Unit 1

시계열분석 소개

7주차. 시계열분석

학습 내용

- 시계열이란?
  - Stationarity
- 시계열 모형
- 시계열 모형의 특징

학습 목표

- 시계열의 정상성에 대해 이해하고 설명할 수 있다.

시계열이란?

시계열의 정의

시계열 시간에 따른 데이터의 series

- 시간 : 독립변수
- 핵심내용 : Serial dependence
  (내재적 의존성, time-dependent structure)
- 용도 : 예측(forecasting)

Date Series는 과거의 값에 **의존성**이 있다.

시계열이란?

시계열의 이론 설명

Stochastic Process

$$dY(w,t)$$

Time

Random Variable

State

- 특정 확률과정의 **실현** 또는 **샘플 함수**
- 시간에 따라 **순차적**으로 생성 된 일련의 **관측 값**
- **시간에 따른 의존성**(무작위 샘플은 개념이 다름)
- 관측값의 **시간 간의 거리**는 **동일**하다고 가정(시간상 더 **가까운 값**들에 대한 의존성이 큼)

시계열이란?

시계열의 이론 설명

Mean Function

$$\mu_t = E[Y_t]$$

Variance Function

$$\sigma_t^2 = \gamma_0 = Var(Y_t) = E(Y_t - \mu_t)^2 = E(Y_t^2) - \mu_t^2 < \infty$$

$$(\gamma_0 \geq 0)$$

- Autocorrelation Function = Normalized Autocovariance (분산으로)

Stationarity

Stationarity 소개

Stationarity(정주성 또는 정상성)

"시계열을 제어하는 확률법칙이 시간에 따라 변하지 않는다."

- 시계열이 통계적으로 균형 상태에 있다는 것을 의미

Stationarity

Strong Stationarity

프로세스가 2개인 경우

$$F_{\boxed{Y_{t_1},Y_{t_2}}}\boxed{(y_1,y_2)} = F_{\boxed{Y_{t_1+k},\cdots,Y_{t_2+k}}}(y_1,y_2) for\ any\ t_1,t_2 and\ k.$$

시간의 거리가 k만큼 변형해도 확률분포가 동일하면 Second-order-Stationary

$$F_{Y_{t_1},\cdots,Y_{t_n}}(y_1,\cdots,y_n) = F_{Y_{t_1+k},\cdots,Y_{t_n+k}}(y_1,\cdots,y_n) for\ any\ t_1,\cdots,t_n and\ k.$$

n개에 대해서 시간의 거리가 달라도 확률분포가 동일하면 nth-order-Stationary

Stationarity

Strong Stationarity

강정상성(Strong Stationarity, 강정주성)에서는

- 평균 › 일정함
- 분산 › 일정함
- Covariance › 시간에 의존함
- Correlation › 시간에 의존함

Stationarity

Weak Stationarity

강한 개념의 Stationary는 거의 관측이 안 됨

- Weak Stationarity라는 개념을 사용해서 가정

Stationarity

Weak Stationarity

Weak Stationarity

시계열은 1차 모멘트와 2차 모멘트가
시간 원점의 변화에 영향을 받지 않는 경우
공분산 안정적(Covariance Stationary)이라고 함

- 평균과 분산 : 동일
- Correlation와 Covariance : 시간의 거리에 의존

시계열 분석에 있어서는 Weak Stationarity가 stationarity의 개념으로 이용

시계열 모형

White Noise

백색잡음(White Noise)

전도체 내부 전자들의 열에 따른 불규칙한 움직임

주파수 스펙트럼, Frequency 도메인으로 전환하면
흰색 빛과 동일한 Frequency를 가져서
백색잡음이라고 불림

시계열 모형

White Noise

백색잡음(White Noise)

- 시계열 모형의 제일 마지막 부분의 잔차 또는 Residual로 묘사
- $E(Y_t) = \mu, Var(Y_t) = \sigma^2$: 고정된 평균과 분산
- $\gamma\_s = 0, s \neq 0$: 자기상관 (Autocorrelation) = 0

시계열 모형

White Noise

Gaussian White Noise

» Normal 분포를 따르면서 평균 0, 분산 σ^2인 Random Variable

Residual

- 모형으로 설명이 안되는 부분
- 제거 불가능
- 모형이 타겟을 완벽하게 설명할 경우 잔차(Residual)은 백색잡음

완벽한 모형은 불가능,
Residual이 백색잡음으로 나타나기 어려움

Residual이 White Noise로 증명이 된다면 매우 훌륭한 모형

시계열 모형

Auto Regressive

AR (Auto Regressive)

$$Y_t = \mu + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$
$$= \mu + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t$$

