Forecasting & Predictive Analytics

ESSEC | CentraleSupélec
Master in Data Sciences and Business Analytics

Guillaume Chevillon & Florens Odendahl

EXAMINATION
December 16, 2019
1:15 PM-4:15 PM

Calculators are NOT authorized (since they are not at all necessary).
Please answer the questions below.
All questions are expecting straightforward answers and their understanding should be clear. If in doubt, however, explain your understanding of the question and answer it accordingly.

The exam consists of four theory exercises, and one empirical study you need to assess.
Exercise 1: 8 points
Exercise 2: 14 points
Exercise 3: 13 points
Exercise 4: 16 points
Empirical Study: 19 points

Total: 70 points

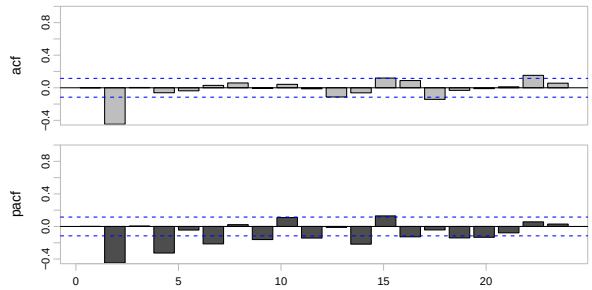
You can answer them in any order
Remember to write legibly and concisely.

Theory exercises

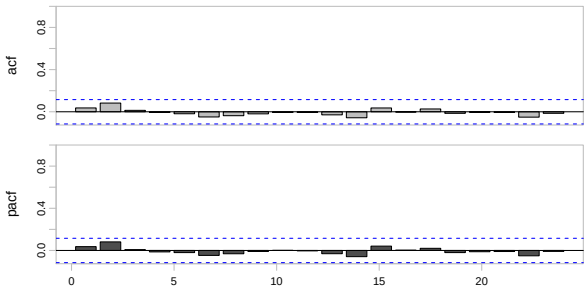
1. In this exercise we use autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs) to determine the lag lengths of ARMA(p, q) models.

   (a) Consider the following four DGPs, 1, 2 ,3 and 4 and plots A, B, C and D (where the plots are based on estimated ACFs and PACFs using a sample of $T = 300$; starting at lag 1). Which DGP generated which plot? Explain, in one sentence (one!) for each of your decisions, why the figure and the DGP belong together. **[8 points in total]**

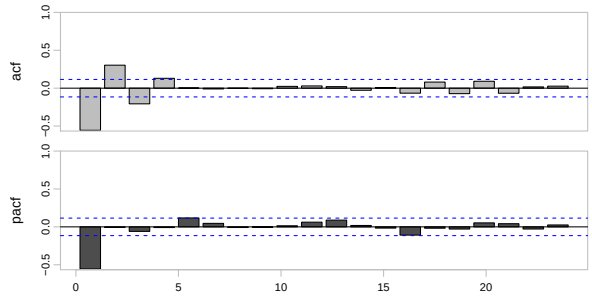
   - 1: $y_t = 0.35y_{t-1} + 0.35y_{t-2} - 0.2y_{t-3} + u_t$, $u_t \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$
   - 2: $y_t = u_t - 0.95u_{t-2}$, $u_t \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$
   - 3: $y_t = 0.99y_{t-1} + u_t - u_{t-1}$, $u_t \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$
   - 4: $y_t = -0.5y_{t-1} + u_t$, $u_t \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$



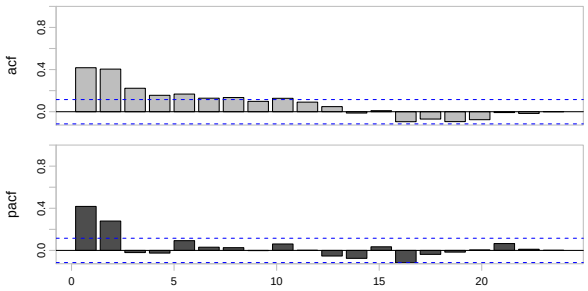
(a) Plot A



(b) Plot B



(c) Plot C



(d) Plot D

2. Consider the process Y_t that follows an ARMA(1,2), such that

$$Y_t = \rho Y_{t-1} + u_t + \theta u_{t-2},$$

where $u_t \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$ and ρ, θ are finite constants.

