Lecture 3: Forecasting with ARIMA models

Geert Mesters (Universitat Pompeu Fabra, Barcelona GSE and VU Amsterdam)

Motivation

Motivation

- In this set of slides we first briefly revisit the class of autoregressive integrated moving average ARIMA models
- Our sole objective is forecasting
- It turns out that ARIMA models (when treated well) are very good at forecasting

Some motivating examples

- Meese and Rogoff puzzle: when forecasting exchange rates it is really hard to outperform the random walk

$$Y_t = Y_{t-1} + \epsilon_t$$

see Rossi (2013) for an overview of the literature

- Inflation forecasting: when forecasting inflation it is really hard to outperform the integrated moving average model

$$Y_t = Y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$$

see Stock & Watson (2008) for an overview of the literature

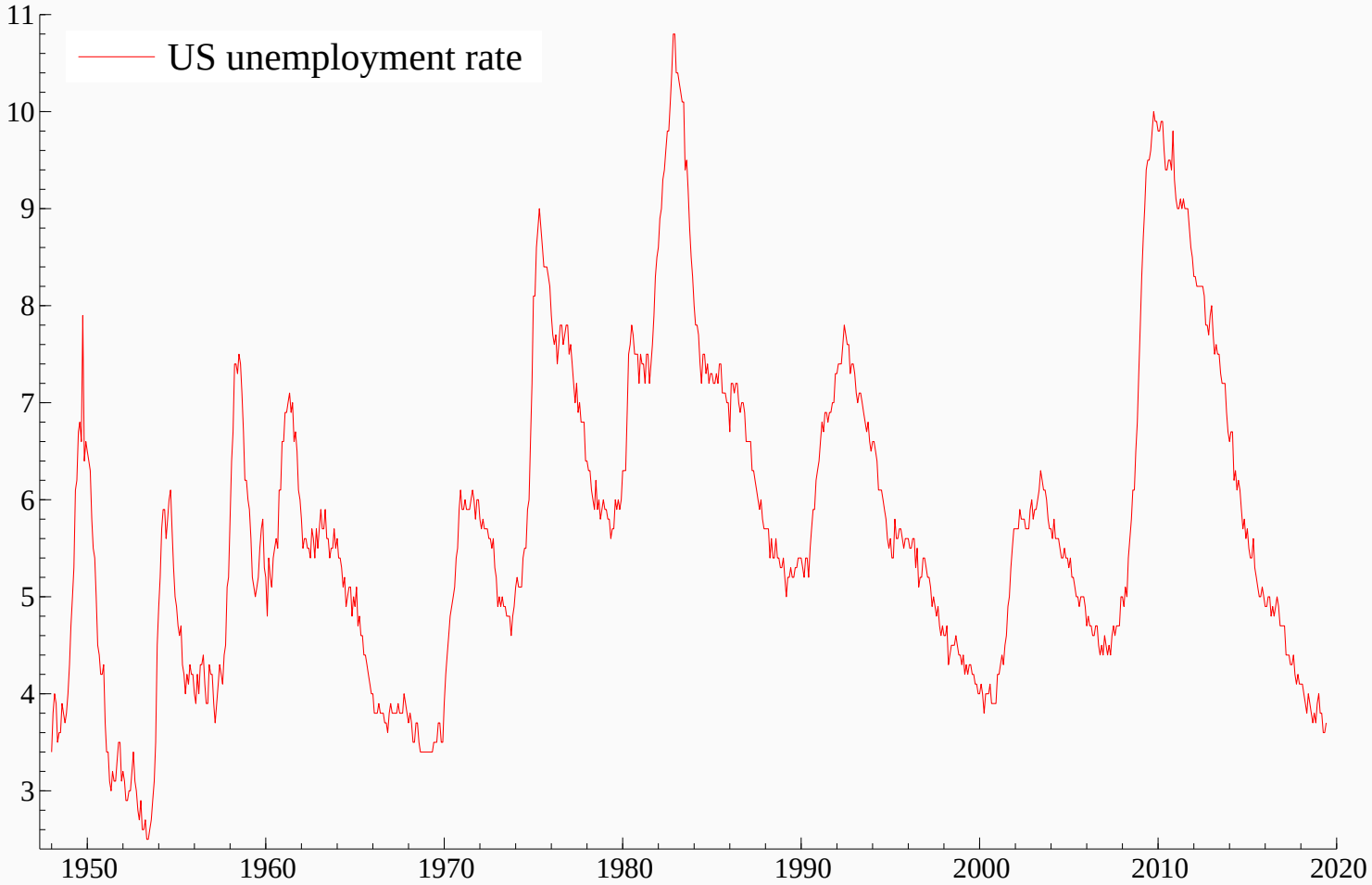
Goals

- Given these examples (there are many more) it seems useful to know how to compute forecasts using ARIMA models?
- Further, the examples imply that forecasts from different models were compared to each other, but how to compare forecasts of competing models?

