Macro ML

Jean-Galaad BARRIERE

05/11/2022

Introduction

In financial econometrics, numerous approaches have been developed to explain the returns of assets. It is often assumed that the excess returns are related to a given set of factors. The exposition of an asset to a factor must be compensated by a "risk premium' '. Therefore, the excess return of an asset depends on those risk premia multiplied by the exposition of the asset to each of the factors.

A key issue of financial factor models resides in the choice of the factors. Various models have been developed, using different sets of factors. For instance, the Fama-French three-factor model is based on market excess return, outperformance of small versus big companies and outperformance of high book-to-market versus low book-to-market companies.

Our article investigates how macro factors can be used in asset pricing models. As already shown in the literature, some macroeconomic variables (such as GDP growth, inflation, unemployment or housing prices) could generate risk premia. Nonetheless, the difficulty lies in the identification of the relevant macroeconomic variables among a very large set of macroeconomic indicators. Some previous papers have arbitrarily chosen one or two macroeconomic variables. Our article innovates by using machine learning techniques so as to construct a few factors out of a large set of macroeconomic variables. The central ML technique used here is **sparse Principal Component Analysis** (PCA). As we will see below, the main advantage of sparse PCA over PCA lies in the interpretability of the factors.

Once the principal components are extracted, we use them as factors in asset pricing models. The goal is to determine whether those factors are relevant and whether they generate significant risk premia. The estimation of the of the risk premia uses the **three-pass methodology** developed by Giglio and Xiu. Their methodology is designed to compute unbiased risk premia estimates under omission of relevant risk factors and measurement error. The concern about factor omission is indeed well founded. If we assume that the asset excess returns are only determined by the macro factors derived from the PCA, we might omit other relevant factors. The three-pass methodology solves this problem.

[reste de l'intro]

PCA and Sparse PCA

Our article performs a sparse PCA on a set of 120 macroeconomic variables from the FRED-MD database. Those variables cover various categories: output and income, labor market, housing, consumption, money and credit, interest and exchanges rates, and prices. Here are some examples of macroeconomic variables: real personal income, industrial production indices, civilian unemployment, number of employees by sector, number of housing permits, M1 money stock, commercial and industrial loans, fed fund rates, consumer price indices.

Before performing the sparse PCA, we need some treatment on the FRED-MD data. We use a csv file on which we reported metadata on the FRED-MD macroeconomic variables, in particular : whether they should be included in the analysis and what transformation should be performed on them (log, log growth, difference). These indications come from ***Table 1*** of the article. After selecting the relevant variables and performing the transformations, we restrict the dataset to the time period considered (1960:01 to 2019:12)

```r
library(dplyr)

file <- "data/2020-11.csv"
data0 <- read.csv(file = file)

x <- data0$sasdate
# we drop the rows which have no date
data1 <- data0[(x!="Transform:" & nchar(x)>2),]
y<-data1[,1]

# extraction of variable names
varnames <- data.frame("FRED_ticker"=colnames(data1)[-1])
write.csv(varnames, "varnames.csv", row.names = F)

##### Keeping only relevant time series
# Importation of csv file with variables metadata
df <- read.csv("data/variables.csv",sep=";")
df <- filter(df,Inclusion==1)
var <- df$FRED_ticker

#on garde la date
var <- c("sasdate", var)

data <- data1[var]

### Transformation of the time series
var_names <- colnames(data)
for(i in 2:length(var_names)){ # exclusion of 1st column (date)
  variable <- var_names[[i]]
  transfo <- df$Transformation[df$FRED_ticker==variable]
  no_lag <- df$No_lag[df$FRED_ticker==variable]
  if(!is.null(transfo)){
    if(transfo=="Log"){
      data[,i]<-log(data[,i])
    }
    if(transfo=="Difference"){
      data[,i]<-c(NA, diff(data[,i])) # length is decreased by 1 when we take the difference
    }
    if(transfo=="Log growth"){
      tmp <- data[,i]
      tmp <- tmp/lag(tmp)
      tmp<-log(tmp)
      data[,i]<-c(tmp) # length is decreased by 1 when we take the difference
    }
    if(no_lag==0){
      data[,i]<-c(0,data[-nrow(data),i])
    }
  }
}
```

```
## Warning in log(tmp): Production de NaN
```

```r
## Time interval
data$sasdate<-as.Date(data$sasdate, format = "%m/%d/%Y") # conversion to date
data <- filter(data, sasdate>="1960-02-01" & sasdate<"2020-01-01")

### Saving to RDS
saveRDS(data, "data/FRED_data.rds")
```

PCA

We first perform of traditional PCA on the 120 variables, and select 9 components. We use the same package as the authors

```r
library(FactoMineR)
library(knitr)

data <- readRDS("data/FRED_data.rds")

data0 <- dplyr::select(data, -1) # we drop the date column
sum(is.na(data0))
```

```
## [1] 2
```

```r
tmp<- data0[,67]
tmp[is.na(tmp)]<-mean(tmp,na.rm = T)
data0[,67]<-tmp
sum(is.na(data0))
```

```
## [1] 0
```

```r
pca <- PCA(data0, ncp=9, graph=F)
table1 <- pca$eig
```

```r
kable(table1[1:9,], caption = "First 9 components of the PCA")
```

Table 1: First 9 components of the PCA

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	21.085387	17.571156	17.57116
comp 2	16.928937	14.107447	31.67860
comp 3	7.718925	6.432438	38.11104
comp 4	6.283974	5.236645	43.34769
comp 5	5.008793	4.173994	47.52168
comp 6	3.743766	3.119805	50.64149
comp 7	3.083516	2.569596	53.21108
comp 8	2.783237	2.319364	55.53045
comp 9	2.620463	2.183719	57.71416

The first nine conventional PCs collectively explain 57.7141648% of the total variation in the macroeconomic variables.

The outcome of our PCA is somewhat different from the results presented in the article. Indeed, the weights of the components are different. This can be explained by modifications of the FRED-MD data between the redaction of the paper on our replication. We noticed that some variables do not have exactly the same name in our version of the FRED data and in the original article. Despite these differences, we are reassured by the fact that in the original article, the first nine PCs collectively explain 57% of the total variation.

