시계열자료분석팀

5팀

김민우
김영호
정승연
조건우
조웅빈

CONTENTS

1

2주차 복습

2주차 복습

선형과정

Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} := \psi(B) Z_t$$

백색잡음들의 선형결합이 선형과정이며,

정상 확률 과정의 선형 결합은 또 다시 정상 확률 과정

2주차 복습

AR, MA, ARMA

정상성과 가역성

	AR(p)	MA(q)	ARMA(p,q)
정상성	조건필요	자체만족	조건필요
가역성	자체만족	조건필요	조건필요

2주차 복습

AR, MA, ARMA

모형의 ACF와 PACF

(감시중)

	AR(p)	MA(q)	ARMA(p,q)
ACF	지수적으로 감소	q+1차부터 절단	지수적으로 감소
PACF	p+1차부터 절단	지수적으로 감소	지수적으로 감소

2

ARIMA

정의

자기회귀 누적이동 평균과정

D차 차분한 시계열 X_t가 ARMA(p,q)를 따름 = ARIMA(p,d,q)를 따름

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t$$

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d X_t = \left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

2 ARIMA

정의

자기회귀 누적이동 평균과정

D차 차분된 ARMA(p,q)를 따름 = ARMA(p,q)를 따름

ARIMA에서 I는 과연 무엇을 의미할까요?

정답은 누적($Integration$)을 의미합니다!

분명 ARMA에 차분(differencing)이 추가된 모형인데, 왜 ARDMA가 아니라, ARIMA일까요?

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t$$

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d X_t = \left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

정의

분명 ARMA에 차분(differencing)이 추가된 모형인데,

자기회귀 누적이동 평균과정

왜 ARDMA가 아니라, ARIMA일까요?

D차 차분한 시계열 X_t가 ARMA(p,q)를 따름

$$\phi(B)(1-B)X_t = \theta(B)Z_t$$

$$Y_t = (1-B)X_t = X_t - X_{t-1}$$

ARIMA(p,d,q)를 따름

$$X_t = X_{t-1} + Y_t = (X_{t-2} + Y_{t-1}) + Y_t = \cdots = X_0 + \sum_{j=1}^{t} Y_j$$

위와 같이, X_t의 관점에서 Y_t의 누적합이 되어

Integration이라는 용어를 사용하는 것!

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d X_t = \left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

