Find Trading Logic by Bayesian Network

Introduction

Trading Logic is very important for traders, especially macro traders. Because only after a correct/effective logic chain is found, trader can focus on the changing of each node in the chain, then make correct decision (long/short).

Usually, to find a logic chain, traders would start with economic fundamental analysis. Then use mathematical tool (correlation or other time series algorithms) to find the most correlated factors for certain period. For example, to predict petroleum price, traders might start with petroleum supply data from OPEC, and estimate demand value from the main petroleum import countries (European, US, China). Then, correlation or other time series algorithms would be used to check which economic indicators from which countries would affect the petroleum price more in certain period. This whole process does work well. However, in the mathematical part, if a factor is not direct correlated with the target feature, it take much time to decide each node in the chain. For example, there are 5 factors, A, B, C, D, and E. A logic chain exists, which is A->B->D->E. A trader only know E is the target. By traditional way, if there is no obvious economic knowledge to support, he/she might have to create a correlation matrix, then test each possible chain and check if this chain is reasonable.

In this case study, I want to introduce Bayesian Network to find trading logic chain in an easier way. I would use Bayesian Network to find trading logic chain for USD index. After getting all the possible factors, I put them into Bayesian Net. The algorithm will return a most possible chain, then further study can be done basing on the chain. I believe this method can shorten much time in the mathematic part of finding a trading strategy.

Data Description

I download all data from Choice Financial Terminal. There are 13 features which represent different aspect of US economy status. They are all monthly data from Oct 1981 to Apr 2017. The reason is some features' historical data start from Oct 1981.

Feature	Data Type	Description
USDindex	numeric	The US Dollar Index is an index (or measure) of the value of the United States dollar relative to a basket of foreign currencies, often referred to as a basket of US trade partners' currencies. It is a weighted geometric mean of the dollar's value relative to other select currencies: Euro (EUR), 57.6% weight Japanese yen (JPY) 13.6% weight Pound sterling (GBP), 11.9% weight Canadian dollar (CAD), 9.1% weight Swedish krona (SEK), 4.2% weight

		Swiss franc (CHF) 3.6% weight
M2_YoY	numeric	M2 is a measure of the money supply of a county. Depends on the theory of Milton Friedman, money supply changing would affect future inflation rate of a country. M2_YoY is the year over year percentage change of M2
CoreCPI_YoY	numeric	Consumer price index (CPI) measures changes in the price level of market basket of consumer goods and services purchased by households. Core CPI is CPI excluding energy and food price. Core CPI Yoy is the year over year percentage change of CPI. It is used as a measure of inflation.
BudgetDeficit_YoY	numeric	A budget deficit is an indicator of financial health in which expenditures exceed revenue. The term budget deficit is most commonly used to refer to government spending rather than business or individual spending, but can be applied to all of these entities. When referring to accrued federal government deficits, the deficits are referred to as the national debt. BudgetDeficit Yoy is the year over year percentage change of budget deficit.
FiscalExpenditure_YoY	numeric	Investment or other money spent by government. This is year over year percentage change.
PMI	numeric	The Purchasing Managers' Index (PMI) is an indicator of the economic health of the manufacturing sector.
NonFarmPayroll_YoY	numeric	Nonfarm payroll is a term used in the U.S. to refer to any job with the exception of farm work, unincorporated self-employment, and employment by private households, the military and intelligence agencies. Proprietors are also excluded. This feature is the year over year percentage change.

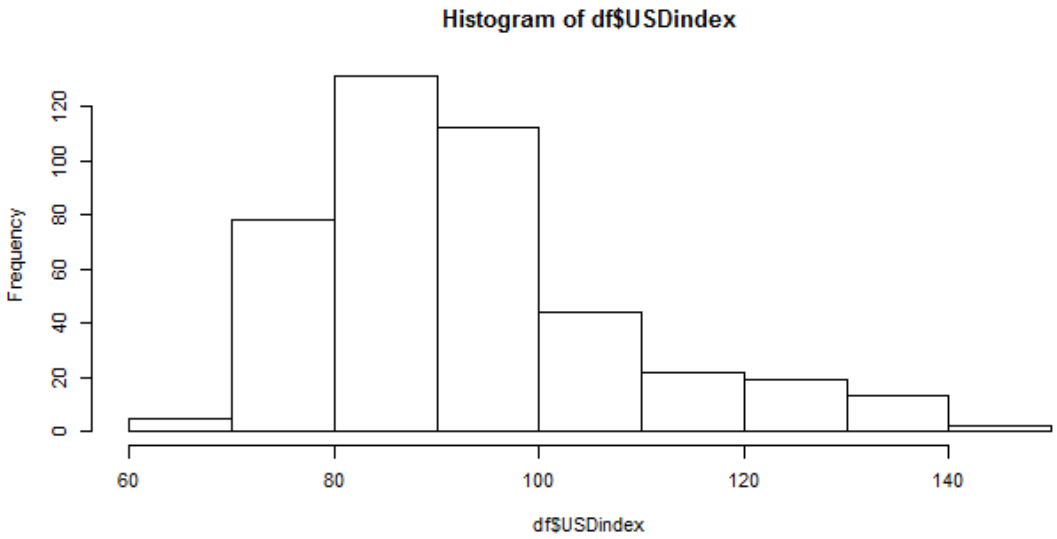
DisposablePersonal Income	numeric	Disposable income, also known as disposable personal income (DPI), is the amount of money that households have available for spending and saving after income taxes have been accounted for. Disposable personal income is often monitored as one of the many key economic indicators used to gauge the overall state of the economy.
DisposablePersonal Income_YoY	numeric	Year over year percentage change of disposable income.
InternationalCapital Flows_MoM	numeric	International capital flows refer to the movement of money into a country from oversea for the purpose of investment, trade or business production, including the flow of capital within corporations in the form of investment capital, capital spending on operations and research and development (R&D). InternationalCapitalFlows_MoM is the month over month percentage change.
LongtermSecurityHolding ByInternationalInvestor	numeric	Long term security holding by international investors. This indicator can be used to check capital flow (inflow/outflow) of secondary market of a country.
10YearBondYield	numeric	Yield of 10 year bond. Usually it would be used to represent currently interest rate.
interestRate	numeric	In the United States, the federal funds rate is the interest rate at which depository institutions (banks and credit unions) lend reserve balances to other depository institutions overnight, on an uncollateralized basis. Reserve balances are amounts held at the Federal Reserve to maintain depository institutions' reserve requirements. Institutions with surplus balances in their accounts lend those balances to institutions in need of larger balances. The federal funds rate is an important benchmark in financial markets.

USDindex is the target value that I am trying to build a logic chain for. Some of the features are similar indicators, for example, 10 year bond yield and interest rate, they are both a kind of benchmark interest rate. In data analysis part, I would check the correlations between features and USD index, if there are several features representing same aspect of economy status, I would only keep the one with highest correlation with USD index.

