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
1. Univariate
2. Multivariate
3. Big Data
Section 1

Univariate
One time series

Sales growth of beer in a store: weekly data

<table>
<thead>
<tr>
<th>Week</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>0.28</td>
</tr>
<tr>
<td>Week 2</td>
<td>-0.39</td>
</tr>
<tr>
<td>Week 3</td>
<td>-0.01</td>
</tr>
<tr>
<td>Week 4</td>
<td>0.15</td>
</tr>
<tr>
<td>Week 5</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 75</td>
<td>-0.87</td>
</tr>
<tr>
<td>Week 76</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Time Series Plot

sales growth of Beer

week

sales growth

-0.5

0.0

0.5

0

20

40

60

0

20

40

60

Univariate
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 \( \gamma \) by using ordinary least squares.
R-code and Output

```r
> 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 $\gamma$ (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 using the formula

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

Can we also forecast at horizon 2? YES

$$\hat{y}_{T+2} = \hat{c} + \hat{\gamma} \hat{y}_{T+1}$$
$T = 76$, forecasts at horizon 1, 2, 3, and 4

sales growth of Beer + forecasts
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 $y_{3,t}$. 
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

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

Forecast at horizon 2 ? NO

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

Why Not? \( y_{2,T+1}, y_{3,T+2} \) not known.
Vector Autoregressive Model

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

- Estimation: OLS equation by equation.
- VAR(1) model
- \( \gamma_{ji} = \text{effect of time series } j \text{ on time series } i \)
- We have \( q \times q = 9 \) autoregressive parameters \( \gamma_{ji} \), 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 ≡ 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

→ 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 \textbf{edge} from node \( i \) to node \( j \) if
  \[
  \hat{\gamma}_{ji} \neq 0
  \]
  - the edge \textbf{width} is the size of the effect
  - the edge \textbf{color} is the sign of the effect
    (blue if positive, red if negative)
price effects on sales

17 product categories
Part II

Basic time series concepts
Outline

1. Stationarity
2. Autocorrelation
3. Differencing
4. AR and MA Models
5. MA-infinity representation
Section 1

Stationarity
Example: souvenirs sold (in dollars)
Frequency: monthly, sample size: \( T = 84 \)
Stationarity

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

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

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 $k$. 
The first 8 autocorrelations are significantly different from zero. There is strong *persistence* 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 $\Delta$ is defined as

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

Linear trends can be eliminated by applying $\Delta$ 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 \( \Delta Y_t \):
Correlograms of $Y_t$ and $\Delta Y_t$: 

random walk with drift

in differences
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 that one loses $s$ observations when differencing.
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 = $y$: 

![Graph showing trend and seasonally differenced series]
Correlogram:

Series y

ACF

Lag

Differencing
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 $\sigma^2$ 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 $\theta$ are unknown parameters.
The autocorrelations of an MA(1) are given by

1. $\rho_0 = 1$
2. $\rho_1 = Corr(Y_t, Y_{t-1}) = -\frac{\theta}{(1+\theta^2)}$
3. $\rho_2 = 0$
4. $\rho_3 = 0$
5. $\vdots$
The correlogram can be used to help us to *specify* an MA(1) process:
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.

```R
cacf(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 $\phi$ are unknown parameters.

The autocorrelations of an AR(1) are given by

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

AR and MA Models
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
If $Y_t$ is a stationary process, then it can be written as an MA($\infty$):

$$Y_t = c + u_t + \sum_{k=1}^{+\infty} \theta_k u_{t-k} \quad \text{for any } t.$$
Example: AR(1)

\[ Y_t = a + \phi Y_{t-1} + u_t \]

\[ = a + \phi (a + \phi Y_{t-2} + u_{t-1}) + u_t \]

\[ = a(1 + \phi) + u_t + \phi u_{t-1} + \phi^2 Y_{t-2} \]

\[ = a(1 + \phi + \phi^2 + \ldots) + u_t + \phi u_{t-1} + \phi^2 u_{t-2} + \ldots \]

Recognize an MA(\(\infty\)) 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 $\theta_k$ (see MA$(\infty)$ representation).

