시계열 데이터의 통계적 분석 방법

이기천 (한양대학교 산업공학과 조교수)

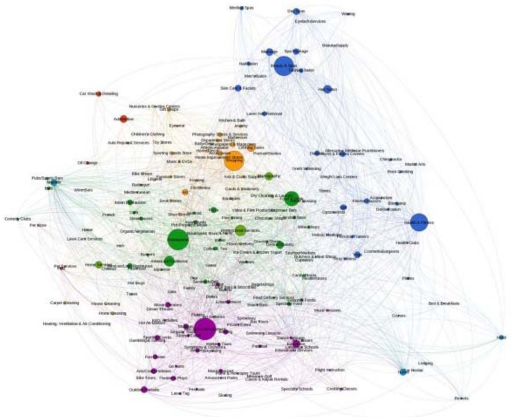
2013. 10. 31 (목)

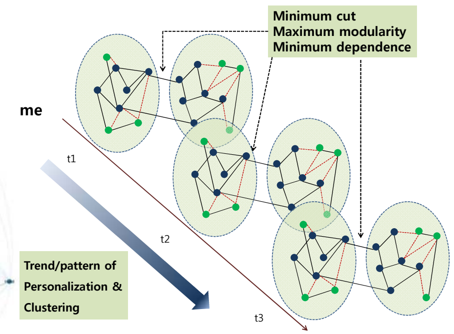
국민대학교 비즈니스IT전문대학원

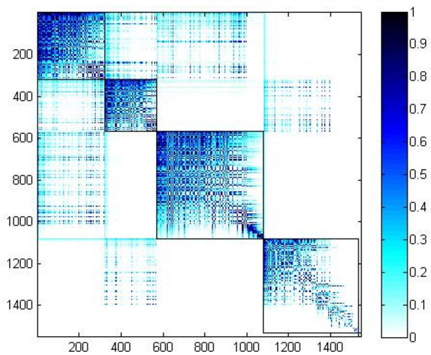
연구분야 Research area

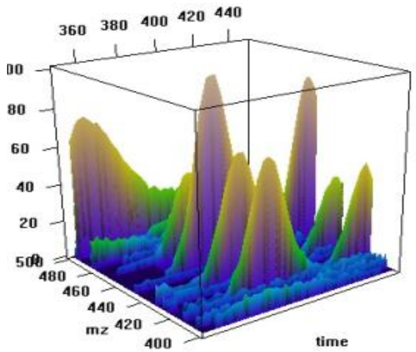
- Data Mining, Text Mining, Machine Learning, Pattern Recognition
- Process Mining, Social Network Analysis, Enterprise Service Computing
- Bioinformatics, Biostatistics, Statistical Computing
- Time Series Analysis, Wavelet Analysis, Frequency Data Analysis

이기천 교수
(한양대학교 산업공학과,
공업센터 705-1호)



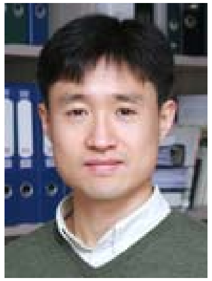






강의과목 Undergraduate course

- Data Mining
- Time Series Analysis
- Database Management
- Design of Experiments
- Applied Probability Models



Work Experiences

- 1998, IBM (인턴)
- 1998-1999, ETRI (위촉연구원)
- 2000-2001, Internet Consulting Group (개발실장)
- 2000, Digital Times IT 컬럼리스트
- 2001-2006, Tmax Soft (연구원)
- 2006, Samsung SDS (연구원)
- 2010-2011, Georgia Tech, Emory Univ. (Post-doctoral Researcher)

Education

•1998, KAIST 산업경영학과 (학사)

부전공: 전자공학

•2000, KAIST 산업공학과 (석사)

세부전공: Human Computer Interaction

•2010, Georgia Institute of Technology 산업공학과 (통계학박사)

부전공: 전자공학

목차

Introduction

- 시계열 데이터 (A time series)는 시간에서 **순차적으로 (sequentially)** 관측한 값들의 집합
  - **Continuous** vs. **discrete time series**

- 그럼, **Q. Discretize** 하는 방법은?
  - 1) continuous time series 로부터 샘플링
  - 2) 특정 기간 동안의 값들을 합치기**(accumulating)**

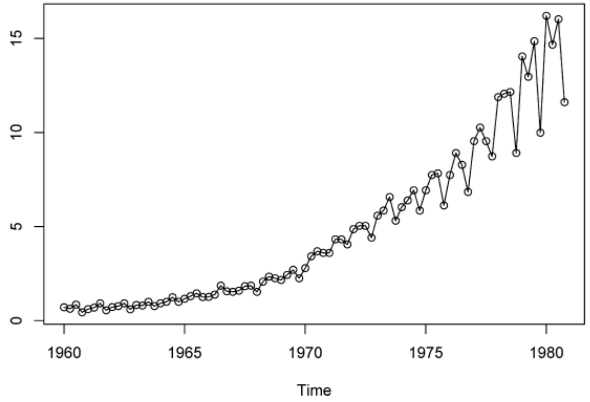
- 고정된 구간 사이의 시간 τ_1, τ_2,…, τ_N 시점에서 관측된 값들을 아래와 같이 표현한다
  - $x(\tau_1)$, $x(\tau_2)$,…, $x(\tau_N)$

Introduction (cont.)

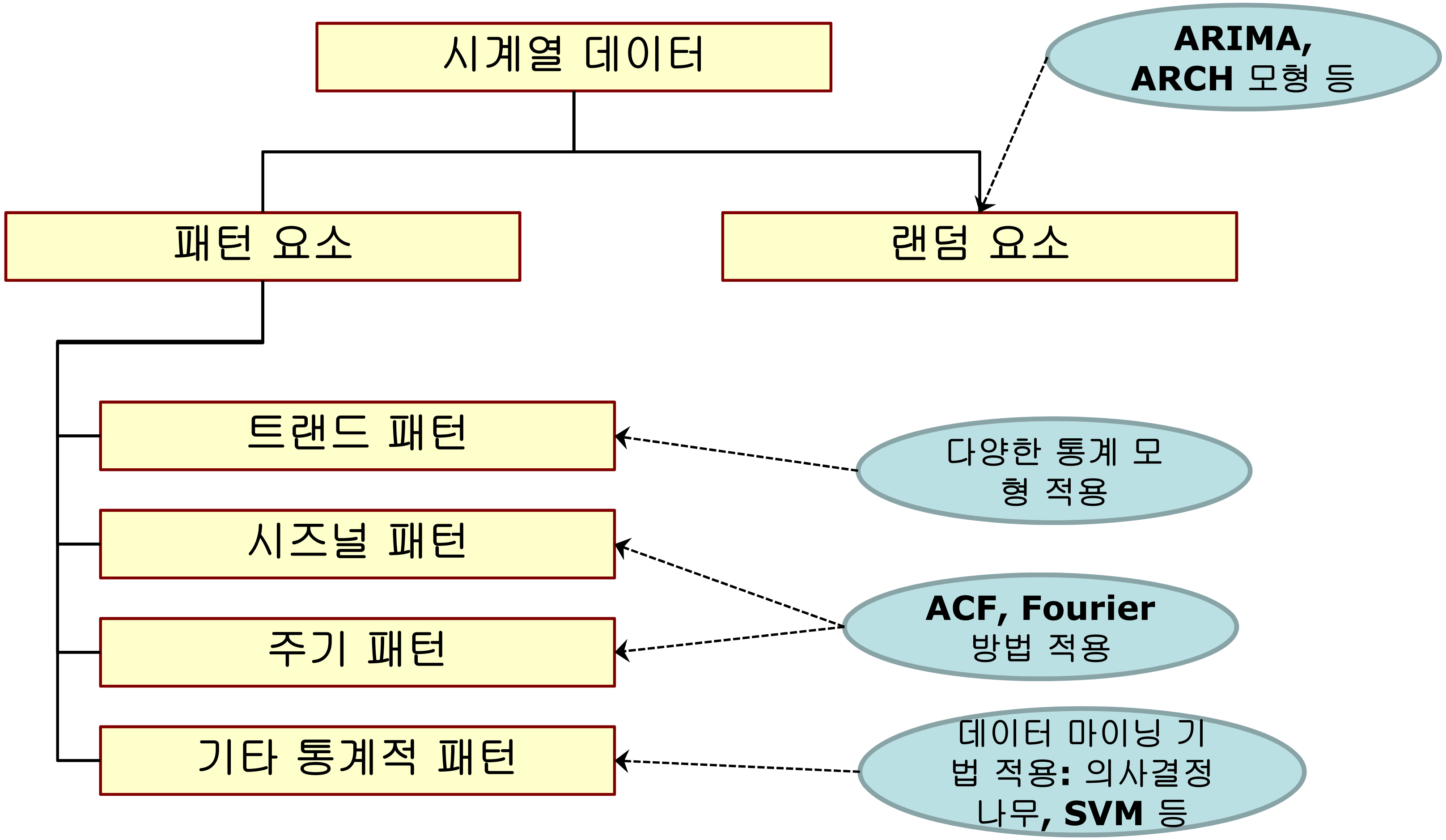
- 특징
  - **Time periods** 는 **equal length**
  - **missing values** 가 없음

- **Q. Missing value**가 있으면 어떻게 처리?
  - Expectation-Maximization 방법으로 처리 (Missing value를 모형을 통해 예측)
  - 일부 데이터만 Missing한 경우 가능