모형적합절차

시계열, ACF 그래프로 정상성 판단

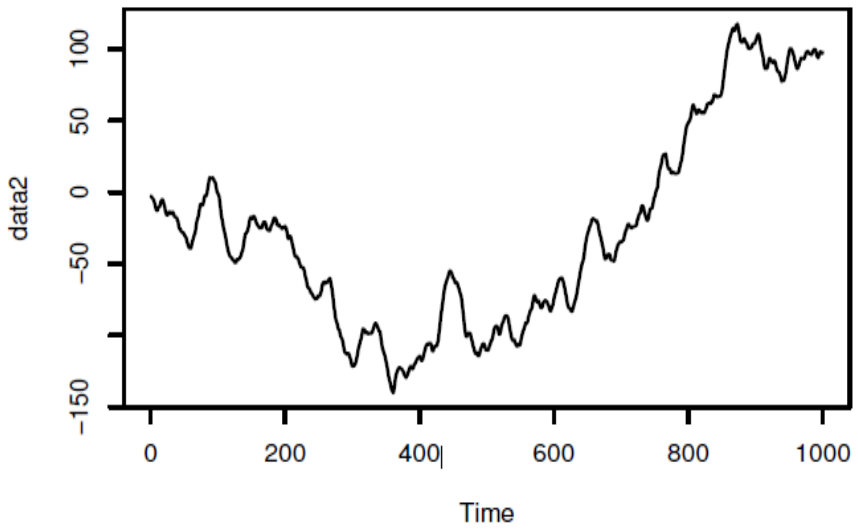
비정상일 경우 차분을 통해 정상화

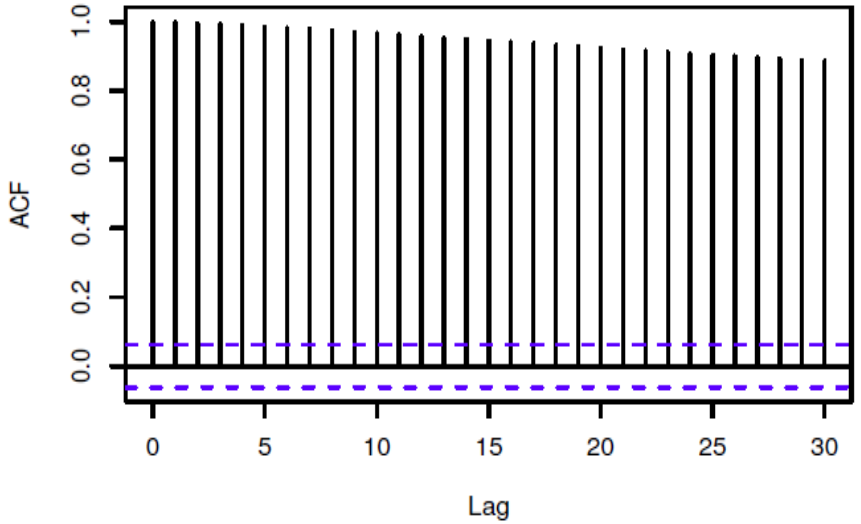
모수 추정 및 모형진단

최종모형 선택 및 **예측** 진행

모형적합절차

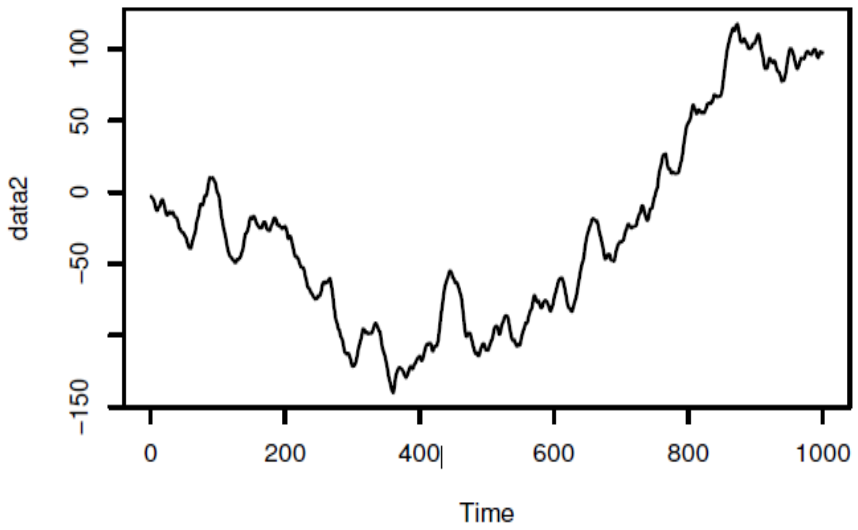
시계열, ACF 그래프로 정상성 판단

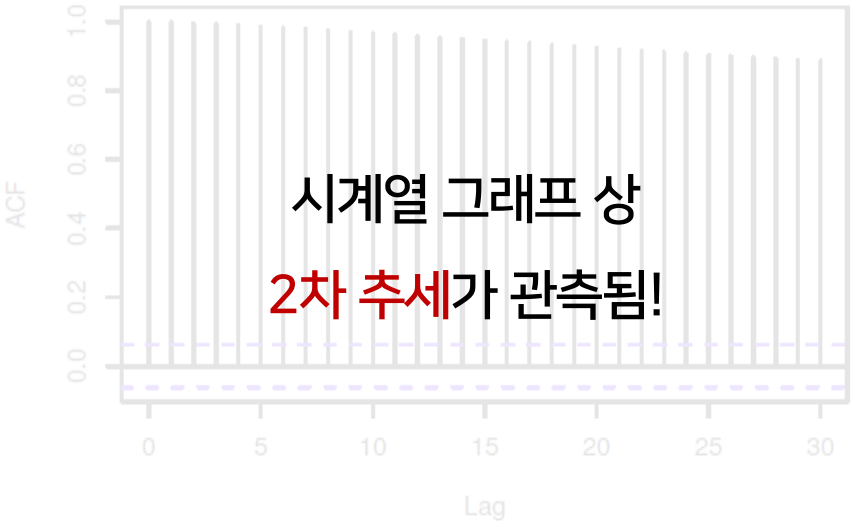