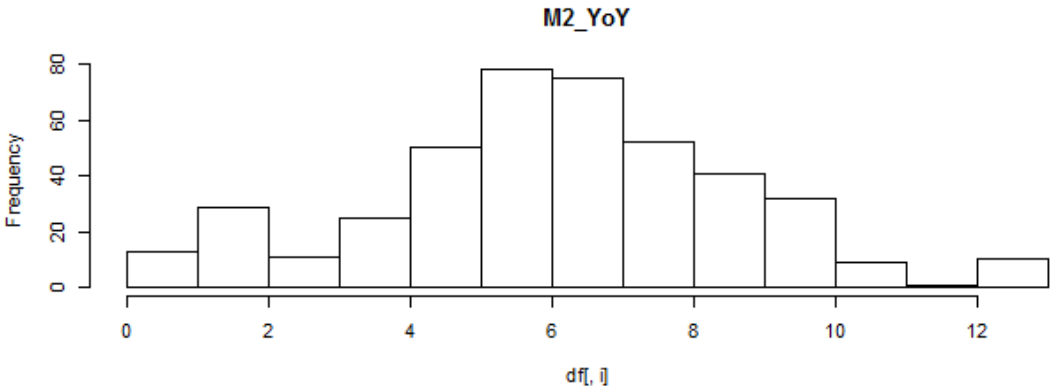
Data Cleaning and Transformation

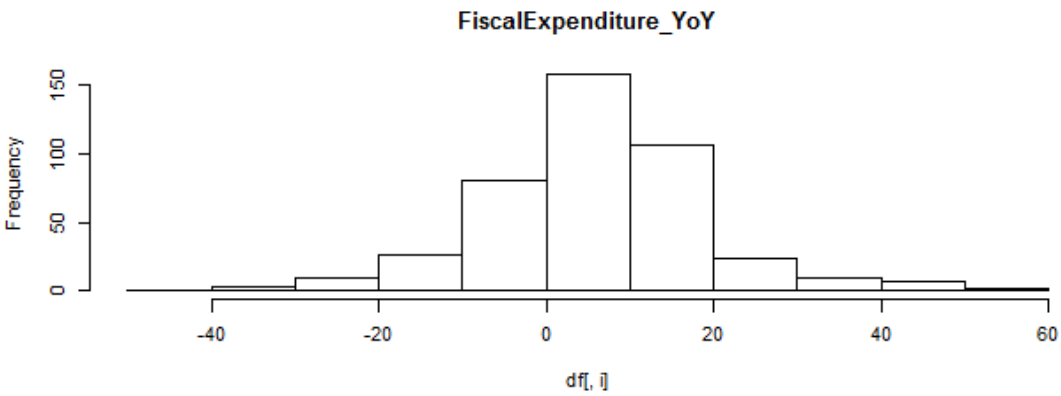
Check the data set, there is no missing data.

Plot all features and target, to see if any of them require transformation.

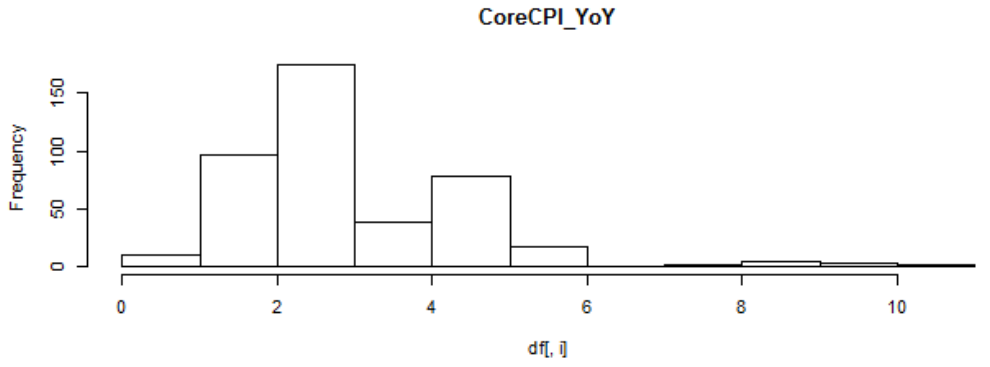


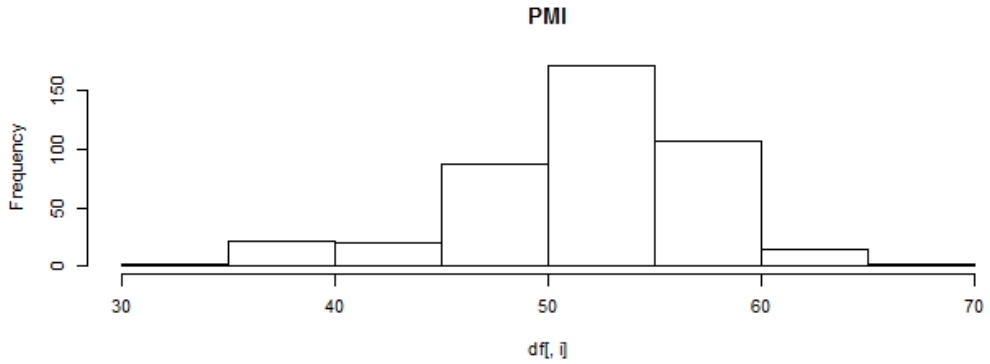
The USDindex looks good.

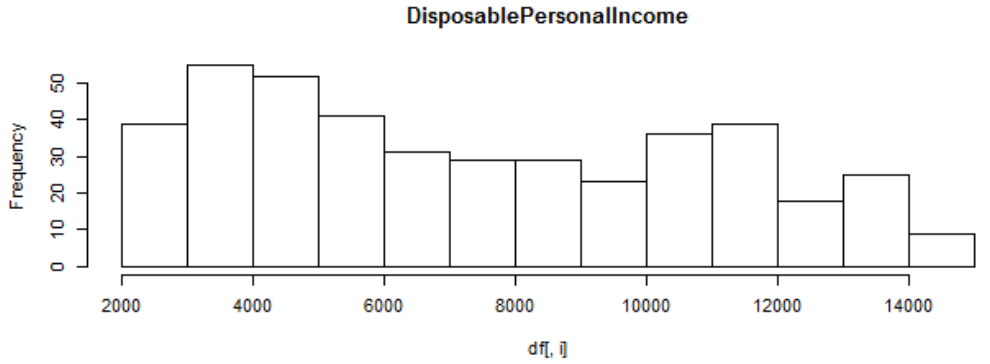


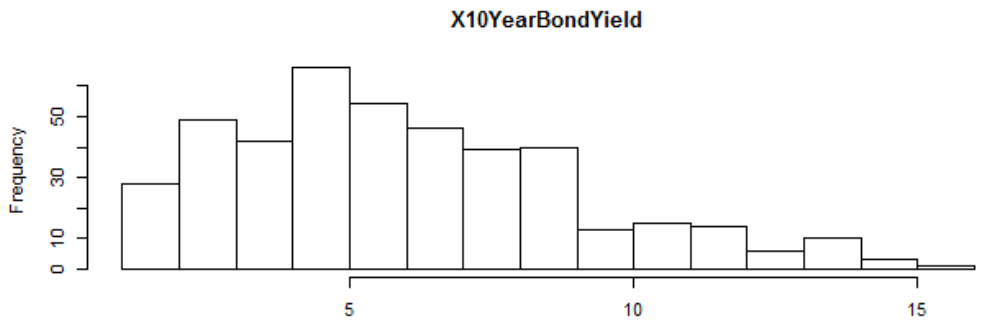


M2_YoY and FiscalExpenditure_YoY look good.