- Johnson & Johnson quarterly earnings per share



시계열 데이터의 요소

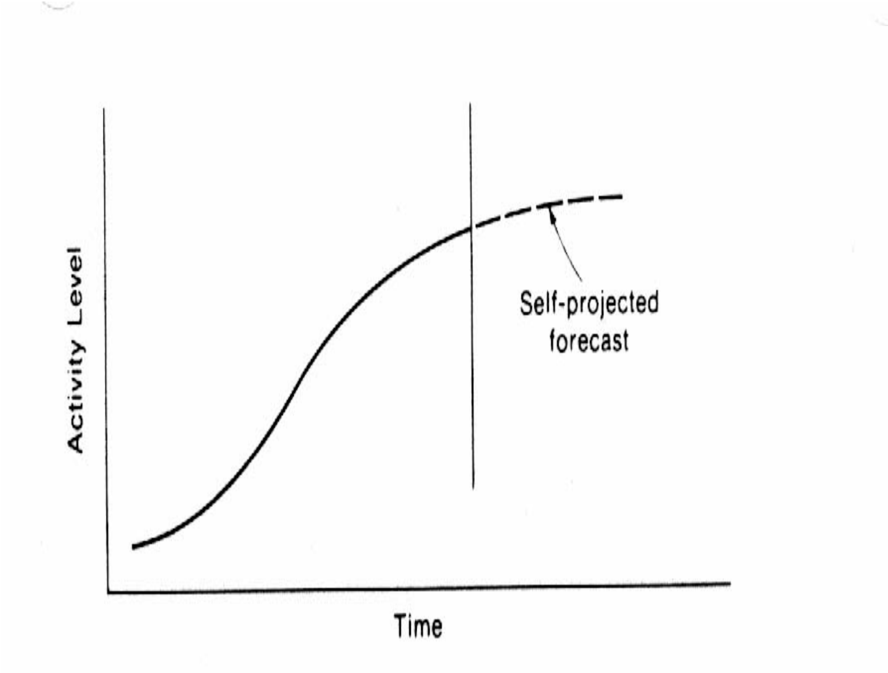

시계열 데이터
ARIMA,
ARCH 모형 등
패턴 요소
랜덤 요소
트랜드 패턴
시즈널 패턴
주기 패턴
기타 통계적 패턴
다양한 통계 모
형 적용
ACF, Fourier
방법 적용
데이터 마이닝 기
법 적용: 의사결정
나무, SVM 등

시계열 데이터 분석의 응용

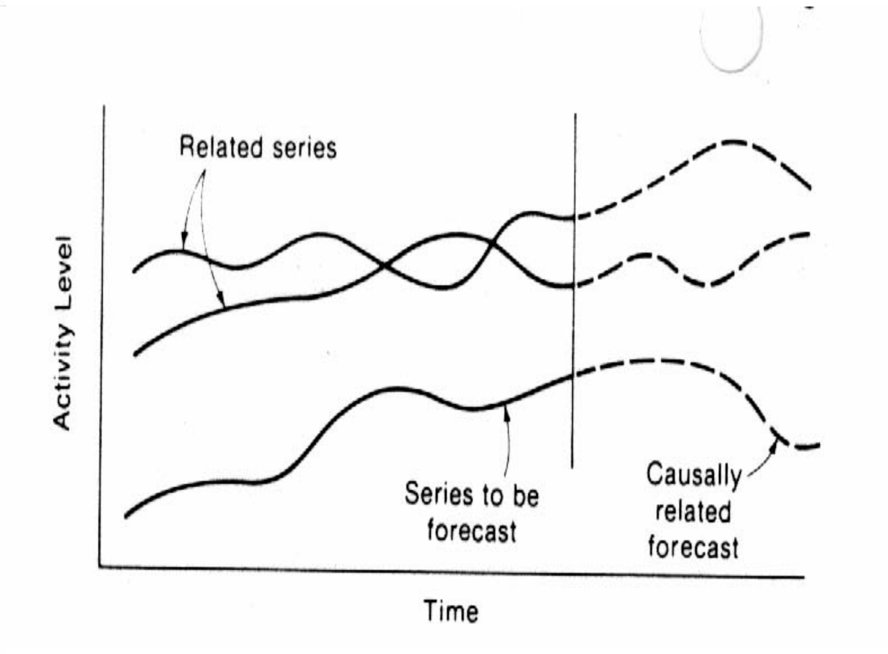
- **기술적**
  - **시계열 사이 관계 분석 (Transfer function 발견)**
    예: $x(\tau_1)$ 와 $y(\tau_1)$의 관계 발견
  - **프로세스 관리/표현 방법**
  - **<u>예측 (Forecasting)</u>**
- **영역**
  - **경제학; 비즈니스 계획; <u>수요계획</u>**
  - **재고 및 생산 관리**
  - **산업 프로세스 관리 및 최적화**
    <u>센서 시그널</u> 분석을 통한 지능화 모니터링

시계열 데이터 분석 접근 방법

- **자체-추정 방법**
  (Self-projecting)



- **원인-결과 방법**
  (Cause-and-effect)



시계열 데이터 분석 접근 방법 (cont.)

- **자체-추정 방법 (Self-projecting)**
  - **장점**

    최소의 데이터로

    빨리, 쉽게 분석

    주로 short-term 예측에 이용

    다른 분석의 초기분석으로 이용
  - **단점**

    long-term 예측에 어려움

    외부요소 고려하지 못함
- **원인-결과 방법 (Cause-and-effect)**
  - **장점**

    많은 정보 이용

    mid-term 예측 가능
  - **단점**

    좀 더 복잡한 과정 필요

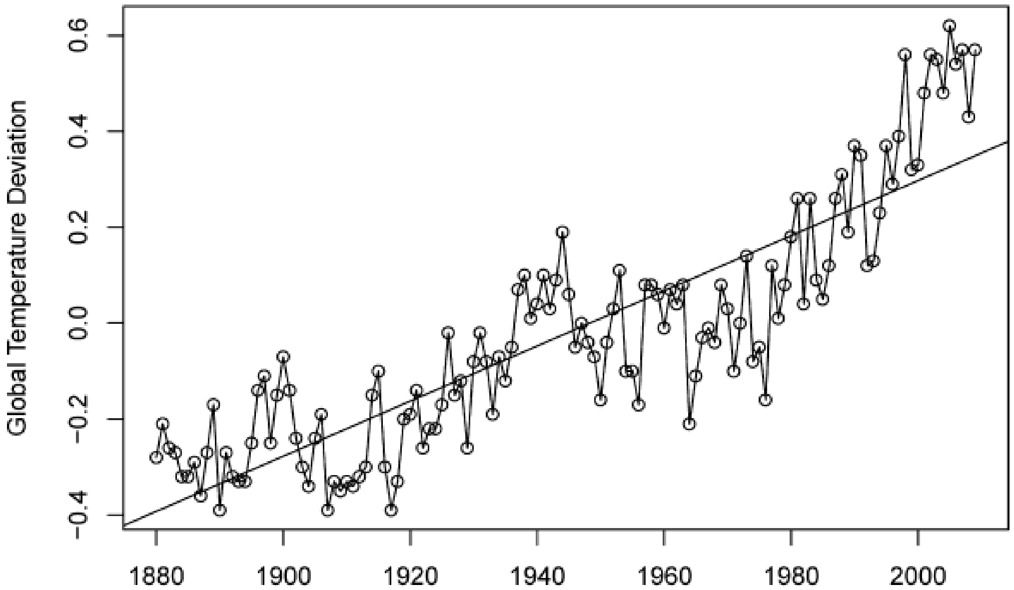
Self-projecting 의 classical 방법

- 전체적 트랜드를 찾아내는 방법들
  - 1차, 2차, ... 트랜드 찾아내기

  예를 들어, 아래와 같은 회귀 모형 적용

기본: 다양한 통계 모형 적용

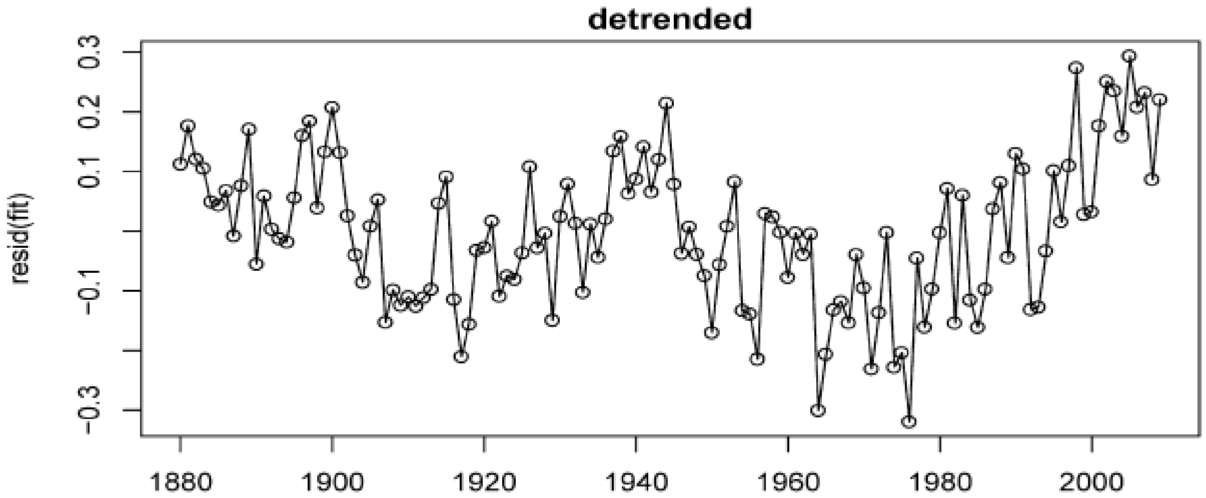
$$x_t = \beta_1 + \beta_2 t + w_t, \quad t = 1880, 1857, \ldots, 2009.$$



Self-projecting 의 classical 방법 (cont.)

