Analysis of pair trading with financial market data

Marcell P. Granát

2021 09 09

Contents

Abstract

If one attends to the extremely large literature of demographic trends in the developed world, then the uncertainty about the effect of economic and human development factors on the fertility rate cannot be covered for a long time. Several empirical studies argue for the existence of the J-shaped effect of the development, but many papers come up with statements to the opposite. The goal of this paper is to contribute to the literature with an advanced panel econometric model based on regional observations. Beyond the human development factors (living standard, education and health) I extend my analysis by using youth unemployment and family benefit indicators as dependent variables. Important to note that statistics about unemployment are available only for a critically short period in the case of many regions. To manage this highly unbalanced nature of the dataset – while not rejecting the possibility to control for youth unemployment – I estimate the model with two different modeling frames: one without youth unemployment and another one with it. As a result, the paper confirms the empirical evidence that increasing human development in developed countries has a positive effect on total fertility rates, and income is the most important component. This finding is robust to the mentioned two frameworks. In contrast, the research come up only with week evidence for the significant effect of expenditure on family on total fertility rates on the long run.

Keywords — fertility rates, human development

List of Tables

List of Figures

Introduction

Literature review

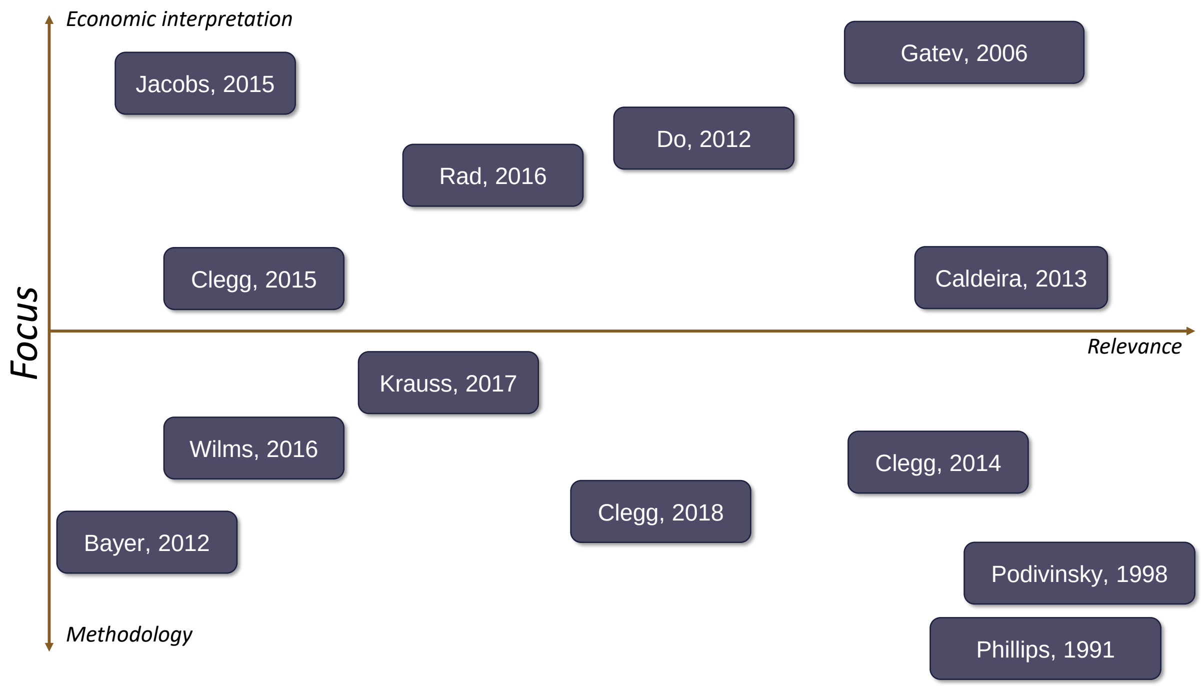

Figure 1: Classification of the core literature.

Empirical usage of traditional econometric tools

Explanatory data analysis

Engle-Granger method is a simple way to test cointegration in the bivariate case. Cointegration is diagnosed if the two tested series are integrated in the same order and a linear combination of them exist, which has an integration order of the original non-stationer series minus one [Kirchgässner and Wolters, 2007]. The most common is when the tested stock prices are I(1) and their linear combination is stationer.

The used stock prices are presented in figure 1. For a first glance, there is a high chance that some cointegrated pairs can be found in this set of series. To commit the tests the first step is to check the time-series integration order. For this purpose, I use ADF-test with a significance level of 5%. As a result, it is concluded that all the series are I(1) if any of their bivariate linear combinations is stationer, then cointegration is diagnosed. The first difference in the stock prices is shown in figure **??**.

Engle-Granger method

The second step is to run OLS with all the possible pairs and check if there is a series of residuals stationer. Just as at the previous step the stationary test is augmented Dickey-Fuller test without constant or trend component in the auxiliary regression and $\alpha = 5\%$.

