Time series models: sparse estimation and robustness aspects

Christophe Croux (KU Leuven, Belgium)

2017 CRonos Spring Course

Limassol, Cyprus, 8-10 April

Based on joint work with Ines Wilms, Ruben Crevits, and Sarah Gelper.

Part I

Autoregressive models: from one time series to many time series

Outline

1. Univariate
2. Multivariate
3. Big Data

Section 1

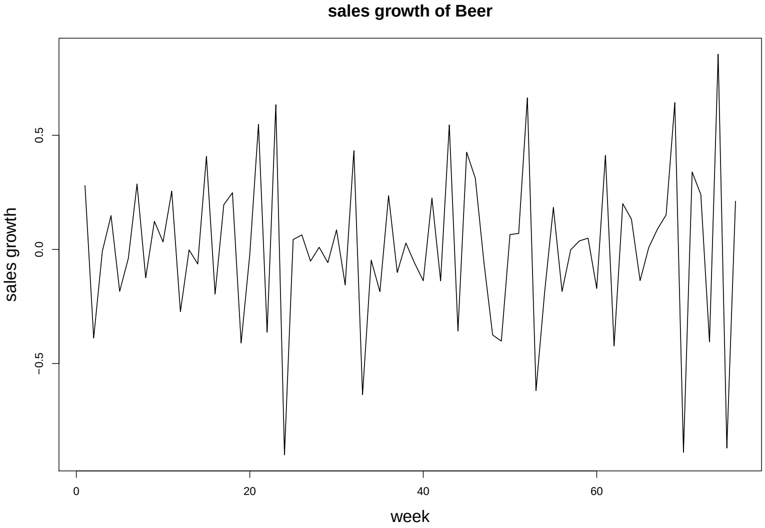
Univariate

One time series

Sales growth of beer in a store: weekly data

	y
Week 1	0.28
Week 2	-0.39
Week 3	-0.01
Week 4	0.15
Week 5	-0.18
⋮	⋮
⋮	⋮
Week 75	-0.87
Week 76	0.21

Time Series Plot



Autoregressive model

Autoregressive model of order 1:

$$y_t = c + \gamma\, y_{t-1} + e_t$$

for every time point $t = 1, \ldots, T$. Here, e_t stands for the error term.

Estimate c and γ by using ordinary least squares.

R-code and Output

```
> plot(y,xlab="week",ylab="sales growth",type="l",main="sales growth of Beer")
> y.actual=y[2:length(y)]
> y.lagged=y[1:(length(y)-1)]
> mylm<-lm(y.actual~ y.lagged)
> summary(mylm)
```

```
Coefficients:
Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.001041   0.034077   0.031    0.976
y.lagged    -0.534737   0.098643  -5.421 7.31e-07 ***


Multiple R-squared:  0.287,Adjusted R-squared:  0.2772
```

Questions: (i) what is the estimate of γ (ii) is it significant (iii) interpret its sign (iv) interpret R^2

Forecasting

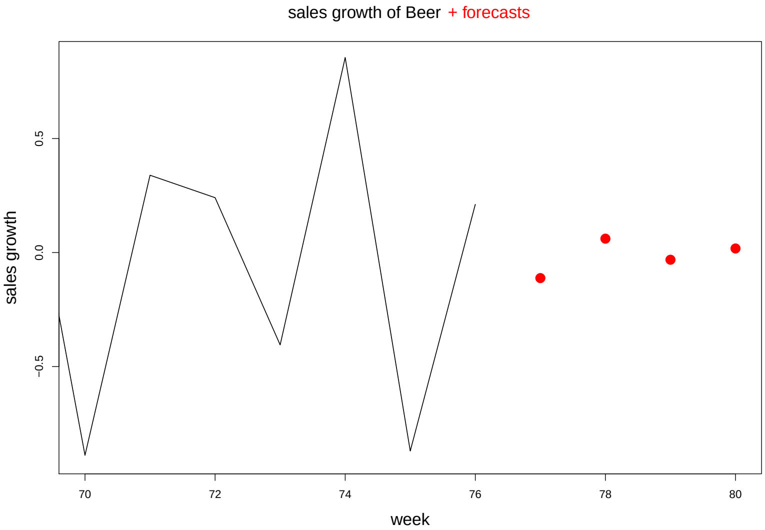
Using the observations $y_1, y_2, \ldots, y_T$ we forecast the next observation uisng the formula

$$\widehat{y}_{T+1} = \hat{c} + \hat{\gamma}\, y_T$$

Can we also forecast at *horizon 2* ? YES

$$\widehat{y}_{T+2} = \hat{c} + \hat{\gamma}\, \hat{y}_{T+1}$$

$T = 76$, forecasts at horizon 1, 2, 3, and 4



Section 2

Multivariate

3 time series

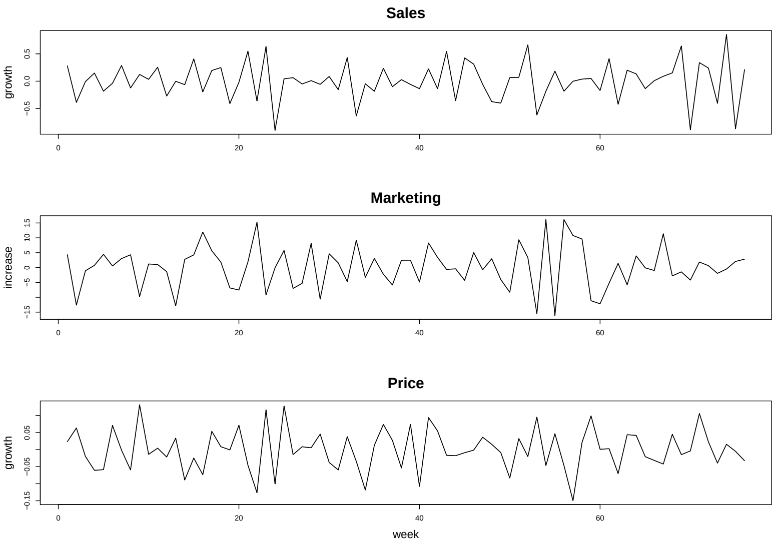
Predict next week sales using the observed

- Sales

+ additional information, like

- Price
- Marketing Effort

Three stationary time series $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$.


Sales
growth
0.5
0.0
-0.5
0
20
40
60
Marketing
increase
15
10
5
0
-5
-15
0
20
40
60
Price
growth
0.05
-0.05
-0.15
0
20
40
60
week

Prediction model for first time series:

$$y_{1,t} = c_1 + \gamma_{11}\, y_{1,t-1} + \gamma_{12}\, y_{2,t-1} + \gamma_{13}\, y_{3,t-1} + e_{1t}$$

Forecast at *horizon 1 ? YES*

$$\widehat{y}_{1,T+1} = \hat{c}_1 + \hat{\gamma}_{11}\, y_{1,T} + \hat{\gamma}_{12}\, y_{2,T} + \hat{\gamma}_{13}\, y_{3,T}$$

Forecast at *horizon 2* ? NO

$$\widehat{y}_{1,T+2} = \hat{c}_1 + \hat{\gamma}_{11}\, \hat{y}_{1,T+1} + \hat{\gamma}_{12}\, y_{2,T+1} + \hat{\gamma}_{13}\, y_{3,T+2}$$

Why Not? $y_{2,T+1}, y_{3,T+2}$ not known.

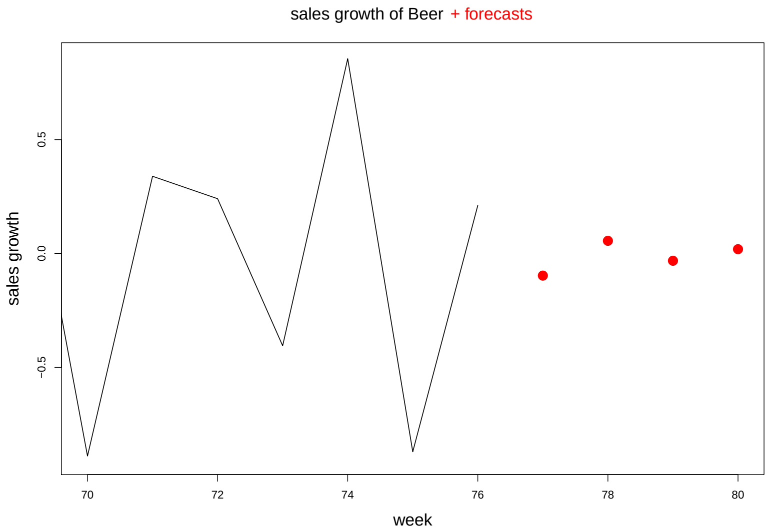
Vector Autoregressive Model

$$\begin{cases} y_{1,t} & = & c_1 + \gamma_{11}\, y_{1,t-1} + \gamma_{12}\, y_{2,t-1} + \gamma_{13}\, y_{3,t-1} + e_{1t} \\ y_{2,t} & = & c_2 + \gamma_{21}\, y_{1,t-1} + \gamma_{22}\, y_{2,t-1} + \gamma_{23}\, y_{3,t-1} + e_{2t} \\ y_{3,t} & = & c_3 + \gamma_{31}\, y_{1,t-1} + \gamma_{32}\, y_{2,t-1} + \gamma_{33}\, y_{3,t-1} + e_{3t} \end{cases}$$

- Estimation: OLS equation by equation.
- VAR(1) model
- γ_{ji} = effect of time series j on time series i
- We have $q \times q = 9$ autoregressive parameters γ_{ji}, with $q = 3$ the number of time series.

Prediction using VAR model

$T = 76$, forecasts at horizon 1, 2, 3, and 4



Section 3

Big Data

Many time series

Predict next week sales of Beer using the observed

- Sales of Beer

$+$

- Price of Beer
- Marketing Effort for Beer

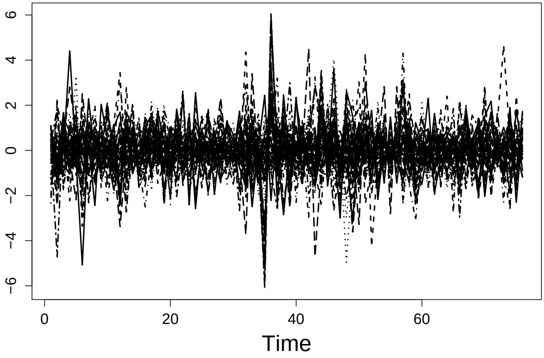
$+$

- Prices of other product categories
- Marketing Effort for other product categories

Vector Autoregressive model $\equiv$ Market Response Model

Market Response Model

Sales, promotion and prices for 17 product categories:
$q = 17 \times 3 = 51$ time series and $T = 77$ weekly observations



VAR model for $q = 3 \times 17 = 51$ time series

- $q \times q = 2601$ autoregressive parameters

$\rightarrow$ Explosion of number of parameters

Use the LASSO instead of OLS.

Question: Why?

Network

Network with q nodes. Each node corresponds with a time series.