- 전체적 트랜드를 찾아내는 방법들
  - 다음으로 트랜드제거(detrended), 즉 잔차(residual)을 구한다

$$\widehat{y}_t = x_t + 11.2 - .006\,t.$$



  - 이 시그널에 대하여 '랜덤요소'분석을 실시한다.
    예를 들어, ARIMA 모형 적용

Self-projecting 의 classical 방법 (cont.)

- **1차 트랜드의 간단한 예로부터 생각해볼 것들**
  - Q. 항상 그림을 그려서 트랜드를 찾아야 하나? (체계적, 자동화된 방법?)

    ☞ 자동상관계수함수(ACF, auto-correlation function) 등을 통계적으로 알 수 있음
    항상, 가능하면, 비주얼 plot을 하는 것이 좋음

  - Q. 1차 트랜드가 아니라 다른 차수 트랜드이면? (일반적 트랜드를 어떻게 해결?)

    ☞ 여러 가지 통계적 smoothing 방법이 있음

Self-projecting 의 classical 방법 (cont.)

- **Autocorrelation (ACs), Autocorrelation Function(ACF)**
  - 자기상관계수 Autocorrelation?
  - 일반적으로

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

  - 의미?

☞ X, Y의 선형적 관계의 정도

  - 시계열 분석에서는

$$\boxed{\rho_k = \frac{E\left[(z_t - \mu)(z_{t+k} - \mu)\right]}{\sigma_z^2}}$$

Self-projecting 의 classical 방법 (cont.)

- **Autocorrelation (ACs), Autocorrelation Function(ACF)**
  - 자기상관계수 Autocorrelation? (cont.)
  - 의미?

☞ 자체 시계열 데이터 내에서 얼마나 선형적 연관성이 있는가

lag k의 의미?

☞ z_t 와 k 만큼 shift 시킨 z_{t-k} 사이의 연관성

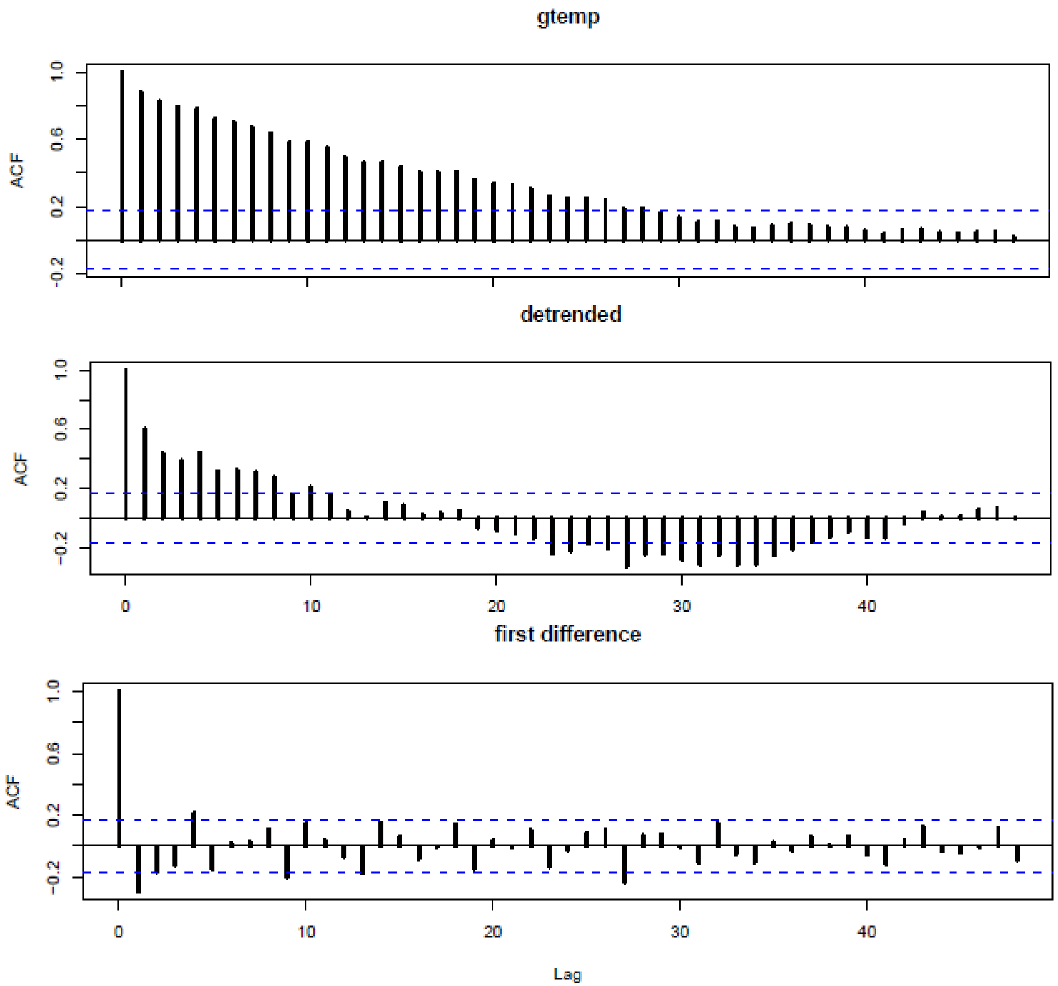
어떻게 계산?

sample autocorrelation

$$r_k = \frac{\sum_{t=1}^{N-k} (z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{N} (z_t - \bar{z})^2} \qquad k = 0,1,2,\ldots k$$

Self-projecting 의 classical 방법 (cont.)

- Global Temperature 데이터의 Autocorrelation Function(ACF)



Original signal Xt

빨리 **decay** 하지 않는다

잔차 **Residual Yt**

빨리 **decay** 한다

1 step 차이 **Xt- Xt-1**

더 빨리 **decay** 한다

ACF의 모양으로 적절한 **ARIMA** 모형 선택 가능

Self-projecting 의 classical 방법 (cont.)

■ 일반적 트렌드 찾아내는 방법

– Polynomial Regression:

$$x_t = f_t + y_t$$

$$f_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p$$

☞ 통계 테스트를 통해 가장 좋은 차수의 모델을 찾아냄

– Moving Average Smoother:

☞ 가중치 $a_{-k}, \ldots, a_0, \ldots, a_k$ 이용하여

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}$$

☞ 5일평균, 10일평균 등

Self-projecting 의 classical 방법 (cont.)

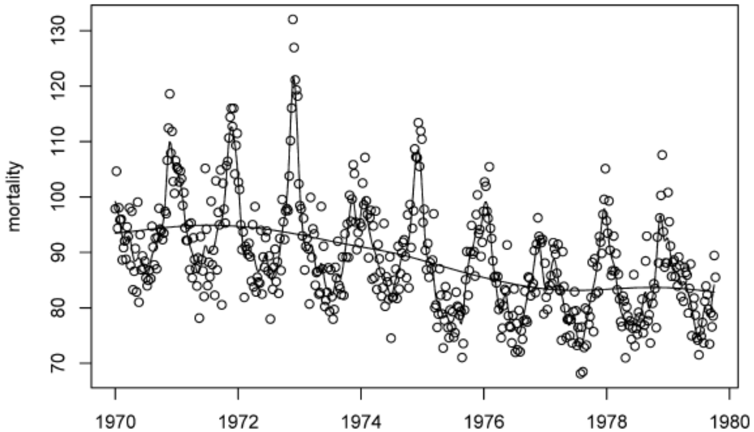
- 일반적 트랜드 찾아내는 방법 (cont.)

  – **Kernel Smoothing:**

  – Moving Average Smoother 보다 진화된 방법으로 가중치 w_i 를 데이터의 근접성에 기반하여 부여하는 방법

$$\hat{f}_t = \sum_{i=1}^{n} w_i(t) x_i \qquad w_i(t) = K\left(\frac{t-i}{b}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{t-j}{b}\right)$$

  – 여기서 K(.) 는 Kernel 함수이고, b 는 smoothing 레벨을 결정하는 파라미터



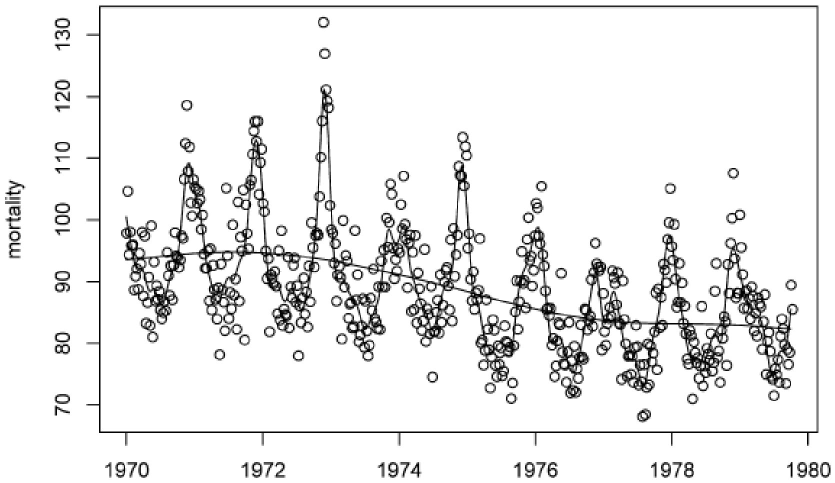
Self-projecting 의 classical 방법 (cont.)