With the described parameters[1] the tests confirm only one cointegrated pair (see Figure 4), and that result holds only if the stock price of Bank of America is in regressor role, but it does not, when that is used as

[1] In my previously mentioned GitHub repository, you may find that I wrote an R function to commit the whole Engle-Granger method with specified parameters. It would be reasonable to see the results with a different stationary test or with a different

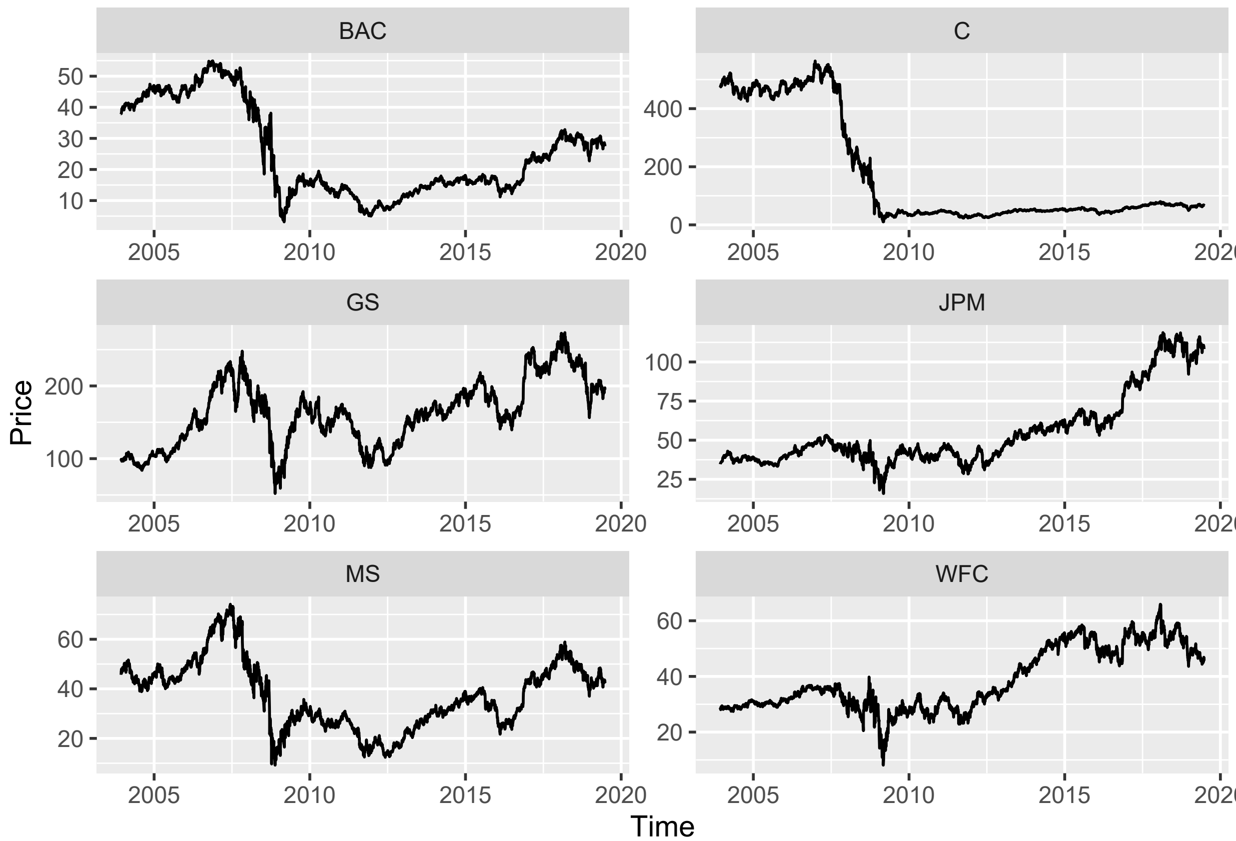

Figure 2: Time-series used in this study

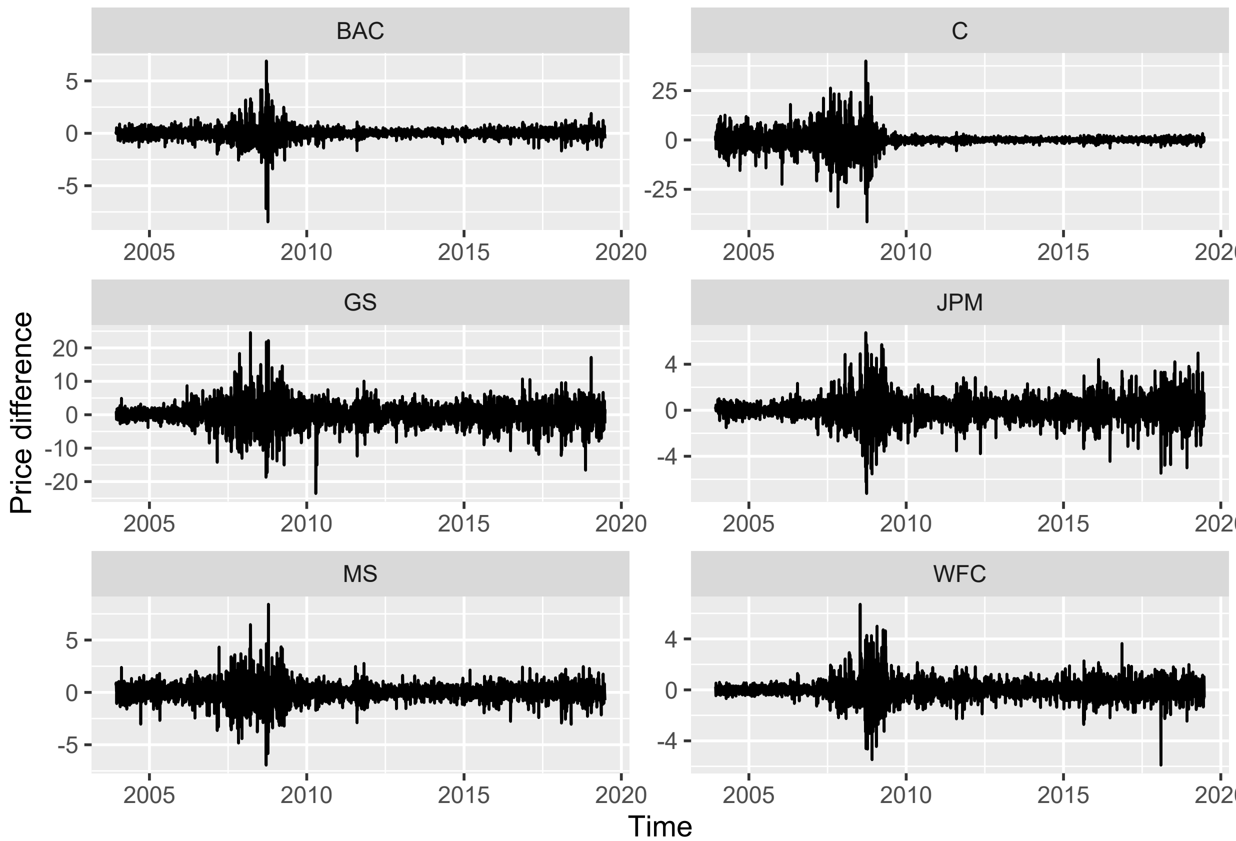

Figure 3: First difference of the time-series

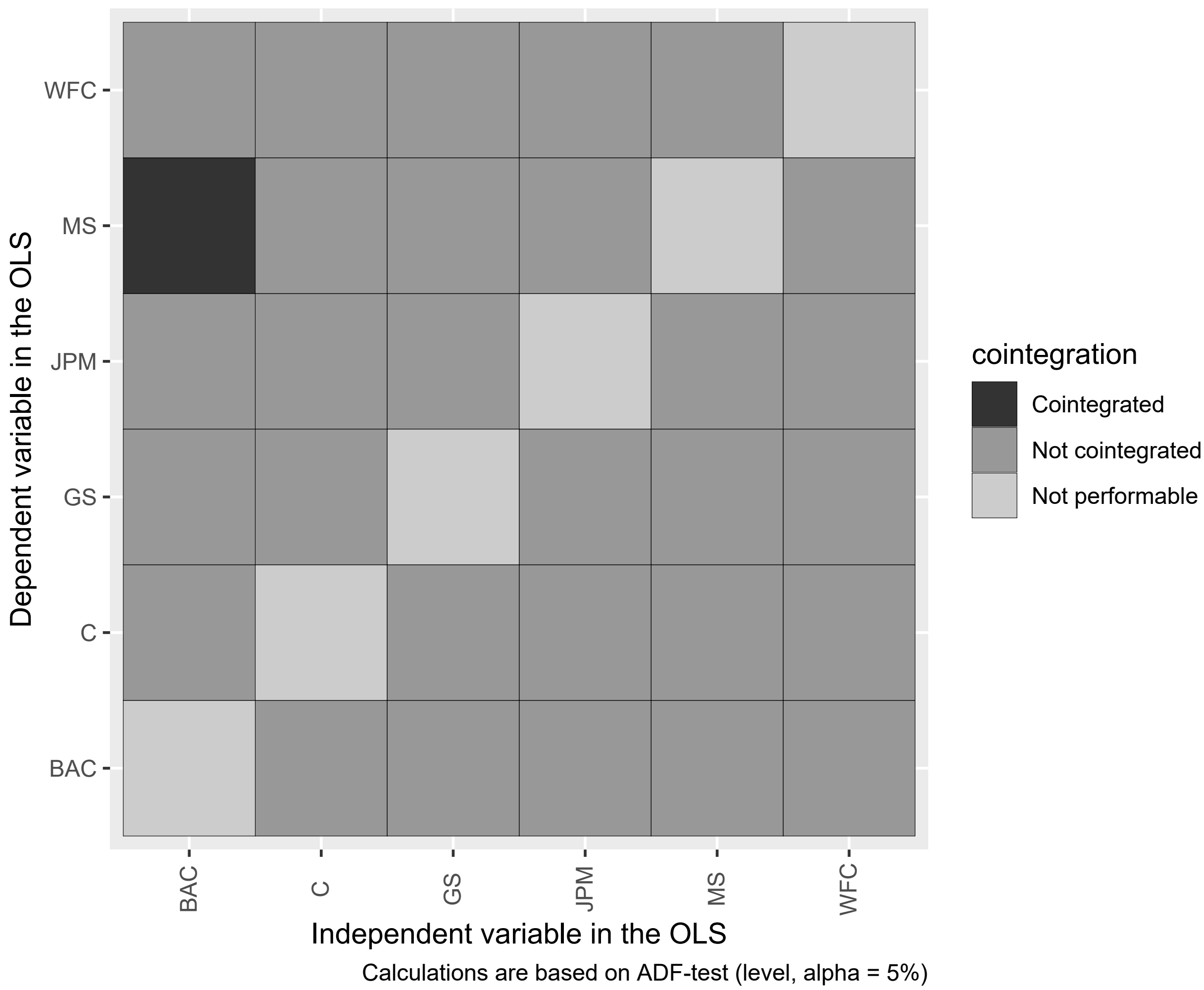

Figure 4: Results of Engle-Granger method

dependent variable[2].