We plot the principal components that we extracted from the 120 FRED-MD macroeconomic variables, as the authors do in **Figure 1** of their article.

```
pca_ts <- ts(data=pca$ind$coord, start = c(1960,1), frequency=12)
par(mfrow = c(3, 3), mar = c(5.1, 4.1, 4.1, 2.1))
for(i in 1:9){
  plot(pca_ts[,i],
       main = paste0("PC",i),
       ylab="")
}
```

```
sd_pc <- sapply(as.data.frame(pca$ind$coord),sd)
```

Sparse PCA

We now perform a sparse PCA, using the same R package as the authors. Before running the `SPC` function, we scale the variables (so that they have a unit variance). In the article, the authors set the shrinkage parameter so that only 108 weights are active. The set the parameter `sumabsv` to 3 to get a similar outcome.

```
library(PMA)

data_unscaled <- as.matrix(data0)
data0<-as.matrix(data0)
data0<-scale(data0) # we scale variables
spca <- SPC(data0,sumabsv = 3, K=9, trace=F)
weights <- spca$v
row.names(weights)<- colnames(data0)
sum(weights!=0)
```

```
## [1] 106
```

```
# Percentage of variance
components <- paste0("comp ", 1:9)
table2 <- data.frame(Component = components,
                     Cumulative_percentage_of_variance = spca$prop.var.explained)
kable(table2, caption = "First 9 components of the SPCA")
```

Table 2: First 9 components of the SPCA

Component	Cumulative_percentage_of_variance
comp 1	0.0708730
comp 2	0.1327347
comp 3	0.1972043
comp 4	0.2585409
comp 5	0.3149324
comp 6	0.3681686
comp 7	0.4044608
comp 8	0.4330433
comp 9	0.4634756

```
#### Identification of active weights
component_names <- c("Yields","Production", "Inflation", "Housing", "Spreads", "Employment", "Costs", "
active_weights<-rep("", 9)
for(i in 1:9){
  active_weights[i] <- paste0(row.names(weights)[weights[,i]!=0], collapse = " ; ")
```

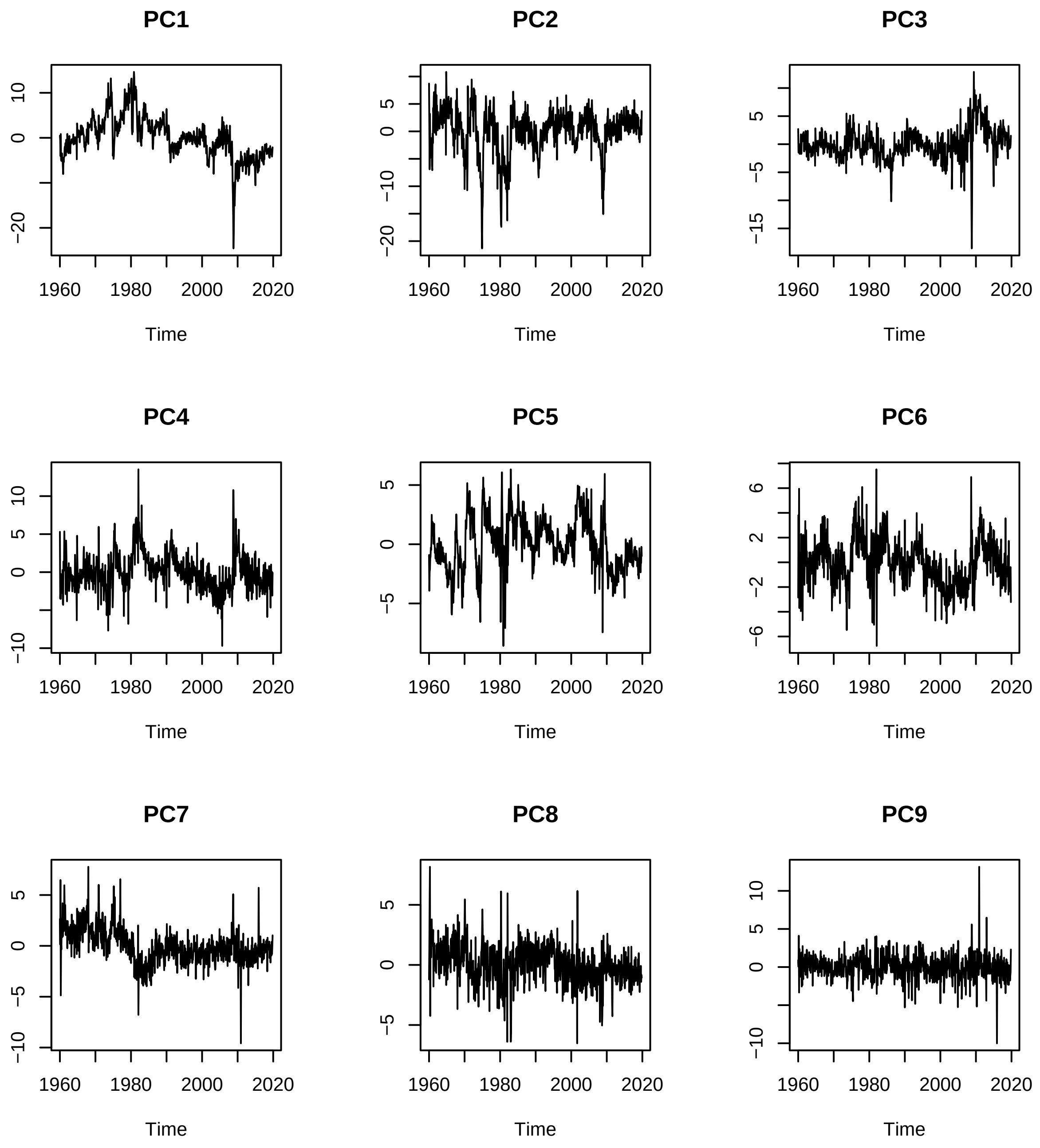

Figure 1: Conventional principal components

```r
}
active_weights_df <- data.frame(Sparse_Component = 1:9,
                                Component_name = component_names,
                                Active_weights = active_weights)
kable(active_weights_df)
```