■ 일반적 트랜드 찾아내는 방법 (cont.)

- **Smoothing Spline:**
- 3차 cubic polynomial 을 사용하여 각 knots 별로 polynomial 연속 되도록 fitting함

$$\sum_{t=1}^{n} [x_t - f_t]^2 + \lambda \int \left(f_t''\right)^2 dt$$

- 파라미터 λ 로 smoothing 수준을 결정함



Self-projecting 의 classical 방법 (cont.)

일반적 트랜드 찾아내는 방법 (cont.)

- Exponential smoothing:
- 최근 관측치와 최근 예측치의 평균으로 다음 예측치를 계산하는 방법
- 나중에 보게될 ARIMA 모형의 특수 경우가 된다, ARIMA(0,1,1)
- 가장 간단한 형태는

$$\tilde{x}_{n+1} = (1-\lambda)x_n + \lambda\tilde{x}_n$$

- Exponentially Weighted Moving Averages (EWMA)라고도 불림

- 이것보다 특정 트랜드(선형, 시즈널 등)을 반영하기 위한 다른 종류의 Exponential smoothing 방법도 있다

Self-projecting 의 classical 방법 (cont.)

- 일반적 트랜드 찾아내는 방법 (cont.)
  - Nearest Neighbor Regression
  - Lowess Regression
  - 등등 여러 방법을 적용할 수 있다
- 일반적으로 기초/중급/고급 통계 & 데이터마이닝 기법들을 적용하여 패턴을 찾아낸다
  - 명확하게 드러나는 시즌영향을 고려하고 (how?)
  - 구간 별로 다른 패턴을 고려하여야 한다

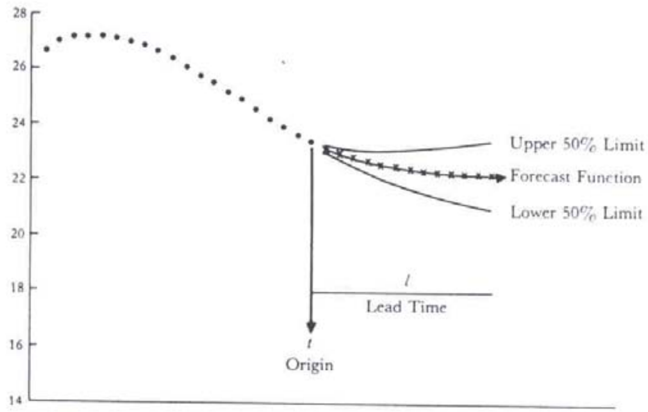
Self-projecting 의 classical 방법 (cont.)

- **Classical 방법의 단점**
  - 모형을 찾는 체계적인 방법이 없음
  - Trial-and-error 로 모형을 찾아야 함
  - 선택된 방법의 좁은 scope 내에서 찾게 됨
  - 잘 작동하는 좋은 모형인지에 대한 (이론적) 검증이 어려움

→ **Box-Jenkins** 접근법의 **ARIMA** 모형 이용

ARIMA models

- Autoregressive Integrated Moving-average
  - 넓은 범위의 시계열 데이터를 표현할 수 있음
  - 미래 관측치에 대한 Confidence Interval 도 구할 수 있음



  - 1960's Box 와 Jenkins가 경제학 관련 예측 연구
    - Time series analysis - forecasting and control, **by George E. P. Box and Gwilym M. Jenkins**
  - Box-Jenkins approach 라고 불림

ARIMA models (cont.)

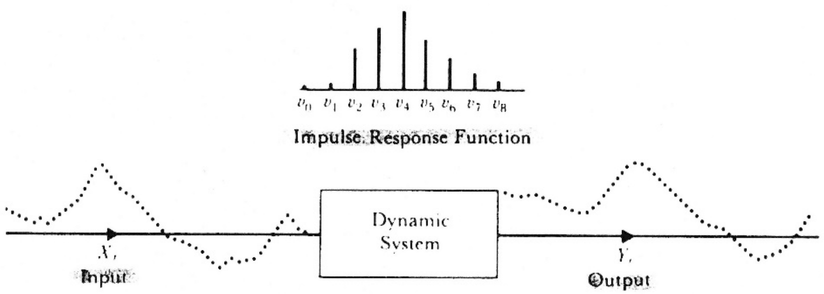
Box-Jenkins models

더 일반적

단일변수

다중변수
(Transfer function)

- **Transfer function model**
  - Lagged regression로 이해할 수 있음

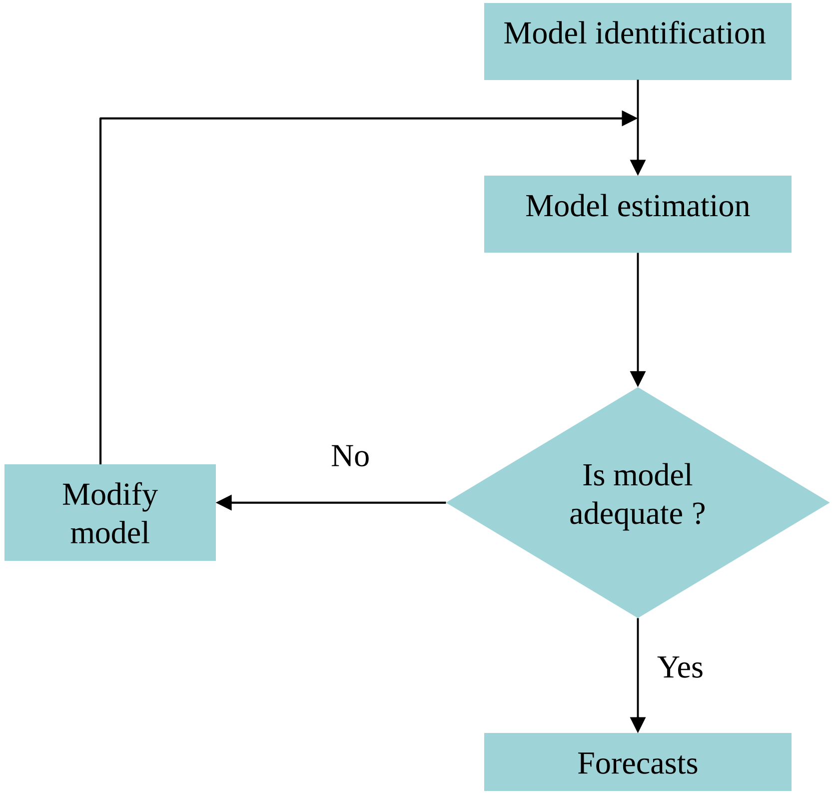


ARIMA models (cont.)

- **Transfer function model (cont.)**
  - **Lagged regression 을 좀 더 자세히 적으면,**
    - $Y_t = \nu(B)X_t$ **where**
    - $\nu(B) = \nu_0 + \nu_1 B + \nu_2 B^2 + \ldots\ldots$
    - **B is the backshift operator**
    - $B^m X_t = X_{t-m}$
  - 프로세스 관리에 많이 사용됨
    - **Control Equation**
    - **Feed-forward** 모델, **Feed-back** 모델 등
  - 설정된 **(Deterministic)** 변화**(**즉**,** 트랜드 부분에 연결**)**와
  - 랜덤한 변화의 부분을 동시에 고려함

ARIMA models (cont.)

- **Transfer function model (cont.)**
  - **모델 구축 과정**



원리가 이런 절차이고 모든 부분을 종합적으로 고려하여 모델 결정하게 된다

ARIMA models (cont.)

- **Model identification**
  - Autocorrelation 함수와
  - Partial-autocorrelation 함수를 사용

→ **Model의 주요 후보군 도출**

- **Model estimation**
  - Sum of squares of errors를 최소가 되게 하는 모델 내의 파라미터 추정
    - Noise가 정규분포를 따른다고 가정한 경우
    - Maximum Likelihood Estimation이 일반적인 방법

→ **Model 내 수치 결정:**
예, xt = 3 + 2xt-1 + at - 4at-1

- **Model validation**
  - 통계 테스트 등을 통하여 모델의 타당성 검증
    - 여러 통계 테스트와 AIC/BIC 값
    - Cross-Validation 개념의 내부예측력 테스트 등을 이용

→ 좋은 **Model** 선정

- **Model forecasting**
  - Future 관측치의 추정치와 confidence interval 계산

주요 개념들

- Normal 프로세스
- Stationary 프로세스/Invertibility/Causality
- AC, Partial AC
- AR, MA 모형
- White 노이즈 프로세스
- 선형 필터 과정

Normal (Gaussian) 프로세스

- 각 관측치 z_t 는 **probability density function** $p(z_t)$ 로부터 관측됨
  - **특히 pdf를 정규분포라고 가정함**
  - **Q. 다른 분포라고 가정할 수 있나?**
    → 가능. 그러나, 복잡 -_-;
- Random 변수의 연속된 관측치가 시계열 데이터이다
- 예를 들어 z_{t_1}, z_{t_2} 는 **joint probability density function** $p(z_{t_1}, z_{t_2})$ 으로부터 주어진다

White Noise 프로세스

- **Box-Jenkins 모델에서는 Xt 가 지금과 그 이전 시점의 uncorrelated된 노이즈들의 영향으로 발생한다고 가정한다**

- White noise et는 다음을 만족하는 프로세스이다

$$E[e_t] = 0 \quad var[e_t] = \sigma_e^2$$

$$\rho_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

- ***e_t, e_{t-1}, e_{t-2},...*** 를 white noise 프로세스라고 한다.

