if the signal is ‘down’, then we short gold; if the signal is ‘unchanged’, we don’t trade.

Last but not least, we tried combinations of economic indexes, different training lengths ranging from 12 months to 180 months, different upper bound percentiles and lower bound percentiles. TABLE II shows grid search results based on the economic indexes of EPU, CPI and PMI. TABLE III shows grid search results based on the economic indexes of EPU and PMI. Win rate shows how many times our predicted returns beat actual returns during the whole trade.
<table>
<thead>
<tr>
<th>Training Length</th>
<th>Upper Bound</th>
<th>Lower Bound</th>
<th>Win Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>55</td>
<td>38</td>
<td>0.56</td>
</tr>
<tr>
<td>168</td>
<td>60</td>
<td>35</td>
<td>0.56</td>
</tr>
<tr>
<td>168</td>
<td>58</td>
<td>35</td>
<td>0.56</td>
</tr>
<tr>
<td>168</td>
<td>55</td>
<td>35</td>
<td>0.57</td>
</tr>
<tr>
<td>156</td>
<td>55</td>
<td>35</td>
<td>0.57</td>
</tr>
<tr>
<td>180</td>
<td>58</td>
<td>35</td>
<td>0.59</td>
</tr>
<tr>
<td>180</td>
<td>55</td>
<td>35</td>
<td>0.59</td>
</tr>
</tbody>
</table>

So, we take the best combination to trade gold in each situation. Fig. 17 shows accumulated returns based on the combinations of EPU, CPI and PMI. Fig. 18 shows the final results based EPU and PMI, which is much better.

Fig. 17. HMM accumulated returns with EPU CPI and PMI

Fig. 18. HMM accumulated returns with EPU and PMI

IV. CONCLUSION

Gold price is very hard to predict using traditional time-series methods because gold price is not just determined by its own historical trends. Gold is closely linked with economic and political conditions. And we should try to predict what would happen in the future economy and politics. Then, predict the probability of gold price rising and falling in the specific situation. That is why HMM model can achieve much better results than other traditional time-series models. Last but not least, it’s important to use data wisely. For example, in linear models and time-series models, we shouldn’t use future data to train models. Therefore, we have to lag macroeconomic data in one or two months and a lot of information has been lost. However, when we use HMM, we don’t have to lag data any more, because we can use transition matrix to predict future economic conditions and then use emission matrix to predict gold performance, which will improve the result.

REFERENCES