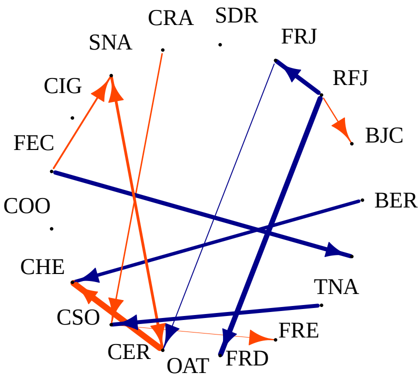
- draw an **edge** from node i to node j if

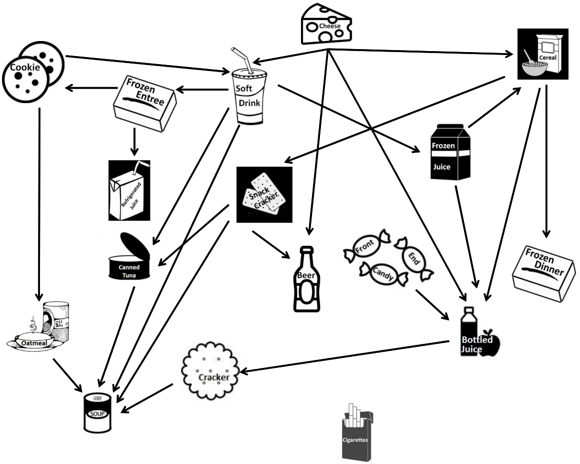
$$\widehat{\gamma}_{ji} \neq 0$$

- the edge **width** is the size of the effect
- the edge **color** is the sign of the effect (blue if positive, red if negative)

price effects on sales

17 product categories




Cheese
Cookie
Soft
Drink
Frozen
Entree
Cereal
Frozen
Juice
Refrigerated
Juice
Snack
Cracker
Canned
Tuna
Beer
Front
End
Candy
Frozen
Dinner
Oatmeal
Bottled
Juice
Cracker
SOUP
Cigarettes

Part II

Basic time series concepts

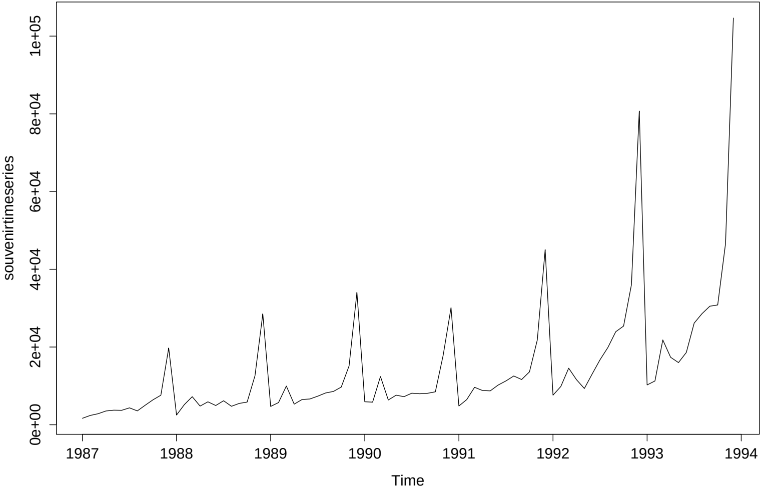
Outline

Section 1

Stationarity

Example: souvenirs sold (in dollars)

Frequency: monthly, sample size: $T = 84$



Stochastic Process

A *Stochastic Process* is a sequence of stochastic variables: $\ldots, Y_1, Y_2, Y_3, \ldots, Y_T \ldots$. We observe the process from $t = 1$ to $t = T$, yielding a sequence of numbers

$$y_1, y_2, y_3, \ldots, y_T$$

which we call a *time series*.

We only treat regularly spaced, discrete time series. Note that the observations in a time series are not independent! We need to rely on the concept of stationarity.

Stationarity

We say that a stochastic process is (weakly) stationary if

1. $E[Y_t]$ is the same for all t
2. $\text{Var}[Y_t]$ is the same for all t
3. $\text{Cov}(Y_t, Y_{t-k})$ is the same for all t, for every $k > 0$.

Section 2

Autocorrelation

Autocorrelations

Then we define the autocorrelation of order k as

$$\rho_k = \mathrm{Corr}(Y_t, Y_{t-k}) = \frac{\mathrm{Cov}(Y_t, Y_{t-k})}{\mathrm{Var}[Y_t]}.$$

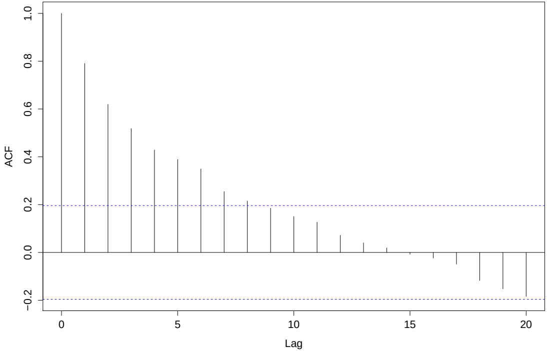
The autocorrelations give insight in the dependency structure of the process.
The autocorrelations can be estimated by

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

The correlogram

A plot of $\hat{\rho}_k$ versus k is called a *correlogram*. On a correlogram, we often see 2 lines, corresponding to the critical values of the test statistic $\sqrt{T}\hat{\rho}_k$ for testing $H_0 : \rho_k = 0$ for a specific value of k.

Example



The first 8 autocorrelations are significantly different from zero. there is strong *persistency* in the series.

Section 3

Differencing

Difference Operators

The "Lag" operator L is defined as

$$LY_t = Y_{t-1}.$$

Note that $L^s Y_t = Y_{t-s}$.

The difference operator Δ is defined as

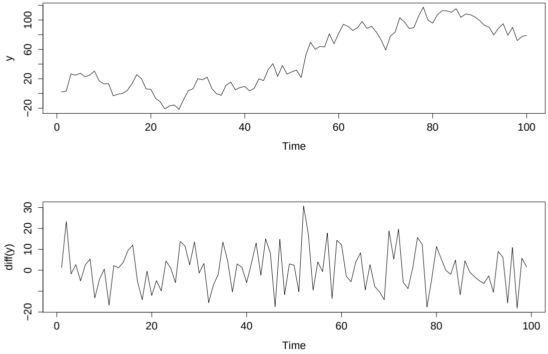
$$\Delta Y_t = (I - L)Y_t = Y_t - Y_{t-1}.$$

Linear trends can be eliminated by applying Δ once. If a stationary process is then obtained, we say that Y_t is integrated of order 1.

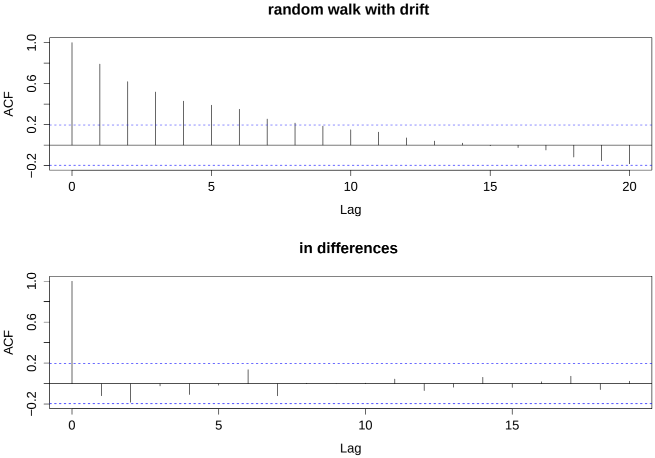
Example

Random Walk with drift: $Y_t = a + Y_{t-1} + u_t$, with u_t i.i.d. white noise.

Plot of Y_t and ΔY_t:



Correlograms of Y_t and ΔY_t:



Seasonality

Seasonal effects of order s can be eliminated by applying the difference operator of order s:

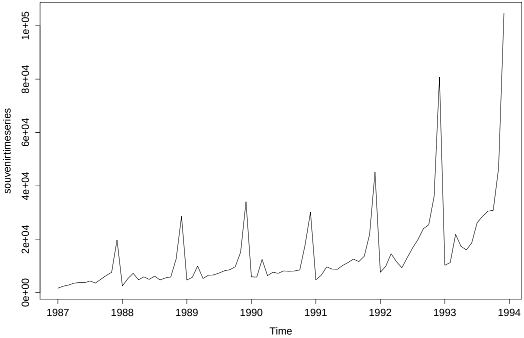
$$\Delta_s Y_t = (I - L^s) Y_t = Y_t - Y_{t-s}$$

- $s = 12$, monthly data
- $s = 4$, quarterly data

Note than one loses s observations when differencing.

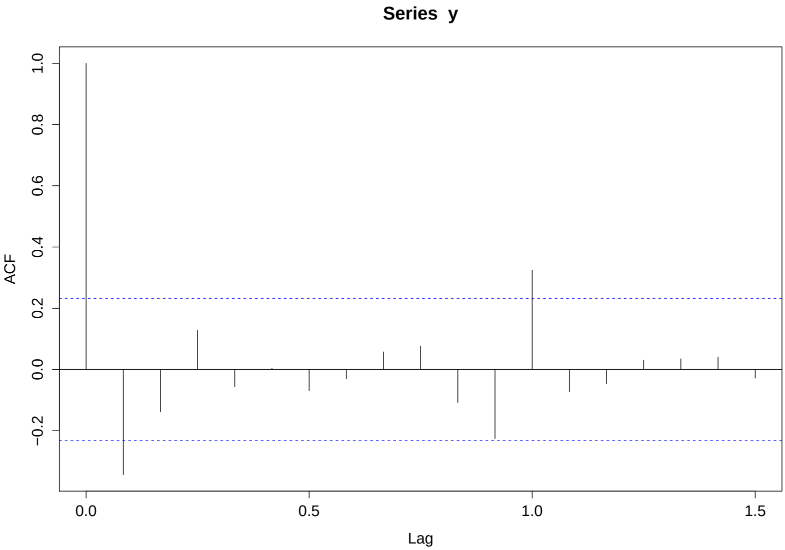
Example in R

```
souvenir <- scan("http://robjhyndman.com/tsdldata/data/fancy.dat")
#declare and plot time series
souvenirtimeseries <- ts(souvenir, frequency=12, start=c(1987,1))
y<-diff(diff(souvenirtimeseries,lag=12))
plot.ts(souvenirtimeseries)
plot.ts(y)
acf(y,plot=T)
```



Trend and seasonally differenced series $=$ y:

Correlogram:



Section 4

AR and MA Models

White Noise

A white noise process is a sequence of i.i.d. observations with zero mean and variance σ^2 and we will denote it by u_t.

It is the building block of more complicated processes:
For example, a random walk (without drift) model Y_t is defined by $\Delta Y_t = u_t$, or $Y_t = Y_{t-1} + u_t$.

u_t is sometimes called the innovation process. It is not predictable.

Why do we need models?

To describe parsimoneously the dynamics of the time series.

For forecasting. For example:

Take a random walk model: $Y_{t+1} = Y_t + u_{t+1}$ for every t. Recall that T is the last observation, then

$$\hat{Y}_{T+h} = Y_T,$$

for every forecast horizon h.

MA model

A stationary stochastic process Y_t is a moving average of order 1, MA(1), if it satisfies

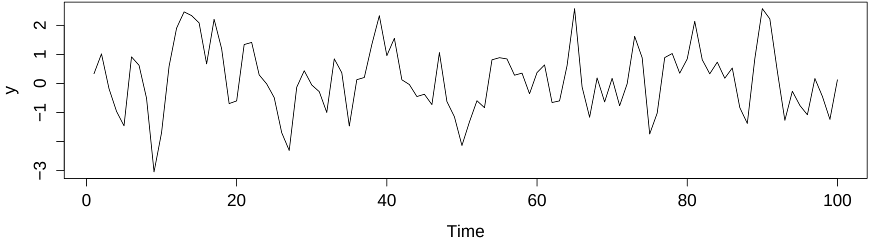
$$Y_t = a + u_t - \theta u_{t-1},$$

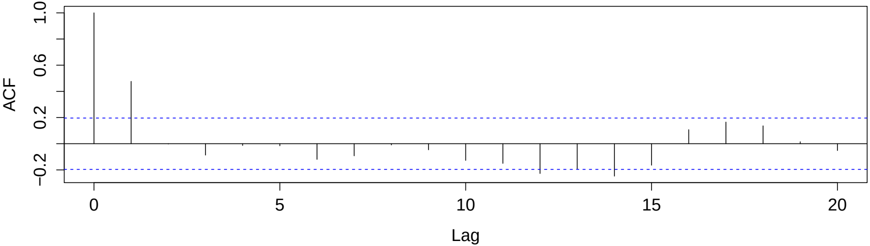
where a, and θ are unknown parameters.

The autocorrelations of an MA(1) are given by

- $\rho_0 = 1$
- $\rho_1 = \mathrm{Corr}(Y_t, Y_{t-1}) = -\frac{\theta}{(1+\theta^2)}$
- $\rho_2 = 0$
- $\rho_3 = 0$
- ...