Johansen-test

Johansen test is adequate cointegration test when there are more than two tested series at the same time. This test is performed to estimate the number of cointegrated vectors (r) in the system. If there is any cointegration in the model then $0 < r < k$, where k is the number of tested time-series. The system decomposition is not unique, so we can only estimate the cointegration rank r [Kirchgässner and Wolters, 2007]. The method can be performed with several tests, in this paper I chose the Lmax test. It gives a vector of the test statistics as a result and that may be compared to critical values. The null hypothesis is that $r \leq x$, where $x = 0, 1, 2, ..., k - 1$. The number of cointegrated vectors is the smallest x, under which the null hypothesis is not rejected. The empirical analysis in this study shows that the r in this system is 1 on the full time-interval[3], which confirms the identical result like the one found with the Engel-Granger method.

Engle-Granger method with rolling window

In this section, I expound the results of the previously presented Engle-Granger method performed with a rolling window. The size of the windows is 250 days. Important to note, it is not sure that a stock price has the same integration order in each window. It can happen that a cointegration test is not performable, because in that period the integration orders do not match. Since this calculation is heavily time-consuming, only three of the six stock will be tested in this paper. This means that the maximum number of cointegrated pairings is 6 ($3 \times 3 - 3$). The test parameters are the same as described before, results are shown in figure 6.

In figure 6 it can be seen that the number of cointegrated pairings reaches the maximum number at the end of 2008, 2012 and in the middle of 2008, 2016. In 2008 there is also a long period when there are 4 cointegrated pairings. This result suggests a pattern that in recession cointegration may be more frequent.

Johansen test with rolling window

Performing the Johansen test with a rolling window is a similar extension as the one presented in the previous chapter. The calculations were performed with the same 250 window size and r is examined at the significance level of 1%, 5% and 10%. The result can be seen in figure **??**.

In figure **??** the period of recession is also visualized. It looks like the $r = 1$ result at that time is more frequent than most of the case when there is no recession, similarly the $r = 2$ result. One deviation from this pattern is at 2018, where $r = 1$ result is extremely frequent.

Looking at the distribution of the results controlling for the period of recession also confirms this hypothesis. During a recession, the proportion of $r = 2$ result (2.19%) is twice as much as the proportion when there is not recession (1.08%) with 10% significance level. Similarly $r = 1$ is the result of 15.31% of the total tests performed with $\alpha = 10\%$ in periods of recession, while 7.34% is when there is expansion. With different significance level, identical results can be concluded.

---

significance level (especially if calculating its profitability is also in focus). With the written function, it is possible to modify the test parameters and see how the results change.

[2] The matrix of the results is not a symmetrical.

[3] Same result is stated on 1%, 5% and 10% significance level.