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

\[
\begin{align*}
\text{Impulse} & \quad \downarrow \\
\text{If } u_t \text{ increases with 1 unit, then} & \\
\text{Response} & \quad \downarrow \\
Y_t \text{ increases with 1} & \\
Y_{t+1} \text{ increases with } \phi & \\
Y_{t+2} \text{ increases with } \phi^2 & \\
Y_{t+k} \text{ increases with } \phi^k. & 
\end{align*}
\]
Part III

Introduction to dynamic models
Outline

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

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)  -45.51     13.94  -3.26  0.0023 **
cons          0.03      0.003  10.33  1.02e-12 ***
---
Signif. codes:  ** 0.001 *** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.40 on 39 degrees of freedom
Multiple R-squared: 0.73, Adjusted R-squared: 0.73
F-statistic: 106 on 1 and 39 DF, p-value: 1.02e-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 \textit{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

\[ \Delta \log(X_t) = \log(X_t) - \log(X_{t-1}) = \log\left(\frac{X_t}{X_{t-1}}\right) \approx \frac{X_t - X_{t-1}}{X_{t-1}}. \]

We get relative differences, or percentagewise increments.

\[ Y = \Delta \log(\text{Industrial Production}) = \text{Growth} \]
\[ X = \Delta \log(\text{Consumption Prices}) = \text{Inflation} \]
Running a regression

```r
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:

![Residual Correlogram]

OLS remains consistent to estimate $\beta_0$ and $\beta_1$, 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

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>33</td>
<td>0.034545</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>35</td>
<td>0.066910</td>
<td>-2</td>
<td>-0.032365</td>
<td>15.459</td>
<td>1.831e-05 ***</td>
</tr>
</tbody>
</table>

---

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

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

- 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

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

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

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

No correlation at *leads and lags* between components of $\mathbf{u}_t$; 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

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

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC(n)</td>
<td>-1.551591e+01</td>
<td>-1.621211e+01</td>
<td>-1.679070e+01</td>
<td>-1.668189e+01</td>
</tr>
<tr>
<td>HQ(n)</td>
<td>-1.542626e+01</td>
<td>-1.606269e+01</td>
<td>-1.658152e+01</td>
<td>-1.641294e+01</td>
</tr>
<tr>
<td>SC(n)</td>
<td>-1.523567e+01</td>
<td>-1.574504e+01</td>
<td>-1.613681e+01</td>
<td>-1.584117e+01</td>
</tr>
<tr>
<td>FPE(n)</td>
<td>1.828564e-07</td>
<td>9.160015e-08</td>
<td>5.193928e-08</td>
<td>5.911318e-08</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>AIC(n)</td>
<td>-1.659127e+01</td>
<td>-1.649105e+01</td>
<td>-1.637611e+01</td>
<td>-1.651495e+01</td>
</tr>
<tr>
<td>HQ(n)</td>
<td>-1.626255e+01</td>
<td>-1.610256e+01</td>
<td>-1.592786e+01</td>
<td>-1.600692e+01</td>
</tr>
<tr>
<td>SC(n)</td>
<td>-1.556373e+01</td>
<td>-1.527668e+01</td>
<td>-1.497491e+01</td>
<td>-1.492692e+01</td>
</tr>
<tr>
<td>FPE(n)</td>
<td>6.692637e-08</td>
<td>7.785562e-08</td>
<td>9.410207e-08</td>
<td>9.110841e-08</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC(n)</td>
<td>-1.629013e+01</td>
<td>-1.645541e+01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ(n)</td>
<td>-1.572234e+01</td>
<td>-1.582785e+01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC(n)</td>
<td>-1.451528e+01</td>
<td>-1.449373e+01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FPE(n)</td>
<td>1.326417e-07</td>
<td>1.393602e-07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VAR example: inflation-growth

> summary(mymodel)

| 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