과거 p 시점의 데이터까지 의존

시계열 모형

Auto Regressive

특성방정식의 모든 해가 단위원 외부에 존재할 경우, AR(p)가 정주성을 지님

$\mathrm{AR}(1)\ |\phi_1| \leq 1$ → Stationarity

- 대부분 1에 가까운, 0.9를 넘는 값
- Serial Defendence가 강하고 Nonstationarity에 근접

시계열 모형

Auto Regressive

AR(p)모형에서 최적의 p값 도출 방법

- ACF(Autcovariance Function)
- PACF(Partial Autocovariance Function)
- 두가지 방법을 통해 시각적으로 파악 가능

시계열 모형

Auto Regressive

? Stationarity가 중요한 이유

Nonstationarity

- 통계적 특성을 활용할 수 없어 분석이 불가능
- Stationarity 데이터로 변환해야 함
- Nonstationarity 할 경우 수렴하지 않고 발산

Stationarity를 확보하는 것이 매우 중요

시계열 모형

Moving Average

MA(Moving Average)

잔차의 값으로 함수를 표현하는 것

- 고정된 평균과 분산을 지님
- 시간격차 q까지는 자기공분산 ≠ 0, 그 이후는 전부 0
- MA(q)는 Weak Stationarity 조건을 충족함

시계열 모형

ARMA(p,q) and ARIMA

ARMA(p,q)

AR(p) + MA(q)

ARIMA(p,d,q)

AR(p) + Integrated + MA(q)

시계열 모형

SARIMA

SARIMA

Seasonal ARIMA(p,d,q)

» 실물경제 관련 데이터는 대부분 계절성을 지님

예 Monthly Pattern, Weekly Pattern 등

시계열 모형

ARMAX

ARMAX

ARMA + 외생적인 설명변수

- 외생적인 설명 변수를 추가하여 모형의 설명력 높여줌
- Multivariate Time-series Analysis 성격을 포함

시계열 모형

State Space Model, Kalman Filter

State Space Model & Kalman Filter

State Evolution 프로세스 도입

- Time-series는 State의 변함에 따라 변동폭이 크거나 또는 안정적인 Series를 Generate함
- 두 가지 경우를 혼합시켜 놓으면 완전히 다른 두 체제의 Series같지만 설명이 가능

> 변동폭이 큰 시계열을 설명하는 모형으로 많이 활용

시계열 모형

GARCH(p,q)

GARCH(p,q)

(Generalized Autoregressive Conditional Heteroskedasticity)

AR모형

이분산성

시간에 따른 변동성을 모형화하기 위해 도입

Conditional Variance

- Stock Return 모형에서 많이 활용
- 수익률 변동성군집현상을 어느 정도 설명
- 통상 p.q의 값은 GARCH(1,1)

시계열 모형

Exponential Moving Average

Exponential Moving Average

(EMA 또는 Exponentially Weighted Average:EWMA)

- Industry에서 Practice가 많이 되는 모형
- 시간이 가까우면 가정치를 많이 주고 시간이 멀면 조금만 반영
- 설명력이 높아 많이 활용

시계열 모형의 특징

Wold's Decomposition Theorem

Wold's Decomposition Theorem

- 정주성 시계열은 상관관계를 가지지 않는 두 개의 확률과정의 합으로 표현 가능
- 한 확률과정은 순전히 비확률적(i.e., 확정적) 요소만으로 구성
- 또 다른 확률 과정은 순전히 확률적 요소만으로 구성