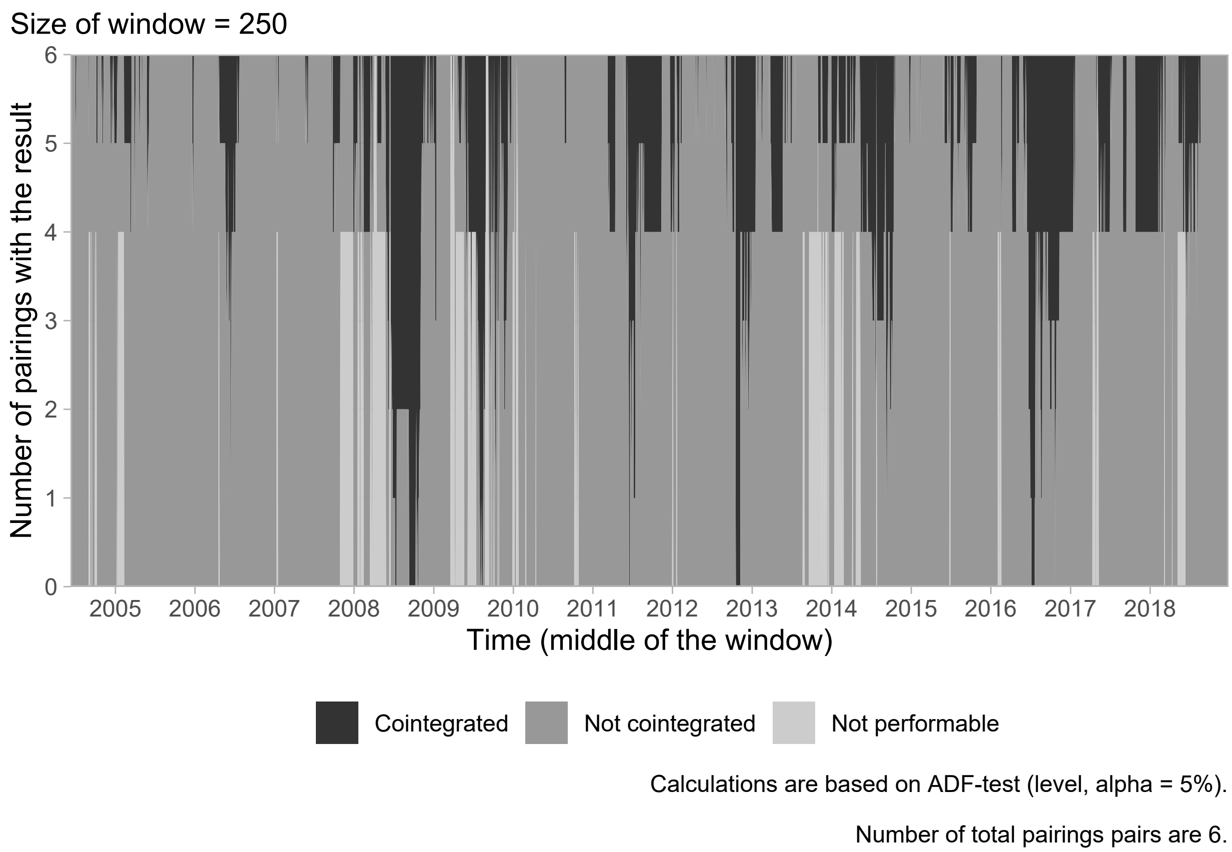

Figure 5: Results of Engle-Granger method with rolling window

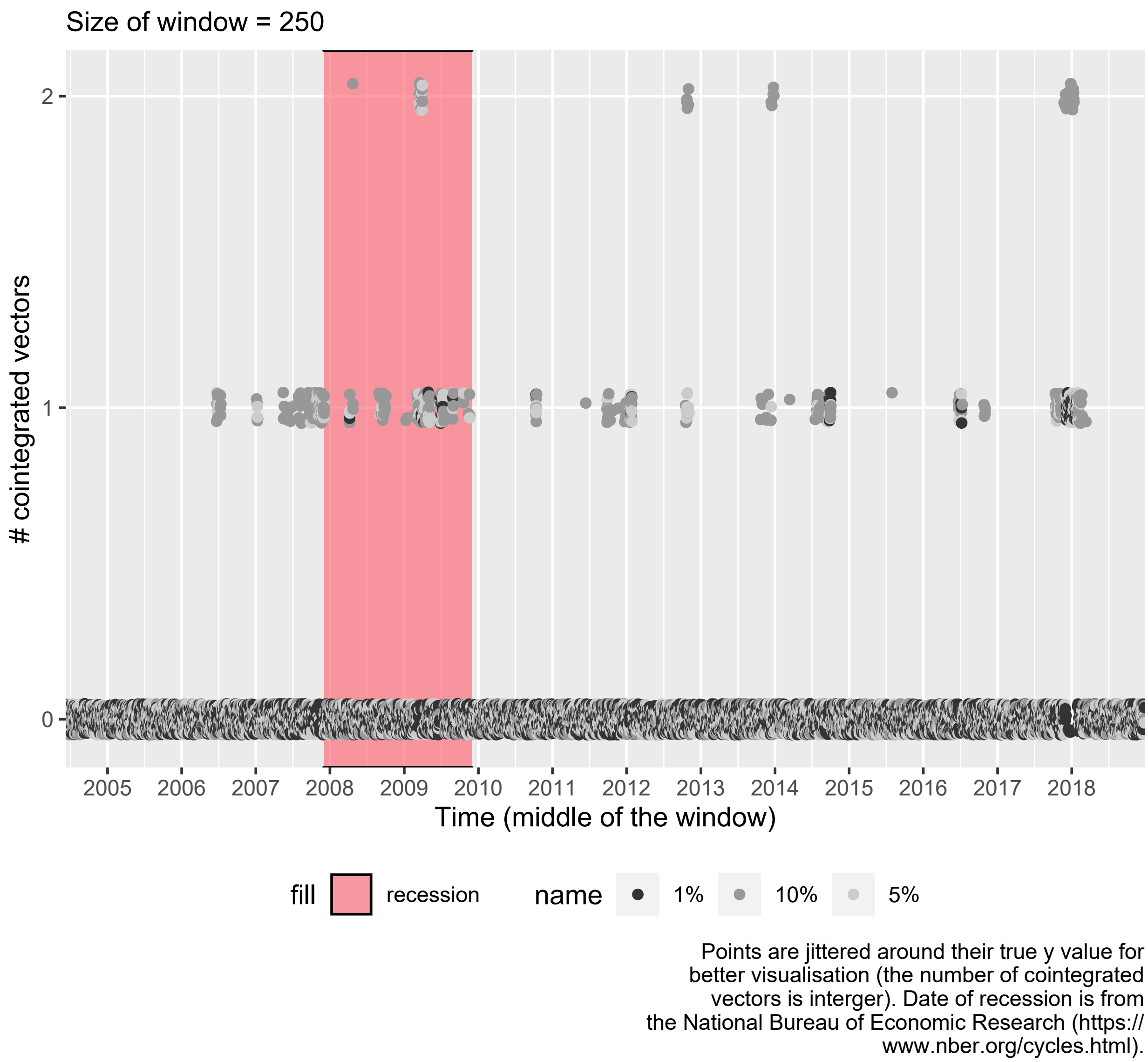

Figure 6: Results of Johansen-test with rolling window across time

References

[Bayer and Hanck, 2013] Bayer, C. and Hanck, C. (2013). Combining non-cointegration tests. *Journal of Time Series Analysis*, 34(1):83–95.