```r
> predict(mymodel, n.ahead=6)
$infl

<table>
<thead>
<tr>
<th></th>
<th>fcst</th>
<th>lower</th>
<th>upper</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02435690</td>
<td>0.007403536</td>
<td>0.041310271</td>
<td>0.01695337</td>
</tr>
<tr>
<td>2</td>
<td>-0.01203683</td>
<td>-0.030790030</td>
<td>0.006716372</td>
<td>0.01875320</td>
</tr>
<tr>
<td>3</td>
<td>0.01626860</td>
<td>-0.003676775</td>
<td>0.036213970</td>
<td>0.01994537</td>
</tr>
<tr>
<td>4</td>
<td>-0.01281687</td>
<td>-0.034638813</td>
<td>0.009005066</td>
<td>0.02182194</td>
</tr>
<tr>
<td>5</td>
<td>0.02049182</td>
<td>-0.002592846</td>
<td>0.043576485</td>
<td>0.02308467</td>
</tr>
<tr>
<td>6</td>
<td>-0.01096528</td>
<td>-0.034561170</td>
<td>0.012630615</td>
<td>0.02359589</td>
</tr>
</tbody>
</table>

$growth

<table>
<thead>
<tr>
<th></th>
<th>fcst</th>
<th>lower</th>
<th>upper</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.013351737</td>
<td>-0.0380387925</td>
<td>0.064742266</td>
<td>0.05139053</td>
</tr>
<tr>
<td>2</td>
<td>-0.056308643</td>
<td>-0.1167153773</td>
<td>0.00498092</td>
<td>0.06040673</td>
</tr>
<tr>
<td>3</td>
<td>0.062663063</td>
<td>-0.007228443</td>
<td>0.126048970</td>
<td>0.06338591</td>
</tr>
<tr>
<td>4</td>
<td>-0.007140285</td>
<td>-0.0727913007</td>
<td>0.058510731</td>
<td>0.06565102</td>
</tr>
<tr>
<td>5</td>
<td>0.021067459</td>
<td>-0.0549650606</td>
<td>0.097099979</td>
<td>0.07603252</td>
</tr>
<tr>
<td>6</td>
<td>-0.045996773</td>
<td>-0.1244454309</td>
<td>0.032451885</td>
<td>0.07844866</td>
</tr>
</tbody>
</table>
Example: impulse-response functions

Impulse Response from infl

95 % Bootstrap CI, 100 runs

Values range from -1.5 to 1.5.
Part IV

Sparse Cointegration
Outline

1. Introduction
2. Penalized Estimation
3. Forecasting Applications
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 $\delta_t$ is stationary.

- $\beta_{11}y_{1,t} + \beta_{21}y_{2,t} = \delta_t$ : Cointegration Equation
- $\beta = (\beta_{11}, \beta_{21})'$ : 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 $y_t$ be a $q$-dimensional multivariate time series, $I(1)$.

Vector Error Correcting Representation:

$$\Delta y_t = \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \Pi y_{t-1} + \epsilon_t, \quad t = 1, \ldots, T$$

where

- $\epsilon_t$ follows $N_q(0, \Sigma)$, denote $\Omega = \Sigma^{-1}$
- $\Gamma_1, \ldots, \Gamma_{p-1}$ $q \times q$ matrices of short-run effects
- $\Pi$ $q \times q$ matrix.
Vector Error Correcting Model (cont.)

\[ \Delta y_t = \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \Pi y_{t-1} + \epsilon_t. \quad t = 1, \ldots, T \]