Running example

US unemployment

Hatzius and Stehn (2012) of Goldman Sachs refer to the unemployment rate as their: **desert island economic indicator**



US unemployment

- Unemployment is a key economic indicator
- Forecasting unemployment is of utmost importance for policy institutions
- Also, for unemployment ARIMA models are hard to outperform
- But, there are more options when compared to inflation, exchange rates, etc

Some preliminaries

Covariance function

- For forecasting the covariances between Y_{T+h} and $Y_T, Y_{T-1}, \ldots$ are very important:

  They capture the dependence structure among the variables

Definition :

If $\{Y_t, t \in \mathbb{Z}\}$ is a process such that $\mathrm{Var}(Y_t) < \infty$ the covariance function $\gamma_Y(\cdot, \cdot)$ for $\{Y_t\}$ is defined as

$$\gamma_Y(r, s) = \mathrm{Cov}(Y_r, Y_s) = \mathrm{E}\left[(Y_r - \mathrm{E}(Y_r))(Y_s - \mathrm{E}(Y_s))\right],$$

for all $r, s \in \mathbb{Z}$ and thus defined as a function $\gamma_Y(\cdot, \cdot) : \mathbb{Z} \to \mathbb{R}$

Covariance Stationary

Definition :

The time series $\{Y_t, t \in \mathbb{Z}\}$ is covariance stationary if

1. $\mathrm{E}\left(|Y_t|^2\right) < \infty$ for all $t \in \mathbb{Z}$
2. $\mathrm{E}(Y_t) = m$ for all $t \in \mathbb{Z}$ (m constant)
3. $\gamma_Y(r, s) = \gamma_Y(r + t, s + t)$ for all $r, s, t \in \mathbb{Z}$