Sparse_Component	Component_name	Active_weights
1	Yields	S.P.div.yield ; FEDFUNDS ; CP3Mx ; TB3MS ; TB6MS ; GS1 ; GS5 ; GS10 ; AAA ; BAA
2	Production	INDPRO ; IPFPNSS ; IPFINAL ; IPCONGD ; IPDCONGD ; IPBUSEQ ; IPMAT ; IPDMAT ; IPMANSICS ; CUMFNS ; MANEMP ; DMANEMP
3	Inflation	WPSFD49207 ; WPSFD49502 ; WPSID61 ; CPIAUCSL ; CPITRNSL ; CUSR0000SAC ; CPIULFSL ; CUSR0000SA0L2 ; CUSR0000SA0L5 ; PCEPI ; DNDGRG3M086SBEA
4	Housing	HOUST ; HOUSTNE ; HOUSTMW ; HOUSTS ; HOUSTW ; PERMIT ; PERMITNE ; PERMITMW ; PERMITS ; PERMITW ; REALLN
5	Spreads	BUSINVx ; COMPAPFFx ; TB3SMFFM ; TB6SMFFM ; T1YFFM ; T5YFFM ; T10YFFM ; AAAFFM ; BAAFFM ; CPIMEDSL ; CUSR0000SAS ; DSERRG3M086SBEA
6	Employment	UNRATE ; PAYEMS ; USGOOD ; USCONS ; MANEMP ; DMANEMP ; NDMANEMP ; SRVPRD ; USTPU ; USWTRADE ; USTRADE ; USFIRE
7	Costs	USFIRE ; BUSINVx ; M2REAL ; S.P.div.yield ; CPIAPPSL ; CPIMEDSL ; CUSR0000SAD ; CUSR0000SAS ; DDURRG3M086SBEA ; DSERRG3M086SBEA ; CES0600000008 ; CES2000000008 ; CES3000000008
8	Money	BUSINVx ; M1SL ; M2SL ; M2REAL ; BOGMBASE ; TOTRESNS ; WPSFD49207 ; WPSFD49502 ; WPSID61 ; WPSID62 ; PPICMM ; MZMSL
9	SPC9	DPCERA3M086SBEA ; CMRMTSPLx ; RETAILx ; IPNCONGD ; IPNMAT ; CE16OV ; CLAIMSx ; USCONS ; CES0600000007 ; AWOTMAN ; AWHMAN ; AMDMNOx ; ISRATIOx

The result of our sparse PCA is quite satisfactory, insofar as they are very similar to those represented in the article. As in the article, the nine components of the PCA explain 46% of the total variation in the 120 macroeconomic variables. By looking at the active weights of each component, we see that they do not exactly match those presented in ***Table 3*** of the article. We can nevertheless give them the same interpretation as in the article, except for the ninth component. The active weights of the ninth component diverge too much from those of the original article. In our results, it is difficult to interpret this component as an index for credit ; we therefore keep the name "SPC 9".

```r
#u <- scale(spca$u)*sd_pc
v<-spca$v
u<-data0%*%v

spca_ts <- ts(data=u, start = c(1960,1), frequency=12)
par(mfrow = c(3, 3), mar = c(5.1, 4.1, 4.1, 2.1))
for(i in 1:9){
  plot(spca_ts[,i],
       main = component_names[i],
       ylab="")
}
```

Even though our sparse components have similar interpretations as those derived by the authors, our plots are very different from those presented in **Figure 2** of the article