The correlogram can be used to help us to *specify* an MA(1) proces:





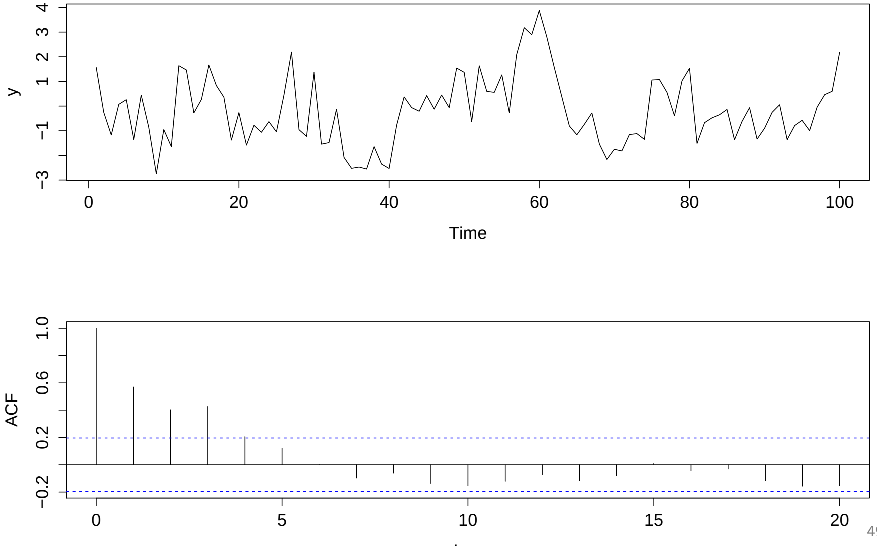
A stationary stochastic process Y_t is a moving average of order q, MA(q), if it satisfies

$$Y_t = a + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \ldots - \theta_q u_{t-q},$$

where a, and $\theta_1, \ldots, \theta_q$ are unknown parameters.

The autocorrelations of an MA(q) process are equal to zero for lags larger than q. If the correlogram shows a strong decline and becomes non significant after lag q, then there is evidence that the series was generated by an MA(q) process

MA(3)



Estimation

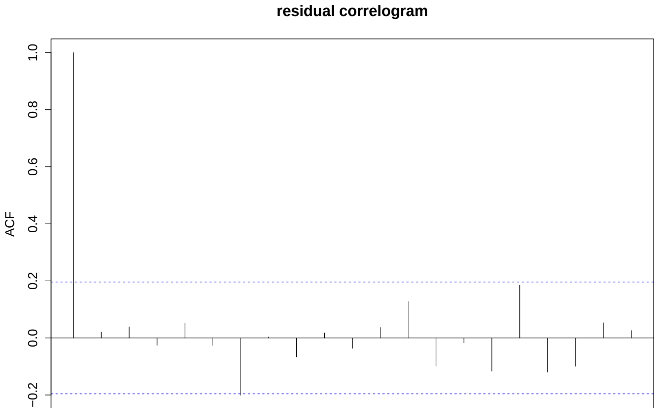
Using Maximum Likelihood (assuming normality of the innovations)

```
mymodel<-arima(y,order=c(0,0,3))
mymodel

Call:
arima(x = y, order = c(0, 0, 3))

Coefficients:
         ma1     ma2     ma3  intercept
      0.5158  0.1880  0.4150     0.0126
s.e.  0.0948  0.0935  0.0936     0.1835

sigma^2 estimated as 0.765:  log likelihood = -128.96,  aic = 267.92
```

Validation

The obtained residuals -after estimation - should be close to a white noise. It is good practice to make a correlogram of the residuals, in order to *validate* an MA(q) model.

```
acf(mymodel$res,plot=T,main="residual correlogram")
```



AR(1) model

A stationary stochastic process Y_t is an autoregressive of order 1, AR(1), if it satisfies

$$Y_t = a + \phi Y_{t-1} + u_t,$$

where a, and ϕ are unknown parameters.

The autocorrelations of an AR(1) are given by

- $\rho_0 = 1$
- $\rho_1 = \mathrm{Corr}(Y_t, Y_{t-1}) = \phi$
- $\rho_2 = \phi^2$
- $\rho_3 = \phi^3$
- ...

A stationary stochastic process Y_t is an autoregressive of order p, AR(p), if it satisfies

$$Y_t = a + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + u_t$$

where a, and $\phi_1, \ldots, \phi_p$ are unknown parameters.

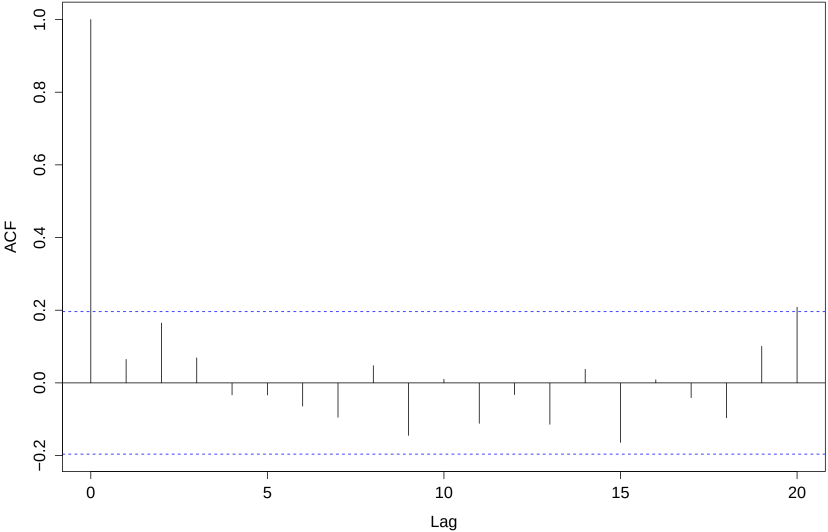
The autocorrelations tend more slowly to zero, and sometimes have a sinusoidal form.

Estimation

Using Maximum Likelihood

```
> mymodel<-arima(y,order=c(2,0,0))
> mymodel

Call:
arima(x = y, order = c(2, 0, 0))

Coefficients:
         ar1     ar2  intercept
      0.5190  0.3538    -0.4266
s.e.  0.0931  0.0940     0.7458

sigma^2 estimated as 1.073:  log likelihood = -146.08,  aic = 300.17
> acf(mymodel$res,plot=T,main="residual correlogram")
```

Residual Correlogram



Section 5

MA-infinity representation

Wold Representation Theorem

If Y_t is a stationary process, then it can be written as an MA(∞):

$$Y_t = c + u_t + \sum_{k=1}^{+\infty} \theta_k u_{t-k} \qquad \text{for any } t.$$

Example: AR(1)

$$
\begin{aligned}
Y_t &= a + \phi Y_{t-1} + u_t \\
&= a + \phi\left(a + \phi Y_{t-2} + u_{t-1}\right) + u_t \\
&= a(1+\phi) + u_t + \phi u_{t-1} + \phi^2 Y_{t-2} \\
&= a(1+\phi+\phi^2+\ldots) + u_t + \phi u_{t-1} + \phi^2 u_{t-2} + \ldots
\end{aligned}
$$

Recognize an MA(∞) with

$$\theta_k = \phi^k$$

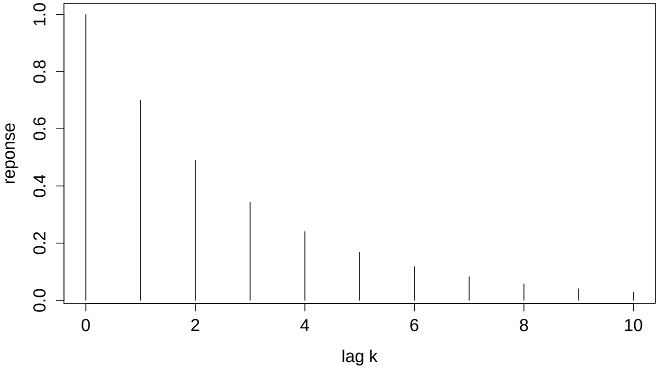
for every k from 1 to $+\infty$ and the constant $c = a/(1-\phi.)$

Impulse Response Function

Given an impulse to u_t of one unit, the response on Y_{t+k} is given by θ_k (see MA(∞) representation)

Example: AR(1):
$k \to \theta_k$ coefficients of .

$$
\begin{array}{cc}
\text{Impulse} & \text{Response} \\
\downarrow & \downarrow \\
\text{If } u_t \text{ increases with 1 unit, then} & Y_t \text{ increases with } 1 \\
 & Y_{t+1} \text{ increases with } \phi \\
 & Y_{t+2} \text{ increases with } \phi^2 \\
 & Y_{t+k} \text{ increases with } \phi^k.
\end{array}
$$


reponse
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
lag k

Part III

Introduction to dynamic models

Outline

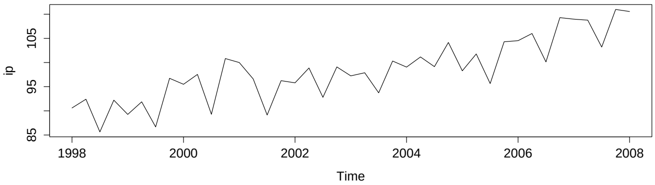
1. Example
2. Granger Causality
3. Vector Autoregressive Model

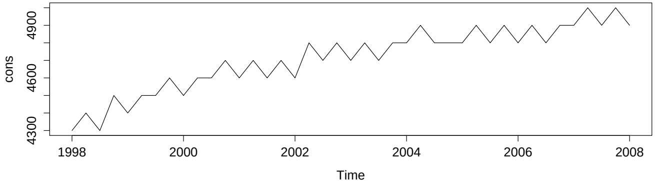
Section 1

Example

Example

Variable to predict: Industrial Production
Predictor: Consumption Prices





Running a regression

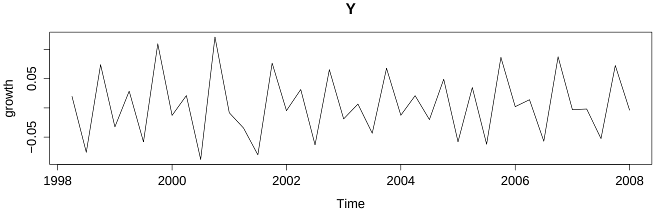
```
lm(formula = ip ~ cons)


Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.507915  13.940872  -3.264  0.00229 **
cons          0.030528   0.002956  10.326 1.02e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.402 on 39 degrees of freedom
Multiple R-squared: 0.7322,Adjusted R-squared: 0.7253
F-statistic: 106.6 on 1 and 39 DF,  p-value: 1.022e-12
```

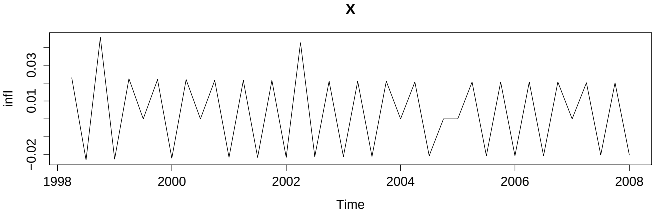
Spurious regression

- We succeed in predicting 73.2% of the variance of Industrial Production.
- This high number is there because both time series are upward trending, and driven by time.
- Regression non-stationary time series on each other is called *spurious regression*.
- Standard Inference requires stationary time series.