Stationary 프로세스

- Stationary: 정적인, 움직이지 않는
- 쉬운 말로, 시계열 데이터의 특징이 (어느 수준 범위에서) 항상 고정되어 있을 때를 말함
- 통계적으로 (간단히): **$p(x_t)$** 에 대해 모든 t에 대해 동일함
  - 아주 강력한 가정임
- Strictly stationary
  - m개의 관측치 **t_1, t_2, …, t_m** 의 **joint pdf**와 **t_{1+k}, t_{2+k}, …, t_{m+k}**의 **joint pdf**가 같음
- Weakly stationary
  - 평균 E[Xt]가 t에 의존하지 않고
  - Xt와 Xt+k의 Correlation/covariance 이 t에 의존하지 않음

이 개념의 **stationarity**를 주로 사용

Stationary 프로세스 (cont.)

- $\mathbf{z_t}$ 가 stationary 프로세스이면 그것의 차이 $\nabla \mathbf{z_t} = \mathbf{z_t} - \mathbf{z_{t-1}}$ (또는 고차원의 차이 $\nabla^d \mathbf{z_t}$) 또한 stationary 프로세스가 된다

- 대부분의 시계열 데이터는 stationary 하지 않다

- *Stationary 한 시계열 데이터 모형으로 만들기 위해 차이를 계산한다*

$$(1^{st}\ \text{order}) \quad \nabla x_t = (1 - B)x_t = x_t - x_{t-1}$$

$$(2^{nd}\ \text{order}) \quad \nabla^2 x_t = (1 - B)^2 x_t = x_t - 2x_{t-1} + x_{t-2}$$

$$\nabla^d = (1 - B)^d$$

"B" 는 **backward shift operator** 라고 부름

Stationary 프로세스 (cont.)

- 단순한 차이가 stationary 프로세스로 연결되지 않을 때도 있다
- 이전의 trend를 제거하는 방법을 사용하는 것도 좋은 방법이다
- 주어진 Xt를 변환 시키는 것도 좋은 방법이다
  - Q. 어떤 변환을 사용하여야 하는가?
  - 1) 통계적으로 Box-Cox 변환 방법을 사용하거나
    - Var[Xt] 의 형태가 시간 t에 따른 형태에 따라 통계적으로 안정적 변환을 찾아냄

$$y_t = \begin{cases} (x_t^{\lambda} - 1)/\lambda & \lambda \neq 0, \\ \log x_t & \lambda = 0. \end{cases}$$

    - 여기서 $\lambda = 1 - \alpha$ 이며 α 는 $\sigma \propto \mu^{\alpha}$

Stationary 프로세스 (cont.)

- 주어진 Xt를 변환 시키는 것도 좋은 방법이다 (cont.)
  - 2) 문제 도메인에 맞게, 해석가능 하게 변환한다

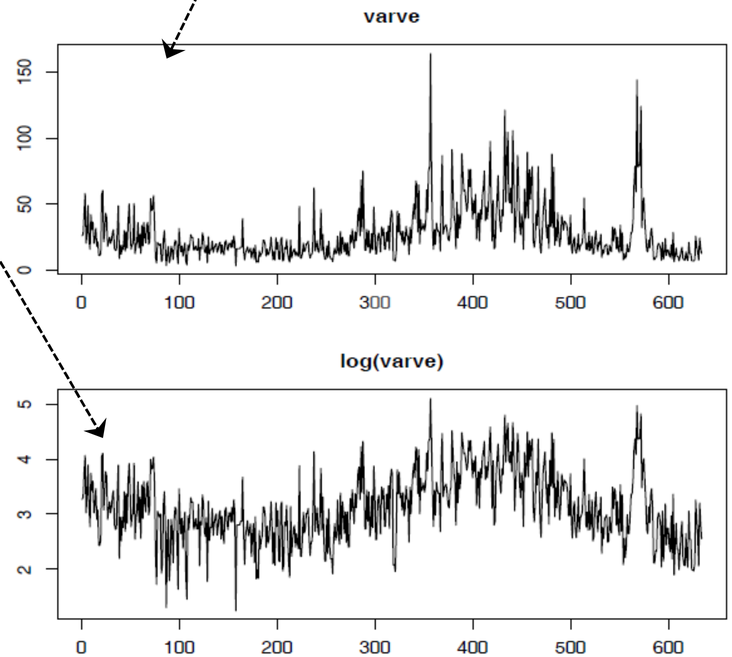
$$\log(x_t) - \log(x_{t-1})$$

$$\mathrm{Sqrt}(x_t)$$

그냥 Xt

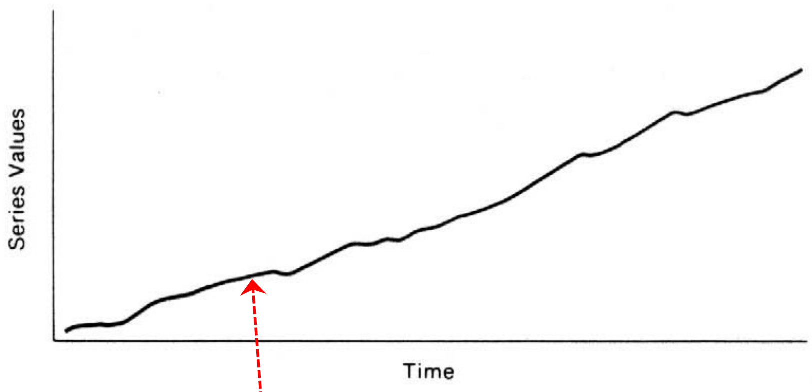
  - 예: England 특정 지역의 연간 퇴적층 증가 x_t
  - 예:
    길이 -> 제곱으로 넓이
    Rate -> 역수로 평균 시간 등등

log Xt

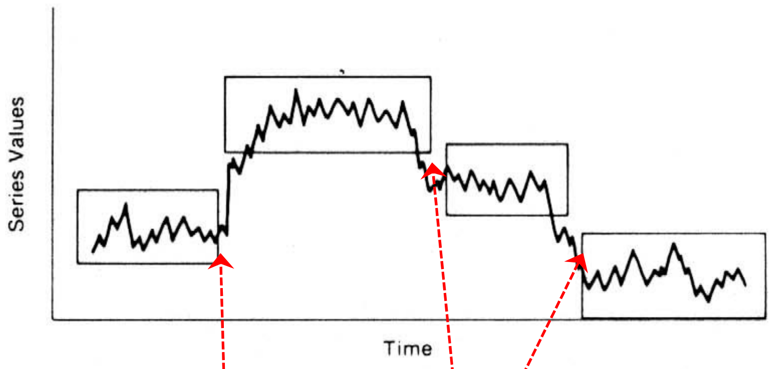


Stationary 프로세스 (cont.)

- Nonstationary 시계열 데이터 형태



A Nonstationary Series: Overall Trend



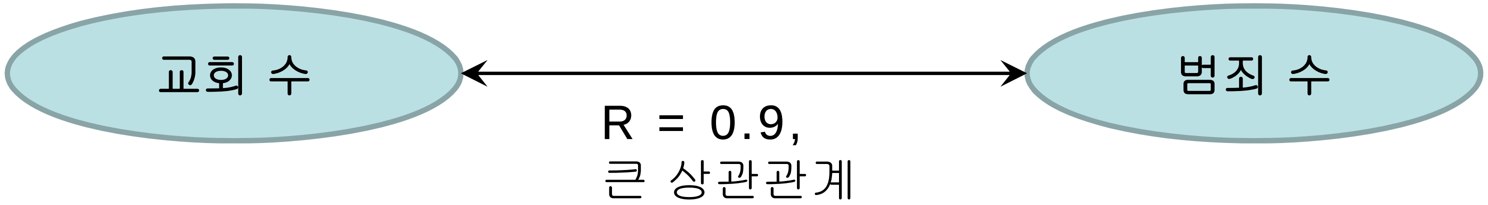
A Nonstationary Series: Random Changes in Level

평균이 **t**에 의존함

일반적으로, **Change-of-point** 방법으로 최적의 구간들을 찾을 수 있음

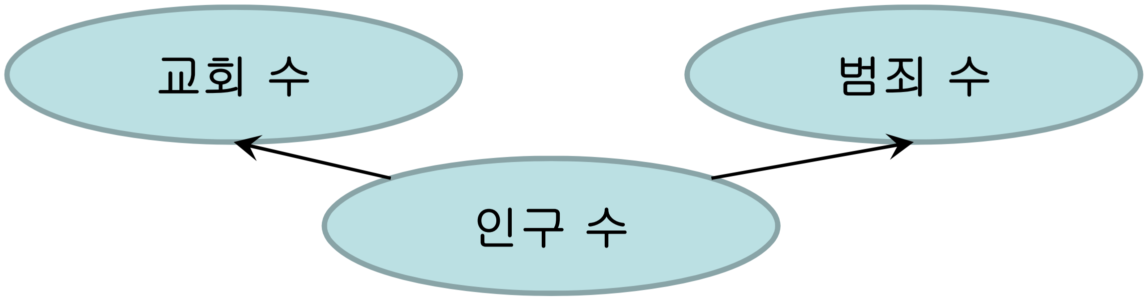
Box-Jenkins 모형에서는 더 간단하게 **ACF, Partial ACF**으로 모형**(**차이 차수 포함**)**의 후보를 도출

ACF, Partial ACF



????