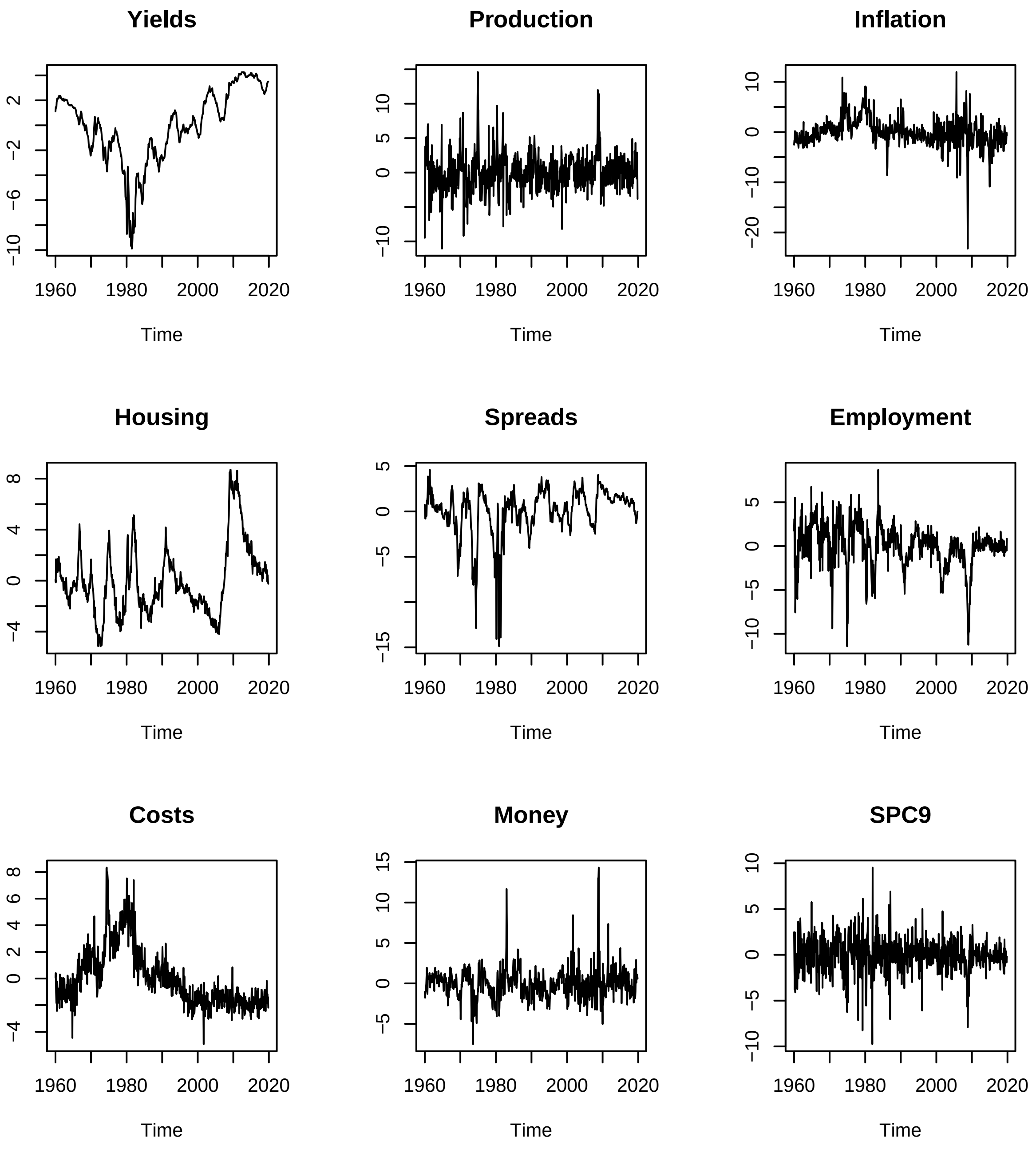

Figure 2: Sparse principal components

Innovations to the PCs

The set of macro factors is composed of the innovations to the principal components which have been extracted by the PCA. The innovations are computed by running a first-order vector autoregression (VAR(1)) on the principal components. For both the conventional and sparse PCAs, we run a VAR(1) on the PCs, we compute the residuals (which correspond to the innovations) and we then compute the correlations between those residuals.

Conventional PCA

We begin with the conventional PCA. `pca$ind$coord` contains the coordinates of each of the 120 macroeconomic variables in the space of the 9 PCs. We use the package `vars` to run the VAR(1).

```
library(vars)
```

```
## Warning: le package 'vars' a été compilé avec la version R 4.2.2
```

```
## Warning: le package 'strucchange' a été compilé avec la version R 4.2.2
```

```
data_pca <- pca$ind$coord
row.names(data_pca) <- data$date
ar_pca <- VAR(data_pca, p=1)
correlations_pca <- round(cor(residuals(ar_pca)),2)
kable(correlations_pca, caption = "Innovation correlations to conventional PCs")
```

Table 4: Innovation correlations to conventional PCs

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9
Dim.1	1.00	0.45	0.79	0.04	0.16	0.02	-0.07	-0.03	-0.04
Dim.2	0.45	1.00	0.17	0.67	-0.08	-0.13	-0.23	0.08	-0.11
Dim.3	0.79	0.17	1.00	-0.13	0.44	-0.15	-0.01	-0.05	0.03
Dim.4	0.04	0.67	-0.13	1.00	-0.17	-0.47	0.18	-0.06	-0.08
Dim.5	0.16	-0.08	0.44	-0.17	1.00	-0.08	0.01	0.09	-0.07
Dim.6	0.02	-0.13	-0.15	-0.47	-0.08	1.00	-0.10	-0.16	-0.16
Dim.7	-0.07	-0.23	-0.01	0.18	0.01	-0.10	1.00	-0.18	-0.03
Dim.8	-0.03	0.08	-0.05	-0.06	0.09	-0.16	-0.18	1.00	-0.02
Dim.9	-0.04	-0.11	0.03	-0.08	-0.07	-0.16	-0.03	-0.02	1.00

The results of this correlation matrix are very close to the one displayed in **Table 4** of the original article.

Sparse PCA

We follow the same method with the sparse PCA. Here, the coordinates of each of the 120 macroeconomic variables in the space of the 9 sparse PCs are stored in `spca$u`.

```
data_spca <- u
row.names(data_spca) <- data$date
colnames(data_spca) <- component_names
ar_spca <- VAR(data_spca, p=1)
correlations_spca <- round(cor(residuals(ar_spca)),2)
kable(correlations_spca, caption = "Innovation correlations to sparse PCs")
```

Table 5: Innovation correlations to sparse PCs

	Yields	Production	Inflation	Housing	Spreads	Employment	Costs	Money	SPC9
Yields	1.00	0.13	-0.15	0.14	0.11	-0.20	-0.17	0.05	-0.20
Production	0.13	1.00	-0.02	0.18	0.07	-0.50	-0.09	0.03	-0.61
Inflation	-0.15	-0.02	1.00	-0.06	-0.03	0.08	0.24	-0.55	0.13
Housing	0.14	0.18	-0.06	1.00	-0.17	-0.20	0.00	-0.04	-0.38
Spreads	0.11	0.07	-0.03	-0.17	1.00	-0.06	-0.11	0.09	0.05
Employment	-0.20	-0.50	0.08	-0.20	-0.06	1.00	0.07	-0.08	0.43
Costs	-0.17	-0.09	0.24	0.00	-0.11	0.07	1.00	-0.09	-0.01
Money	0.05	0.03	-0.55	-0.04	0.09	-0.08	-0.09	1.00	-0.10
SPC9	-0.20	-0.61	0.13	-0.38	0.05	0.43	-0.01	-0.10	1.00

Once again, our results look quite similar to those of the original article, except for the ninth sparse PC. However, for some correlations, the reported sign is the opposite of the one indicated in the original article.

Risk premia estimates

We now turn to the estimation of the risk premia of the sparse macro factors. The objective is to determine whether some of the macro factors generate some significant risk premia.