- this definition is also referred to as second order stationary

Autocovariance function

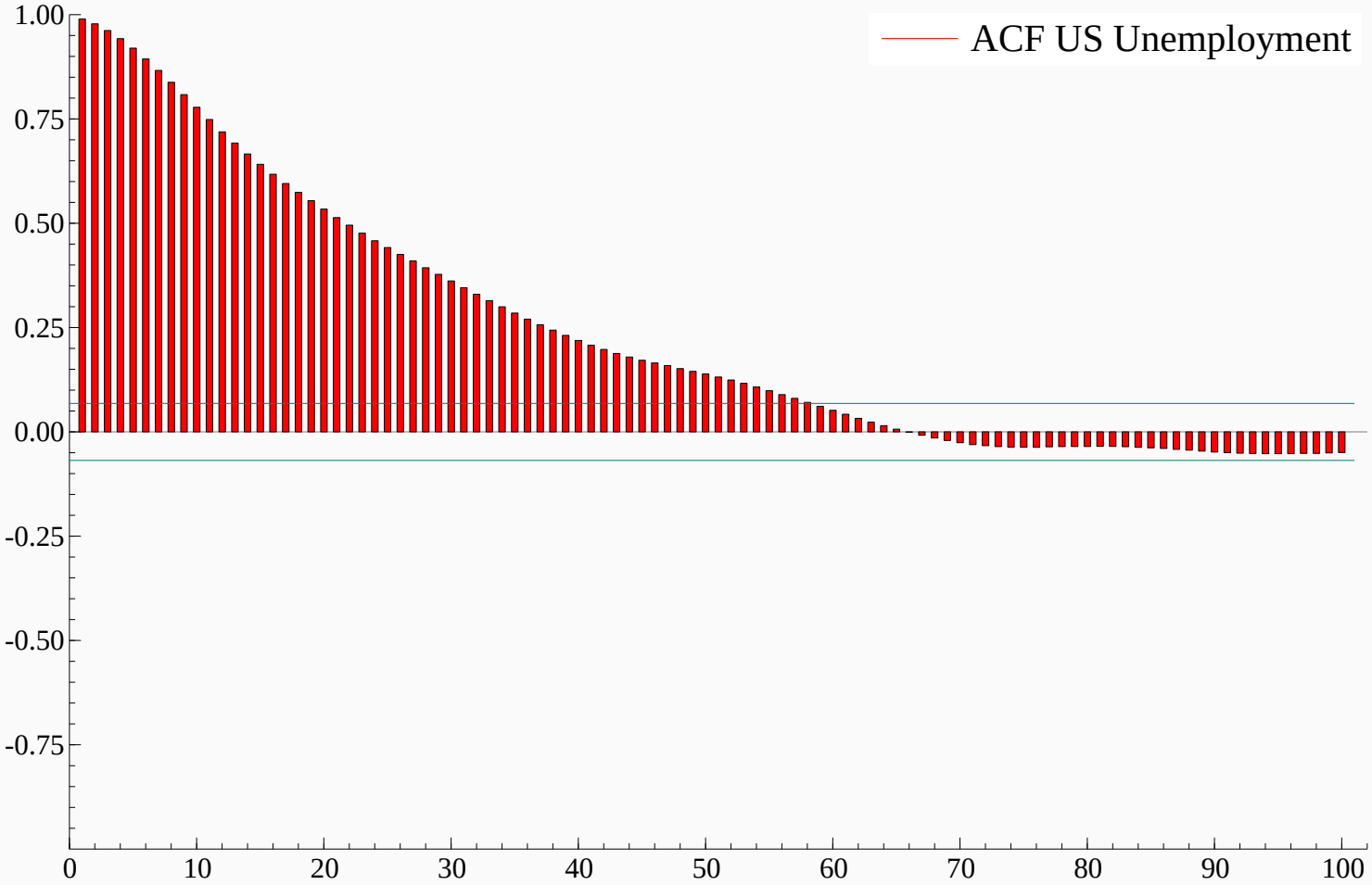
- The covariance stationary conditions imply that $\mathrm{Cov}(Y_r, Y_s)$ does not depend on r or s, but only on $|r - s|$, such that we can conveniently define the autocovariance function as

$$\gamma_Y(h) = \mathrm{Cov}(Y_{t+h}, Y_t), \qquad t, h \in \mathbb{Z}$$

- We also define the autocorrelation function (ACF) as

$$\rho_Y(h) = \gamma_Y(h)/\gamma_Y(0) = \mathrm{Corr}(Y_{t+h}, Y_t), \qquad t, h \in \mathbb{Z}$$