- 이유는 인구 수가 모두에 영향을 주기 때문이다



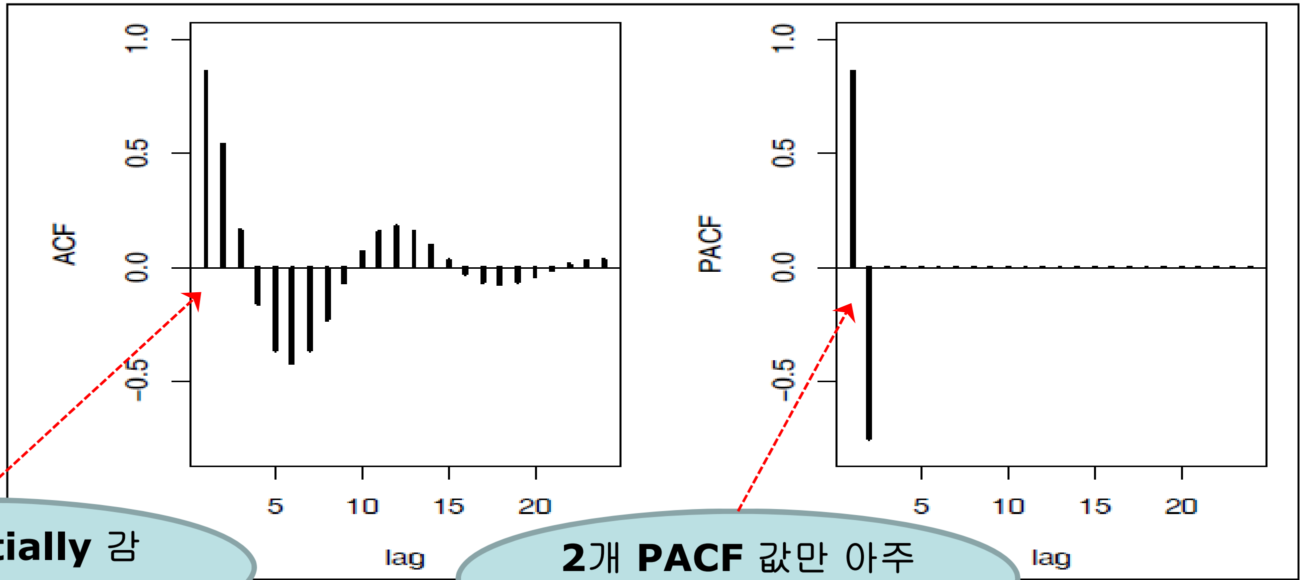
- Partial auto-correlation 은 x와 y의 상관 관계를 다른 변수들 z의 영향을 빼고 계산한다

ACF, Partial ACF (cont.)

- 더 정확하게 시계열에서 lag k에 대한 Partial auto-correlation 아래와 같은 회귀 식의 $\boldsymbol{\theta_{kk}}$ 을 계산하는 것이다.

$$X_t = \theta_{k1}X_{t-1} + \theta_{k2}X_{t-2} + \dots + \theta_{kk}X_{t-k} + \varepsilon_t$$

  - 그 사이의 관측치의 영향을 제외하고 상관 계수를 계산하게 됨
- 예: $X_t = 1.5\ X_{t-1} - 0.75\ X_{t-2} + a_t$ 인 AR(2) 모형의 시계열 데이터



Exponentially 감소

2개 **PACF** 값만 아주 큼

AR, MA 모형

- 모델 구성을 위한 요소들
  - Autoregressive (AR) models
  - Moving-average (MA) models
  - (Mixed) ARMA models
  - Non stationary models (ARIMA models)
  - The mean parameter
  - The trend parameter

AR, MA 모형

- 차수 p의 Autoregressive 시계열 모형은 다음과 같다

$$\boxed{x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + e_t}$$

$$\boxed{\phi(B)x_t = e_t}$$

$$\boxed{\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p}$$

- 위 모형은 $\phi^{-1}(B)$ 의 transfer function이 **white noise e_t** 에 대해 적용된 것과 같은 모형이다.
- $\phi^{-1}(B)$ 이 구해질 수 있을 때 **invertible** 한 시계열 모형이 되고 또한 causal 한 시계열 모형이 된다.
- $\phi(B) = 0$ 을 특성식이라고 부른다.

AR, MA 모형 (cont.)

- 차수 q의 Moving-average 시계열 모형은 다음과 같다

$$\boxed{x_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \ldots - \beta_q e_{t-q}}$$

$$\boxed{x_t = \theta(B) e_t}$$

$$\boxed{\theta(B) = 1 - \beta_1 B - \beta_2 B^2 - \ldots - \beta_q B^q}$$

- 위 모형은 **θ(B)** 의 transfer function이 **white noise e_t** 에 대해 적용된 것과 같은 모형이다.
- **θ(B) = 0** 을 특성식이라고 부른다.

AR, MA 모형 (cont.)

- AR, MA 모형은 서로 연결되어 있다
- 예로 MA(1)의 경우를 생각해보자

MA(1) 모형

$$X_t = (1-\theta_1 B)a_t$$

$$\frac{1}{(1-\theta_1 B)}X_t = a_t$$

$$(1+\theta_1 B+\theta_1^2 B^2+\theta_1^3 B^3+\ldots)x_t = \mathrm{a}_t$$

AR(∞) 모형

$$x_t = -\theta_1 x_{t-1} - \theta_1^2 x_{t-2} - \theta_1^3 x_{t-3} - \ldots + a_t$$

- 즉, 여러 가능한 모형 중에 단순한 모형을 찾는 것이 필요하다

AR, MA 모형 (cont.)

- 단순한 모형을 찾기 위해 AR과 MA를 합친 모형을 고려한다

- ARMA(p,q) 모형

$$\phi(B)x_t = \theta(B)a_t$$

$$x_t = \frac{\theta(B)}{\phi(B)} a_t$$

- ARMA(p,q) 모형은 **white noise** a_t에 transfer function **θ(B)/ϕ(B)**이 적용된 것과 같은 모형이다.

Model Identification

Stationarity 조건 만족 시킴

Transformation, differencing 등을 통해

↓

Sample AC, PAC계산

↓

이론적 ACs and PACs 와 비교

↓

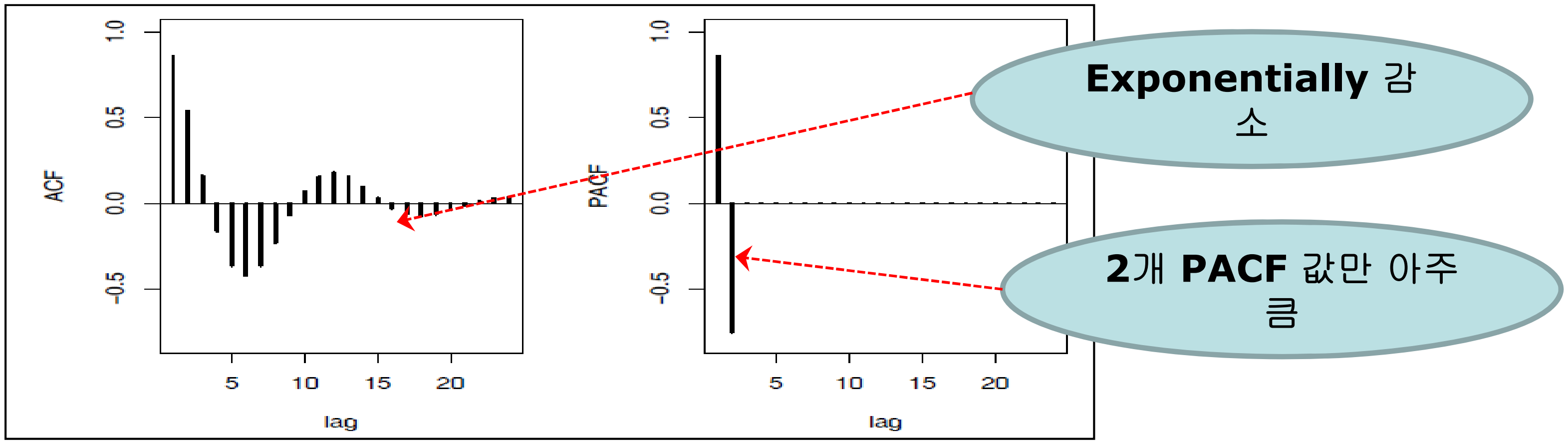
모델 (후보) 도출

Model Identification (cont.)

- ARMA 모형에 대한 이론적 AC와 PAC의 형태

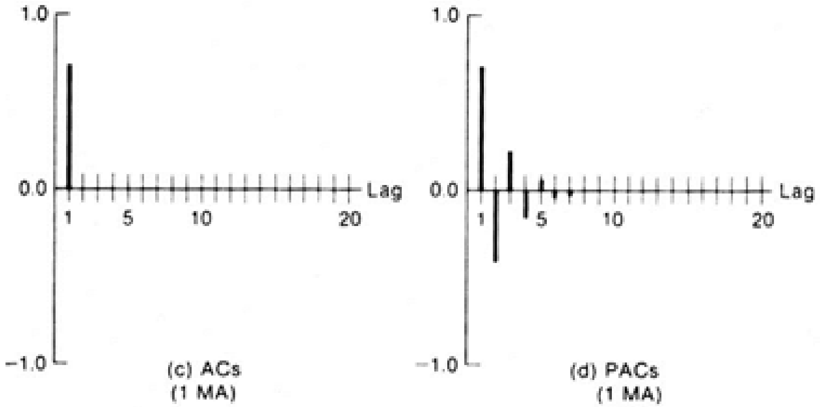
	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

- 아래와 같은 ACF, PACF는 AR(2)를 나타내는 것임

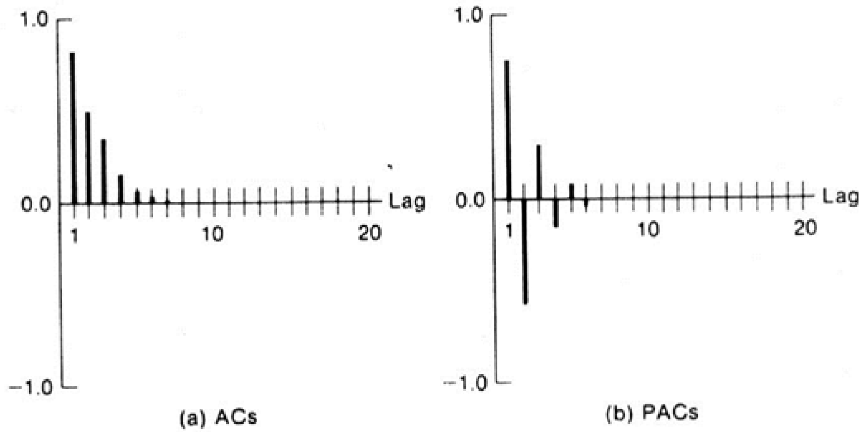


Model Identification (cont.)