Going in log-differences

$Y = \Delta \log(\text{Industrial Production})$
$X = \Delta \log(\text{Consumption Prices})$





Note that

$$\Delta \log(X_t) = \log(X_t) - \log(X_{t-1}) = \log(X_t / X_{t-1}) \approx \frac{X_t - X_{t-1}}{X_{t-1}}.$$

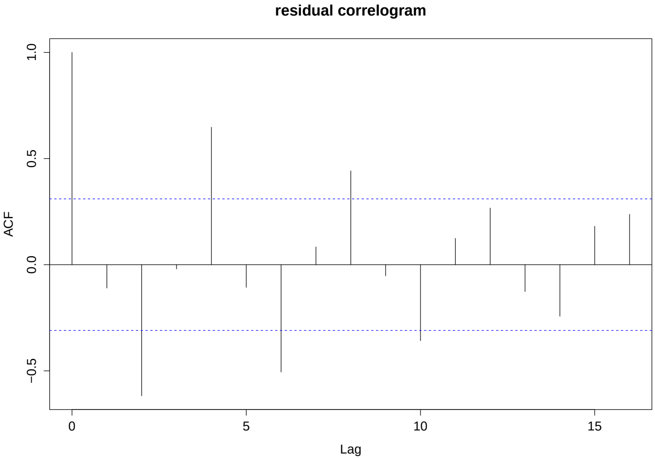
We get relative differences, or percentagewise increments.

$Y = \Delta$ log(Industrial Production) = Growth
$X = \Delta$ log(Consumption Prices) = Inflation

Running a regression

```
lm(formula = growth ~ infl)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000934   0.006408  -0.146    0.885
infl          1.810329   0.299670   6.041    5e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.04005 on 38 degrees of freedom
Multiple R-squared: 0.4899,Adjusted R-squared: 0.4765
F-statistic: 36.49 on 1 and 38 DF,  p-value: 5e-07
```

The regression in log-differences is a regression on stationary variables, but:



OLS remains consistent to estimate β_0 and β_1, but use *Newey-West* Standard Errors.

Section 2

Granger Causality

Granger Causality

Econometrics/Statistics can never proof that a causal relationship between X and Y exists.
Consider the equation

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \ldots + \alpha_k Y_{t-k} +$$

$$\beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \varepsilon_t.$$

We say that X *Granger causes* Y if it provides incremental predictive power for predicting Y.

Remark: select the lag k to have a valid model.

Test for no Granger Causality

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \ldots + \alpha_k Y_{t-k} +$$
$$\beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \varepsilon_t.$$

Test

$$H_0 : \beta_1 = \ldots = \beta_k = 0$$

using an F-statistics.

If we reject H_0, then there is significant Granger Causality.

Example: Does inflation Granger Causes growth?

R-code

```
lag=2;T=length(infl)
x=infl[(lag+1):T]
x.1=infl[(lag):(T-1)]
x.2=infl[(lag-1):(T-2)]
y.1=growth[(lag):(T-1)]
y.2=growth[(lag-1):(T-2)]

model<-lm(y~y.1+y.2+x.1+x.2)
model.small<-lm(y~y.1+y.2)
anova(model,model.small)
```

```
Analysis of Variance Table

Model 1: y ~ y.1 + y.2 + x.1 + x.2
Model 2: y ~ y.1 + y.2
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     33 0.034545
2     35 0.066910 -2 -0.032365 15.459 1.831e-05 ***
---
```

We strongly reject the hypothesis of no Granger Causality.

Comment: Interchanging the roles of X and Y yields an $F = 3.12 (P = 0.056)$. Hence there is also some evidence for Granger Causality in the other direction.

Section 3

Vector Autoregressive Model

VAR(1) for 3 series

$$
\begin{cases}
x_t & = & c_1 + a_{11}x_{t-1} + a_{12}y_{t-1} + a_{13}z_{t-1} + u_{x,t} \\
y_t & = & c_2 + a_{21}x_{t-1} + a_{22}y_{t-1} + a_{23}z_{t-1} + u_{y,t} \\
z_t & = & c_3 + a_{31}x_{t-1} + a_{32}y_{t-1} + a_{33}z_{t-1} + u_{z,t}.
\end{cases}
$$

- A VAR is estimated by OLS, equation by equation.
- The components of a VAR(p) do not follow AR(p) models
- The lag length p is selected using information criteria

The error terms are serially uncorrelated, with covariance matrix

$$\mathrm{Cov}(\vec{u}_t) = \begin{pmatrix} \mathrm{Var}(u_{x,t}) & \mathrm{Cov}(u_{x,t}, u_{y,t}) & \mathrm{Cov}(u_{x,t}, u_{z,t}) \\ \mathrm{Cov}(u_{x,t}, u_{y,t}) & \mathrm{Var}(u_{y,t}) & \mathrm{Cov}(u_{y,t}, u_{z,t}) \\ \mathrm{Cov}(u_{x,t}, u_{z,t}) & \mathrm{Cov}(u_{y,t}, u_{z,t}) & \mathrm{Var}(u_{z,t}) \end{pmatrix}.$$

We assume that $\vec{u}_t$ is a multivariate white noise:

- $E[\vec{u}_t] = 0$
- $\mathrm{Cov}(\vec{u}_t, \vec{u}_{t-k}) = 0$ for $k > 0$
- $\mathrm{Cov}(\vec{u}_t) := \Sigma$

No correlation at *leads and lags* between components of $\vec{u}_t$; only instantaneous correlation is allowed.

Impulse-response functions:

If component i of the innovation $\vec{u}_t$ changes with one-unit, then component j of $\vec{y}_{t+k}$ changes with $(B_k)_{ji}$ (other things equal).

The function

$$k \to (B_k)_{ji}$$

is called the impulse-response function.

There are k^2 impulse response functions.

[There exists many variants of the impulse response functions.]

VAR example: inflation-growth

```
> library(vars)
> mydata<-cbind(infl,growth)
> VARselect(mydata)
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     3      3      3      3

$criteria
                  1             2             3             4
AIC(n) -1.551591e+01 -1.621211e+01 -1.679070e+01 -1.668189e+01
HQ(n)  -1.542626e+01 -1.606269e+01 -1.658152e+01 -1.641294e+01
SC(n)  -1.523567e+01 -1.574504e+01 -1.613681e+01 -1.584117e+01
FPE(n)  1.828564e-07  9.160015e-08  5.193928e-08  5.911318e-08
                  5             6             7             8
AIC(n) -1.659127e+01 -1.649105e+01 -1.637611e+01 -1.651495e+01
HQ(n)  -1.626255e+01 -1.610256e+01 -1.592786e+01 -1.600692e+01
SC(n)  -1.556373e+01 -1.527668e+01 -1.497491e+01 -1.492692e+01
FPE(n)  6.692637e-08  7.785562e-08  9.410207e-08  9.110841e-08
                  9            10
AIC(n) -1.629013e+01 -1.645541e+01
HQ(n)  -1.572234e+01 -1.582785e+01
SC(n)  -1.451528e+01 -1.449373e+01
FPE(n)  1.326417e-07  1.393602e-07
```

VAR example: inflation-growth

```
> summary(mymodel)


Estimation results for equation infl:
=====================================
infl = infl.l1 + growth.l1 + infl.l2 + growth.l2 + infl.l3 + growth.l3 + const

           Estimate Std. Error t value Pr(>|t|)
infl.l1   -0.360218   0.162016  -2.223  0.03387 *
growth.l1 -0.078241   0.052501  -1.490  0.14660
infl.l2   -0.207708   0.164635  -1.262  0.21680
growth.l2  0.058537   0.040690   1.439  0.16061
infl.l3   -0.254240   0.161223  -1.577  0.12530
growth.l3 -0.100953   0.052114  -1.937  0.06219 .
const      0.006283   0.001869   3.361  0.00213 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


Residual standard error: 0.00865 on 30 degrees of freedom
Multiple R-Squared:  0.85,Adjusted R-squared: 0.8201
F-statistic: 28.34 on 6 and 30 DF,  p-value: 4.38e-11
```

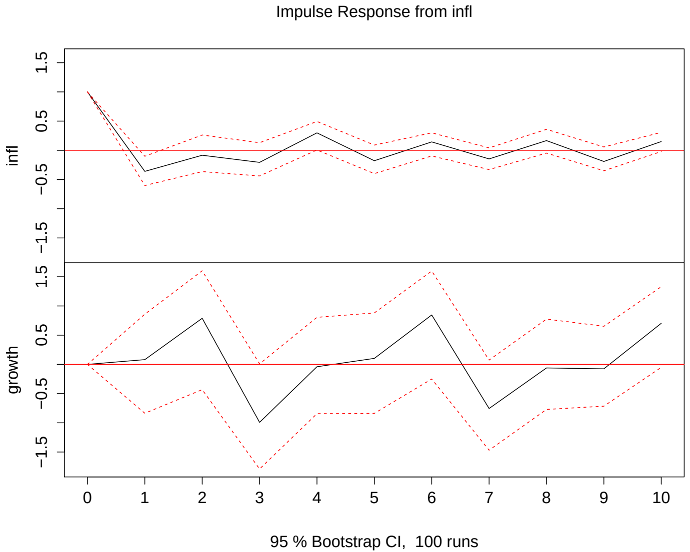
VAR example: inflation-growth

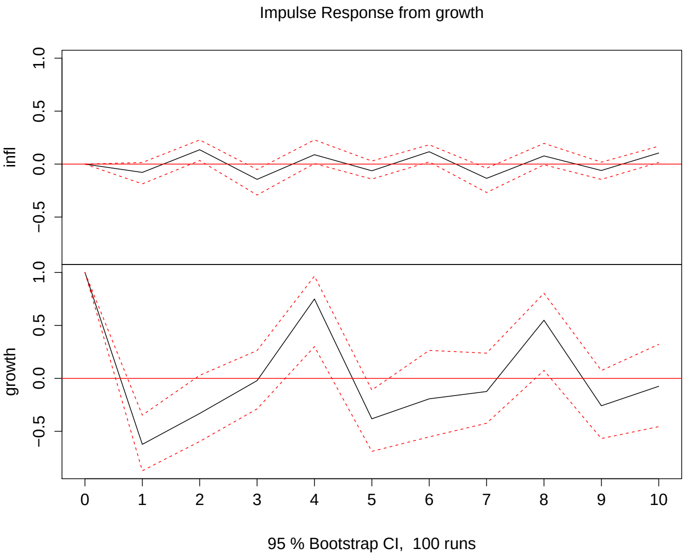
```
Estimation results for equation growth:
=======================================
growth = infl.l1 + growth.l1 + infl.l2 + growth.l2 + infl.l3 + growth.l3 + const

           Estimate Std. Error t value Pr(>|t|)
infl.l1    0.081107   0.491116   0.165 0.869935
growth.l1 -0.623130   0.159147  -3.915 0.000481 ***
infl.l2    0.868815   0.499055   1.741 0.091946 .
growth.l2 -0.713378   0.123342  -5.784 2.56e-06 ***
infl.l3   -0.122685   0.488712  -0.251 0.803497
growth.l3 -0.615628   0.157972  -3.897 0.000506 ***
const      0.012084   0.005667   2.132 0.041267 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


Residual standard error: 0.02622 on 30 degrees of freedom
Multiple R-Squared: 0.8089,Adjusted R-squared: 0.7707
F-statistic: 21.17 on 6 and 30 DF,  p-value: 1.513e-09
```

VAR example: prediction

```
> predict(mymodel,n.ahead=6)
$infl
            fcst        lower       upper         CI
[1,]  0.02435690  0.007403536 0.041310271 0.01695337
[2,] -0.01203683 -0.030790030 0.006716372 0.01875320
[3,]  0.01626860 -0.003676775 0.036213970 0.01994537
[4,] -0.01281687 -0.034638813 0.009005066 0.02182194
[5,]  0.02049182 -0.002592846 0.043576485 0.02308467
[6,] -0.01096528 -0.034561170 0.012630615 0.02359589

$growth
             fcst         lower       upper         CI
[1,]  0.013351737 -0.0380387925 0.064742266 0.05139053
[2,] -0.056308643 -0.1167153773 0.004098092 0.06040673
[3,]  0.062663063 -0.0007228443 0.126048970 0.06338591
[4,] -0.007140285 -0.0727913007 0.058510731 0.06565102
[5,]  0.021067459 -0.0549650606 0.097099979 0.07603252
[6,] -0.045996773 -0.1244454309 0.032451885 0.07844866
```

Example: impulse-response functions




Impulse Response from growth
infl
growth
1.0
0.5
0.0
−0.5
0 1 2 3 4 5 6 7 8 9 10
95 % Bootstrap CI, 100 runs

Part IV

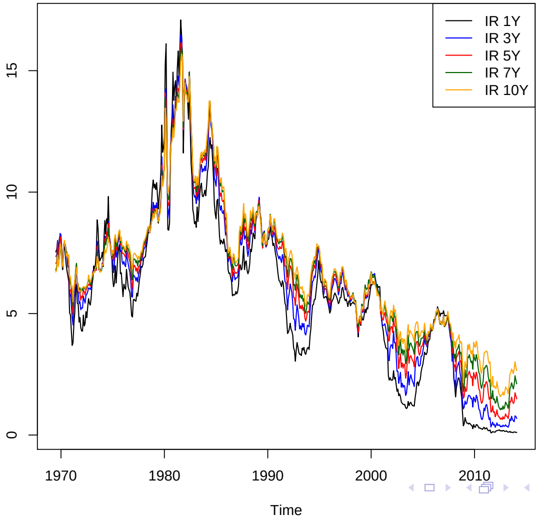
Sparse Cointegration

Outline

Section 1

Introduction

Interest Rate Example



Bivariate cointegration

Consider two time series $y_{1,t}$ and $y_{2,t}$, $I(1)$.