[Caldeira and Moura, 2013a] Caldeira, J. and Moura, G. V. (2013a). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy.

[Caldeira and Moura, 2013b] Caldeira, J. and Moura, G. V. (2013b). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *SSRN Electronic Journal*.

[Clegg, 2014] Clegg, M. (2014). On the persistence of cointegration in pairs trading. *SSRN Electronic Journal*.

[Clegg and Krauss, 2018] Clegg, M. and Krauss, C. (2018). Pairs trading with partial cointegration. *Quantitative Finance*, 18(1):121–138.

[Do and Faff, 2012] Do, B. and Faff, R. (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2):261–287.

[Gonzalo and Lee, 1998] Gonzalo, J. and Lee, T.-H. (1998). Pitfalls in testing for long run relationships. *Journal of Econometrics*, 86(1):129–154.

[Huck, 2015] Huck, N. (2015). Pairs trading: does volatility timing matter? *Applied Economics*, 47(57):6239–6256.

[Huck and Afawubo, 2014] Huck, N. and Afawubo, K. (2014). Pairs trading and selection methods: is cointegration superior? *Applied Economics*, 47(6):599–613.

[Jacobs and Weber, 2015] Jacobs, H. and Weber, M. (2015). On the determinants of pairs trading profitability. *Journal of Financial Markets*, 23:75–97.

[Kirchgässner and Wolters, 2007] Kirchgässner, G. and Wolters, J. (2007). *Introduction to modern time series analysis*. Springer, Berlin and New York.

[Krauss and Herrmann, 2017] Krauss, C. and Herrmann, K. (2017). On the power and size properties of cointegration tests in the light of high-frequency stylized facts. *Journal of Risk and Financial Management*, 10(1):7.

[Neely et al., 2014] Neely, C. J., Rapach, D. E., Tu, J., and Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators. *Management Science*, 60(7):1772–1791.

[Rad et al., 2016] Rad, H., Low, R. K. Y., and Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 16(10):1541–1558.