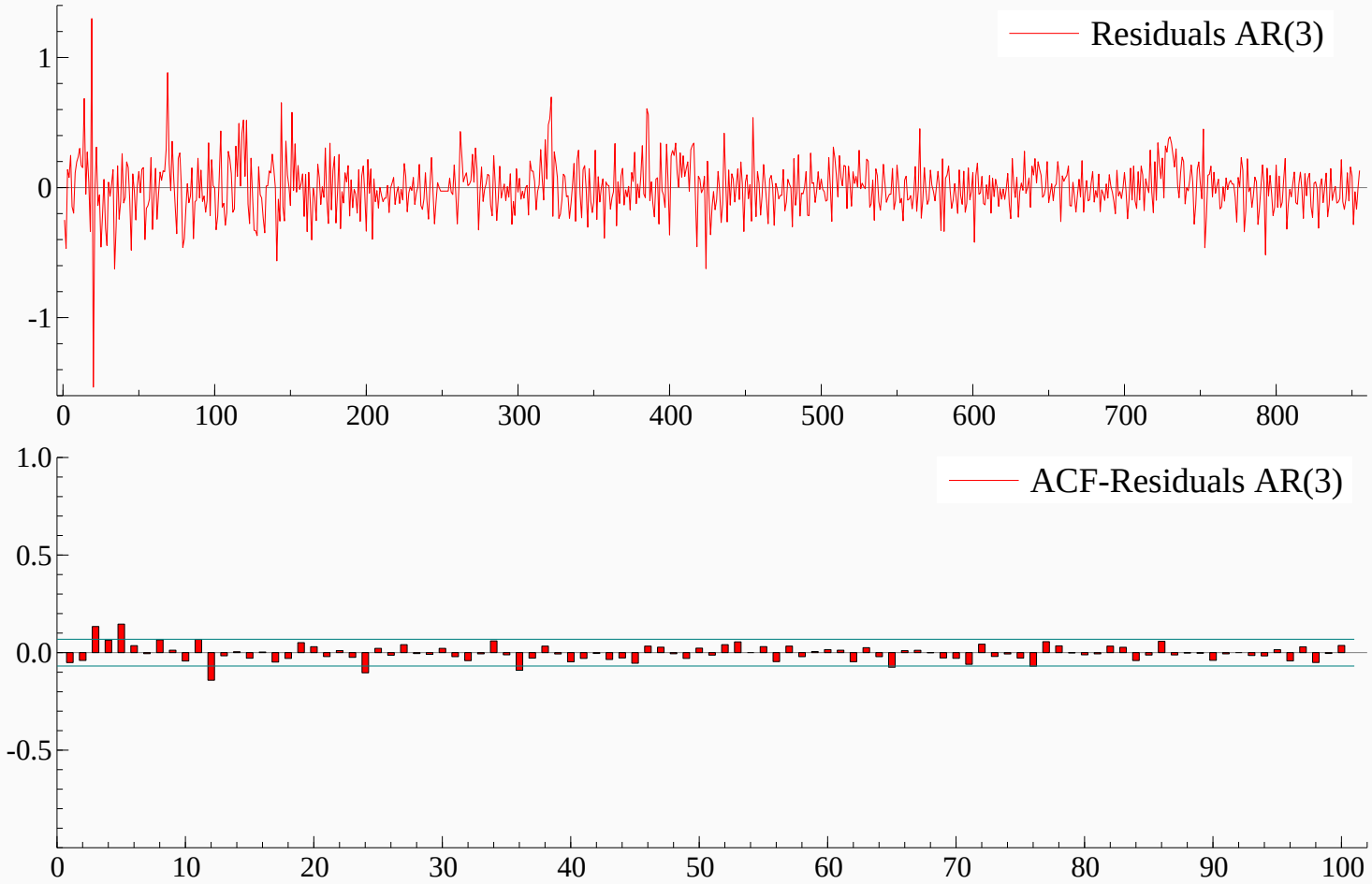
Sample autocorrelation US Unemployment



Implications

- The ACF indicates that current and past values are correlated with each other: there should be some predictive ability from past observations
- The question is how to formalize a model that is "optimal" for forecasting
- Classical regression $Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \epsilon_t$ is often insufficient for explaining all of the interesting dynamics of a time series
- To illustrate take $p = 3$ and look at the residuals

Autocorrelation AR(3) residuals



Picking up serial correlation

- We need to extend the class of prediction models to explicitly capture serial correlation
- Picking up the extra serial correlation typically really improves forecasts
- This motivates considering ARMA models, where the MA component captures the serial correlation
- If the data are non-stationary we first need to take differences which motivates ARIMA models

AR(I)MA models

ARMA(p,q)

Definition :

The process $\{Y_t, t \in \mathbb{Z}\}$ is said to be an ARMA(p,q) process if $\{Y_t\}$ is stationary and for every t and

$$Y_t - \underbrace{\phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p}}_{\text{autoregressive}} = \epsilon_t + \underbrace{\theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}}_{\text{moving average}}$$

where $\epsilon_t \sim WN(0, \sigma^2)$ and the polynomials
$(1 - \phi_1 z^1 - \phi_2 z^2 - \ldots - \phi_p z^p)$ and
$(1 + \theta_1 z^1 + \theta_2 z^2 + \ldots + \theta_q z^q)$ have no common factors.

We say $\{Y_t\}$ is an ARMA(p,q) process with mean μ if $\{Y_t - \mu\}$ is an ARMA(p,q) process.

ARMA(p,q)

More compactly we can write for $\{Y_t\}$ being an ARMA(p,q),

$$\phi(L)Y_t = \theta(L)\epsilon_t, \qquad t \in \mathbb{Z},$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p$$

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_p L^q$$

are referred to as the auto-regressive and moving-average polynomials, respectively.