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

Then, \( \beta' y_t \) stationary

- \( y_t \) cointegrated with cointegration rank \( r \)
- \( \beta \): cointegrating vectors
- \( \alpha \): adjustment coefficients.
Rewrite the VECM in matrix notation:

\[ \Delta Y = \Delta Y_L \Gamma + Y \Pi' + E, \]

where

- \( \Delta Y = (\Delta y_{p+1}, \ldots, \Delta y_T)' \)
- \( \Delta Y_L = (\Delta X_{p+1}, \ldots, \Delta X_T)' \) with \( \Delta X_t = (\Delta y'_{t-1}, \ldots, \Delta y'_{t-p+1})' \)
- \( Y = (y_p, \ldots, y_{T-1})' \)
- \( \Gamma = (\Gamma_1, \ldots, \Gamma_{p-1})' \)
- \( E = (\varepsilon_{p+1}, \ldots, \varepsilon_T)' \).
Maximum likelihood estimation (Johansen, 1996) (cont.)

Negative log likelihood

\[ \mathcal{L}(\Gamma, \Pi, \Omega) = \frac{1}{T} \text{tr} \left( \left( \Delta Y - \Delta Y_L \Gamma - Y \Pi' \right) \Omega \left( \Delta Y - \Delta Y_L \Gamma - Y \Pi' \right)' \right) - \log |\Omega|. \]

Maximum likelihood estimator:

\[ \left( \hat{\Gamma}, \hat{\Pi}, \hat{\Omega} \right) = \arg \min_{\Gamma, \Pi, \Omega} \mathcal{L}(\Gamma, \Pi, \Omega), \]

subject to \( \Pi = \alpha \beta' \).

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

$$y_i = x'_i \beta + e_i,$$

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

Penalized estimate of $$\beta$$:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x'_i \beta)^2 + n\lambda P(\beta),$$

with $$\lambda$$ a penalty parameter and $$P(\beta)$$ a penalty function.
Penalized Regression (cont.)

Choices of penalty functions:

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

Penalized negative log likelihood

\[ \mathcal{L}_P(\Gamma, \Pi, \Omega) = \frac{1}{T} \text{tr} \left( (\Delta Y - \Delta Y_L \Gamma - Y \Pi') \Omega (\Delta Y - \Delta Y_L \Gamma - Y \Pi')' \right) + \log |\Omega| \]
\[ + \lambda_1 P_1(\beta) + \lambda_2 P_2(\Gamma) + \lambda_3 P_3(\Omega), \]

with \( P_1, P_2 \) and \( P_3 \) three penalty functions.

Penalized maximum likelihood estimator:

\[ (\hat{\Gamma}, \hat{\Pi}, \hat{\Omega}) = \arg\min_{\Gamma, \Pi, \Omega} \mathcal{L}_P(\Gamma, \Pi, \Omega), \]

subject to \( \Pi = \alpha / \beta' \).
Algorithm

Iterative procedure:

1. Solve for \( \Pi \) conditional on \( \Gamma, \Omega \)
2. Solve for \( \Gamma \) conditional on \( \Pi, \Omega \)
3. Solve for \( \Omega \) conditional on \( \Gamma, \Pi \)
1. Solving for $\Pi$ conditional on $\Gamma, \Omega$

Solve

$$(\hat{\alpha}, \hat{\beta})|\Gamma, \Omega = \arg\min_{\alpha, \beta} \frac{1}{T} \text{tr} \left( (G - Y\beta'\alpha')\Omega(G - Y\beta'\alpha')' \right) + \lambda_1 P_1(\beta).$$

subject to $\alpha'\Omega\alpha = I_r$

with

- $G = \Delta Y - \Delta Y_L \Gamma$

→ Penalized reduced rank regression (e.g. Chen and Huang, 2012)
1.1 Solving for $\alpha$ conditional on $\Gamma, \Omega, \beta$

Solve

$$\hat{\alpha}|_{\Gamma, \Omega, \beta} = \arg\min_{\alpha} \frac{1}{T} \text{tr}\left( (G - B\alpha')\Omega(G - B\alpha')' \right)$$

subject to $\alpha'\Omega\alpha = I_r$,

with

- $G = \Delta Y - \Delta Y_L\Gamma$
- $B = Y\beta$

→ Weighted Procrustes problem
1.2 Solving for $\beta$ conditional on $\Gamma$, $\Omega$, $\alpha$