Appendix: R codes

```r
# Setup ----------------------------------------------------------------------------------

library(tidyverse)
library(urca)

WD <- getwd() %>% # root directory
  gsub(pattern = "PairsTrading.*", replacement = "PairsTrading")

load(str_c(WD, "/data.RData")) # financial assets data

theme_set(theme_light() + theme(
  legend.title = element_blank(),
  plot.title.position = "plot",
  plot.tag.position = "topright",
  plot.caption.position = "plot"
))

# EDA ------------------------------------------------------------------------------------

Bankdata %>%
  pivot_longer(-1) %>%
  ggplot(aes(x = Date, y = value)) +
  geom_line() +
  facet_wrap(vars(name), nrow = 3, scales = "free") +
  labs(
    x = "Time", y = "Price"
  )

Bankdata %>% select(-1) %>% cor() %>% data.frame() %>% rownames_to_column() %>%
  pivot_longer(-1) %>% mutate(
  value = ifelse(rowname == name, NA, value)
) %>%
  ggplot(aes(rowname, name, fill = value)) + geom_tile(color = "black") +
      scale_fill_gradient2(
      low = "#00A3AB", high = "#FF5B6B", space = "Lab", na.value = "grey50",
      guide = "legend", midpoint = 0, aesthetics = "fill", limits = c(-1,1)
    ) + labs(
      x = "", y = "", title = "Correlation-matrix", tag = "Not included"
    ) + theme(
      panel.border = element_blank()
)

# Engle-Granger method -------------------------------------------------------------------

Bankdata %>%
  select(-1) %>%
  apply(2, function(x) {
    # number of differences required for stationarity to each series
    forecast::ndiffs(x, test = "adf", alpha = 0.05, type = "level")
  })

Bankdata %>%
```

```r
  select(-1) %>%
  apply(2, function(x) {
    diff(x)
  }) %>%
  data.frame() %>%
  mutate(
    Date = tail(Bankdata$Date, -1)
  ) %>%
  pivot_longer(-Date) %>%
  ggplot(aes(x = Date, y = value)) +
  geom_line() +
  facet_wrap(vars(name), nrow = 3, scales = "free") +
  labs(
    x = "Time", y = "Price difference"
  )

cointegration_tests <- function(df, test, type, alpha) {
  # test cointegrity for all combination in a df
  ndiff_df <- df %>%
    select(-1) %>%
    apply(2, function(x) { # # of differences required for stationarity to each series
      forecast::ndiffs(x, test = test, alpha = alpha, type = type)
    })

  v <- df %>% select(-1) %>% # remove year ---> IT MUST BE IN THE INPUT DF !
    names(.)
  df2 <- expand.grid(v, v) %>%
    rename_all(funs(c("y", "x"))) %>%
    mutate(
      y = as.character(y),
      x = as.character(x),
      ndiff = ifelse(ndiff_df[y] == ndiff_df[x], ndiff_df[y], 0),
      ndiff = ifelse(y == x, 0, ndiff) # if series are the same, put 0
    )

  v <- vector()
  for (i in seq(nrow(df2))) {
    if (df2[i, 3] != 0) {
      if (lm(y ~ x, data = rename_all(data.frame(y = df[df2[i, 1]], x = df[df2[i, 2]]),
                                      funs(c("y", "x")))) %>%
        broom::augment() %>% .$.resid %>%
        forecast::ndiffs(test = test, alpha = alpha, type = type) == df2[i, 3] - 1) {
        v[i] <- 2 # 2 ---> series are cointegrated
      } else {
        v[i] <- 1 # 1 ---> not cointegrated, but test is commitable
      }
    } else {
      v[i] <- 0 # 0 ---> test is not performable [I(0) OR not the same I() order OR
      # series are the same]
    }
  }
  df2 %>%
    mutate(
```

```r
      cointegration = v
    ) %>%
    select(y, x, cointegration)
}

cointegration_tests_results <- cointegration_tests(df = Bankdata, test = "adf",
                                                   type = "level", alpha = 0.05)

cointegration_tests_results %>%
  mutate(
    cointegration = case_when(
      cointegration == 0 ~ "Not performable",
      cointegration == 1 ~ "Not cointegrated",
      cointegration == 2 ~ "Cointegrated"
    ),
    cointegration = factor(cointegration, levels = c("Cointegrated", "Not cointegrated",
                                                     "Not performable"))
  ) %>%
  ggplot() +
  geom_tile(aes(x = x, y = y, fill = cointegration), color = "black") +
  scale_fill_grey() +
  theme(
    axis.text.x = element_text(angle = 90, vjust = 0.45),
  ) +
  labs(
    y = "Dependent variable in the OLS",
    x = "Independent variable in the OLS",
    caption = "Calculations are based on ADF-test (level, alpha = 5%)"
  ) + theme(
  panel.border = element_blank()
)

# Johansen-test -------------------------------------------------------------------------

Bankdata %>%
  select(-1) %>%
  ca.jo(type = "eigen", K = 5, ecdet = "none", spec = "longrun") %>%
  summary() # Number of cointegrated vectors = 1

## Engle-Granger method with rolling window --------------------------------------------

for (i in 1:(nrow(Bankdata) - 249)) {
  if (i == 1) {
    cointegration_tests_rw <- mutate(
      cointegration_tests(df = Bankdata[i:(i + 249), 1:4], test = "adf", type = "level",
                          alpha = 0.05),
      t = i
    )
  } else {
    cointegration_tests_rw <- rbind(cointegration_tests_rw, mutate(
      cointegration_tests(df = Bankdata[i:(i + 249), 1:4], test = "adf", type = "level",
                          alpha = 0.05),
      t = i
```

```
    ))
  }
}

cointegration_tests_rw %>%
  filter(y != x) %>%
  ggplot(aes(x = t, y = cointegration)) +
  geom_point() +
  facet_grid(cols = vars(x), rows = vars(y)) +
  scale_y_continuous(breaks = c(0, 1, 2),
                     labels = c("Not performable", "Not cointegrated", "Cointegrated")) +
  labs(
    subtitle = "Size of window = 250",
    y = "Result of the test",
    x = "# window",
    caption = "Calculations are based on ADF-test (level, alpha = 5%)\n
    Dependent variables (in the OLS) are placed horizontal, independents are vertical."
  )

cointegration_tests_rw %>%
  filter(cointegration == 2) %>%
  mutate(cointegration = factor(cointegration)) %>%
  group_by(y, x) %>%
  tally() %>%
  arrange(x) %>%
  mutate(
    n = n / max(cointegration_tests_rw$t),
    n = scales::percent(n, accuracy = .01)
  ) %>%