$y_{1,t}$ and $y_{2,t}$ are cointegrated if there exists a linear combination

$$\beta_{11} y_{1,t} + \beta_{21} y_{2,t} = \delta_t$$

such that δ_t is stationary.

- $\beta_{11} y_{1,t} + \beta_{21} y_{2,t} = \delta_t$: *Cointegration Equation*
- $\boldsymbol{\beta} = (\beta_{11}, \beta_{21})'$: *Cointegrating vector*

Bivariate cointegration (cont.)

Vector Error Correcting Representation:

$$\begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} = \begin{bmatrix} \gamma_{11,1} & \gamma_{12,1} \\ \gamma_{21,1} & \gamma_{22,1} \end{bmatrix} \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + \ldots + \begin{bmatrix} \gamma_{11,p-1} & \gamma_{12,p-1} \\ \gamma_{21,p-1} & \gamma_{22,p-1} \end{bmatrix} \begin{bmatrix} \Delta y_{1,t-p+1} \\ \Delta y_{2,t-p+1} \end{bmatrix} + \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}.$$

Note: $\Delta y_{1,t} = y_{1,t} - y_{1,t-1}$

Vector Error Correcting Model

Let $\mathbf{y}_t$ be a q-dimensional multivariate time series, $I(1)$.

Vector Error Correcting Representation:

$$\boldsymbol{\Delta}\mathbf{y}_t = \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \boldsymbol{\Delta}\mathbf{y}_{t-i} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T$$

where

- ε_t follows $N_q(\mathbf{0}, \boldsymbol{\Sigma})$, denote $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$
- $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{p-1}$ $q \times q$ matrices of short-run effects
- $\boldsymbol{\Pi}$ $q \times q$ matrix.

Vector Error Correcting Model (cont.)

$$\boldsymbol{\Delta}\mathbf{y}_t = \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \boldsymbol{\Delta}\mathbf{y}_{t-i} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \varepsilon_t. \qquad t = 1, \ldots, T$$

If $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ $q \times r$ matrices of full column rank r $(r < q)$

Then, $\boldsymbol{\beta}'\mathbf{y}_t$ stationary

- $\mathbf{y}_t$ cointegrated with cointegration rank r
- $\boldsymbol{\beta}$: cointegrating vectors
- $\boldsymbol{\alpha}$: adjustment coefficients.

Maximum likelihood estimation (Johansen, 1996)

Rewrite the VECM in matrix notation:

$$\mathbf{\Delta Y} = \mathbf{\Delta Y_L \Gamma} + \mathbf{Y \Pi}' + \mathbf{E},$$

where

- $\mathbf{\Delta Y} = (\mathbf{\Delta y}_{p+1}, \ldots, \mathbf{\Delta y}_T)'$
- $\mathbf{\Delta Y_L} = (\mathbf{\Delta X}_{p+1}, \ldots, \mathbf{\Delta X}_T)'$ with $\mathbf{\Delta X}_t = (\mathbf{\Delta y}'_{t-1}, \ldots, \mathbf{\Delta y}'_{t-p+1})'$
- $\mathbf{Y} = (\mathbf{y}_p, \ldots, \mathbf{y}_{T-1})'$
- $\mathbf{\Gamma} = (\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_{p-1})'$
- $\mathbf{E} = (\varepsilon_{p+1}, \ldots, \varepsilon_T)'$.

Maximum likelihood estimation (Johansen, 1996) (cont.)

Negative log likelihood

$$\mathcal{L}(\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}) = \frac{1}{T}\mathrm{tr}\Big((\mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma} - \mathbf{Y\Pi}')\Omega(\mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma} - \mathbf{Y\Pi}')'\Big) - \log|\mathbf{\Omega}|.$$

Maximum likelihood estimator:

$$(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Pi}}, \widehat{\mathbf{\Omega}}) = \underset{\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}}{\mathrm{argmin}} \;\; \mathcal{L}(\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}),$$

$$\text{subject to} \quad \mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'.$$

Problems:

- When $T \approx pq$: ML estimator has low precision
- When $T < pq$: ML estimator does not exist

Section 2

Penalized Estimation

Penalized Regression

Given standard regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i,$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$.

Penalized estimate of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + n\lambda P(\boldsymbol{\beta}),$$

with λ a penalty parameter and $P(\boldsymbol{\beta})$ a penalty function.

Penalized Regression (cont.)

Choices of penalty functions:

- $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$: Lasso - regularization and sparsity
- $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$: Ridge - regularization
- ...

Penalized ML estimation

Penalized negative log likelihood

$$\mathcal{L}_P(\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}) = \frac{1}{T}\mathrm{tr}\Big((\mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma} - \mathbf{Y\Pi}')\Omega(\mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma} - \mathbf{Y\Pi}')'\Big) - \log|\mathbf{\Omega}| + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\mathbf{\Gamma}) + \lambda_3 P_3(\mathbf{\Omega}),$$

with P_1, P_2 and P_3 three penalty functions.

Penalized maximum likelihood estimator:

$$(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Pi}}, \widehat{\mathbf{\Omega}}) = \underset{\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}}{\mathrm{argmin}} \;\; \mathcal{L}_P(\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}),$$

$$\text{subject to} \quad \mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'.$$

Algorithm

Iterative procedure:

1. Solve for $\mathbf{\Pi}$ conditional on $\mathbf{\Gamma}, \mathbf{\Omega}$
2. Solve for $\mathbf{\Gamma}$ conditional on $\mathbf{\Pi}, \mathbf{\Omega}$
3. Solve for $\mathbf{\Omega}$ conditional on $\mathbf{\Gamma}, \mathbf{\Pi}$

1. Solving for $\mathbf{\Pi}$ conditional on $\mathbf{\Gamma}, \mathbf{\Omega}$

Solve

$$(\hat{\alpha}, \hat{\beta})|\mathbf{\Gamma}, \mathbf{\Omega} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{argmin}} \frac{1}{T} \text{tr}\Big((\mathbf{G} - \mathbf{Y}\beta\alpha')\mathbf{\Omega}(\mathbf{G} - \mathbf{Y}\beta\alpha')'\Big) + \lambda_1 P_1(\beta).$$

$$\text{subject to } \boldsymbol{\alpha}'\mathbf{\Omega}\boldsymbol{\alpha} = \mathbf{I}_r$$

with

- $\mathbf{G} = \mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma}$

→ Penalized reduced rank regression (e.g. Chen and Huang, 2012)

1.1 Solving for α conditional on $\mathbf{\Gamma}, \mathbf{\Omega}, \beta$

Solve

$$\hat{\alpha}|\mathbf{\Gamma}, \mathbf{\Omega}, \beta = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\mathbf{G} - \mathbf{B}\alpha')\mathbf{\Omega}(\mathbf{G} - \mathbf{B}\alpha')'\Big)$$

$$\text{subject to } \boldsymbol{\alpha}'\mathbf{\Omega}\boldsymbol{\alpha} = \mathbf{I}_r,$$

with

- $\mathbf{G} = \mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma}$
- $\mathbf{B} = \mathbf{Y}\beta$

$\rightarrow$ Weighted Procrustes problem

1.2 Solving for β conditional on $\mathbf{\Gamma}, \mathbf{\Omega}, \boldsymbol{\alpha}$

Solve

$$\hat{\beta}|\mathbf{\Gamma}_{,}\mathbf{\Omega}, \boldsymbol{\alpha} \;=\; \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \;\; \frac{1}{T}\mathrm{tr}\Big((\mathbf{R} - \mathbf{Y}\beta)(\mathbf{R} - \mathbf{Y}\beta)'\Big) + \lambda_1 P_1(\beta).$$

with

- $\mathbf{R} = \mathbf{G}\Omega\alpha = (\mathbf{\Delta Y} - \mathbf{\Delta Y_L \Gamma})\Omega\alpha$

$\rightarrow$ Penalized multivariate regression

Choice of penalty function

Our choice:

- Lasso penalty: $P_1(\boldsymbol{\beta}) = \sum_{i=1}^{q} \sum_{j=1}^{r} |\beta_{ij}|$

Other penalty functions are possible:

- Adaptive Lasso: $P_1(\boldsymbol{\beta}) = \sum_{i=1}^{q} \sum_{j=1}^{r} w_{ij} |\beta_{ij}|$
  - Weights : $w_{ij} = 1/|\hat{\beta}_{ij}^{initial}|$
- ...

2. Solving for $\mathbf{\Gamma}$ conditional on $\mathbf{\Pi}, \mathbf{\Omega}$

Solve

$$\widehat{\mathbf{\Gamma}}|\mathbf{\Pi}, \mathbf{\Omega} = \underset{\mathbf{\Gamma}}{\text{argmin}} \quad \frac{1}{T}\text{tr}\Big((\mathbf{D} - \mathbf{\Delta Y_L \Gamma})\mathbf{\Omega}(\mathbf{D} - \mathbf{\Delta Y_L \Gamma})'\Big) + \lambda_2 P_2(\mathbf{\Gamma}).$$

with

- $\mathbf{D} = \mathbf{\Delta Y} - \mathbf{Y\Pi}'$
- Lasso penalty: $P_2(\mathbf{\Gamma}) = \sum_{i=1}^{q} \sum_{j=1}^{q} \sum_{k=1}^{p-1} |\gamma_{ijk}|$

$\rightarrow$ Penalized multivariate regression

3. Solving for $\mathbf{\Omega}$ conditional on $\mathbf{\Gamma}, \mathbf{\Pi}$

Solve

$$\widehat{\mathbf{\Omega}}|\mathbf{\Gamma},\mathbf{\Pi} = \underset{\mathbf{\Omega}}{\operatorname{argmin}} \ \frac{1}{T}\operatorname{tr}\Big((\mathbf{D}-\mathbf{\Delta Y_L \Gamma})\mathbf{\Omega}(\mathbf{D}-\mathbf{\Delta Y_L \Gamma})'\Big)-\log|\mathbf{\Omega}|+\lambda_3 P_3(\mathbf{\Omega}).$$

with

- $\mathbf{D} = \mathbf{\Delta Y} - \mathbf{Y\Pi}'$
- Lasso penalty: $P_3(\mathbf{\Omega}) = \sum_{k\neq k\prime} |\Omega_{kk'}|$

$\rightarrow$ Penalized inverse covariance matrix estimation (Friedman et al., 2008)

Selection of tuning parameters

λ_1 and λ_2: Time series cross-validation (Hyndman, 2014)

λ_3: Bayesian Information Criterion

Time series cross-validation

Denote the response by $\mathbf{z}_t$

For $t = S, \ldots, T-1$, (with $S = \lfloor 0.8T \rfloor$)

1. Fit model to $\mathbf{z}_1, \ldots, \mathbf{z}_t$
2. Compute $\widehat{\mathbf{e}}_{t+1} = \mathbf{z}_{t+1} - \hat{\mathbf{z}}_{t+1}$

Select value of tuning parameter that minimizes

$$\text{MMAFE} = \frac{1}{T-S} \frac{1}{q} \sum_{t=S}^{T-1} \sum_{i=1}^{q} \frac{|\hat{e}_{t+1}^{(i)}|}{\hat{\sigma}^{(i)}},$$

with

- $\hat{e}_t^{(i)}$ the i^{th} component of $\widehat{\mathbf{e}}_t$
- $\hat{\sigma}^{(i)} = \text{sd}(z_t^{(i)})$.