Solve

$$\hat{\beta}|_{\Gamma, \Omega, \alpha} = \arg\min_{\beta} \frac{1}{T} \text{tr} \left( (R - Y\beta)(R - Y\beta)' \right) + \lambda_1 P_1(\beta).$$

with

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

→ Penalized multivariate regression
Choice of penalty function

Our choice:

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

Other penalty functions are possible:

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

- \( \ldots \)
2. Solving for $\Gamma$ conditional on $\Pi$, $\Omega$

Solve

$$\hat{\Gamma}|\Pi, \Omega = \arg\min_{\Gamma} \frac{1}{T} \text{tr}\left( (D - \Delta Y_L \Gamma) \Omega (D - \Delta Y_L \Gamma)' \right) + \lambda_2 P_2(\Gamma).$$

with

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

→ Penalized multivariate regression
Solve

\[ \hat{\Omega} | \Gamma, \Pi = \arg \min_{\Omega} \frac{1}{T} \text{tr} \left( (D - \Delta Y L \Gamma) \Omega (D - \Delta Y L \Gamma)' \right) - \log |\Omega| + \lambda_3 P_3(\Omega). \]

with

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

→ Penalized inverse covariance matrix estimation (Friedman et al., 2008)
Selection of tuning parameters

$\lambda_1$ and $\lambda_2$: Time series cross-validation (Hyndman, 2014)

$\lambda_3$: Bayesian Information Criterion
Time series cross-validation

Denote the response by $z_t$

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

1. Fit model to $z_1, \ldots, z_t$
2. Compute $\hat{e}_{t+1} = z_{t+1} - \hat{z}_{t+1}$

Select value of tuning parameter that minimizes

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

with

- $\hat{e}^{(i)}_t$ the $i^{th}$ component of $\hat{e}_t$
- $\hat{\sigma}(i) = \text{sd}(z^{(i)}_t)$.
Determination of cointegration rank

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

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

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

with

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

→ 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(\(\Pi\))
- Low computational cost
Simulation Study

VECM(1) with dimension $q$:

$$\Delta y_t = \alpha \beta' y_{t-1} + \Gamma_1 \Delta y_{t-1} + e_t, \quad (t = 1, \ldots, T),$$

where $e_t$ follows $N_q(0, I_q)$.

**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

![Graph showing the average angle between estimated and true cointegration space for different adjustment coefficients. The graph compares ML and PML (sparse) methods.](image)
Non-sparse Low-dimensional design

Adjustment coefficients $a$

Average angle

ML

PML (sparse)
Sparse High-dimensional design

![Graph showing average angle vs. adjustment coefficients a for ML and PML (sparse) methods.](graph.png)
Non-sparse High-dimensional design

![Graph showing the relationship between average angle and adjustment coefficients a for ML and PML (sparse).]
Section 3

Forecasting Applications
Rolling window forecast with window size $S$

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

\[
\hat{\Delta y}_{t+h} = \sum_{i=1}^{p-1} \hat{\Gamma}_i \Delta y_{t+1-i} + \hat{\Pi} y_t,
\]

for forecast horizon $h$.