   (a) Under which additional assumption is the process stationary? **[1 points]**

   (b) *Assume from now on (here and also in questions c) to e) below) that the process Y_t is stationary.* Rewrite the process Y_t in the Wold-decomposition and find the expression for parameters ψ_j, for $j = 0, 1, 2, 3$ as a function of ρ and θ, in $Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$. **[4 points]**

   (c) Derive the optimal one-, two-, three-, and h-step ahead ($h > 3$) forecasts under the MSFE loss function. **[3 points]**

   (d) Compute the h-step ahead ($h > 3$) forecast error, denoted by $e_{t,h} = Y_{t+h} - Y_{t+h|t}$, where $Y_{t+h|t}$ denotes the optimal h-step ahead forecast. **[3 points]**

   (e) Show that $\mathsf{var}(e_{t,h}) \geq \mathsf{var}(e_{t,h-1})$. Explain your intuition for this result.**[3 points]**

3. *Recall that the Diebold-Mariano (DM) test is a test for equal predictive ability of two models, say model A and model B, i.e. it tests the null hypothesis of*

$$\mathsf{H}_0 : \mathsf{E}\big(e^2_{t+1,A} - e^2_{t+1,B}\big) = 0$$

vs. the alternative

$$\mathsf{H}_A : \mathsf{E}\big(e^2_{t+1,A} - e^2_{t+1,B}\big) \neq 0,$$

where $e^2_{t+1,A}, e^2_{t+1,B}$ denote the forecast errors of model A and model B. The DM test statistic takes the form of a t-test on the mean of $d_j = e^2_{j,A} - e^2_{j,B}$:

$$\sqrt{P}\frac{(\bar{d}_P - 0)}{\sqrt{\widehat{\mathsf{var}}(d)}},$$

where $P = T - R$ is the size of the evaluation sample (R denotes the number of observations used in estimating the models*),*

$$\bar{d}_P = \frac{1}{P}\sum_{j=R+1}^{T} d_j = \frac{1}{P}\sum_{j=R+1}^{T}\big(e^2_{j,A} - e^2_{j,B}\big)$$

and

$$\widehat{\mathsf{var}}(d) = \frac{1}{P}\sum_{j=R+1}^{T}\big[d_j - \bar{d}_P\big]^2 = \frac{1}{P}\sum_{j=R+1}^{T}\big[(e^2_{j,A} - e^2_{j,B}) - \bar{d}_P\big]^2.$$

Also note that the theoretical counterpart to the estimator $\widehat{\mathsf{var}}(d)$ is the variance that we compute in questions c) and d) below

$$\mathsf{var}(e^2_{j,A} - e^2_{j,B}) = \mathsf{E}\left[(e^2_{j,A} - e^2_{j,B})^2\right] - \big[\mathsf{E}(e^2_{j,A} - e^2_{j,B})\big]^2.$$

Assume that the DGP of the random variable Y_t is $Y_t = \varepsilon_t$, where $\varepsilon_t \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$.

Assume that there are two competing forecasting models, with the following one-step-ahead forecasts:

$$\begin{aligned} &\text{Model A: } \hat{Y}_{t+1,A} = 0 \\ &\text{Model B: } \hat{Y}_{t+1,B} = \hat{\alpha}_t, \end{aligned} \tag{1}$$

where $\hat{\alpha}_t = \frac{1}{R}\sum_{j=t-R+1}^{t} Y_j$, where R is a finite constant. We define the two resulting forecast errors as

$$\begin{aligned} &\text{Model A: } e_{t+1,A} = Y_{t+1} - \hat{Y}_{t+1,A} = Y_{t+1} - 0 \\ &\text{Model B: } e_{t+1,B} = Y_{t+1} - \hat{Y}_{t+1,B} = Y_{t+1} - \hat{\alpha}_t, \end{aligned} \tag{2}$$