Determination of cointegration rank

Iterative procedure based on Rank Selection Criterion (Bunea et al., 2011):

- Set $r_{\text{start}} = q$
- For $\hat{r} = r_{\text{start}}$, obtain $\widehat{\mathbf{\Gamma}}$ using the sparse cointegrating algorithm
- Update $\hat{r}$ to

$$\hat{r} = \max\{r : \lambda_r(\widetilde{\mathbf{\Delta Y}}' \mathbf{P} \widetilde{\mathbf{\Delta Y}}) \geq \mu\},$$

with

- $\widetilde{\mathbf{\Delta Y}} = \mathbf{\Delta Y} - \mathbf{\Delta Y_L}\widehat{\mathbf{\Gamma}}$
- $\mathbf{P} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-}\mathbf{Y}'$
- $\mu = 2S^2(q + l)$ with
  - $l = \text{rank}(\mathbf{Y})$
  - $S^2 = \frac{||\widetilde{\mathbf{\Delta Y}} - \mathbf{P}\widetilde{\mathbf{\Delta Y}}||_F^2}{Tq - lq}$

$\rightarrow$ Iterate until $\hat{r}$ does not change in two successive iterations

Determination of cointegration rank (cont.)

Properties of Rank Selection Criterion :

- Consistent estimate of rank($\mathbf{\Pi}$)
- Low computational cost

Simulation Study

VECM(1) with dimension q:

$$\Delta \mathbf{y}_t = \alpha\beta' \mathbf{y}_{t-1} + \mathbf{\Gamma}_1 \Delta \mathbf{y}_{t-1} + \mathbf{e}_t, \qquad (t = 1, \ldots, T),$$

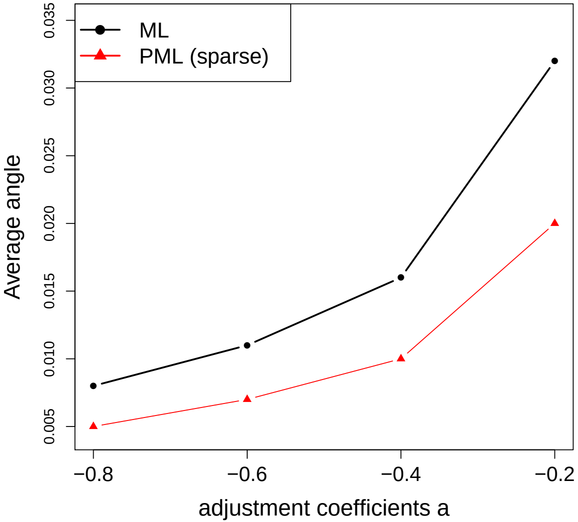
where $\mathbf{e}_t$ follows $N_q(\mathbf{0}, \mathbf{I_q})$.

Designs:

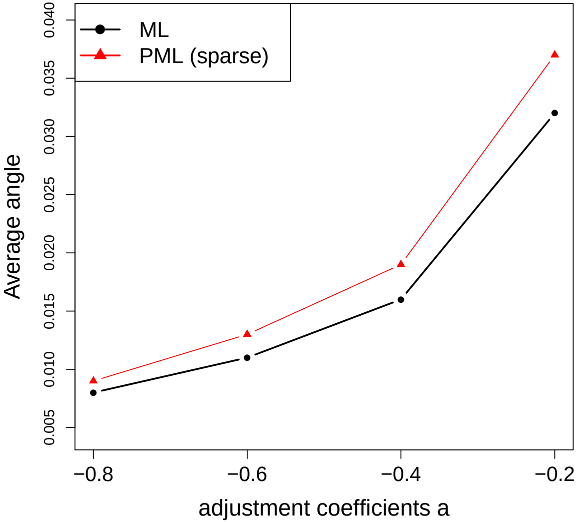
1. Low-dimensional design: $T = 500, q = 4, r = 1$
   - Sparse cointegrating vector
   - Non-sparse cointegrating vector
2. High-dimensional design: $T = 50, q = 11, r = 1$
   - Sparse cointegrating vector
   - Non-sparse cointegrating vector

Sparse Low-dimensional design

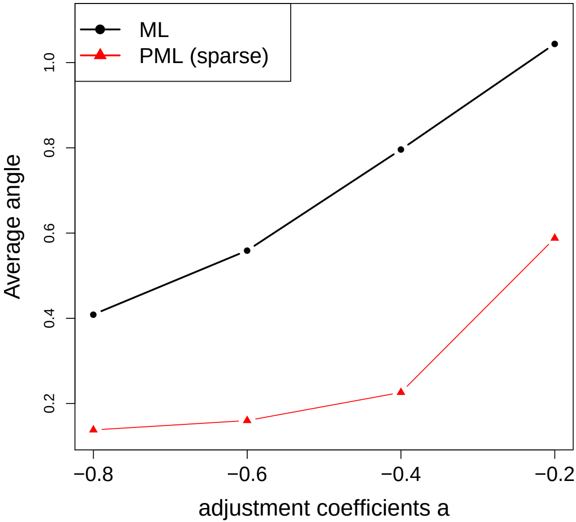
Average angle between estimated and true cointegration space



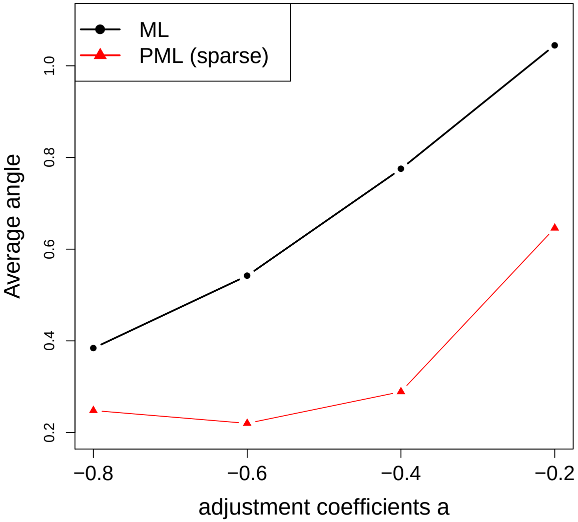
Non-sparse Low-dimensional design



Sparse High-dimensional design



Non-sparse High-dimensional design



Section 3

Forecasting Applications

Rolling window forecast with window size S

Estimate the VECM at $t = S, \ldots, T - h$

$$\widehat{\Delta \mathbf{y}}_{t+h} = \sum_{i=1}^{p-1} \widehat{\boldsymbol{\Gamma}}_i \boldsymbol{\Delta} \mathbf{y}_{t+1-i} + \widehat{\boldsymbol{\Pi}} \mathbf{y}_t,$$

for forecast horizon h.

Obtain h-step-ahead multivariate forecast errors

$$\widehat{\mathbf{e}}_{t+h} = \Delta \mathbf{y}_{t+h} - \widehat{\boldsymbol{\Delta} \mathbf{y}}_{t+h}.$$

Forecast error measures

Multivariate Mean Absolute Forecast Error:

$$\text{MMAFE} = \frac{1}{T-h-S+1}\sum_{t=S}^{T-h}\frac{1}{q}\sum_{i=1}^{q}\frac{|\Delta y_{t+h}^{(i)} - \widehat{\Delta y}_{t+h}^{(i)}|}{\widehat{\sigma}_{(i)}},$$

where $\widehat{\sigma}_{(i)}$ is the standard deviation of the i^{th} time series in differences.

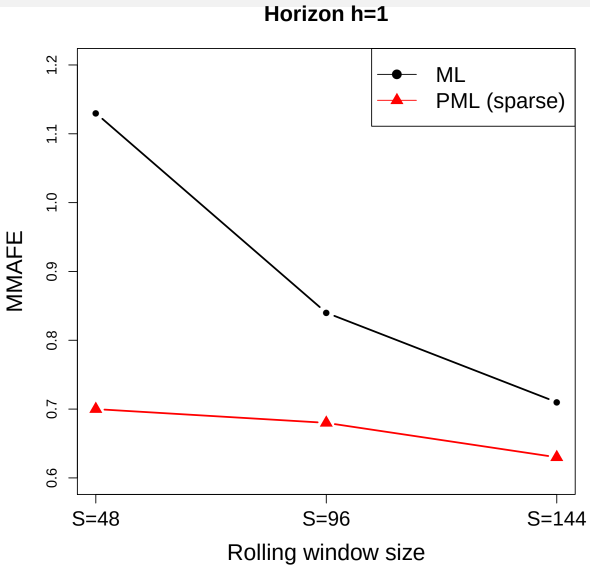
Interest Rate Growth Forecasting

Interest Data: $q = 5$ monthly US treasury bills

- Maturity: 1, 3, 5, 7 and 10 year
- Data range: January 1969 - June 2015

Methods: PML (sparse) versus ML

Multivariate Mean Absolute Forecast Error



Consumption Growth Forecasting

Data: $q = 31$ monthly consumption time series

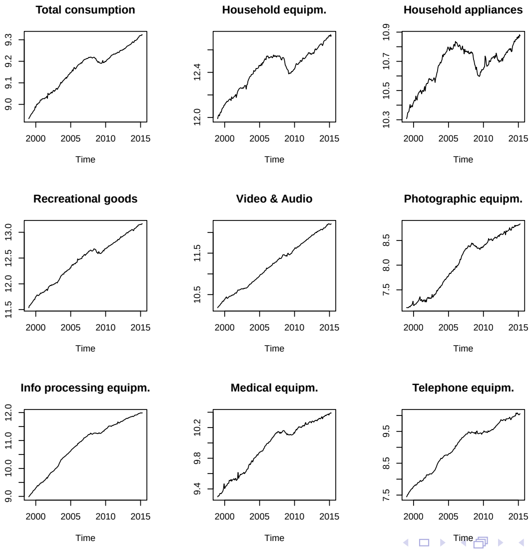
- Total consumption and industry-specific consumption
- Data range: January 1999-April 2015

Forecast: Rolling window of size $S = 144$

Methods:

- Cointegration: PML (sparse), ML, Factor Model
- No cointegration: PML (sparse), ML, Factor Model, Bayesian, Bayesian Reduced Rank

Consumption Time Series



Multivariate Mean Absolute Forecast Error

	Method	$h = 1$	$h = 3$	$h = 6$	$h = 12$
Cointegration	**PML (sparse)**	0.79	**0.62**	**0.63**	**0.61**
	ML	0.74	0.78	0.82	0.72
	Factor Model	**0.66**	0.66	0.67	0.65
No Cointegration	PML (sparse)	0.94	0.67	0.67	0.66
	ML	5.40	4.81	4.84	5.22
	Factor Model	0.72	0.75	0.77	0.72
	Bayesian	0.69	0.71	0.74	0.72
	Bayesian Reduced Rank	0.69	0.71	0.74	0.72

References

1. Wilms, I. and Croux, C. (2016), "Forecasting using sparse cointegration," *International Journal of Forecasting*, 32(4), 1256-1267.

2. Bunea, F.; She, Y. and Wegkamp, M. (2011), "Optimal selection of reduced rank estimators of high-dimensional matrices," *The Annals of Statistics*, 39, 1282-1309.
3. Chen, L. and Huang, J. (2012), "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*,107, 1533-1545.
4. Friedman, J., Hastie, T. and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432-441.
5. Johansen (1996), *Likelihood-based inference in cointegrated vector autoregressive models*, Oxford: Oxford University Press.