- 아래와 같은 ACF, PACF는 MA(1)을 나타내는 것임

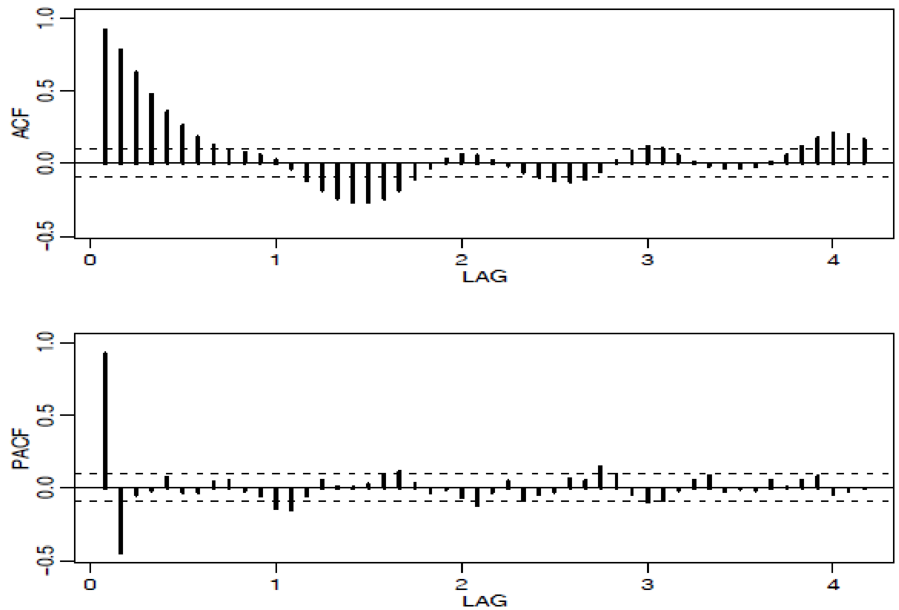


- 아래와 같은 ARMA(p,q)를 나타내는 것임



Model Identification (cont.)

- 보통 white noise 의 경우라고 간주할 수 있는 95% Confidence Interval을 계산하면 도움이 된다



- 명확하게 하나의 모델로 귀결되는 것이 아니라 대략적 모델의 후보를 추천할 수 있게 된다
  - 위의 경우, AR(2) 또는 ARMA(p,q)를 추천할 수 있다

Model Identification (cont.)

- Q. 여러 후보들의 모델 중에 좋은 것을 찾아내는 방법은?

  - Golden rule: 예측력이 좋은 모형을 찾는 것이다

- 1) 이론적 예측력이 좋은 것을 골라내는 방법
  - AIC: Akaike's Information Criterion

$$\mathrm{AIC} = \log\ \hat{\sigma}_k^2 + \frac{n+2k}{n}$$

$\hat{\sigma}^2$는 model estimation에서 구해지는 error 레벨, k는 추측해야 할 변수 개수

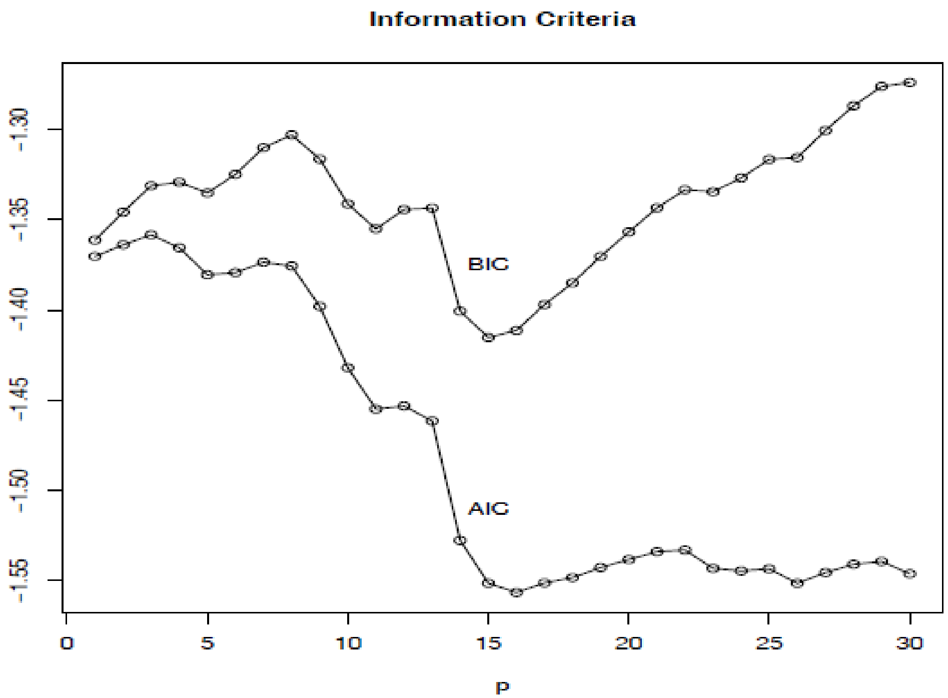
  - BIC: Bayesian Information Criterion

$$\mathrm{BIC} = \log\ \hat{\sigma}_k^2 + \frac{k \log n}{n}$$

  - AIC나 BIC값이 적게 되는 모형을 찾아낸다
  - AIC, BIC 정의는 저자, SW구현마다 다를 수 있는데, 내부 비교에는 문제가 없다
  - 이 과정을 model selection이라고 따로 구분할 수도 있다

Model Identification (cont.)

- 예: 만약 AR(p)모형을 사용해야 한다고 정했다고 가정하자

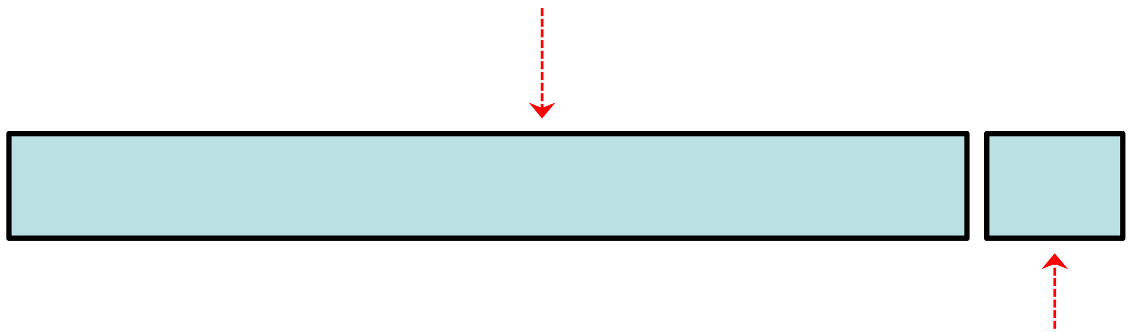


- AR(15)또는 AR(16)의 모형을 추천
- BIC 기준을 사용할 때 좀 더 단순한 모형을 얻을 수 있다

Model Identification (cont.)

- 2) 실질 (내부) 예측력이 좋은 것을 골라내는 방법
  - Cross-validation과 같은 방법을 사용하여 좋은 모형을 찾아낸다
  - FPE (Forward Prediction Error) 지수가 이것의 예이다
  - 단점: 계산 과정이 길다
  - 장점: 예측력 입장에서 가장 좋은 방법
  - 예:

전체 시계열 데이터의 일부를 가져와서 **Model Build**에 사용



구성된 **Model**로 앞 **10** 개의 데이터에 대해 예측

위 과정을 여러 partition에 대해 반복

Model Validation

- 위의 방법으로 얻어진 Model이 가정을 잘 만족시키고 있는지 타당성 검사를 한다
- Q. 검사하는 방법은?