Some properties

- Causality: if $\phi(z) \neq 0$ for all $|z| < 1$ then we can write

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

- Invertibility: if $\theta(z) \neq 0$ for all $|z| < 1$ the we can write

$$\epsilon_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}$$

Getting causal (invertible) coefficients

We find the coefficients $Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ by noting $\psi(z) = \chi(z)\theta(z)$ where $\chi(z) = 1/\phi(z)$. This implies that we must solve

$$\phi(z)\psi(z) = \theta(z)$$

which is

$$(1 - \phi_1 z - \ldots - \phi_p z^p)(\psi_0 + \psi_1 z + \psi_2 z^2 + \ldots) = (1 + \theta_1 z + \ldots + \theta_q z^q)$$

and we find by equating the coefficients of z^j

$$\begin{aligned} 1 &= \psi_0 \\ \theta_1 &= \psi_1 - \psi_0\phi_1 \\ \theta_2 &= \psi_2 - \psi_1\phi_1 - \psi_0\phi_2 \\ &\vdots \end{aligned}$$

Same can be done to find coefficients of $\epsilon_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}$.

Autocovariances ARMA

- To do forecasting we need the autocovariances of Y_t.

A general recipe:

Pre-multiply both sides of

$$Y_t - \phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

by Y_{t-k} and take expectations to get

$$\gamma_Y(k) - \phi_1 \gamma_Y(k-1) - \ldots - \phi_p \gamma_Y(k-p) = \sigma^2 \sum_{j=k}^{q} \theta_j \psi_{j-k}$$

for $0 \leq k < \max(p, q+1)$ and

$$\gamma_Y(k) - \phi_1 \gamma_Y(k-1) - \ldots - \phi_p \gamma_Y(k-p) = 0$$

for $k \geq \max(p, q+1)$. Solve these equations for autocovariances

Trends and differencing

In many instances the time series include some trends, for instance

$$Y_t = \underbrace{\beta_0 + \beta_1 t}_{\text{linear trend}} + X_t, \qquad Y_t = \underbrace{Y_{t-1}}_{\text{stochastic trend}} + X_t$$

Typically, differencing allows to remove the trend, define difference operator

$$\Delta Y_t = Y_t - Y_{t-1}$$

In general, we may need to difference multiple times to make the process stationary, e.g. $\Delta^d = \Delta \cdot \Delta \cdot \ldots \cdot \Delta$, e.g.

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + X_t \qquad \Delta^2 Y_t = \mu + W_t$$

where W_t is stationary.

ARIMA(p,d,q)

Definition :

The process $\{Y_t, t \in \mathbb{Z}\}$ is said to be an ARIMA(p,d,q) process if the process

$$\Delta^d Y_t$$

is an ARMA(p,q) process.

Forecasting

Conditional expectation

Suppose that we are interested in forecasting Y_{T+h} given X_T, where X_T is a vector of random variables whose realization is observed at time T. Let *a* forecast be denoted by $\hat{Y}_{T+h} = g(\boldsymbol{X}_T)$ for some function $g(\cdot)$. We evaluate the quality of this forecast using the mean-squared error loss function

$$\mathrm{E}\left((Y_{T+h} - g(\boldsymbol{X}_T))^2\right).$$

Result: Taking $g(\boldsymbol{X}_T)$ as the conditional expectation $\mathrm{E}(Y_{T+h}|\boldsymbol{X}_T)$ minimizes the mean-squared error loss function.

Linear prediction

Computing $\mathrm{E}(Y_{T+h}|\boldsymbol{X}_T)$ is hard if the random variables do not follow Gaussian distributions. Therefore we often consider the forecasts to be a linear function of $\boldsymbol{X}_T$. The best linear predictor is given by

$$\mathrm{Proj}(Y_{T+h}|\boldsymbol{X}_T) = a_0 + \boldsymbol{a}'\boldsymbol{X}_T,$$

where $\boldsymbol{a}$ is a vector of coefficients. That is chosen such that

$$\mathrm{E}\left((Y_{T+h} - a_0 - \boldsymbol{a}'\boldsymbol{X}_T)^2\right)$$

is minimized.

Forecasting linear time series

Now we want to apply the projection in a time series context.

- For a stationary time series $\{Y_t, t \in \mathcal{Z}\}$ we are interested in forecasting Y_{T+h} using $\boldsymbol{X}_T = (Y_T, Y_{T-1}, \ldots, Y_1)'$.
- We assume that the mean $\mathrm{E}(Y_t) = \mu_Y$ and autocovariance function $\gamma_Y(\cdot)$ are known. For ARMA models these can be computed as discussed above.