6. Gelper, S., Wilms, I. and Croux, C. (2016), "Identifying demand effects in a large network of product categories," *Journal of Retailing*, 92(1), 25-39.

Part V

Robust Exponential Smoothing

Outline

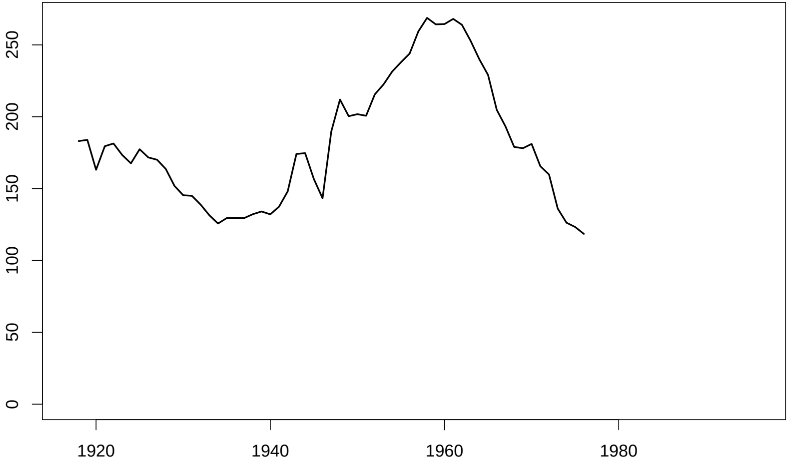
Section 1

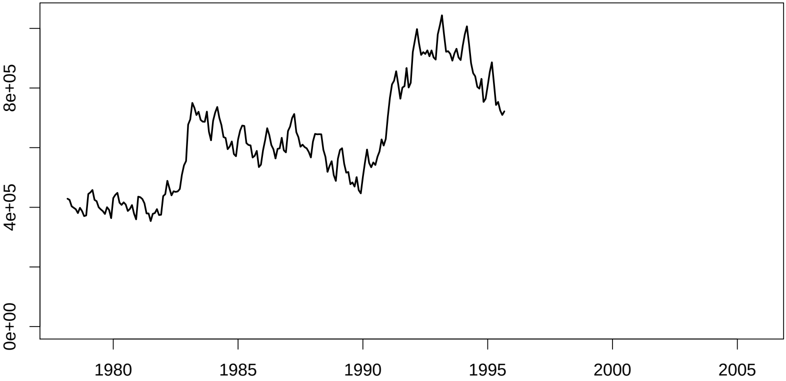
Introduction

Goal

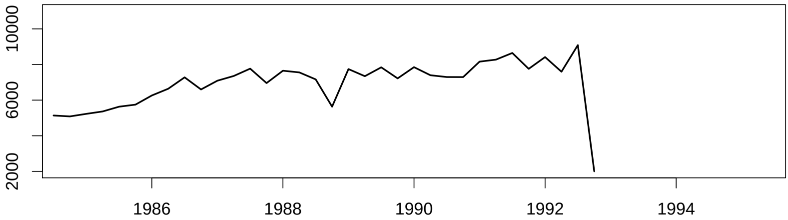
Forecast many kinds of univariate time series

Usually rather short time series





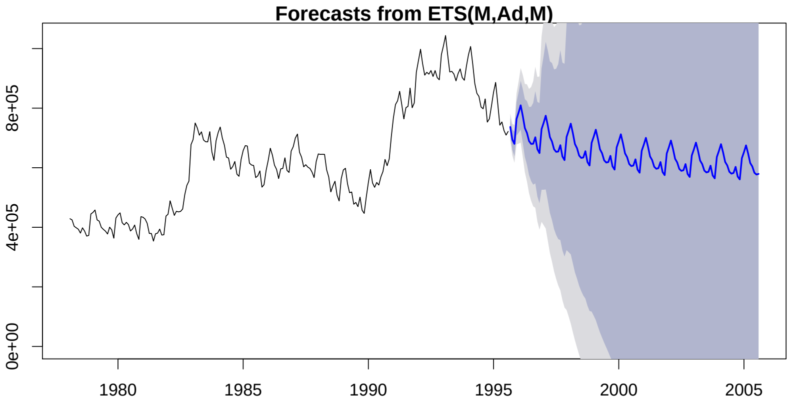
Monthly number of unemployed persons in Australia, from February 1978 till August 1995



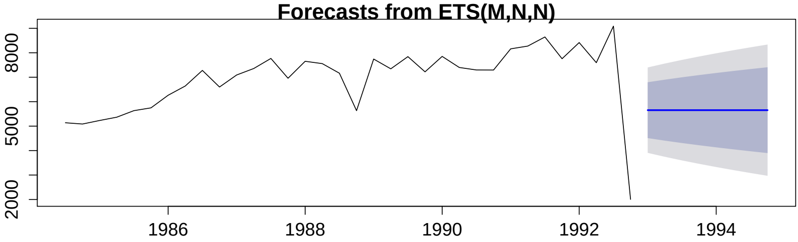
A quarterly microeconometric time series

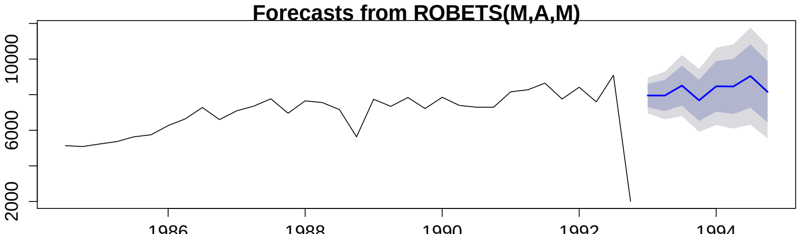
Additive or Multipicative noise? No trend, damped trend or trend? No seasonality? Additive or multiplicative seasonality?

```
model <- ets(y)      # Hyndman and Khandakar (2008)
plot(forecast(model, h = 120))
```



```
model1 <- ets(y)      # Hyndman and Khandakar (2008)
model2 <- robets(y)   # our procedure
plot(forecast(model1, h = 8)) # first plot
plot(forecast(model2, h = 8)) # second plot
```





Section 2

Exponential Smoothing

Simple exponential smoothing

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t \\
\ell_t &= \ell_{t-1} + \alpha(y_t - \ell_{t-1})
\end{aligned}
$$

y_t: univariate time series
$\hat{y}_{t+h|t}$: *h*-step ahead prediction
ℓ_t: level
α: smoothing parameter; in [0,1]

Exponential smoothing with Trend and Seasonality (ETS)

E, underlying error model: A (additive) or M (multiplicative),

T, type of trend: N (none), A (additive) or A_d (damped) and

S, type of seasonal: N (none), A (additive) or M (multiplicative).

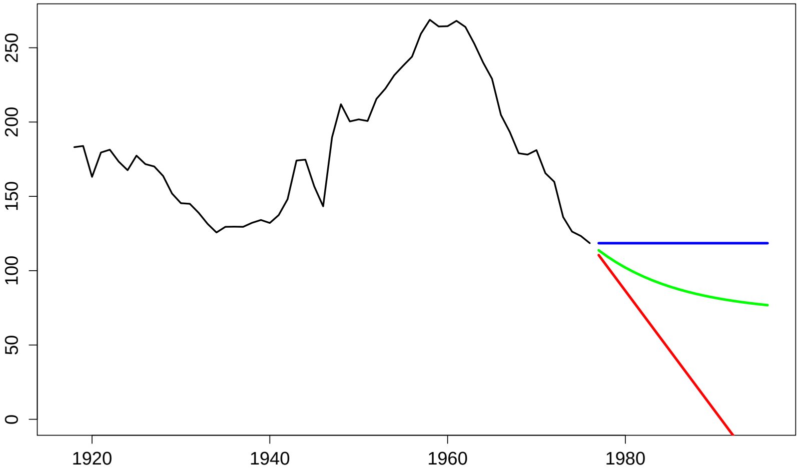
Example 1: additive damped trend without seasonality

Model: $\mathrm{AA_dN}$ / $\mathrm{MA_dN}$

$$\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + \sum_{j=1}^{h} \phi^j b_t \\
\ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1}
\end{aligned}$$

ℓ_t: the level
b_t: the (damped) trend
α, β: smoothing parameters
ϕ: damping parameter

Births per 10,000 of 23 year old women, USA



No trend ($\phi = 0$, ANN-model) , Full trend ($\phi = 1$, AAN-model)
$\rightarrow$ damped trend (AA_dN)

Example 2: additive damped trend and multiplicative seasonality (MA_dM)

$$
\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + \sum_{j=1}^{h} \phi^j b_t) s_{t+h_m^+ - m} \\
\ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \\
s_t &= \gamma \frac{y_t}{\ell_{t-1} + \phi b_{t-1}} + (1-\gamma) s_{t-m}.
\end{aligned}
$$

ℓ_t: the level

b_t: the (damped) trend

s_t: the seasonal components

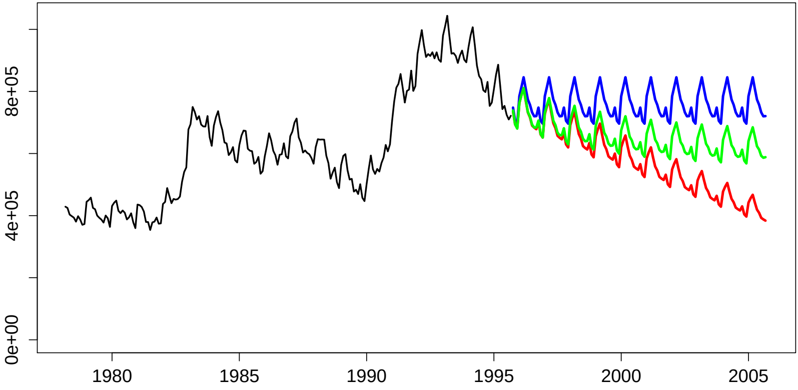
α, β, γ: smoothing parameters

ϕ: damping parameter

$h_m^+ = (h-1) \bmod m + 1$

m: number of seasons per period

Monthly number of unemployed persons in Australia, from February 1978 till August 1995



No trend ($\phi = 0$, MNM-model) , Full trend ($\phi = 1$, MAM-model)
$\rightarrow$ damped trend (MA_dM)

How ets and robets work

$$\begin{tikzcd}
\text{ANN} \arrow[d] & \text{AAN} \arrow[d] & \text{AA}_\text{d}\text{N} \arrow[d, "\text{estimate parameters}"] & \cdots & \text{MA}_\text{d}\text{M} \arrow[d] \\
\alpha \arrow[d] & \alpha, \beta \arrow[d] & \alpha, \beta, \phi \arrow[d, "\text{compute AICC}"] & & \alpha, \beta, \phi, \gamma \arrow[d] \\
\text{AICC} & \text{AICC} & \text{AICC} & \cdots & \text{AICC} \arrow[dll, "\text{select model with lowest AICC}"] \\
 & & \text{MA}_\text{d}\text{M} & &
\end{tikzcd}$$

AICC: selection criterion that penalizes models with a lot of parameters

Section 3

Robust approach

ETS is not robust

Model: $\mathrm{AA_dA}$ / $\mathrm{MA_dA}$

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + \sum_{j=1}^{h} \phi^j b_t + s_{t-m+h_m^+} \\
\ell_t &= \alpha(\boxed{y_t} - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \\
s_t &= \gamma(\boxed{y_t} - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}.
\end{aligned}
$$

$$\Downarrow$$

linear dependence of y_t on all forecasts $\hat{y}_{t+h|t}$
(for this variant, but also for all other variants)

How to make it robust?