  – Residual 이 normal 분포를 따르는가?
  ☞ QQ plot 이 직선으로 보이는지 확인

  – Residual 이 white noise 처럼 uncorrelated 되어 있는가?
  ☞ Ljung-Box-Pierce Q 통계치

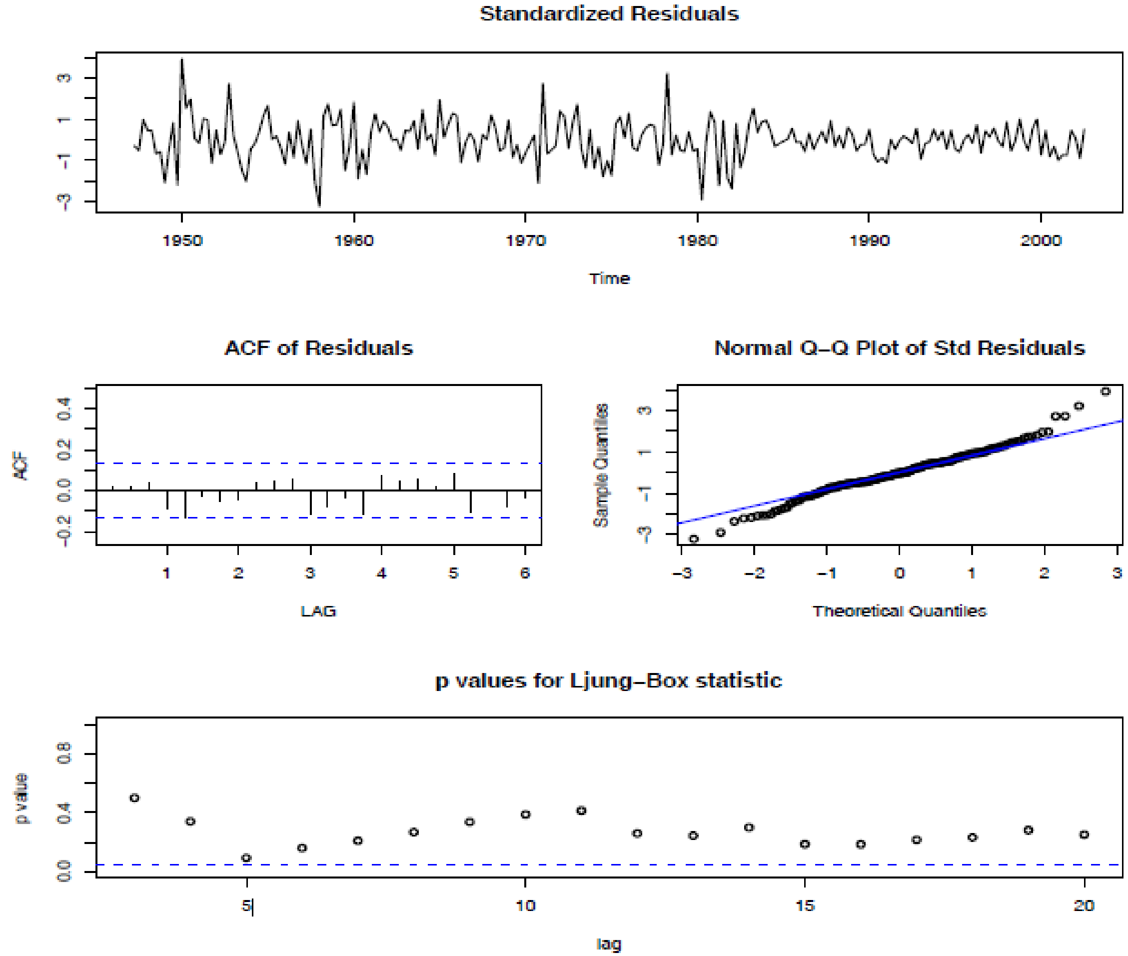
$$Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}_{e}^{2}(h)}{n-h}$$

  - Q 값이 크면 (p-value가 작으면) 지금 모형이 별로 좋지 않다는 뜻

  – ☞ Residual 의 ACF 을 그려서 white noise의 CI 내에 존재하는지

Model Validation (cont.)

- 예: GNP 성장률 log 차이의 시계열 데이터를 MA(2) 로 fitting한 후



Model Estimation (모수 예측)

- Model 이 정해지면 (예를 들어 ARMA(1,1)으로) 보통 SW에서 사용되는 모수(Parameter)를 추정해 준다
- 내부적으로 아래의 방식으로 진행됨
  - White noise가 정규분포를 따른다고 가정하고 Maximum Likelihood Estimation 방법을 적용

    ☞ Residual의 sum of squares가 최소가 되는 모수를 찾는 방법과 같은 개념

    ☞ likelihood가 크게 되는 모수를 computational 방법으로 찾아야 하는데,
    보통 Newton-Raphson 방법으로 찾아낸다

  - ACF에 대한 Yule-Walker 공식을 이용하여 모수 추정

  - Durbin-Levinson 의 순차적 알고리즘을 이용하여 모수 추정

- 특별한 모델을 가정할 경우 위와 같은 방식으로 추정한다

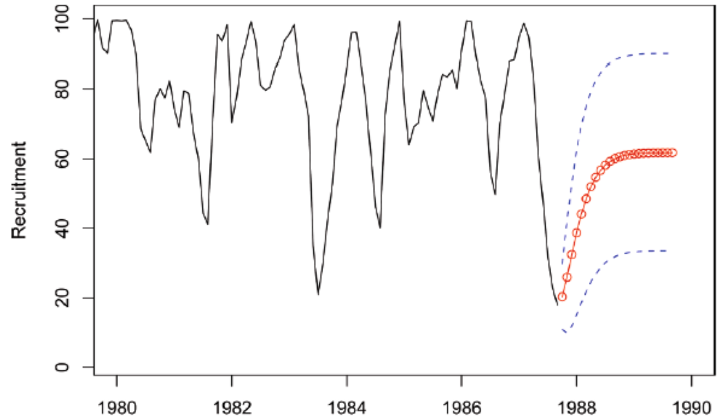
Forecasting (미래 값 예측)

- 최적의 Model 이 정해지 과거 관측치가 주어진 상황에서 미래 관측치를 예측한다

- 여러 방법 중에 Square 손실 함수를 최소화 하기 위해 평균으로 예측한다

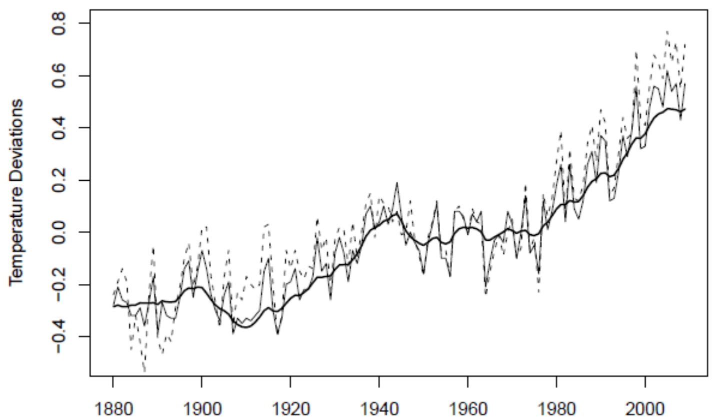
$$E[X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \ldots]$$

- 주로 (미래 10개 정도의) short-term 예측을 하고, long-term 예측인 경우 모델에서 주어지는 평균으로 수렴하게 된다

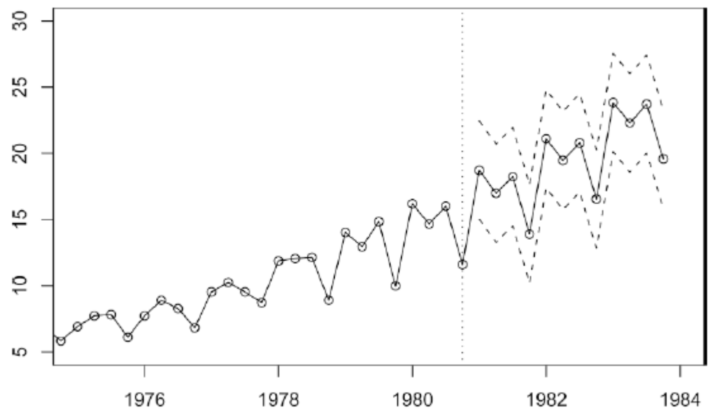


다른 방법들

- State-space 모델 방식
  - ARIMA 모형의 일반화 하여 transfer function을 구할 수 있으며 smoothing과 forecasting 모두 가능하다



Smoothing



Forecasting

  - ARIMA를 기본 모형으로 하여 일반적으로 확장한 것이다. 훨씬 복잡하다.

다른 방법들 (cont.)

- ARFIMA 모형
  - ARIMA 모형에서 차이 difference 차수 d가 정수가 아니라 일반 유리수
  - 랜덤 노이즈 같은 프로세스의 특징을 찾아낼 수도 있다
- Fourier 변환에 의한 주기 방식
  - 주기성을 가진 데이터에 대하여 잘 작동한다
  - Periodogram을 구하여 주기를 찾아낼 수 있고 smoothing도 가능하다
  - 특이점이 있는 데이터에 대하여 잘 작동하지 않는다
- Wavelet 변환 방식 분석
  - 특이점이 있는 데이터에 대해서도 smoothing을 할 수 있고
  - 랜덤 노이즈 같은 시그널의 특징을 찾아낼 수 있다
  - 시간 도메인에서 찾기 어려운 feature들을 찾을 수 있다

결론

- 시계열 분석을 위한 기초 통계적 접근 방법 설명함
  - Box-Jenkins 방법의 ARIMA 방법을 중심으로 설명함

- 개인적 결론적 메세지
  - 기초(기본 통계, 데이터 마이닝)가 중요함
  - 합리적으로 융합하는 것이 필요함
  - SW를 이용하여 처리하는 것이 필요함
  - 특별한 모형을 만들어 접근할 때 모델링 및 구현을 할 수 있어야 함
  - 고차원 모델일 수록 할 수 있는 것은 많고 또한 해야 할 것도 많음

감사합니다!

수고하셨습니다!