Forecasting linear time series

Our goal is to find a linear combination of $1, Y_1, \ldots, Y_T$ that forecasts Y_{T+h} with minimum mean squared error. The best linear predictor is given by

$$\text{Proj}(Y_{T+h}|Y_T, \ldots, Y_1) = a_0 + a_1 Y_T + \ldots + a_T Y_1$$

where we determine the coefficients $a_0, \ldots, a_T$ by minimizing

$$\text{E}\left[(Y_{T+h} - a_0 - a_1 Y_T - \ldots - a_T Y_1)^2\right]$$

which is a quadratic function that is bounded by zero.

Forecasting linear time series

Taking the derivatives with respect to $a_0, \ldots, a_T$ gives

$$\mathrm{E}\left[Y_{T+h} - a_0 - \sum_{i=1}^{T} a_i Y_{T+1-i}\right] = 0$$

and

$$\mathrm{E}\left[(Y_{T+h} - a_0 - \sum_{i=1}^{T} a_i Y_{T+1-i}) Y_{T+1-j}\right] = 0, \qquad j = 1, \ldots, T,$$

which can be solved for $a_0, \ldots, a_T$.

Forecasting linear time series

These derivatives can be written in the following illuminating form

$$a_0 = \mu_Y \left(1 - \sum_{i=1}^{T} a_i\right), \qquad \boldsymbol{\Gamma}_{Y,T}\boldsymbol{a}_T = \boldsymbol{\gamma}_T(h),$$

where $\boldsymbol{a}_T = (a_1, \ldots, a_T)'$,
$\boldsymbol{\gamma}_T(h) = (\gamma_Y(h), \gamma_Y(h+1), \ldots, \gamma_Y(h+T-1))'$ and

$$\boldsymbol{\Gamma}_{Y,T} = \begin{bmatrix} \gamma_Y(0) & \gamma_Y(1) & \ldots & \gamma_Y(T-1) \\ \gamma_Y(1) & \gamma_Y(0) & \ldots & \gamma_Y(T-2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_Y(T-1) & \gamma_Y(T-2) & \ldots & \gamma_Y(0) \end{bmatrix}$$

Forecasting linear time series

It follows that

$$\text{Proj}(Y_{T+h}|Y_T, \ldots, Y_1) = \mu_Y + \sum_{i=1}^{T} a_i(Y_{T+1-i} - \mu_Y)$$

where $\boldsymbol{\Gamma}_{Y,T}\boldsymbol{a}_T = \boldsymbol{\gamma}_T(h)$ holds for $\boldsymbol{a}_T$

Forecasting linear time series

Summarizing

- For any ARMA(p,q) model we may compute the autocovariance function $\gamma_Y(h)$
- Using $\gamma_Y(h)$ we may compute the optimal prediction coefficients $\boldsymbol{a}_T$ and compute the prediction
- Several algorithm exists that compute the coefficients $\boldsymbol{a}_T$ recursively, e.g. Durbin-Levinson and Innovations algorithm (see Brockwell & Davis 1991 for more details)

Parameter estimation

Parameter estimation

- The autocovariances typically depend on unknown parameters $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$
- If we have ARMA(p,0) then we can use OLS to estimate the parameters
- For general ARMA(p,q) we typically estimate the parameters by maximum likelihood estimation

Parameter estimation

- Assume that $\epsilon_t \sim NID(0, \sigma^2)$

This implies that $\boldsymbol{Y} = (Y_1, \ldots, Y_T)'$ is a Gaussian vector and we can write down the log likelihood function

$$\ell(\phi, \theta; Y) = -\frac{T}{2} \log 2\pi - \frac{1}{2} \log |\Gamma_{Y,T}| - \frac{1}{2} \boldsymbol{Y}' \Gamma_{Y,T}^{-1} \boldsymbol{Y}$$

where we recall that

$$\boldsymbol{\Gamma}_{Y,T} = \begin{bmatrix} \gamma_Y(0) & \gamma_Y(1) & \ldots & \gamma_Y(T-1) \\ \gamma_Y(1) & \gamma_Y(0) & \ldots & \gamma_Y(T-2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_Y(T-1) & \gamma_Y(T-2) & \ldots & \gamma_Y(0) \end{bmatrix}$$

Maximum likelihood estimates

The maximum likelihood estimates are defined as

$$\{\hat{\phi}, \hat{\theta}\} = \arg\max_{\phi,\theta} \ell(\phi, \theta; Y)$$

- The maximum is typically found by numerical optimization methods
- Standard errors can be computed by inverting the negative Hessian matrix