(a) Compute the theoretical value of the mean squared forecast error (MSFE), i.e. $\mathsf{E}(e^2_{t+1,A} - e^2_{t+1,B})$ as a function of σ^2 and R. **[3 points]**

(b) Is one model better than the other in terms of MSFE? **[1 points]**

(c) Compute the variance of the squared forecast error difference **[4 points]**

$$\mathsf{var}(e^2_{j,A} - e^2_{j,B}) = \mathsf{E}\big[(e^2_{j,A} - e^2_{j,B})^2\big] - \big[\mathsf{E}(e^2_{j,A} - e^2_{j,B})\big]^2$$

(d) Consider an alternative forecast model, Model B*, defined as

$$\text{Model B}^*\text{: } \hat{Y}_{t+1,B} = \tilde{\alpha}_t, \tag{3}$$

where $\tilde{\alpha}^*_t = \frac{1}{t}\sum_{j=1}^{t} Y_j$, i.e. a recursive-window estimation scheme. Compute $\mathsf{var}(e^2_{j,A} - e^2_{j,B^*})$. What happens to the variance when $t \to \infty$, i.e. when the in-sample estimation size gets large. Can you think of a problem that this might cause for the Diebold-Mariano test? Does this problem also occur in the rolling window estimation scheme? Explain. **[5 points]**

4. Consider the model for the stock price, p_t, and dividend, d_t, of a specific company that follows vector autoregressive process of order 1

$$\begin{bmatrix} p_t \\ d_t \end{bmatrix} = \begin{bmatrix} 0 \\ \mu_d \end{bmatrix} + \begin{bmatrix} 1 & \beta \\ 0 & \rho \end{bmatrix} \begin{bmatrix} p_{t-1} \\ d_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{d,t} \end{bmatrix} \tag{4}$$

where

$$\begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{d,t} \end{bmatrix} \overset{i.i.d.}{\sim} \mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_p^2 & \sigma_{dp} \\ \sigma_{dp} & \sigma_d^2 \end{bmatrix} \right).$$

(a) Define the concept of Granger-causality and explain whether p_t and d_t Granger-cause each other. **[2 points]**

(b) Under which conditions is the vector $(p_t, d_t)'$ stationary, integrated and possibly cointegrated? Explain. **[3 points]**

(c) At time t, the company wants to forecast its future stock price p_{t+1} and considers the scenario where its future divend $d_{t+1} = \delta$ (some value). What is, under this scenario, the predicted value $p_{t+1|t}$, assuming an MSFE loss function? **[4 points]**

(d) Would it not be easier for the company to estimate directly one of the following equations to forecast its price p_{t+1} given a scenario for d_{t+1}? **[3 points]**

$$p_t = \alpha_0 + \alpha_1 d_t + u_t, \text{ or}$$
$$\Delta p_t = a_0 + a_1 d_t + v_t$$

(e) Now assume that (p_t, d_t) do not denote prices and dividends for a specific company but represent the market average and assume that they follow the same data generating process (DGP) as before. We assume that the analyst only observes p_t but not d_t. Suggest (simply, without too much detail) how the analyst can use the model above to forecast prices. **[4 points]**

Empirical Study

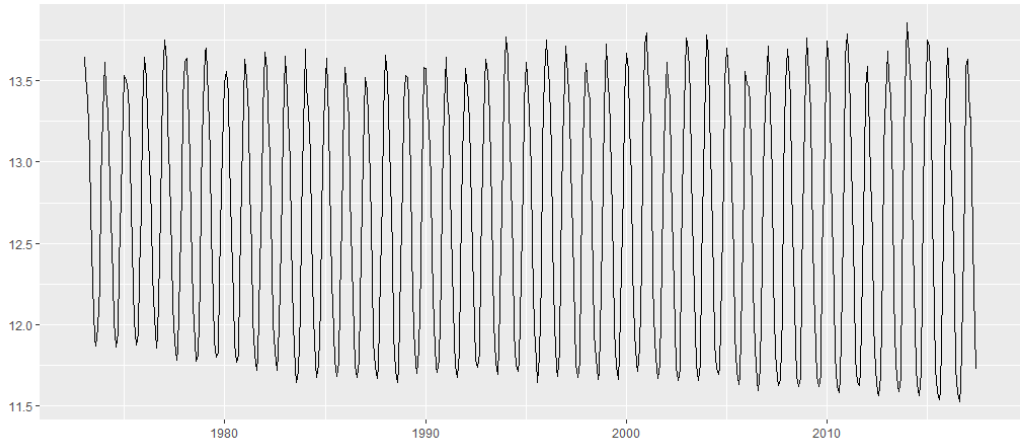
In this section, we go through a small empirical study consisting of forecasting monthly U.S. gas consumption using time series models and climate variables (temperatures, precipitations, number of very hot days within the month, the montly minimum and maximum of the latter variables — their "anomalies" compared to historical observations are denoted "Ano"...) as well as the number of Google searches for the four keywords: "heatwave", "extreme weather", "snow storm" and "school closed".