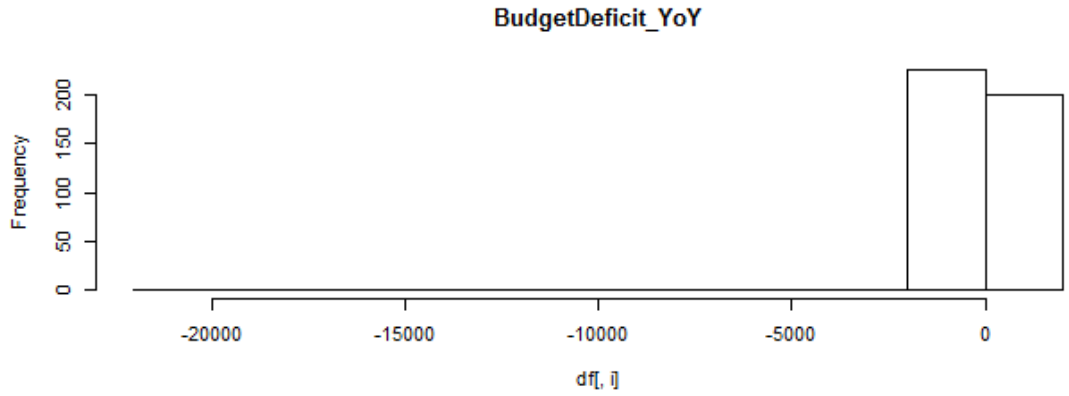


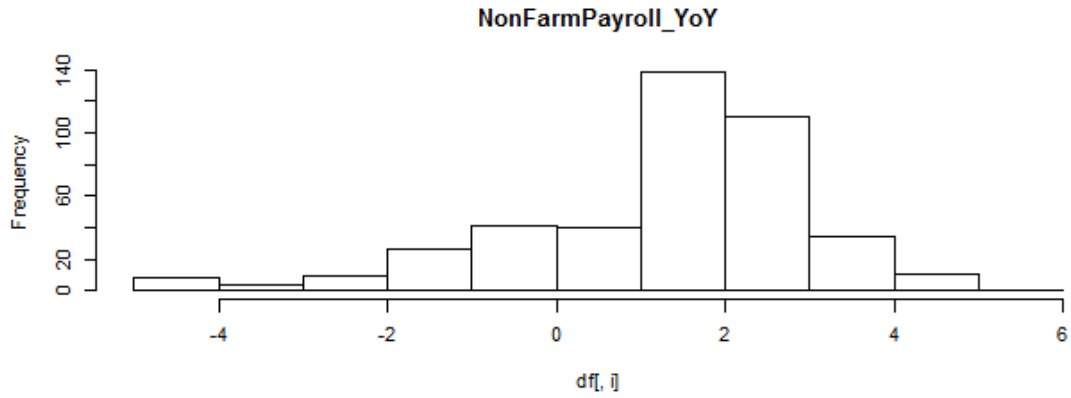


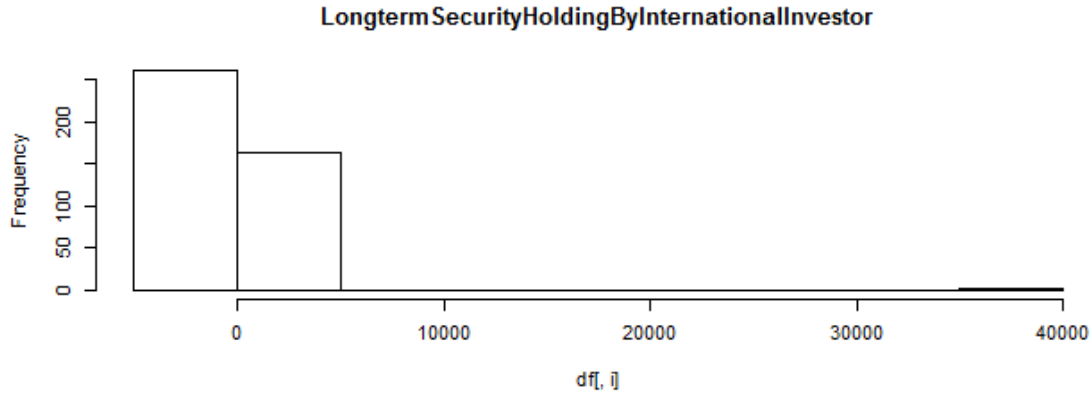


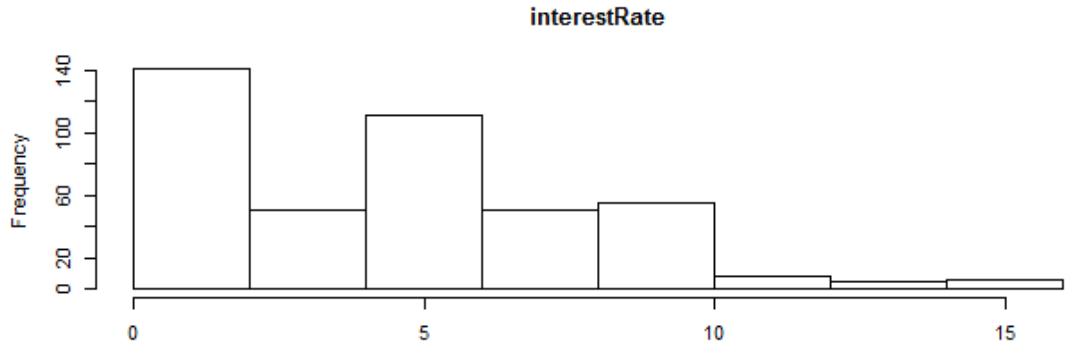
CoreCPI_YoY, PMI, DisposablePersonalIncome, and 10YearBondYield do not require any transformation.