Return to unemployment forecasting

Unemployment parameter estimates

ARMA	(1,0)	(1,1)	(2,0)	(2,1)	(2,2)	(3,0)	(3,1)	(3,2)
ϕ_1	0.99	0.99	1.11	1.89	1.84	1.08	1.59	2.61
ϕ_2			-0.12	-0.89	-0.85	0.19	-0.37	-2.34
ϕ_3						-0.28	-0.23	0.73
θ_1		0.08		-0.75	-0.85		-0.59	-1.60
θ_2					0.22			0.81
ℓ	127	132	133	173	191	168	190	194
AIC	-249	-255	-259	-337	-371	-327	-368	-374

The MA component matters a lot for the in-sample fit!!!

Unemployment forecasting

- Next, we study the out-of-sample forecasting performance of the different ARMA models
- This requires making some choices
  - Horizon: $h = 1$
  - Loss function: $L(\hat{Y}_{T+h}) = (Y_{t+h} - \hat{Y}_{T+h})^2$
  - Sample split: January 2000
  - Window choice: rolling
- In practice all choices ask for a thorough robustness check!

Unemployment forecasting results

We report for each model average mean squared forecast error

$$\bar{L} = \sum_{T=\text{sample split}}^{\text{last period}} (Y_{T+h} - \hat{Y}_{T+h})^2$$

ARMA	(1,0)	(1,1)	(2,0)	(2,1)	(2,2)	(3,0)	(3,1)	(3,2)
$\bar{L}$	25.97	25.28	24.88	22.47	21.34	23.05	21.07	21.28

The MA component matters for out-of-sample forecasts!!!

Forecast evaluation

Forecast evaluation

- So far we can compute forecasts and losses for different models
- But are the documented differences significant?
- This requires forecast comparison tests
- We discuss
  - Diebold-Mariano test
  - Model confidence sets

Diebold & Mariano

Let

- Y_j realized value for period $j = 1, \ldots, J$
- $\hat{Y}_{1,j}$ forecast for period $j = 1, \ldots, J$ by procedure/model 1
- $\hat{Y}_{2,j}$ forecast for period $j = 1, \ldots, J$ by procedure/model 2

The goal is to compare $\{\hat{Y}_{1,j}\}$ to $\{\hat{Y}_{2,j}\}$

More specifically

$$H_0 : \text{The forecasts are equally good}$$

Diebold & Mariano

Define the forecast errors

$$e_{i,j} = Y_j - \hat{Y}_{i,j} \qquad i = 1, 2 \quad j = 1, \ldots, J$$

and the loss

$$L_{i,j} = g(e_{i,j})$$

where $g(\cdot)$ is such that

- takes the value zero when no error is made
- is never negative
- is increasing in size as the errors become larger in magnitude.

Popular choices include $g(e_{i,j}) = e_{i,j}^2$ and $g(e_{i,j}) = |e_{i,j}|$

Diebold & Mariano

To compare the losses we define the loss difference

$$d_j = L_{1,j} - L_{2,j}$$

- The forecasts are deemed equal if the loss difference is equal to zero in expectation for all j

$$H_0 : \mathbb{E} d_j = 0 \qquad j = 1, \ldots, J$$

versus

$$H_1 : \mathbb{E} d_j \neq 0$$

Diebold & Mariano

Define the test statistic

$$DM = \frac{\bar{d}}{\sqrt{\hat{\sigma}_d^2 / J}}$$

where

- $\bar{d} = \frac{1}{J} \sum_{j=1}^{J} d_j$
- $\hat{\sigma}_d^2$ is a consistent estimate for the long run variance $\lim_{J \to \infty} J\mathbb{V}ar(\bar{d})$
- In practice one could use a HAC estimator for the long run variance

Diebold & Mariano

Theorem

Assume that $\{d_j\}$ *is* α*-mixing of size* $-r/(r-2)$ *for* $r > 2$ *such that* $\mathbb{E}|d_j|^{r+\delta} < \Delta < \infty$ *for all* t *and some* $\delta > 0$*. Further, assume that* $\hat{\sigma}_d^2$ *is a consistent estimate for the long run variance* $\lim_{J\to\infty} J\mathbb{V}ar(\bar{d})$*, then*

$$DM \xrightarrow{d} N(0,1)$$

Proof, follows immediately from the α-mixing CLT of lecture 1.