Here is the actual data description:

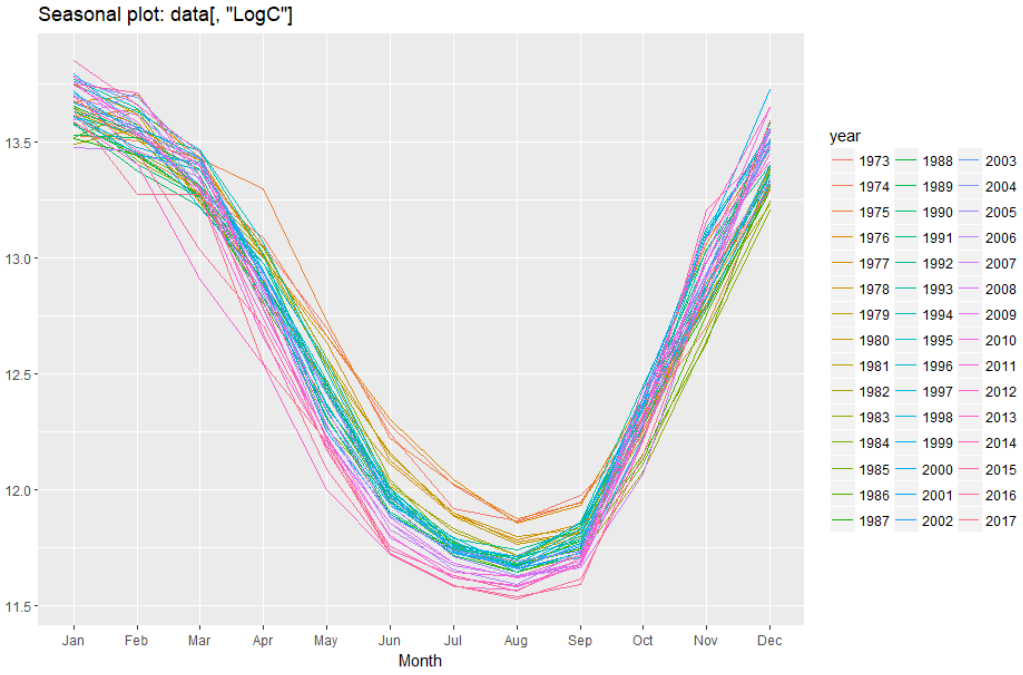
```
# 18 variables, from three sources :
# source 1 :  Gas consumption
# Period :  from jan 1973 to june 2017
# variables :  [gas_con, gas_con_log]
# source 2 :  Weather
# Period :  from jan 1973 to feb 2017
# variables :  [Temperature_Value, Temperature_Ano, Temperature_Value_Min,
# Temperature_Ano_Min, Temperature_Value_Max,
# Temperature_Ano_Max, Precipitation_Value, Precipitation_Ano,
# Cooling_Degree_Days_Value, Cooling_Degree_Days_Ano, Heating_Degree_Days_Value,
# Heating_Degree_Days_Ano]
# source 3 :  Google trend
# Period :  from jan 2004 to dec 2017 (from Google)
# Variables :  [Heatwave, Extreme_weather, Snow_storm, School_closed]
```

You will find below some elements of the study and you are asked to comment on them.