Case 3 Yiming Wang



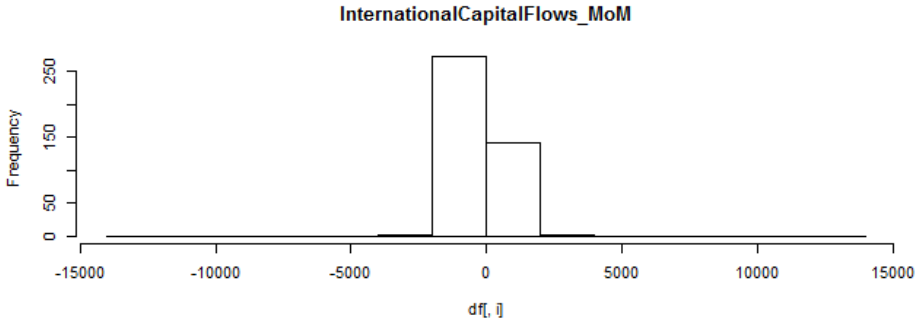






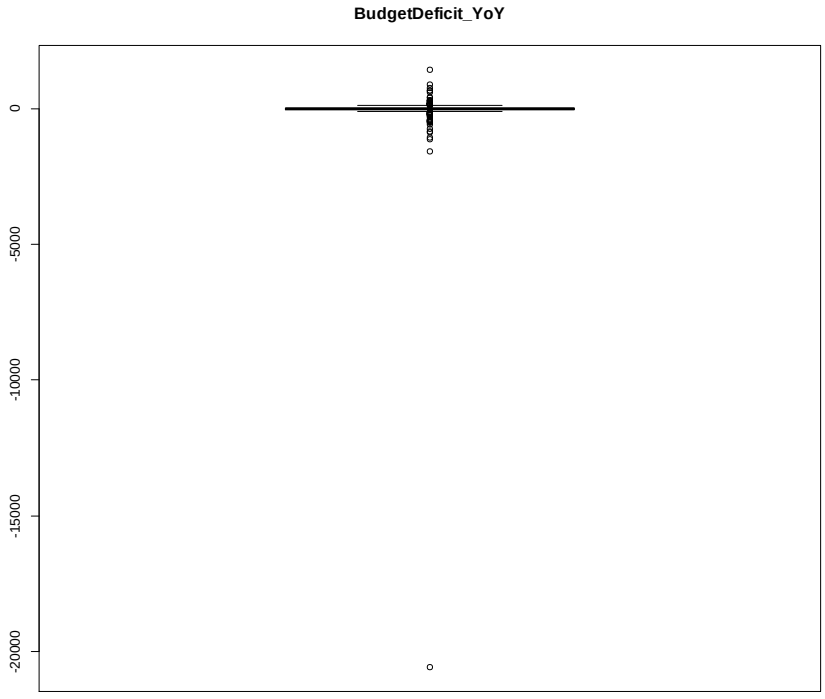
Require more analysis on BudgetDeficit_YoY and LongtermSecurityHoldingByInternational Investor.

Case 3 Yiming Wang



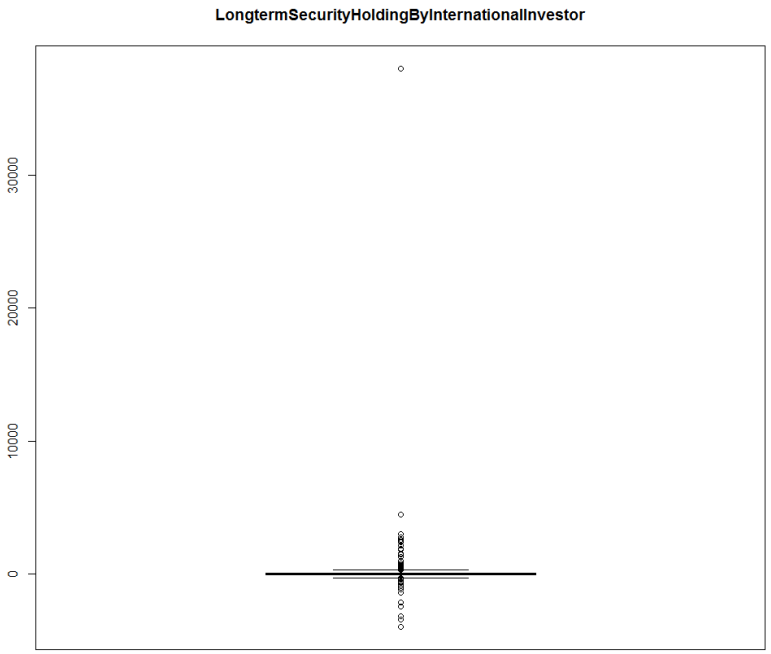
Require more analysis on InternationalCapitalFlows_MoM.

If LongtermSecurityHoldingByInternationalInvestor, InternationalCapitalFlows_MoM or Budget Deficit_YoY look skew, I would normalize it. Probably try logarithm.

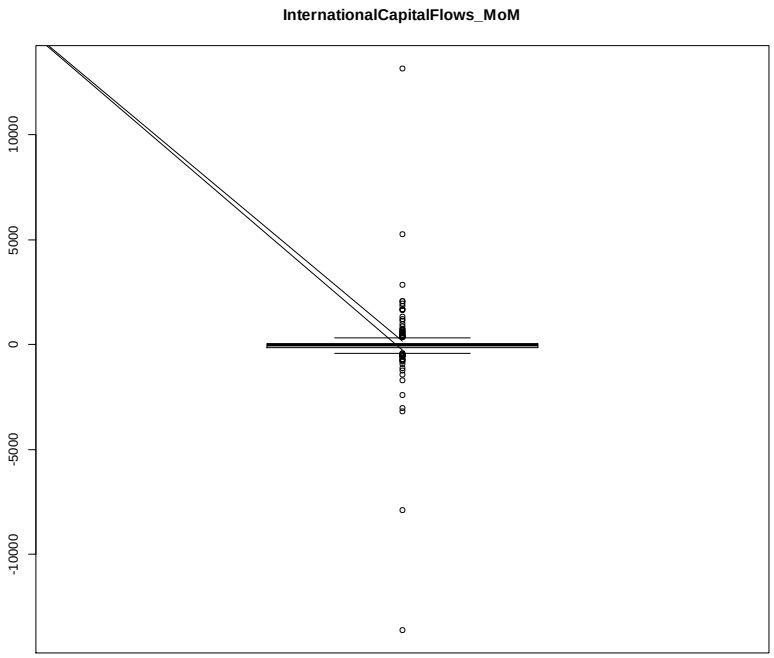


From the boxplot, we can see that most of the points are near 0, except one, -2000. It looks like an outlier, but, since this is economic data, probably some extremely thing happened. I would keep this point.