We import the data on portfolio returns and keep the same time period as the authors (1963:07 to 2019:12).

```r
R <- readRDS("data/portfolios.rds")
R <- filter(R, date<='2019-12-01')
dates <- R$date
R<-dplyr::select(R,-1)
```

We need to compute the excess returns of each portfolios. This requires data on the risk-free rate at every period in time. The authors use the CRSP risk-free return. However, as these data are not freely available, we replace the risk-free rate by TB3MS variable from FREDMD (3-Month Treasury Bill Secondary Market Rate, Discount Basis).

```r
data_rf <- read.csv(file = "data/TB3MS.csv")
data_rf <- dplyr::select(data_rf, -1) # we remove the date
for (i in 1:ncol(R)){
  R[,i] <- as.numeric(R[,i]) - data_rf[,1]
}
```

We demean the excess returns of each portfolio

```r
R_d <- R-t(as.matrix(colMeans(R))) # the result is != 0 due to approx errors
```

We run a PCA of the excess returns of our portfolios, to estimated the rotated fundamental factors (denoted `ksi`)

```r
t <- nrow(R_d)
n <- ncol(R_d)
R_d <- t(as.matrix(R_d))
mat <- (t(R_d) %*% R_d)/(t*n)
r_pca <- PCA(mat, ncp=15,graph = F)

ksi <- t(r_pca$var$coord) #eigenvectors
V <- sqrt(t)*t(r_pca$var$coord)

# estimator of beta (exposure to factors)
```

```r
beta <- (1/t)*R_d%*%t(V)

r_mean <- colMeans(R) #average return
gamma <- solve(t(beta)%*%beta) %*% t(beta) %*% as.matrix(r_mean) #OLS


# alternative : with OLS
lm1 <- lm(r_mean~-1+beta)
summary(lm1)
```

```
##
## Call:
## lm(formula = r_mean ~ -1 + beta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62699 -0.07578  0.00926  0.08367  0.36106
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## betaDim.1  -0.0285316  0.0001884 -151.410  < 2e-16 ***
## betaDim.2  -0.1425057  0.0029320  -48.604  < 2e-16 ***
## betaDim.3  -0.2647945  0.0115596  -22.907  < 2e-16 ***
## betaDim.4   0.0522575  0.0230164    2.270  0.02382 *
## betaDim.5   0.4417834  0.0281623   15.687  < 2e-16 ***
## betaDim.6  -0.2261978  0.0369536   -6.121 2.64e-09 ***
## betaDim.7   0.5512277  0.0657879    8.379 1.56e-15 ***
## betaDim.8  -1.0062995  0.0770880  -13.054  < 2e-16 ***
## betaDim.9   0.0949227  0.0760880    1.248  0.21308
## betaDim.10  0.9460650  0.1055167    8.966  < 2e-16 ***
## betaDim.11  1.7915206  0.1258126   14.240  < 2e-16 ***
## betaDim.12  0.1629277  0.1537956    1.059  0.29020
## betaDim.13  0.9970662  0.2213628    4.504 9.26e-06 ***
## betaDim.14  0.6403207  0.2170883    2.950  0.00341 **
## betaDim.15 -0.3667756  0.2448742   -1.498  0.13514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1397 on 330 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 1.501e+04 on 15 and 330 DF,  p-value: < 2.2e-16
```

```r
# R² proche de l'article avec intercept, mais erronné sans intercept (calcul du R² dans un modèle sans
```

The last step is to run a time-series regression of the observed factors on the rotated fundamental factors.

```r
# we restrict the observed factors to the good time period
dates_pca <- data$sasdate
indices_dates <- dates_pca>="1963-07-01" & dates_pca<= "2019-12-01"

# residuals of the VAR(1)
res <- residuals(ar_pca)
G <- res[indices_dates[-1],] # we drop the first element of res (ar(1) has one obs less)
G <- t(G)
```

```
eta <- G %*% t(V) %*% solve(V %*% t(V))


gamma_g <- eta %*% gamma
df_pca <- data.frame(Factor = paste0("PC ", 1:9),
                     gamma_g=gamma_g)
kable(df_pca, caption = "Estimators of the risk premia for the conventional PCA")
```

Table 6: Estimators of the risk premia for the conventional PCA

	Factor	gamma_g
Dim.1	PC 1	1.1415494
Dim.2	PC 2	2.2633176
Dim.3	PC 3	0.8242541
Dim.4	PC 4	1.0621275
Dim.5	PC 5	0.1873770
Dim.6	PC 6	-0.0343401
Dim.7	PC 7	-0.2785029
Dim.8	PC 8	-0.2525590
Dim.9	PC 9	-0.2533587

```
# with tslm
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
G_ts <- ts(t(G))
ksi_ts <- ts(t(ksi))
lm3 <- tslm(G_ts~0+ksi_ts)

####### same for sparse PCA :

# residuals of the VAR(1)
res_spca <- residuals(ar_spca)
G_spca <- res_spca[indices_dates[-1],] # we drop the first element of res (ar(1) has one obs less)
G_spca <- t(G_spca)

eta_spca <- G_spca %*% t(V) %*% solve(V %*% t(V))


gamma_g_spca <- eta_spca %*% gamma
df_spca <- data.frame(Factor = component_names,
                     gamma_g=gamma_g_spca)
kable(df_spca, caption = "Estimators of the risk premia for the sparse PCA")
```

Table 7: Estimators of the risk premia for the sparse PCA

	Factor	gamma_g
Yields	Yields	-0.0563535
Production	Production	-1.9270851
Inflation	Inflation	0.3885636

	Factor	gamma_g
Housing	Housing	-0.1804507
Spreads	Spreads	-0.0255368
Employment	Employment	1.1800340
Costs	Costs	0.4265595
Money	Money	0.3305398
SPC9	SPC9	1.2334070

Three-pass methodology, with notations of Zhou-Rapach

J'ai essayé de répliquer le résultat en utilisant les notations de notre article (Zhou-Rapach).

Importation of asset returns

```r
R <- readRDS("data/portfolios.rds")
R <- filter(R, date<='2019-12-01')
dates <- R$date
R<-dplyr::select(R,-1)

data_rf <- read.csv(file = "data/TB3MS.csv")
data_rf <- dplyr::select(data_rf, -1) # we remove the date
for (i in 1:ncol(R)){
  R[,i] <- as.numeric(R[,i]) - data_rf[,1]
}
R <- t(R)
```

First, we stationarize and normalise our data

```r
#the stationarize version
R_sta <- matrix(0,nrow=nrow(R),ncol = (ncol(R)-1))
for (i in 1:nrow(R_sta)){
  R_sta[i,] <- diff(R[i,])
}

#the demeanded version
R_d <- R_sta
for (i in 1:nrow(R_d)){
  R_d[i,] <- R_sta[i,] - mean(R_sta[i,])
}
```