  pivot_wider(id_cols = y, values_from = n, names_from = x, names_prefix = "x = ") %>%
  arrange(y)

merge(expand.grid(1:(nrow(Bankdata) - 249), c(0, 1, 2)) %>%
        rename_all(funs(c("t", "cointegration"))),
  cointegration_tests_rw %>% filter(y != x) %>%
    group_by(t, cointegration) %>%
    summarise(n = n()),
  all.x = T
) %>%
  mutate(
    n = ifelse(is.na(n), 0, n),
    cointegration = case_when(
      cointegration == 0 ~ "Not performable",
      cointegration == 1 ~ "Not cointegrated",
      cointegration == 2 ~ "Cointegrated"
    ),
    cointegration = factor(cointegration, levels = c("Cointegrated", "Not cointegrated",
                                                      "Not performable")),
    t = as.Date(Bankdata$Date)[t + 125]
  ) %>%
  ggplot() +
  geom_area(aes(x = t, y = n, fill = cointegration)) +
  scale_y_continuous(expand = c(0, 0)) +
```

```r
  scale_x_date(expand = c(0, 0), date_breaks = "1 year", date_labels = "%Y") +
  theme(
    legend.position = "bottom"
  ) +
  labs(
    subtitle = "Size of window = 250",
    y = "Number of pairings with the result",
    x = "Time (middle of the window)",
    caption = "Calculations are based on ADF-test (level, alpha = 5%).\n
    Number of total pairings pairs are 6."
  ) +
  scale_fill_grey()

# Johansen test with rolling window ------------------------------------------------------

johansen_tests_rw <- data.frame(t = 1:(nrow(Bankdata) - 249)) %>% mutate(
  pct10 = NA, pct5 = NA, pct1 = NA
)

for (i in 1:(nrow(Bankdata) - 249)) {
  if (i == 1) {
    johansen_critical_values <- ca.jo(
      x = Bankdata[i:(i + 249), 2:4], type = "eigen",
      K = 5, ecdet = "none", spec = "longrun"
    )@cval
  }
  johansen_tests_rw[i, 2] <- which.max(rev(ca.jo(
    x = Bankdata[i:(i + 249), 2:4], type = "eigen",
    K = 5, ecdet = "none", spec = "longrun"
  )@teststat) < rev(johansen_critical_values[, 1])) - 1
  johansen_tests_rw[i, 3] <- which.max(rev(ca.jo(
    x = Bankdata[i:(i + 249), 2:4], type = "eigen",
    K = 5, ecdet = "none", spec = "longrun"
  )@teststat) < rev(johansen_critical_values[, 2])) - 1
  johansen_tests_rw[i, 4] <- which.max(rev(ca.jo(
    x = Bankdata[i:(i + 249), 2:4], type = "eigen",
    K = 5, ecdet = "none", spec = "longrun"
  )@teststat) < rev(johansen_critical_values[, 3])) - 1
}

ggplot() +
  geom_ribbon(aes(
    x = c(as.Date("2007-12-01"), as.Date("2009-12-01")),
    ymin = -Inf,
    ymax = Inf,
    fill = "recession"), color = "black", alpha = .6) +
  geom_jitter(data = johansen_tests_rw %>%
                pivot_longer(-1) %>%
                mutate(
                  name = case_when(
                    name == "pct1" ~ "1%",
                    name == "pct5" ~ "5%",
                    name == "pct10" ~ "10%"
```

```r
                  ),
                  t = as.Date(Bankdata$Date)[t + 125]
                ),
              aes(x = t, y = value, color = name),width = 0, height = 0.05) +
  scale_color_grey() +
  theme(
    legend.position = "bottom"
  ) +
  scale_y_continuous(breaks = c(0, 1, 2)) +
  scale_x_date(expand = c(0, 0), date_breaks = "1 year", date_labels = "%Y") +
  labs(
    subtitle = "Size of window = 250",
    y = "# cointegrated vectors",
    x = "Time (middle of the window)",
    caption = str_wrap(str_c(
      "Points are jittered around their true y value for better ",
       "visualisation (the number of cointegrated vectors is interger). ",
        "Date of recession is from the National Bureau of Economic Research ",
        "(https://www.nber.org/cycles.html)."), 50)
  ) +
  theme(
    panel.grid.minor.y = element_blank()
  ) +
  scale_fill_manual(values = c("recession" = "#FF5B6B"))

johansen_tests_rw %>%
  select(-1) %>%
  gather() %>%
  mutate(
    key = case_when(
      key == "pct1" ~ "1%",
      key == "pct5" ~ "5%",
      key == "pct10" ~ "10%"
    ),
    key = factor(key, levels = c("10%", "5%", "1%"))
  ) %>%
  group_by(key, value) %>%
  tally() %>%
  ggplot() +
  geom_bar(aes(x = key, y = n, fill = factor(value, levels = 2:0)), position = "fill",
           stat = "identity", color = "black") +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1), expand = c(0, 0),
                     breaks = seq(from = 0, to = 1, by = .1)) +
  scale_fill_grey() +
  labs(
    title = "Distribution of the Johansen-test results with rolling window",
    x = "Alpha",
    y = "Proportion",
    fill = "Number cointegrated vectors (r)",
    subtitle = "Size of window = 250"
  ) +
  theme(
    legend.title = element_text(),
```

```
    legend.position = "bottom"
  )

johansen_tests_rw %>%
  pivot_longer(-1) %>%
  mutate(
    name = factor(name, levels = c("pct1", "pct5", "pct10")),
    t = as.Date(Bankdata$Date)[t + 125],
    t = ifelse(t > as.Date("2007-12-01") & t < as.Date("2009-12-01"), "recession",
               "expansion")
  ) %>% filter(t == "expansion") %>% group_by(name) %>% count(value) %>% pivot_wider(
    id_cols = value, values_from = n, names_from = name
  )  %>% mutate(
    pct1 = scales::percent(pct1/sum(pct1, na.rm = T), accuracy = .01),
    pct5 = scales::percent(pct5/sum(pct5, na.rm = T), accuracy = .01),
    pct10 = scales::percent(pct10/sum(pct10, na.rm = T), accuracy = .01)
  ) %>% rename_all(funs(c("# cointegrated vectors", "1%", "5%", "10%")))

johansen_tests_rw %>%
  pivot_longer(-1) %>%
  mutate(
    name = factor(name, levels = c("pct1", "pct5", "pct10")),
    t = as.Date(Bankdata$Date)[t + 125],
    t = ifelse(t > as.Date("2007-12-01") & t < as.Date("2009-12-01"), "recession",
               "expansion")
  ) %>% filter(t == "recession") %>% group_by(name) %>% count(value) %>%
  pivot_wider(
    id_cols = value, values_from = n, names_from = name
  )  %>% mutate(
    pct1 = scales::percent(pct1/sum(pct1, na.rm = T), accuracy = .01),
    pct5 = scales::percent(pct5/sum(pct5, na.rm = T), accuracy = .01),
    pct10 = scales::percent(pct10/sum(pct10, na.rm = T), accuracy = .01)
  ) %>% rename_all(funs(c("# cointegrated vectors", "1%", "5%", "10%")))
```