Obtain $h$-step-ahead multivariate forecast errors

\[
\hat{e}_{t+h} = \Delta y_{t+h} - \hat{\Delta y}_{t+h}.
\]
Forecast error measures

Multivariate Mean Absolute Forecast Error:

\[
\text{MMAFE} = \frac{1}{T - h - S + 1} \sum_{t=S}^{T-h} \frac{1}{q} \sum_{i=1}^{q} \left| \frac{\Delta y_{t+h}^{(i)} - \hat{\Delta y}_{t+h}^{(i)}}{\hat{\sigma}(i)} \right|
\]

where \( \hat{\sigma}(i) \) 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

Horizon h=1

Rolling window size

ML
PML (sparse)
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

- **Total consumption**
  - 2000: 9.0
  - 2005: 9.1
  - 2010: 9.2
  - 2015: 9.3

- **Household equipm.**
  - 2000: 12.0
  - 2005: 12.4

- **Household appliances**
  - 2000: 10.3
  - 2005: 10.5
  - 2010: 10.7
  - 2015: 10.9

- **Recreational goods**
  - 2000: 11.5
  - 2005: 12.0
  - 2010: 12.5
  - 2015: 13.0

- **Video & Audio**
  - 2000: 10.5
  - 2005: 11.5

- **Photographic equipm.**
  - 2000: 7.5
  - 2005: 8.0
  - 2010: 8.5

- **Info processing equipm.**
  - 2000: 9.0
  - 2005: 10.0
  - 2010: 11.0
  - 2015: 12.0

- **Medical equipm.**
  - 2000: 9.4
  - 2005: 9.8
  - 2010: 10.2
  - 2015: 10.7

- **Telephone equipm.**
  - 2000: 7.5
  - 2005: 8.5
  - 2010: 9.5
## Multivariate Mean Absolute Forecast Error

<table>
<thead>
<tr>
<th>Method</th>
<th>$h = 1$</th>
<th>$h = 3$</th>
<th>$h = 6$</th>
<th>$h = 12$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cointegration</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PML (sparse)</td>
<td>0.79</td>
<td><strong>0.62</strong></td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td>ML</td>
<td>0.74</td>
<td>0.78</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>Factor Model</td>
<td><strong>0.66</strong></td>
<td>0.66</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>No Cointegration</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PML (sparse)</td>
<td>0.94</td>
<td>0.67</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>ML</td>
<td>5.40</td>
<td>4.81</td>
<td>4.84</td>
<td>5.22</td>
</tr>
<tr>
<td>Factor Model</td>
<td>0.72</td>
<td>0.75</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td>Bayesian</td>
<td>0.69</td>
<td>0.71</td>
<td>0.74</td>
<td>0.72</td>
</tr>
<tr>
<td>Bayesian Reduced Rank</td>
<td>0.69</td>
<td>0.71</td>
<td>0.74</td>
<td>0.72</td>
</tr>
</tbody>
</table>
References


Part V

Robust Exponential Smoothing
Outline

1. Introduction
2. Exponential Smoothing
3. Robust approach
4. Simulations
5. ROBETS on Real Data
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 Multiplicative 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))

Forecasts from ETS(M,Ad,M)

0e+00 4e+05 8e+05
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

Forecasts from ETS(M,N,N)

Forecasts from ROBETS(M,A,M)
Section 2

Exponential Smoothing
Simple exponential smoothing

\[ \hat{y}_{t+h|t} = \ell_t \]
\[ \ell_t = \ell_{t-1} + \alpha (y_t - \ell_{t-1}) \]

\( y_t \): univariate time series
\( \hat{y}_{t+h|t} \): \( h \)-step ahead prediction
\( \ell_t \): level
\( \alpha \): 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: $\text{AA}_d\text{N} / \text{MA}_d\text{N}$

\[
\hat{y}_{t+h|t} = \ell_t + \sum_{j=1}^{h} \phi_j b_t
\]

\[
\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})
\]

\[
b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)\phi b_{t-1}
\]

$\ell_t$: the level

$b_t$: the (damped) trend

$\alpha, \beta$: smoothing parameters

$\phi$: damping parameter
Births per 10,000 of 23 year old women, USA

No trend ($\phi = 0$, ANN-model), Full trend ($\phi = 1$, AAN-model) → damped trend ($AA_dN$)
Example 2: additive damped trend and multiplicative seasonality (MA$_d$M)

\[
\hat{y}_{t+h|t} = (\ell_t + \sum_{j=1}^{h} \phi^j b_t)s_{t+h^+_m-m}
\]

\[
\begin{align*}
\ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta (\ell_t - \ell_{t-1}) + (1 - \beta) \phi b_{t-1} \\
s_t &= \gamma \frac{y_t}{\ell_{t-1} + \phi b_{t-1}} + (1 - \gamma) s_{t-m}.
\end{align*}
\]