1- Apply conventional PCA to demeaned excess returns for the N test assets to estimate the rotated fundamental factors STEP 1

```r
t <- ncol(R)
n <- nrow(R)
mat <- (t(R_d) %*% R_d)/(t*n)
r_pca <- PCA(mat, ncp=15, graph=F, scale.unit = TRUE)

#ksi <- t(r_pca$var$coord) #eigenvectors
V <- sqrt(t)*t(r_pca$var$coord)

#normalize Vt
for (i in 1:nrow(V)){
  V[i,] <- (1/sqrt(t))*(V[i,])/sqrt(var(V[i,]))
}
```

```
#V %*% t(V) #it's now eq to the identity of dim 15 = p_hat

# estimator of beta (exposure to factors)
beta <- (1/t)*R_d%*%t(V)
```

2- Run time-series regressions of r on ksi to estimate beta Run a cross-sectional regression of r_bar on the columns of beta to estimate gamma Pb : time-regressions de r ou de r demeaned? Je pense que c'est r demeaned (mais ça équivaut normalement à régresser r avec constante)

```
r_mean <- matrix(rowMeans(R_sta)) #average return
gamma <- solve(t(beta)%*%beta) %*% t(beta) %*% r_mean #OLS

#Now, we calculate Rf2 associated with this cross-sectional reg
r_mean_mean <- mean(r_mean) #almost 0 by construction
Rf2 <- sum((beta %*% gamma - r_mean_mean)**2) / sum((r_mean-r_mean_mean)**2)
Rf2
```

```
## [1] 0.4945828
```

STEP 2 - variant avec OLS

```
# alternative : with OLS
lm1 <- lm(r_mean~beta)
summary(lm1)
```

```
##
## Call:
## lm(formula = r_mean ~ beta)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.016690 -0.001599  0.000168  0.001830  0.009673
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.005176   0.002672   1.937  0.05360 .
## betaDim.1   -0.015336   0.012588  -1.218  0.22400
## betaDim.2   -0.028100   0.006543  -4.295 2.31e-05 ***
## betaDim.3    0.040245   0.015867   2.536  0.01166 *
## betaDim.4    0.028457   0.012974   2.193  0.02898 *
## betaDim.5    0.023143   0.010850   2.133  0.03366 *
## betaDim.6   -0.068836   0.016343  -4.212 3.27e-05 ***
## betaDim.7    0.028611   0.015865   1.803  0.07223 .
## betaDim.8   -0.043887   0.015614  -2.811  0.00524 **
## betaDim.9   -0.005115   0.016989  -0.301  0.76354
## betaDim.10   0.074361   0.016447   4.521 8.59e-06 ***
## betaDim.11  -0.041440   0.015991  -2.591  0.00998 **
## betaDim.12   0.026547   0.019276   1.377  0.16939
## betaDim.13  -0.051394   0.022219  -2.313  0.02134 *
## betaDim.14  -0.041297   0.018465  -2.237  0.02599 *
## betaDim.15   0.003620   0.022158   0.163  0.87032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002994 on 329 degrees of freedom
## Multiple R-squared:  0.4803, Adjusted R-squared:  0.4566
## F-statistic: 20.27 on 15 and 329 DF,  p-value: < 2.2e-16
```

```
#on trouve le même R² à la main avec l'étape d'avant
```

3. Run time-series regressions of g on ksi to estimate theta

```
# we restrict the observed factors to the good time period
dates_pca <- data$sasdate
# we drop the first element of res (ar(1) has one obs less)
indices_dates <- dates_pca>="1963-08-01" & dates_pca<= "2019-12-01"
#indices_dates <- dates_pca>="1963-07-01" & dates_pca<= "2019-12-01"

# residuals of the VAR(1)
res <- residuals(ar_pca)
G <- res[indices_dates[-1],]
G <- t(G)
#shall we normalize? I don't think

#to use the OLS, we delet the constant using the time average
G_d <- G - rowMeans(G)
eta <- G_d %*% t(V) %*% solve(V %*% t(V))
gamma_g <- eta %*% gamma
rg <- 1:9
for (i in 1:9){
  rg[i] <- (sum(((eta %*% V)[i,] - rowMeans(G_d)[i])**2))/(sum((G_d[i,] - rowMeans(G_d)[i])**2))
}
df_pca <- data.frame(Factor = paste0("PC ", 1:9),
                     gamma_g=gamma_g, Rg = rg)
kable(df_pca, caption = "Estimators of the risk premia for the conventional PCA")
```

Table 8: Estimators of the risk premia for the conventional PCA

	Factor	gamma_g	Rg
Dim.1	PC 1	-0.1842100	0.0119212
Dim.2	PC 2	0.2820189	0.0186665
Dim.3	PC 3	-0.4492634	0.0142294
Dim.4	PC 4	0.1359776	0.0174903
Dim.5	PC 5	-0.2797554	0.0229127
Dim.6	PC 6	-0.4810276	0.0356222
Dim.7	PC 7	-0.0330074	0.0120626
Dim.8	PC 8	0.5912853	0.0199767
Dim.9	PC 9	0.0958569	0.0272220

```
# with tslm
#library(forecast)
#G_ts <- ts(t(G))
#ksi_ts <- ts(t(ksi))
#lm3 <- tslm(G_ts~0+ksi_ts)

####### same for sparse PCA :

# residuals of the VAR(1)
res_spca <- residuals(ar_spca)
G_spca <- res_spca[indices_dates[-1],] # we drop the first element of res (ar(1) has one obs less)
```

```r
G_spca <- t(G_spca)
G_spca_d <- G_spca - rowMeans(G_spca)
eta_spca <- G_spca_d %*% t(V) %*% solve(V %*% t(V))
gamma_g_spca <- eta_spca %*% gamma
rgs <- 1:9
for (i in 1:9){
  rgs[i] <- (sum(((eta_spca %*% V)[i,] - rowMeans(G_spca_d)[i])**2))/(sum((G_spca_d[i,] - rowMeans(G_spc
}
df_spca <- data.frame(Factor = component_names,
                     gamma_g=gamma_g_spca, Rg = rgs)

kable(df_spca, caption = "Estimators of the risk premia for the sparse PCA")
```