Case 3 Yiming Wang



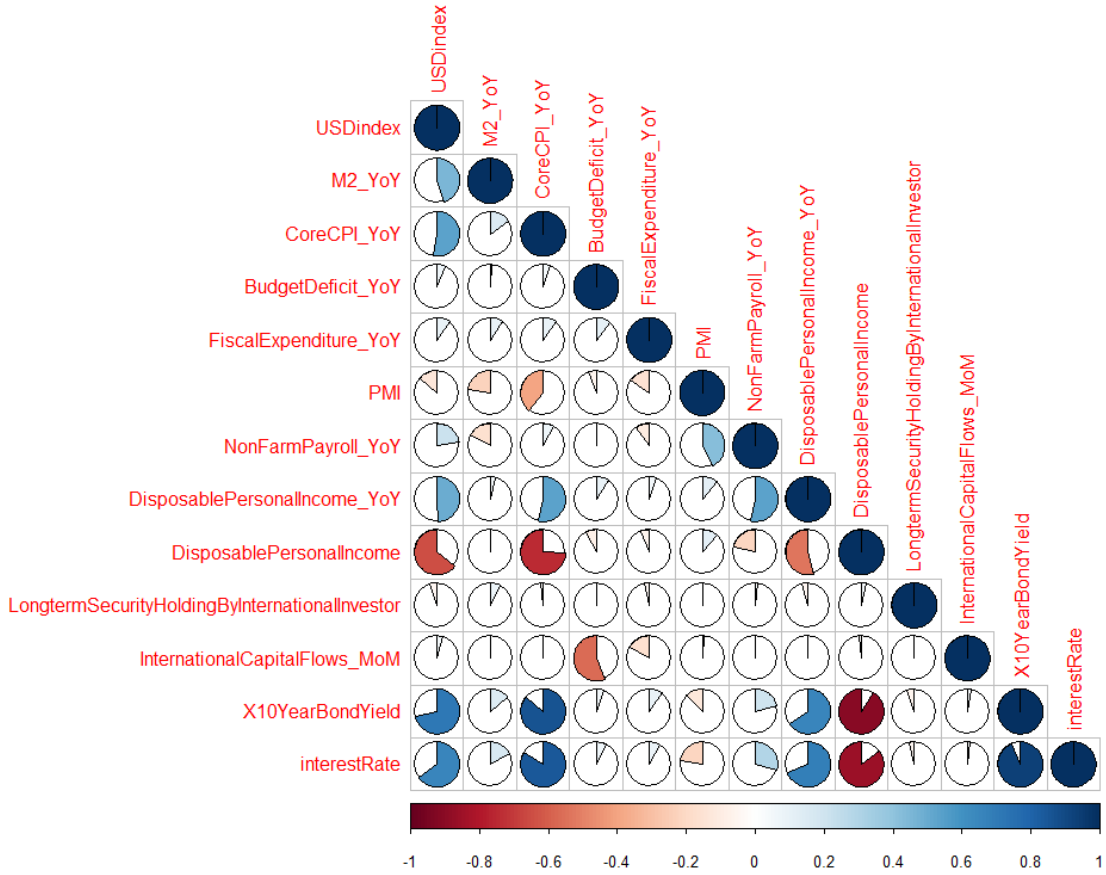
Same reason, I would also keep the point, 4000.



Most of the points are near 0, same reason, would not do any transformation.

Data Analysis

As I pointed out in data description part, I need to delete some similar indicators. So, I plot a correlation matrix.



10YearBondYield and interestRate are both indicator of interest rate benchmark. From the chart, it looks like 10 year bond is higher correlated with USDindex, so, I would keep it and drop interestRate.

DisposablePersonalIncome and DisposablePersonalIncome_YoY are obviously similar factor. I keep DisposablePersonalIncome due to its higher correlation with USDindex.

For LongtermSecurityHodlingByInternationalInvestor and InternationalCapitalFlows_MoM, they both evaluate the international capiral flows. Depending on economic knowledge, InternationalCapitalFlows_MoM might be a better indicator since LongtermSecurityHodling ByInternationalInvestor is more focus on secondary market capital flow. What is more, from the chart above, InternationalCapitalFlows_MoM has high correlation with BudgetDeficit_YoY (might have causal relationship), so, I would keep InternationalCapitalFlows_MoM.

After deleting 3 features, I got a data set with 1+9 features.

Experimental Results

I use the bnlearn package in R to build Bayesian network. Since we are trying to find a Bayesian network as logic chain, normal data mining processes: training + testing and k-folder cross-validation are not appropriate. I would use the whole data set, and test different score method to get a best network structure for this data set.

Case 3 Yiming Wang

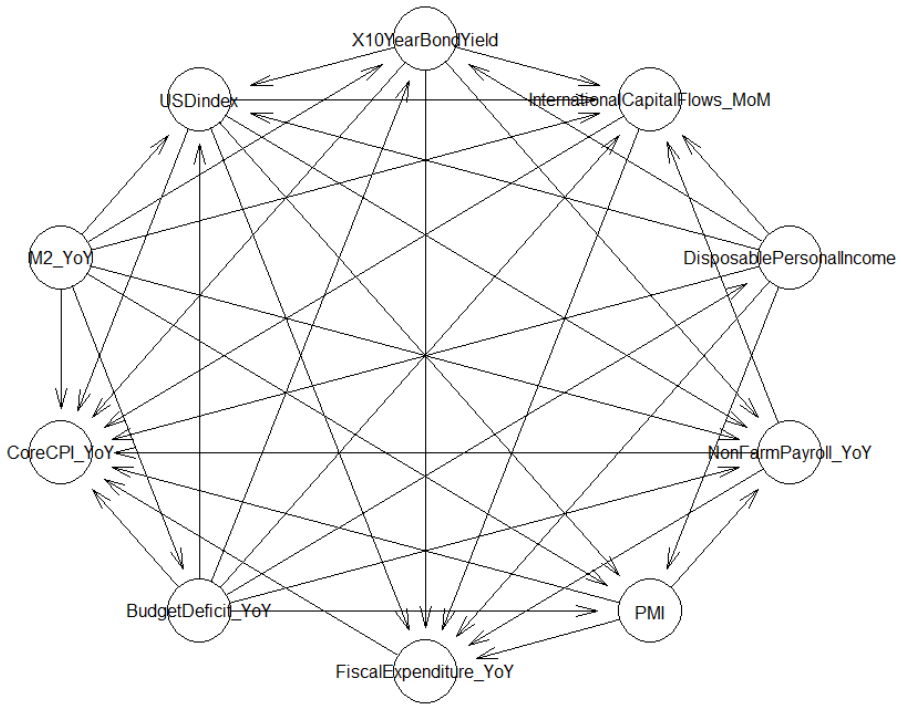
In bnlearn package, there are 2 Bayesian network building function: hc() (hill-climbing) and tabu() (Tabu search). What is more, for continuous data, 4 kinds of score methods could be used:

- the multivariate Gaussian log-likelihood (loglik-g) score
- the corresponding Akaike Information Criterion score (aic-g)
- the corresponding Bayesian Information Criterion score (bic-g)
- a score equivalent Gaussian posterior density (bge)

I would test all combination, then choose the one with highest score to be the trading logic chain of USD index.

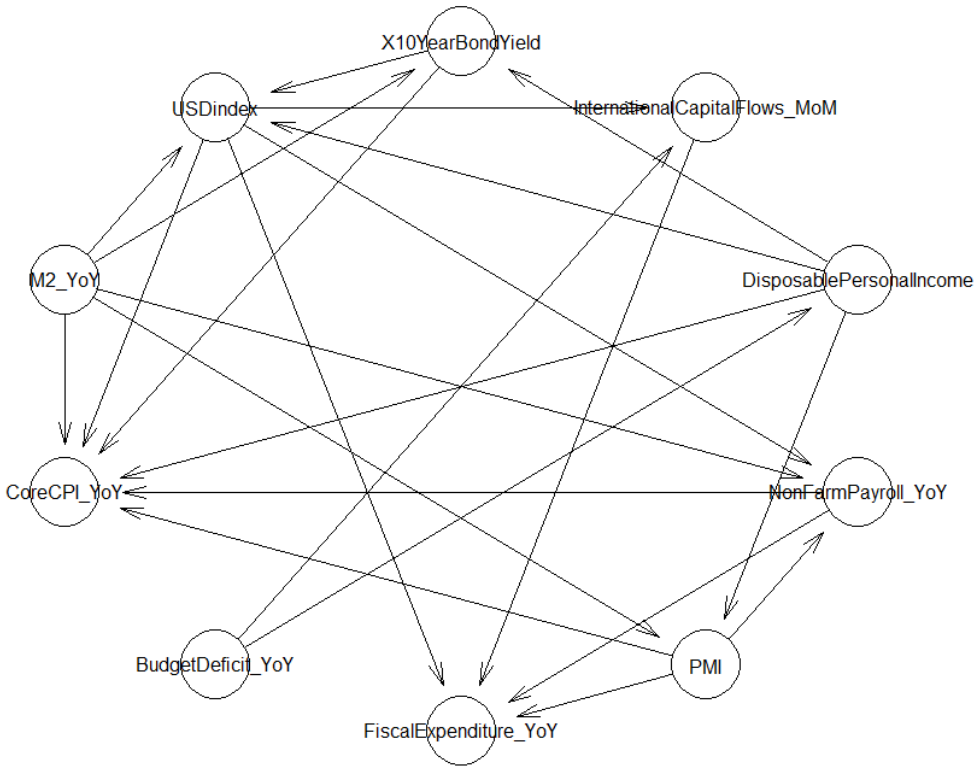
Hill-climbing:

i. multivariate Gaussian log-likelihood score: -18680.76



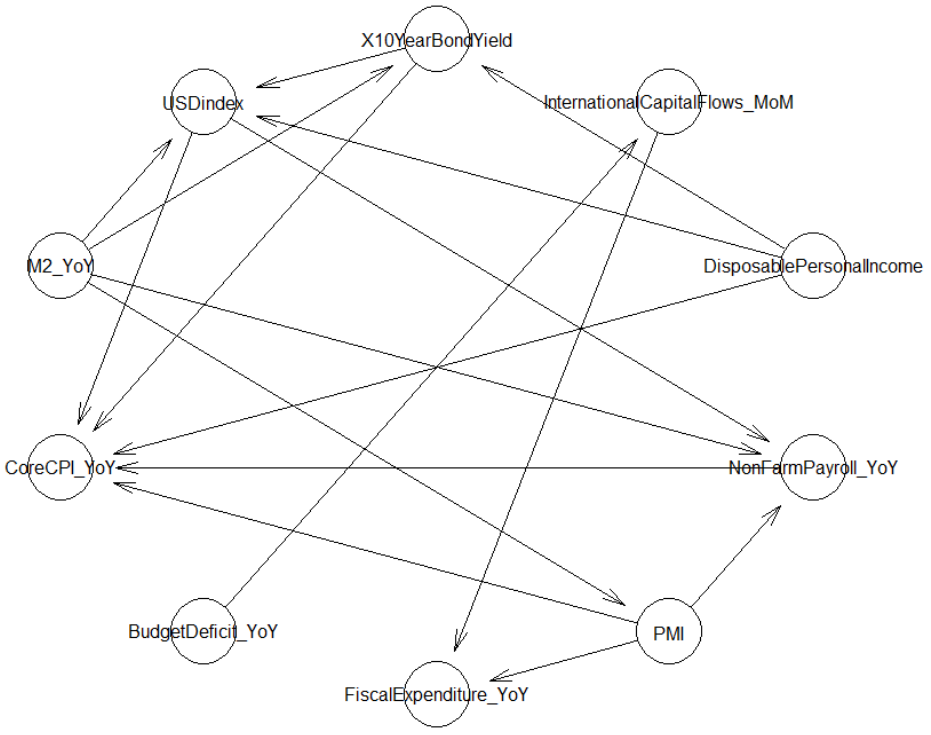
It looks so complicated. Almost every node connects to all the rest of nodes. I would prefer not to use this kind of logic chain. It might be overfitting.

ii. corresponding Akaike Information Criterion score: -18637.64



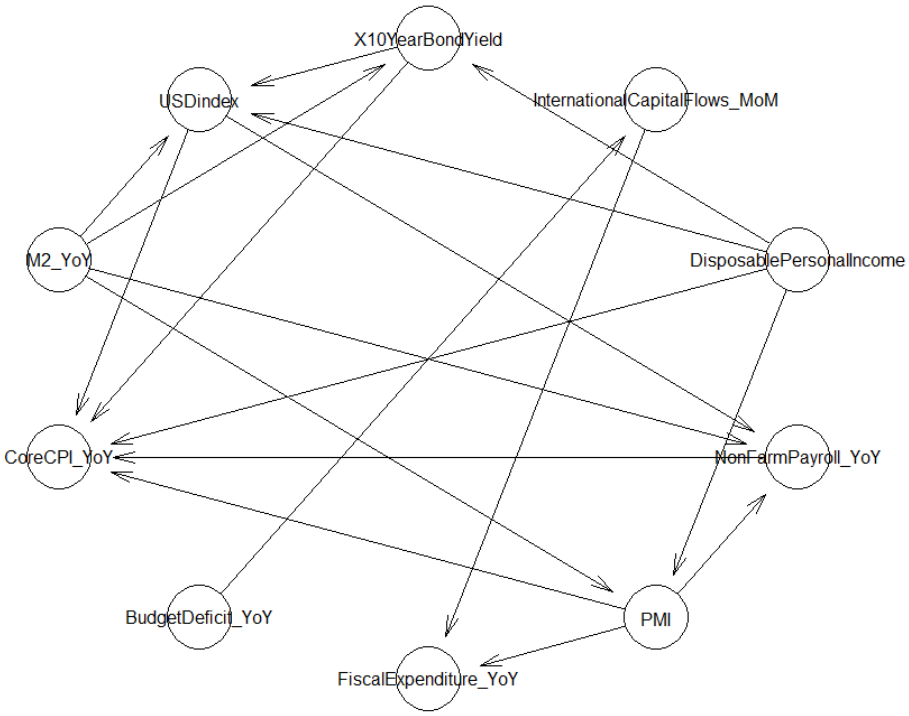
It returns much cleaner network with higher score.

iii. corresponding Bayesian Information Criterion score: -18629.13



Higher score than the network build by corresponding Akaike Information Criterion.

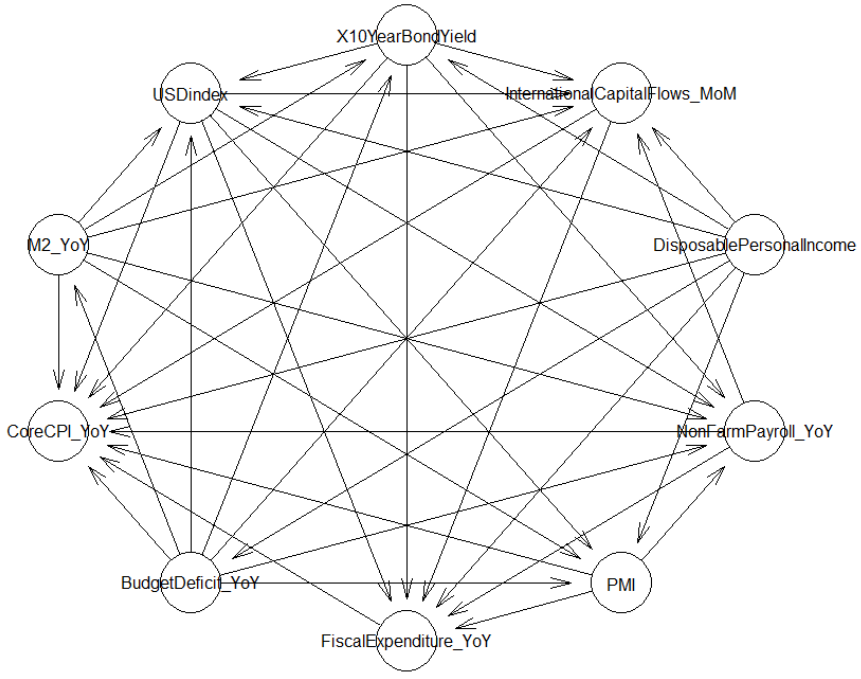
iv. a score equivalent Gaussian posterior density: -18629.34



The score is slightly lower than network build by corresponding Bayesian Information Criterion score.