\(\ell_t\): the level
\(b_t\): the (damped) trend
\(s_t\): the seasonal components
\(\alpha, \beta, \gamma\): smoothing parameters
\(\phi\): damping parameter
\(h^+_m = (h - 1) \mod m + 1\)
\(m\): number of seasons per period
Monthly number of unemployed persons in Australia, from February 1978 till August 1995

No trend \((\phi = 0, \text{MNM-model})\), Full trend \((\phi = 1, \text{MAM-model})\)
→ damped trend \((\text{MA}_d \text{M})\)
AICCC: selection criterion that penalizes models with a lot of parameters
Section 3

Robust approach
ETS is not robust

Model: AA_dA / MA_dA

\[
\hat{y}_{t+h|t} = \ell_t + \sum_{j=1}^{h} \phi^j b_t + s_{t-m+h_m^+}
\]

\[
\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})
\]

\[
b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)\phi b_{t-1}
\]

\[
s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1 - \gamma)s_{t-m}.
\]

\[
\downarrow
\]

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

How to make it robust?

\[
\hat{y}_{t+h|t} = \ell_t + \sum_{j=1}^{h} \phi^j b_t + s_{t-m+h_m^+}
\]

\[
\ell_t = \alpha(y^*_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})
\]

\[
b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)\phi b_{t-1}
\]

\[
s_t = \gamma(y^*_t - \ell_{t-1} - \phi b_{t-1}) + (1 - \gamma)s_{t-m}.
\]

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 \( \psi \), the Huber function and \( \hat{\sigma}_t \) an online scale estimator
Cleaning the time series

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

Huber \( \psi \)
Scale estimator

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

with \( \rho \) Biweight function:
Estimating parameters

Parameter vector $\theta$: $\alpha, \beta, \gamma, \phi$.

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


- additive error model

\[ y_t = \hat{y}_{t|t-1} + \epsilon_t \]
\[ \ell_t = \hat{y}_{t|t-1} + \alpha \epsilon_t \]

\[ \hat{\theta} = \arg\max_{\theta} - \frac{T}{2} \log \left( \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_{t|t-1}(\theta))^2 \right) \]

- multiplicative error model

\[ y_t = \hat{y}_{t|t-1}(1 + \epsilon_t) \]
\[ \ell_t = \hat{y}_{t|t-1}(1 + \alpha \epsilon_t). \]

\[ \hat{\theta} = \arg\max_{\theta} - \frac{T}{2} \log \left( \frac{1}{T} \sum_{t=1}^{T} \left( \frac{y_t - \hat{y}_{t|t-1}(\theta)}{\hat{y}_{t|t-1}(\theta)} \right)^2 \right) - \sum_{t=1}^{T} \log \left| \hat{y}_{t|t-1}(\theta) \right| \]
Robust approach

Robust estimation of parameters

Replace mean sum of squares by $\tau^2$-scale

- for additive error model:

$$\hat{\theta} = \arg\max_\theta \text{roblik}_A(\theta)$$

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

$$s_T(\theta) = 1.4826 \med_{t} |y_t - \hat{y}_{t|t-1}(\theta)|$$
for multiplicative error model:

\[
\hat{\theta} = \text{argmin}_{\theta} \frac{s_T^2(\theta)}{T} \sum_{t=1}^{T} \rho \left( \frac{y_t - \hat{y}_{t|t-1}(\theta)}{\hat{y}_{t|t-1}(\theta)s_T(\theta)} \right)
\]