시계열 모형의 특징

Partial Autocovariance

Partial Autocovariance

- 두 변수간의 상관관계 측정 시
  모든 변수의 영향을 제거 후 상관관계 측정

예 시간 간격이 3인 PACF일 경우

Y_{t-1} 과 Y_{t-2} 의 영향을 제거한 뒤
Y_t 와 Y_{t-3} 간 상관관계를 측정

- 시간간격이 1인 경우 ACF 와 PACF는 일치

시계열 모형의 특징

Invertibility Condition

Invertibility Condition

다음 조건 충족 시 MA(q) 모형을 AR(∞)로 표현 가능

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

- 이 방정식의 모든 해가 단위원 외부에 존재할 때 변환 가능

Stationarity 조건을 만족하면
변환이 가능하다는 것을 의미

시계열 모형의 특징

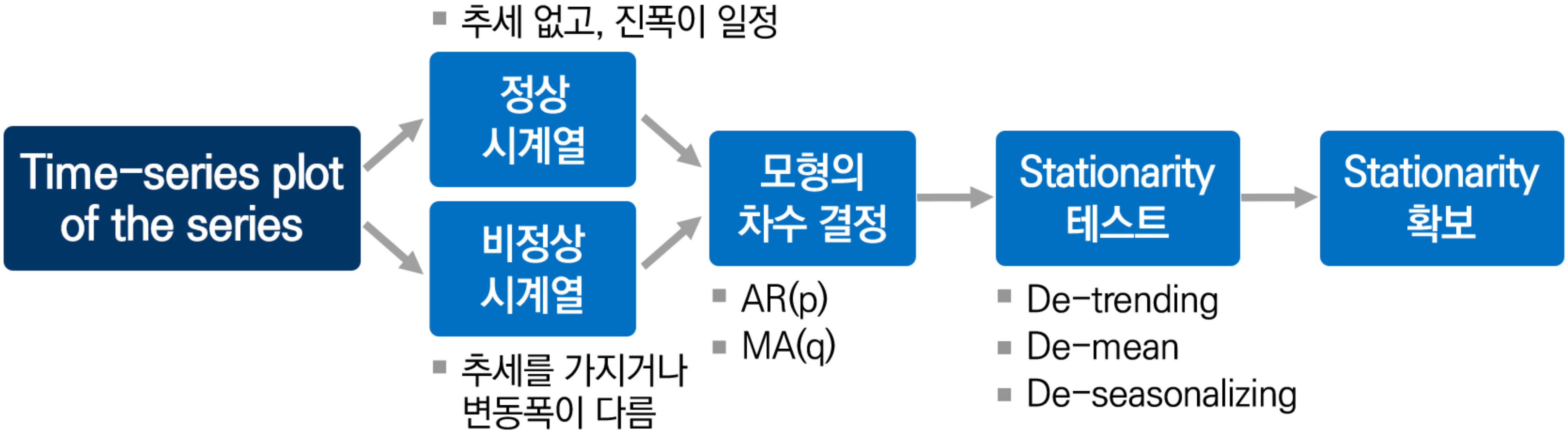
ACF, PACF의 시계열별 특성

	ACF	PACF
AR(p)	기하급수적으로 감소	$PACF(s) \neq 0,\ s \leq q$ $\boxed{PACF(s) = 0,\ s > q}$
MA(q)	$ACF(s) \neq 0,\ s \leq q$ $\boxed{ACF(s) = 0,\ s > q}$	기하급수적으로 감소
ARMA(p,q)	기하급수적으로 감소	기하급수적으로 감소

시계열 모형의 특징

시계열 분석의 순서

1 모형 식별(Model Identification)



시계열 모형의 특징

시계열 분석의 순서

2 모형 Specify

- AR,MA,또 다른 GARCH 모형 등

3 추정(Estimation)

- 최대우도법
  (Maximum Likelihood Estimation, MLE)
- 적률추정법
  (Method of Moments Estimation, MME)

4 모형진단(Diagnostic Checking)

- Residual error terms
- Independency of error terms
- Constant error variance (Homoscedasticity)

- Error Term이 White Noise이면 완벽하고 훌륭한 모형

시계열 모형의 특징

시계열 분석의 순서

5 예측(Forecasting)

- 트레이닝셋, 테스트셋, 데이터셋을 구분 후 **테스트셋을 통해 예측**
- 여러 가지 척도를 통해 평가

시계열 모형의 특징

Parsimonious Model

절약형 모델(Parsimonious Model)

하기 쉬운 실수 : Feature의 수를 늘리는 것

" 좋은 모형이란, 기본적인 절약형 모형 "

시계열 모형의 특징

Parsimonious Model

절약형 모델(Parsimonious Model)

Overfitting 문제 해결 가능

- Time-series에서 적절한 lag 수를 정하지 않으면 Overfitting 문제 야기
- 적절한 lag수 결정시 ACF나 PACF 등을 사용해도 효과 미비

시계열 모형의 특징

모형 선택 기준

모형 선택 기준

잔차제곱합의 함수

- 비용 함수

패널티 항

- Lag수 증가시 비용이 발생하게끔 그 항을 Specify

위 기준을 가지고 Parameter를 추정하면 **최적의 래그 수** 결정 가능

시계열 모형의 특징

모형 선택 기준

- Akaike's Information Criteria
- Schwarz's Bayesian Information Criteria
- Hannan-Quinn Information Criteria