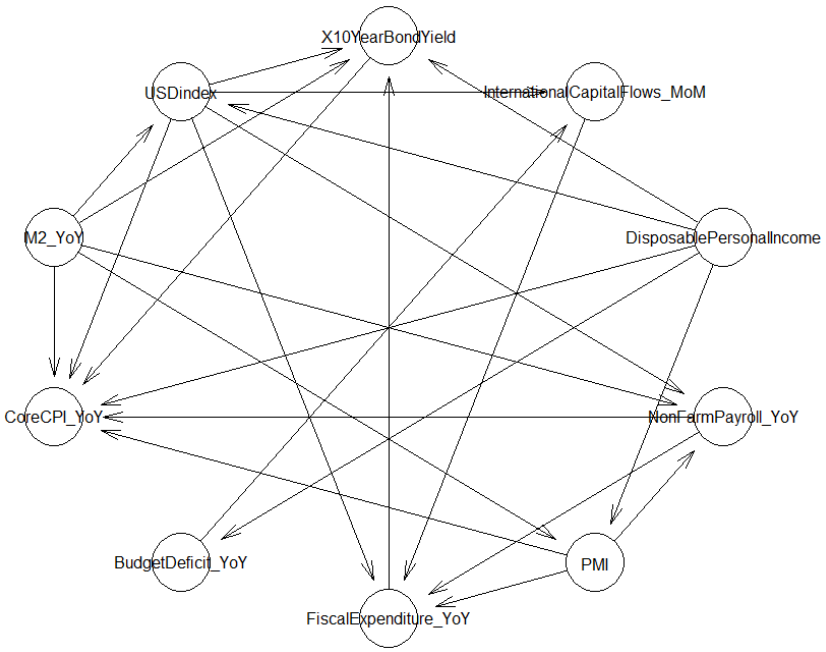
Tabu search：

i. the multivariate Gaussian log-likelihood (loglik-g) score: -18680.76



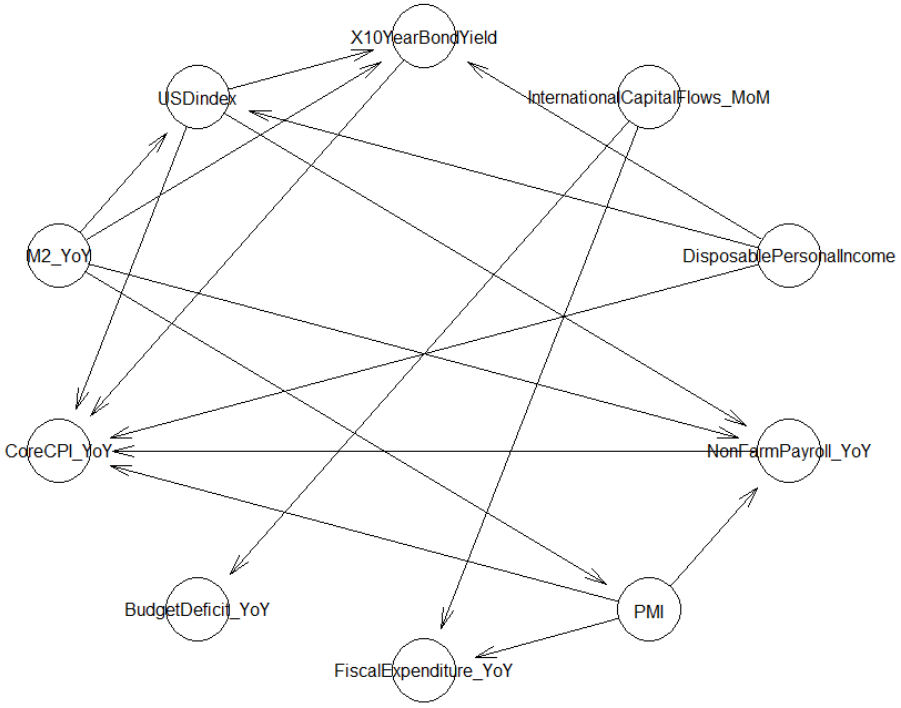
It returns same reason with hill-climbing.

ii. the corresponding Akaike Information Criterion score (aic-g): -18639.51



The score is lower than the score from hill-climbing.

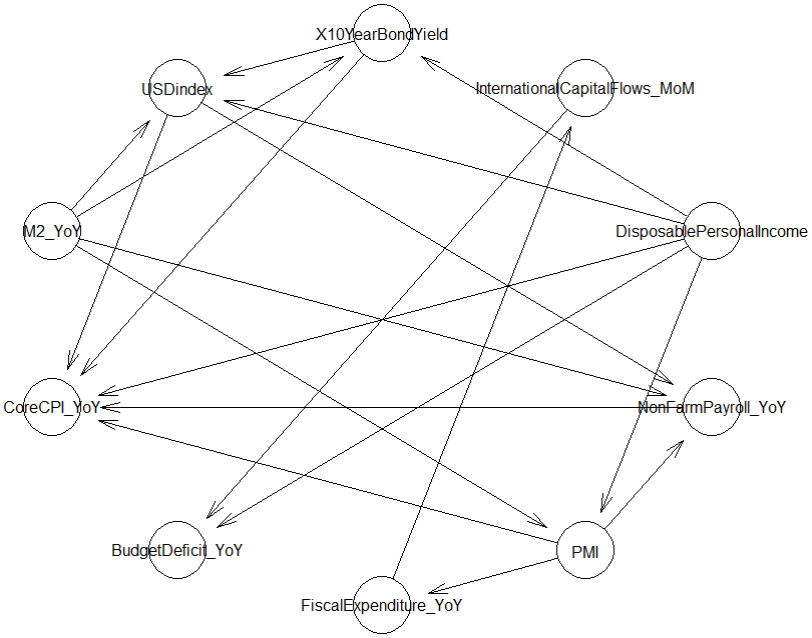
iii. the corresponding Bayesian Information Criterion score (bic-g): -18629.13



Same result with hill-climbing.