1. We start by an analysis of the data for c_t, the (natural) logarithm of Gas consumption at time t, which is reported in Figure 2. Would you argue that the distribution of c_t is constant over time, explain. **[2 points]**



(a) Time series graph of c_t



(b) Seasonal graph, each curve represents a different year

Figure 2: Data on the logarithm of U.S. monthly consumption of gas

2. The Augmented Dickey-Fuller test applied to c_t reports a p-value of 0.01. What do you conclude at the 5% significance level? Explain the null and alternative hypotheses. **[3 points]**

3. We decide to start with a VAR(p) model where we keep all variables (except the "anomaly" variables which are dropped for now) and use the R code:

```
ts_mtx_train <- window(ts_mtx, start=c(2004,1), end=c(2016,2))
var_selection <- VARselect(ts_mtx_train, lag.max = 12, type = c("const"), season
= 12)
```

where `ts_mtx` denotes a dataset comprising the union of all time series and `season` is a control that includes seasonal dummy variables. This code aims to help selecting the lag order for the VAR and provides the following output:

```
>var_selection

$selection
AIC(n) HQ(n) SC(n) FPE(n)
12     12     1     12
$criteria
        1         2         3         4         5         6         7
AIC(n) 1.68e+01 1.73e+01 1.75e+01 1.79e+01 1.81e+01 1.83e+01 1.83e+01
HQ(n)  1.85e+01 1.97e+01 2.06e+01 2.17e+01 2.26e+01 2.36e+01 2.42e+01
SC(n)  2.09e+01 2.31e+01 2.51e+01 2.72e+01 2.92e+01 3.12e+01 3.29e+01
FPE(n) 2.16e+07 3.65e+07 4.86e+07 8.12e+07 1.23e+08 2.25e+08 3.33e+08
        8         9         10        11        12
AIC(n) 1.83e+01 1.75e+01 1.63e+01 1.27e+01 6.27e+00
HQ(n)  2.49e+01 2.49e+01 2.44e+01 2.15e+01 1.57e+01
SC(n)  3.40e+01 3.56e+01 3.62e+01 3.43e+01 2.96e+01
FPE(n) 6.69e+08 7.84e+08 9.58e+08 2.11e+08 1.12e+07
```

where AIC stands for Akaike Information Criterion, HQ for Hannan-Quinn (information criterion), SC for the Schwarz (also called Bayesian) information criterion and FPE means Final Prediction Error.
An alternative for selecting the lag order p of the VAR consists in assessing the differences in Root Mean-Square Error (RMSE) between the *training* and *testing* subsamples. Below is the outcome when the testing subsample consists of the forecasts over horizons $h = 1, ..., 12$.

```
p=1, RMSE train:43572, RMSE test:75825
p=2, RMSE train:42180, RMSE test:70327
p=3, RMSE train:41260, RMSE test:68758
p=4, RMSE train:40249, RMSE test:67016
p=5, RMSE train:38661, RMSE test:69702
```

```
p=6, RMSE train:36827, RMSE test:63768
p=7, RMSE train:34256, RMSE test:46631
p=8, RMSE train:31880, RMSE test:55441
p=9, RMSE train:29991, RMSE test:58338
p=10, RMSE train:26872, RMSE test:64419
p=11, RMSE train:23353, RMSE test:80241
p=12, RMSE train:18452, RMSE test:219007
```

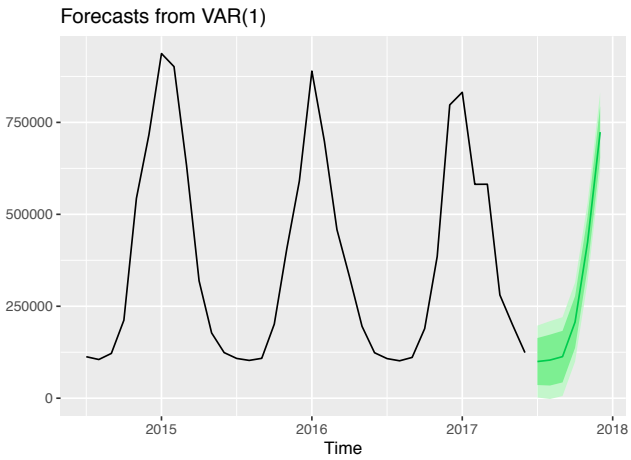
What is your recommendation regarding the choice of lag order p for the VAR? Please justify your recommandation by weighing the pros and cons of the analysis above and suggesting alternative techniques if you think the analysis could be improved **[6 points]**

4. We use four different forecasting models for the U.S. monthly consumption of gas: a VAR(1), a VAR(7), a SARIMA model, and an ETS model, i.e. a double Exponential smoother with Trend and Seasonals. We report them in Figure 3.