Table 9: Estimators of the risk premia for the sparse PCA

	Factor	gamma_g	Rg
Yields	Yields	0.0810399	0.0439517
Production	Production	-0.2850333	0.0161897
Inflation	Inflation	-0.5058385	0.0174907
Housing	Housing	-0.1213422	0.0243281
Spreads	Spreads	-0.1787040	0.0213622
Employment	Employment	0.2261529	0.0101166
Costs	Costs	-0.2396293	0.0234005
Money	Money	-0.0192897	0.0128196
SPC9	SPC9	0.0018723	0.0136646

Modifs

STEP1- Apply conventional PCA to demeaned excess returns for the N test assets to estimate the rotated fundamental factors STEP 1

```r
r_t <- t(R) # excess returns, one row per date

r_pca <- PCA(r_t, ncp=15, graph=F, scale.unit = TRUE)
ksi <- r_pca$ind$coord #rotated factors
```

STEP 2- Run time-series regressions of r on ksi to estimate beta. Run a cross-sectional regression of r_bar on the columns of beta to estimate gamma Pb : time-regressions de r ou de r demeaned? Je pense que c'est r demeaned (mais ça équivaut normalement à régresser r avec constante)

```r
lm_step2 <- tslm(ts(r_t)~ts(ksi)) #without constant

beta <- t(lm_step2$coefficients)
beta <- beta[,-1] # we drop the constant

r_bar <- colMeans(t(R)) #average return
lm_step2_CS <- lm(r_bar~beta+0) # no intercept in the model (euation 3.3)
summary(lm_step2_CS)
```

```
##
## Call:
## lm(formula = r_bar ~ beta + 0)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.43017 -0.06579  0.00426  0.06979  0.38689
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## betats(ksi)Dim.1  -10.39620    0.01910 -544.271  < 2e-16 ***
## betats(ksi)Dim.2   -1.48255    0.01914  -77.474  < 2e-16 ***
## betats(ksi)Dim.3    0.24949    0.01860   13.411  < 2e-16 ***
## betats(ksi)Dim.4   -0.57035    0.01842  -30.967  < 2e-16 ***
## betats(ksi)Dim.5   -0.12294    0.01733   -7.096 7.89e-12 ***
## betats(ksi)Dim.6   -0.26862    0.01817  -14.780  < 2e-16 ***
## betats(ksi)Dim.7    0.22274    0.01823   12.221  < 2e-16 ***
## betats(ksi)Dim.8   -0.06708    0.01732   -3.874 0.000129 ***
## betats(ksi)Dim.9    0.03106    0.01880    1.652 0.099492 .
## betats(ksi)Dim.10  -0.14594    0.01827   -7.990 2.28e-14 ***
## betats(ksi)Dim.11   0.17223    0.01766    9.751  < 2e-16 ***
## betats(ksi)Dim.12   0.20244    0.01816   11.148  < 2e-16 ***
## betats(ksi)Dim.13  -0.08104    0.01609   -5.036 7.82e-07 ***
## betats(ksi)Dim.14   0.05553    0.01815    3.060 0.002399 **
## betats(ksi)Dim.15   0.04418    0.01889    2.339 0.019917 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1205 on 330 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 2.019e+04 on 15 and 330 DF,  p-value: < 2.2e-16
```

```r
gamma <- matrix(coefficients(lm_step2_CS)) # without the constant

# R² très proche de l'article!
```

STEP 3. Run time-series regressions of g on ksi to estimate theta

```r
# we restrict the observed factors to the good time period
dates_pca <- data$sasdate
# we drop the first element of res (ar(1) has one obs less)
indices_dates <- dates_pca>="1963-07-01" & dates_pca<= "2019-12-01"

# residuals of the VAR(1)
g_t_conv <- ts(residuals(ar_pca)[indices_dates[-1],])
View(g_t_conv)

lm_factors_conv <- tslm(g_t_conv~ts(ksi)) # without constant (equation 3.2)
theta_conv <- t(coefficients(lm_factors_conv))
theta_conv <- theta_conv[,-1]

r_squared_g_conv <- vector()
# Computation of the R²
for(i in 1:9){
  lm_tmp <- lm(g_t_conv[,i]~ksi)
  r_squared_g_conv<-c(r_squared_g_conv, summary(lm_tmp)$r.squared)
}
r_squared_g_conv <- round(100*r_squared_g_conv,2)
```

Conclusion :

```r
gamma_g_conv <- theta_conv%*% gamma
df <- data.frame(Factor = paste0("PC",1:9),
                 gamma_g = round(gamma_g_conv,3),
                 R_g_squared = paste0(r_squared_g_conv,"%"))
kable(df, caption = "Estimators of the risk premia for the conventional PCA", row.names = F)
```

Table 10: Estimators of the risk premia for the conventional PCA

Factor	gamma_g	R_g_squared
PC1	0.145	3.14%
PC2	-0.006	3.92%
PC3	0.051	2.11%
PC4	0.001	3.38%
PC5	-0.005	1.18%
PC6	0.058	3.18%
PC7	0.012	2.75%
PC8	-0.052	4.48%
PC9	-0.077	2.86%

On n'est pas trop loin des résultats de l'article!

Same method for SPCA

```r
g_t_sparse <- ts(residuals(ar_spca)[indices_dates[-1],])

View(g_t_sparse)

lm_factors_sparse <- tslm(g_t_sparse~ts(ksi)) # without constant (equation 3.2)
theta_sparse <- t(coefficients(lm_factors_sparse))
theta_sparse <- theta_sparse[,-1]
View(theta_sparse)

r_squared_g_sparse <- vector()
# Computation of the R²
for(i in 1:9){
  lm_tmp <- lm(g_t_sparse[,i]~ksi)
  r_squared_g_sparse<-c(r_squared_g_sparse, summary(lm_tmp)$r.squared)
}
r_squared_g_sparse <- round(100*r_squared_g_sparse,2)


gamma_g_sparse <- theta_sparse%*% gamma
df <- data.frame(Factor = paste0("PC",1:9),
                 gamma_g = round(gamma_g_sparse,4),
                 R_g_squared = paste0(r_squared_g_sparse,"%"))
kable(df, caption = "Estimators of the risk premia for the sparse PCA", row.names = F)
```

Table 11: Estimators of the risk premia for the sparse PCA

Factor	gamma_g	R_g_squared
PC1	-0.0668	13.83%
PC2	0.0193	3.39%
PC3	0.0918	2.02%

Factor	gamma_g	R_g_squared
PC4	-0.0137	4.26%
PC5	-0.0246	2.18%
PC6	0.0559	2.36%
PC7	0.0524	1.72%
PC8	-0.0353	4.63%
PC9	-0.0158	5.21%

Biases without the three-pass methodology

What happens if we do not use the 3-pass methodology?