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

\[
s_T(\theta) = 1.4826 \text{ med}_t \left| \frac{y_t - \hat{y}_{t|t-1}(\theta)}{\hat{y}_{t|t-1}(\theta)} \right|
\]
Model selection

\[
\text{robust AICC} = -2 \text{ rob lik} + 2 \frac{p^T}{T - p - 1}
\]

\(p\): 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

<table>
<thead>
<tr>
<th>generating model</th>
<th>no outliers</th>
<th>outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>R</td>
</tr>
<tr>
<td>ANN</td>
<td>5</td>
<td>4.98</td>
</tr>
<tr>
<td>ANA</td>
<td>5.39</td>
<td>5.39</td>
</tr>
<tr>
<td>AAN</td>
<td>4.94</td>
<td>4.98</td>
</tr>
<tr>
<td>AA\textsubscript{d}N</td>
<td>5.38</td>
<td>5.48</td>
</tr>
<tr>
<td>MNN</td>
<td>5.29</td>
<td>5.27</td>
</tr>
<tr>
<td>MNA</td>
<td>5.02</td>
<td>5.16</td>
</tr>
<tr>
<td>MAN</td>
<td>17.15</td>
<td>16.99</td>
</tr>
<tr>
<td>MAA</td>
<td>17.27</td>
<td>17.72</td>
</tr>
<tr>
<td>MA\textsubscript{d}N</td>
<td>7.78</td>
<td>7.71</td>
</tr>
<tr>
<td>MNM</td>
<td>5.03</td>
<td>5.15</td>
</tr>
<tr>
<td>MAM</td>
<td>16.05</td>
<td>17.22</td>
</tr>
</tbody>
</table>
## RMSE: Unknown model, unknown parameters

<table>
<thead>
<tr>
<th>generating model</th>
<th>no outliers</th>
<th></th>
<th>outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>R</td>
<td>C</td>
</tr>
<tr>
<td>ANN</td>
<td>5.18</td>
<td>5.38</td>
<td>12.55</td>
</tr>
<tr>
<td>ANA</td>
<td>5.51</td>
<td>5.58</td>
<td>14.12</td>
</tr>
<tr>
<td>AAN</td>
<td>5.1</td>
<td>5.46</td>
<td>14.77</td>
</tr>
<tr>
<td>AAA</td>
<td>5.56</td>
<td>5.75</td>
<td>19.53</td>
</tr>
<tr>
<td>AAdN</td>
<td>5.72</td>
<td>5.91</td>
<td>18.36</td>
</tr>
<tr>
<td>MNN</td>
<td>5.46</td>
<td>5.68</td>
<td>19.39</td>
</tr>
<tr>
<td>MNA</td>
<td>5.18</td>
<td>5.32</td>
<td>10.67</td>
</tr>
<tr>
<td>MAN</td>
<td>17.42</td>
<td>17.87</td>
<td>54.06</td>
</tr>
<tr>
<td>MAA</td>
<td>17.46</td>
<td>17.69</td>
<td>46.58</td>
</tr>
<tr>
<td>MAdN</td>
<td>7.95</td>
<td>8.16</td>
<td>23.61</td>
</tr>
<tr>
<td>MNM</td>
<td>5.03</td>
<td>5.44</td>
<td>10.2</td>
</tr>
<tr>
<td>MAM</td>
<td>16.08</td>
<td>17.01</td>
<td>42.3</td>
</tr>
</tbody>
</table>

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 $h = 1, \ldots, 18$:

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

$t_i$: length of estimation period of time series $i$
Table: The median absolute percentage error for all data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecasting horizon $h$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ets</td>
<td>3.0</td>
</tr>
<tr>
<td>robets</td>
<td>3.0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>time series length</th>
<th>non-robust method</th>
<th>robets</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>75</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>200</td>
<td>14</td>
<td>35</td>
</tr>
</tbody>
</table>

**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: `values <- 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(values, freq = 12)` and `mod2 <- robets(values)` 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