iv. a score equivalent Gaussian posterior density (bge): -18630.05

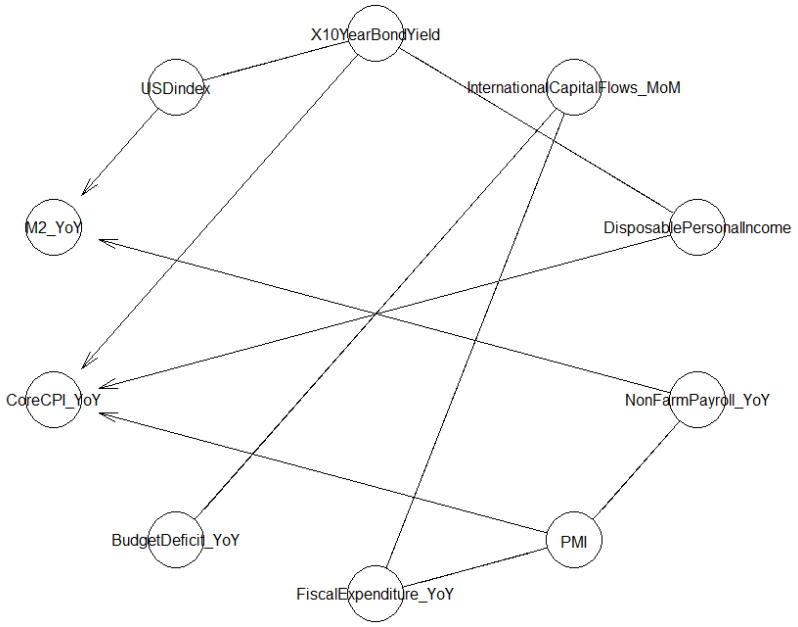
Case 3 Yiming Wang



	Hill-climbing				Tabu search			
	loglik-g	aic-g	bic-g	bge	loglik-g	aic-g	bic-g	bge
score	-18680.8	-18637.6	-18629.1	-18629.3	-18680.8	-18639.5	-18629.1	-18630.1

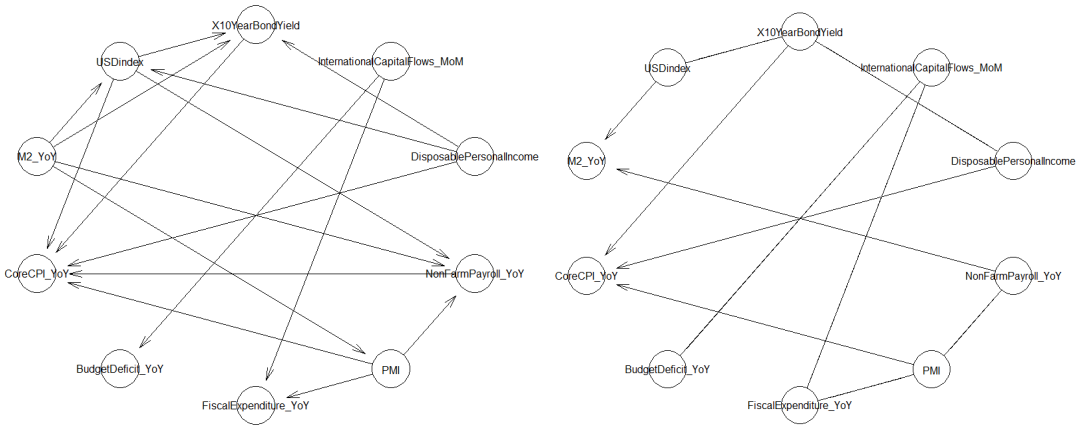
From the table, we can see the bic-g score method always return best network (same network).

I also build a bayesian network by Constraint-based structure (gs, iamb, fast.iamb, inter.iamb). Since all Constraint-based structure method would return same network, I only give the result from gs() (Grow-Shrink algorithm) :



I would descuss both of this 2 network (bic-g score and gs) in the next part, and pick one as the final result.

Experimantal Analysis



Above is the 2 networks we get from previous part. In left network (score-based structure), M2_YoY and Disposable Personal Income are the most important factor for the value of USDindex. However, due to the design of score-based method, undirected edges are not allowed, which makes some possible 2 way causal relationship missed. For example, depending on economic knowledge, USD index and 10 year bond yield (the interest rate of US) is bi-causal relationship. It does not show on the left network, only on the right network (Constraint-based structure). On the other hand, in the right network, it misses the potential influence to USD index from Disposable Personal Income (Yellen mentioned in March this year, to decide the timing of next hiking, the income is one of the factors she would consider). Either one is not a perfect solution.

So, I would say, use the network by Constraint-based structure as the base network, and do some adjustment (according to score-based network). Then a trading logic chain can be built. Which means, for USD index trading, the most important factor should be M2 year over year, the interest rate and disposable personal income. The second level factor should be Non-farm payroll yoy (it affect M2 yoy). For the rest of factors, they might be the third level factor depending on the network. It is quite true for USD index trading.

In next step, I might try to split USD index into different time period, such as bull and bear. Then rebuild the networks, to see if the network would be still the same.

Conclusion

In this case, I use Bayesian Network to build a network to be trading logic chain. It did work but looks not perfect. The reason is, the best fit network built by different method (score-base, or constraint-based) could have different result. I would recommend traders to use the constraint-based network as base network, then add some more edges according to the best fit score-based network. Then a new network would be created which can be used as a trading logic chain. In the USD index example, the final logic chain can work great.

Although this method is not perfect yet, I believe it can save much time comparing with the traditional “correlation searching” method.

In the future, I would try to find if there exist another algorithm which can be used to build trading logic chain easier than Bayesian network.

Appendix

```
#install.packages("bnlearn")
library(bnlearn)

df <- read.csv("e:/Case3Data.csv")
df <- df[2:14]
df

hist(df$USDindex)

par(mfrow=c(4,3))
for(i in 2:13) {
   hist(df[,i], main=names(df)[i],cex=5)
}

par(mfrow=c(1,1))
boxplot(df$BudgetDeficit_YoY,main="BudgetDeficit_YoY")
boxplot(df$LongtermSecurityHoldingByInternationalInvestor,main="LongtermSecurityHoldingB
yInternationalInvestor")
boxplot(df$InternationalCapitalFlows_MoM,main="InternationalCapitalFlows_MoM")

#install.packages("corrplot")
library(corrplot)
M<-cor(df)
corrplot(M, method="pie", type = "lower")

df1=cbind(df[1:7],df[9],df[11:12])
df1
sum(is.na(df1))



hc1 <-hc(df1,score="loglik-g")
plot(hc1)
score(hc1,df1)

hc2 <-hc(df1,score="aic-g")
plot(hc2)
score(hc2,df1)

hc3 <-hc(df1,score="bic-g")
plot(hc3)
score(hc3,df1)
```

```
hc4 <-hc(df1,score="bge")
plot(hc4)
score(hc4,df1)



tabu1 <-tabu(df1,score="loglik-g")
plot(tabu1)
score(tabu1,df1)

tabu2 <-tabu(df1,score="aic-g")
plot(tabu2)
score(tabu2,df1)

tabu3 <-tabu(df1,score="bic-g")
plot(tabu3)
score(tabu3,df1)

tabu4 <-tabu(df1,score="bge")
plot(tabu4)
score(tabu4,df1)



help(gs)
bngs <-gs(df1)
plot(bngs)

bn2 <- iamb(df1)
plot(bn2)

bn3 <- fast.iamb(df1)
plot(bn3)

bn4 <- inter.iamb(df1)
plot(bn4)
```