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + \sum_{j=1}^{h} \phi^j b_t + s_{t-m+h_m^+} \\
\ell_t &= \alpha(\boxed{y_t^*} - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \\
s_t &= \gamma(\boxed{y_t^*} - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}.
\end{aligned}
$$

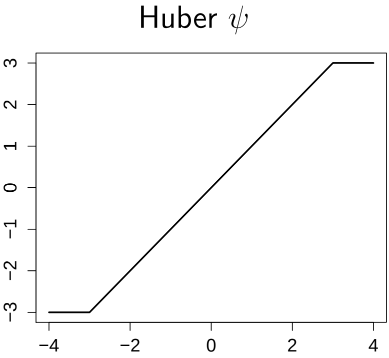
replace y_t by y_t^*, resulting in forecasts $\hat{y}_{t+h|t}$.

$$
y_t^* = \psi\left[\frac{y_t - \hat{y}_{t|t-1}}{\hat{\sigma}_t}\right]\hat{\sigma}_t + \hat{y}_{t|t-1}
$$

with ψ, the Huber function and $\hat{\sigma}_t$ an online scale estimator

Cleaning the time series

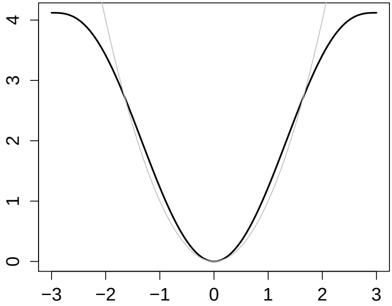
$$y_t^* = \psi\left[\frac{y_t - \hat{y}_{t|t-1}}{\hat{\sigma}_t}\right]\hat{\sigma}_t + \hat{y}_{t|t-1}$$



Scale estimator

$$\hat{\sigma}_t^2 = 0.1\,\rho\left(\frac{y_t - \hat{y}^*_{t|t-1}}{\hat{\sigma}_{t-1}}\right)\hat{\sigma}^2_{t-1} + 0.9\hat{\sigma}^2_{t-1}$$

with ρ Biweight function:



Estimating parameters

Parameter vector $\boldsymbol{\theta}$: α, β, γ, ϕ.

ℓ_0, b_0, $s_{-m+1}, \dots, s_0$ are estimated in a short startup period.

Estimating parameters

Hyndman and Khandakar (2008): maximum likelihood (MLE):

- additive error model

$$\begin{aligned} y_t &= \hat{y}_{t|t-1} + \epsilon_t \\ \ell_t &= \hat{y}_{t|t-1} + \alpha\epsilon_t \end{aligned}$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} -\frac{T}{2}\log\left(\frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})\right)^2\right)$$

- multiplicative error model

$$\begin{aligned} y_t &= \hat{y}_{t|t-1}(1 + \epsilon_t) \\ \ell_t &= \hat{y}_{t|t-1}(1 + \alpha\epsilon_t). \end{aligned}$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} -\frac{T}{2}\log\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta})}\right)^2\right) - \sum_{t=1}^{T}\log\left|\hat{y}_{t|t-1}(\boldsymbol{\theta})\right|$$

Robust estimation of parameters

Replace mean sum of squares by τ^2-scale

- for additive error model:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \operatorname{roblik}_{\mathsf{A}}(\boldsymbol{\theta})$$

$$\operatorname{roblik}(\boldsymbol{\theta}) = -\frac{T}{2} \log \left(\frac{s_T^2(\boldsymbol{\theta})}{T} \sum_{t=1}^{T} \rho \left(\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{s_T(\boldsymbol{\theta})} \right) \right)$$

$$s_T(\boldsymbol{\theta}) = 1.4826 \ \underset{t}{\operatorname{med}} |y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})|$$

- for multiplicative error model:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \boxed{\frac{s_T^2(\boldsymbol{\theta})}{T} \sum_{t=1}^{T} \rho\left(\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta}) s_T(\boldsymbol{\theta})}\right)}$$

$$\operatorname{roblik}(\boldsymbol{\theta}) = -\frac{T}{2} \log\left(\frac{s_T^2(\boldsymbol{\theta})}{T} \sum_{t=1}^{T} \rho\left(\frac{y_t - \hat{y}^*_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta}) s_T(\boldsymbol{\theta})}\right)\right) - \sum_{t=1}^{T} \log\left|\hat{y}_{t|t-1}(\boldsymbol{\theta})\right|$$

$$s_T(\boldsymbol{\theta}) = 1.4826 \operatorname*{med}_t \left|\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta})}\right|$$

Model selection

$$\text{robust AICC} = -2\,\text{rob lik} + 2\frac{pT}{T-p-1}$$

p: number of parameters to be estimated

Section 4

Simulations

Simulations

Generate time series of length $T+1=61$ from models of different types

Add contamination

Use the non-robust method (C) and the robust method (R) to predict value at $T+1$ from first T observations.

Compute over 500 simulation runs

$$\text{RMSE} = \sqrt{\frac{1}{500}\sum_{i=1}^{500}(y_{61,i} - \hat{y}_{61|60,i})^2}$$

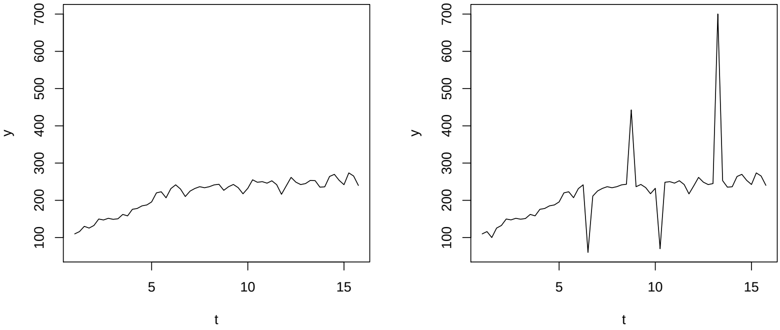


Figure: Left: clean simulation, right: outlier contaminated

4 seasons, 15 years $\rightarrow T = 60$

RMSE: Known model, unknown parameters

generating model	no outliers C	no outliers R	outliers C	outliers R
ANN	5	4.98	8.5	5.22
ANA	5.39	5.39	12.44	5.49
AAN	4.94	4.98	12.6	5.65
AA_dN	5.38	5.48	12.56	5.41
MNN	5.29	5.27	14.43	5.37
MNA	5.02	5.16	9.44	5.78
MAN	17.15	16.99	55.82	17.5
MAA	17.27	17.72	46.12	18.1
MA_dN	7.78	7.71	22.59	8.67
MNM	5.03	5.15	12.54	5.78
MAM	16.05	17.22	44.06	19.25

RMSE: Unknown model, unknown parameters

generating model	no outliers C	no outliers R	outliers C	outliers R
ANN	5.18	5.38	12.55	5.7
ANA	5.51	5.58	14.12	5.56
AAN	5.1	5.46	14.77	6.09
AAA	5.56	5.75	19.53	6.29
AA_dN	5.72	5.91	18.36	6.03
MNN	5.46	5.68	19.39	5.59
MNA	5.18	5.32	10.67	5.94
MAN	17.42	17.87	54.06	17.99
MAA	17.46	17.69	46.58	18.41
MA_dN	7.95	8.16	23.61	9.07
MNM	5.03	5.44	10.2	5.83
MAM	16.08	17.01	42.3	18.22

slightly larger error

Section 5

ROBETS on Real Data

Real data

- 3003 time series from M3-competition.
- Yearly, quarterly and monthly.
- Microeconometrics, macroeconometrics, demographics, finance and industry.
- Length of series: from 16 to 144.

Length out-of-sample period: for yearly: 6, quarterly 8 and monthly 18

Median Absolute Percentage Error

For every $h = 1, \ldots, 18$:

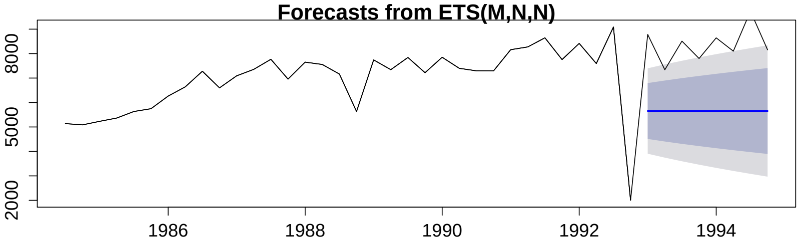
$$\text{MedAPE}_h = 100\% \text{ median} \left| \frac{y_{t_i+h,i} - \hat{y}_{t_i+h|t_i,i}}{y_{t_i+h,i}} \right| .$$

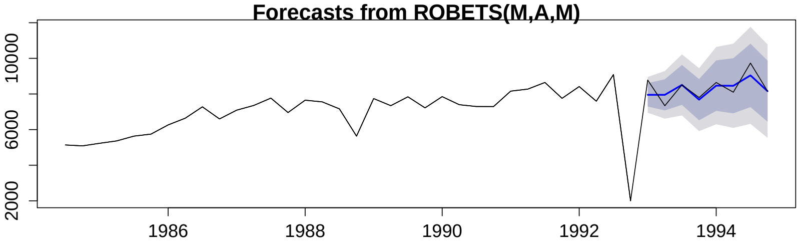
t_i: length of estimation period of time series i

Method	Forecasting horizon h					
	1	2	3	6	12	18
`ets`	3.0	3.8	4.6	6.6	7.0	9.8
`robets`	3.0	3.8	4.7	7.0	7.0	10.1

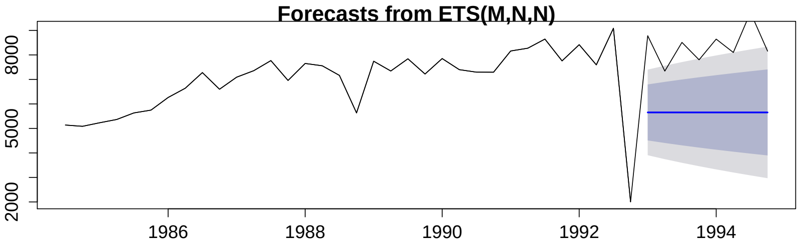
Table: The median absolute percentage error for all data.

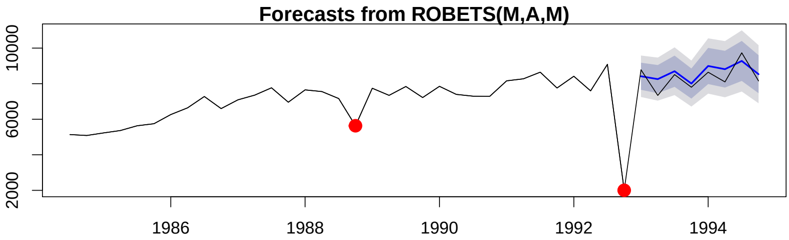
Sometimes `ets` is better, sometimes `robets` is better





A quarterly microeconometric time series where `robets` has better out-of-sample forecasts





Outlier detection.

Computing time

4 seasons, including model selection.

time series length	non-robust method	robets
25	5	7
50	7	10
75	9	15
100	9	20
200	14	35

Table: Average computation time in milliseconds.

Exercise

Data: `http://stats.stackexchange.com/questions/146098`

- Type in the time series: `values <- c(27, 27, 7, 24, 39, 40, 24, 45, 36, 37, 31, 47, 16, 24, 6, 21, 35, 36, 21, 40, 32, 33, 27, 42, 14, 21, 5, 19, 31, 32, 19, 36, 29, 29, 24, 42, 15, 24, 21)`
- Create a time series object: `valuests <- ts(values, freq = 12)`. Plot the series. Are there outliers in the time series?
- Install and load the `forecast` and the `robets` package
- Estimate two models: `mod1 <- ets(valuests)`

  and `mod2 <- robets(valuests)`and look at the output.
- Make a forecast with both models: `plot(forecast(mod, h = 24))`
- Compare both forecasts, and use `plotOutliers(mod2)` to detect outliers.

References

Crevits R. and Croux C. (2016), “Forecasting Time Series with Robust Exponential Smoothing with Damped Trend and Seasonal Components,” `https://rcrevits.wordpress.com/research`

[1] Gelper, S., Fried, R. & Croux, C. (2010). Robust Forecasting with Exponential and Holt-Winters Smoothing *Journal of Forecasting*, **29**, 285–300.

[2] Hyndman, R. J. & Khandakar, Y. (2008). Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software*, **27**, 3.