The motivation for using the three-pass methodology is to avoid potential omitted factors bias. We now study whether this concern is relevant, i.e. whether there is evidence of such biases. To achieve this, we estimate the risk premia with a simple two-pass methodology, and then compare our results to the outcome of the three-pass methodology.

Let us therefore assume that the true model for asset returns only depends on our macro factors :

$\boldsymbol{r_t = \alpha + \beta' \gamma + \beta' g_t + \varepsilon_t}$

Where $\boldsymbol{g_t}$ are the macro factors (i.e. the innovations to the PCs) and $\boldsymbol{\gamma}$ the risk premia.

If this assumption is true, then we can derive unbiased estimates of the risk premia with a two-pass methodology. This methodology consists in two steps :

1. Time series regression of the demeaned asset excess returns on the innovations to the macro factors, to estimate the risk exposures of each asset ($\boldsymbol{\beta}$)
2. Cross-sectional regression of the average returns of each asset on the asset' risk exposures to estimate the risk premia ($\boldsymbol{\gamma}$)

We run this estimation on the macro factors obtained with the conventional PCA, and then on the sparse macro factors.

Conventional PCA

Importation of returns

`g_t_conv`correspond to the conventional macro factors, i.e. the innovatiopns to the conventional PCs.

```r
R <- readRDS("data/portfolios.rds")
R <- filter(R, date<='2019-12-01')
dates <- R$date
R<-dplyr::select(R,-1)

data_rf <- read.csv(file = "data/TB3MS.csv")
data_rf <- dplyr::select(data_rf, -1) # we remove the date
for (i in 1:ncol(R)){
  R[,i] <- as.numeric(R[,i]) - data_rf[,1]
}

r_t <- R

r_t <- ts(R) # excess returns

#R_d <- R-t(as.matrix(colMeans(R))) # excess returns
```

```
## TS regression
g_t_conv <- ts(residuals(ar_pca)[indices_dates[-1],])

lm_pca <- tslm(r_t~g_t_conv)
beta <- t(lm_pca$coefficients)

beta <- beta[,-1]#we drop the constants

# CS regression
r_bar <- colMeans(R)

lm_pca_2 <- lm(r_bar~beta)
summary(lm_pca_2)
```

```
##
## Call:
## lm(formula = r_bar ~ beta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88551 -0.08092  0.02000  0.11239  0.45891
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.9151     0.1073 -36.479  < 2e-16 ***
## betag_t_convDim.1  -0.3671     0.2175  -1.688 0.092400 .
## betag_t_convDim.2  -0.2742     0.3017  -0.909 0.364151
## betag_t_convDim.3   0.2308     0.2397   0.963 0.336401
## betag_t_convDim.4  -0.5678     0.2159  -2.630 0.008943 **
## betag_t_convDim.5   0.6723     0.1720   3.909 0.000112 ***
## betag_t_convDim.6   0.5698     0.1773   3.214 0.001435 **
## betag_t_convDim.7   0.1833     0.1706   1.074 0.283384
## betag_t_convDim.8   0.2010     0.1594   1.261 0.208029
## betag_t_convDim.9  -0.4839     0.1433  -3.377 0.000819 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1886 on 335 degrees of freedom
## Multiple R-squared:  0.2229, Adjusted R-squared:  0.2021
## F-statistic: 10.68 on 9 and 335 DF,  p-value: 1.383e-14
```

```
kable(coefficients(lm_pca_2))
```

	x
(Intercept)	-3.9151195
betag_t_convDim.1	-0.3670601
betag_t_convDim.2	-0.2741985
betag_t_convDim.3	0.2307868
betag_t_convDim.4	-0.5677929
betag_t_convDim.5	0.6723108
betag_t_convDim.6	0.5697628
betag_t_convDim.7	0.1832690
betag_t_convDim.8	0.2010384

	x
betag_t_convDim.9	-0.4838797

```
g_t_sparse <- ts(residuals(ar_spca)[indices_dates[-1],])

lm_spca <- tslm(r_t~g_t_sparse)
beta_s <- t(lm_spca$coefficients)
beta_s <- beta_s[,-1]#we drop the constants


lm_spca_2 <- lm(r_bar~beta_s)
summary(lm_spca_2)
```

Sparse PCA

```
##
## Call:
## lm(formula = r_bar ~ beta_s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94511 -0.07988  0.01465  0.10877  0.47096
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -4.33752    0.12564 -34.525  < 2e-16 ***
## beta_sg_t_sparseYields       0.10781    0.02365   4.558 7.24e-06 ***
## beta_sg_t_sparseProduction  0.76683    0.26115   2.936  0.00355 **
## beta_sg_t_sparseInflation   0.49544    0.24659   2.009  0.04532 *
## beta_sg_t_sparseHousing    -0.26202    0.04995  -5.245 2.77e-07 ***
## beta_sg_t_sparseSpreads     0.20484    0.12108   1.692  0.09161 .
## beta_sg_t_sparseEmployment -0.08645    0.21118  -0.409  0.68253
## beta_sg_t_sparseCosts       0.18844    0.13196   1.428  0.15420
## beta_sg_t_sparseMoney      -0.21235    0.14413  -1.473  0.14160
## beta_sg_t_sparseSPC9       -0.12579    0.18180  -0.692  0.48945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1794 on 335 degrees of freedom
## Multiple R-squared:  0.2968, Adjusted R-squared:  0.2779
## F-statistic: 15.71 on 9 and 335 DF,  p-value: < 2.2e-16
```

Even though those estimates are biased, we find that the sparse components 1 and 4 (yield and housing) generate significant risk premia. This result is consistent with the result of the original article.