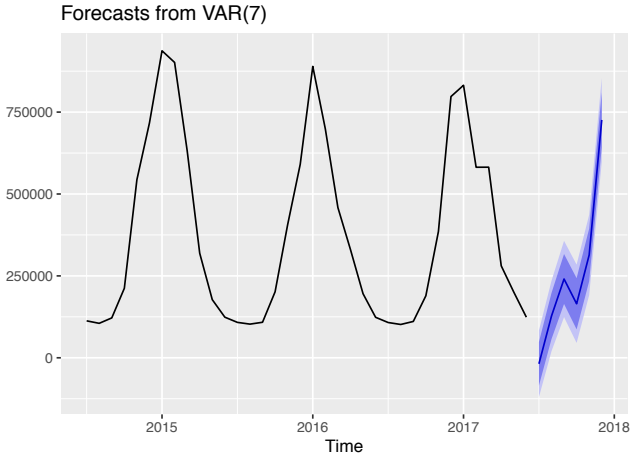
   (a) Can you explain why the VAR(1) and VAR(7) present much wider forecast confidence intervals than the other two univariate techniques? **[2 points]**

   (b) Does the pattern for the evolution of forecast uncertainty as a function of the horizon h correspond to your intuition? **[4 points]**

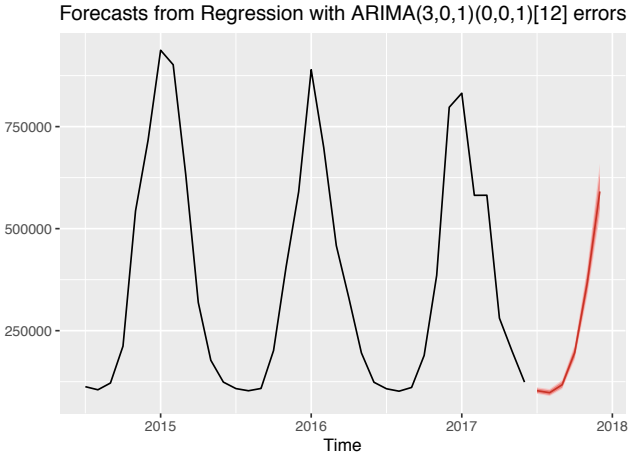
   (c) Also can you think of an explanation for the forecast patterns generated by the VAR(7)? **[2 points]**



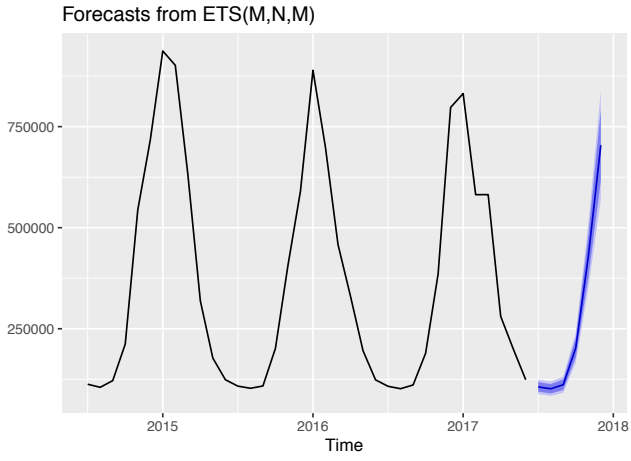
(a) Forecasts based on a VAR(1)



(b) Forecasts based on a VAR(7)



(c) Forecasts based on a SARIMA



(d) Forecasts from an ETS

Figure 3: 6-step ahead forecasts of monthly U.S. gas consumption