Unemployment forecasting results

We compare the AR(3) model to the ARMA(3,1) model

$$DM = 2.805 \qquad p = 0.005$$

which implies that we reject the null that the forecasts are equally good

The MA component matters for out-of-sample forecasts!!!

Some comments

- Be careful when using Diebold & Mariano when comparing nested models
  - Use only with rolling windows
  - Otherwise Clark & McCracken (2001) provide adjustment for test statistic
- Clark & McCracken (2013) have a nice handbook chapter on forecast evaluation

Model confidence sets

What to do when you want to compare many models?

- Answer 1: Do many Diebold & Mariano tests $\rightarrow$ you need to do a correction for multiple hypothesis testing and you run the risk that the results are conflicting
- Answer 2: construct Model Confidence Sets (MCS)

Model confidence sets

- MSC is the set of models that contains the best performing models for a given level of confidence

Denote $\mathcal{M}^0 = \mathcal{M}$ set of initial model, we define the MCS for this set of models as

$$\mathcal{M}^* \equiv \{l \in \mathcal{M}^0 : \mathbb{E}(d_{lk,t}) \leq 0 \qquad \text{for all} \qquad k \in \mathcal{M}^0\},$$

where

$$d_{lk,t} = L_{l,t} - L_{k,t}$$

loss differential between models l and k

Model confidence sets

- The MCS procedure finds estimate $\hat{\mathcal{M}}^*_{1-\alpha}$
- Asymptotically consistent estimator of $\mathcal{M}^*$ with confidence $1-\alpha$
- For this iterate between
  - Test equivalence based on $\{H_{0,\mathcal{M}} : \mu^{lk} = 0 \text{ for all } l, k \in \mathcal{M} \subseteq \mathcal{M}^0\}$
  - Elimination rule $e_{\mathcal{M}}$ identifies the model that should be removed from $\mathcal{M}$ if H_0 is rejected.

Model confidence sets

To test

$$\{H_{0,\mathcal{M}} : \mu^{lk} = 0 \text{ for all } l, k \in \mathcal{M} \subseteq \mathcal{M}^0\}$$

Use the test statistic

$$T_{\mathcal{M}} \equiv \max_{l,k \in \mathcal{M}} |t_{lk}|, \qquad t_{lk} = \frac{\bar{d}^{lk}}{\sqrt{\hat{Var}(\bar{d}^{lk})}}.$$

with corresponding rule

$$e_{\mathcal{M}} = \arg\max_{l \in \mathcal{M}} \sup_{k \in \mathcal{M}} t_{lk}$$

The asymptotic distribution of this test statistic is non-standard as it depends on nuisance parameters. However, the distribution can be computed using bootstrap methods which implicitly deal with the nuisance parameter problem, see for details Hansen, Lunde & Nason (2011)

Unemployment forecasting results

Model confidence set:

	Rank	MCS-p
ARMA(1,0)	8	0.5220
ARMA(1,1)	9	0.4336
ARMA(1,2)	10	0.3546
ARMA(1,3)	12	0.2238
ARMA(2,0)	11	0.3070
ARMA(2,1)	6	1.0000
ARMA(2,2)	4	1.0000
ARMA(2,3)	5	1.0000
ARMA(3,0)	7	0.9994
ARMA(3,1)	3	1.0000
ARMA(3,2)	1	1.0000
ARMA(3,3)	2	1.0000

Final comments

Some comments

Thinking hard often helps when forecasting

- Unemployment: exploit labor market flows (Barnichon & Nekarda 2012), exploit Google searches for unemployment (D'Amuri & Marcucci 2017)
- Exchange rates: exploit Taylor rule fundamentals (Molodtsova and Papell 2009), Net-foreign asset positions (Gourinchas and Rey 2007)
- Inflation: ???

References & Material

Reference & Material

- References:
  R.H. Shumway & D.S. Stoffer, "Time Series Analysis and Its Applications with R examples", Chapter 4
  F. Diebold & R, Mariano, "Comparing Predictive Accuracy", Journal of Business and Economic Statistics, 1995 and 2012
  P.R. Hansen, A. Lunde & J.M. Nason, "The Model Confidence Set", Econometrica, 2011
- Code:
  `CodeLecture3.R`