모형적합절차

시계열, ACF 그래프로 정상성 판단





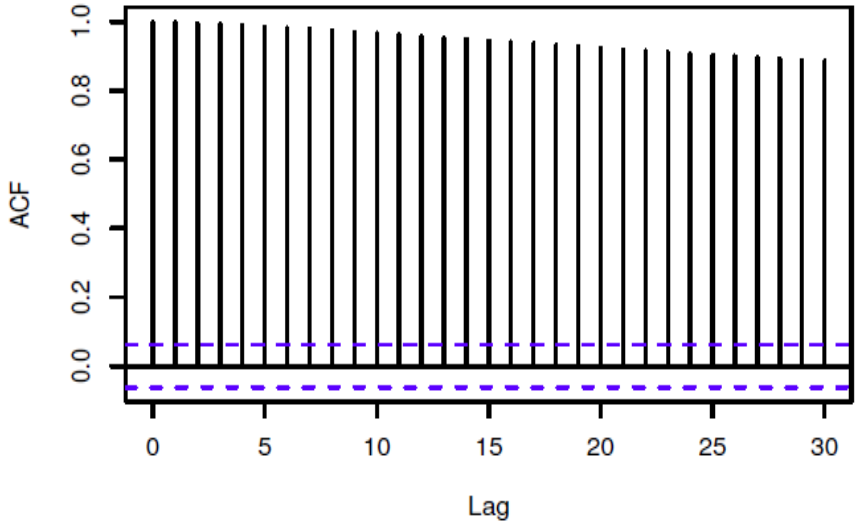
시계열 그래프 상

2차 추세가 관측됨!

모형적합절차

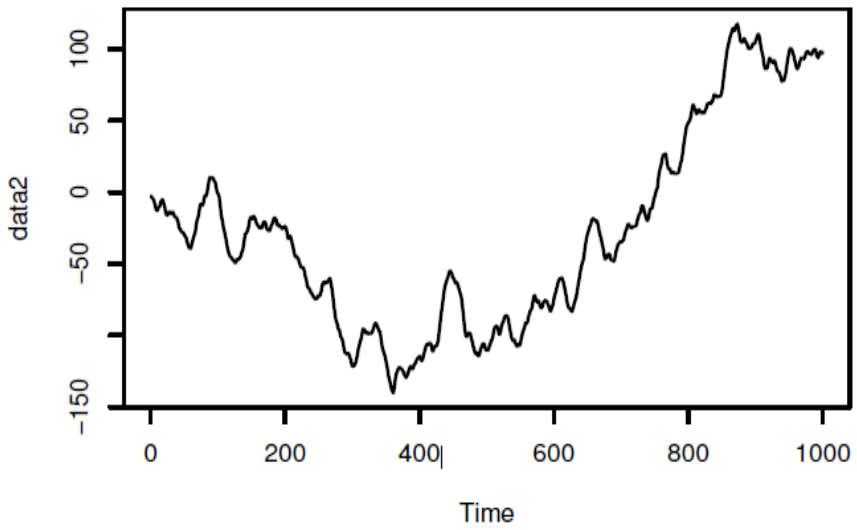
시계열, ACF 그래프로 정상성 판단

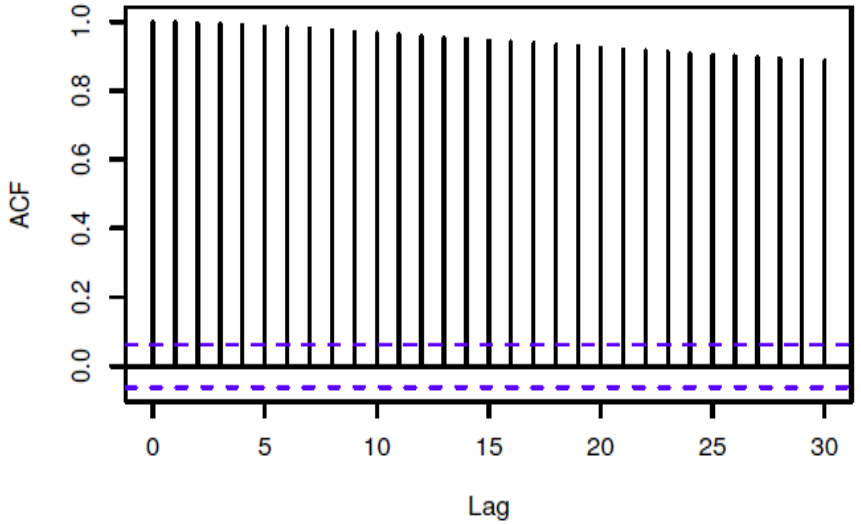
ACF가 linear decay!



모형적합절차

시계열, ACF 그래프로 정상성 판단

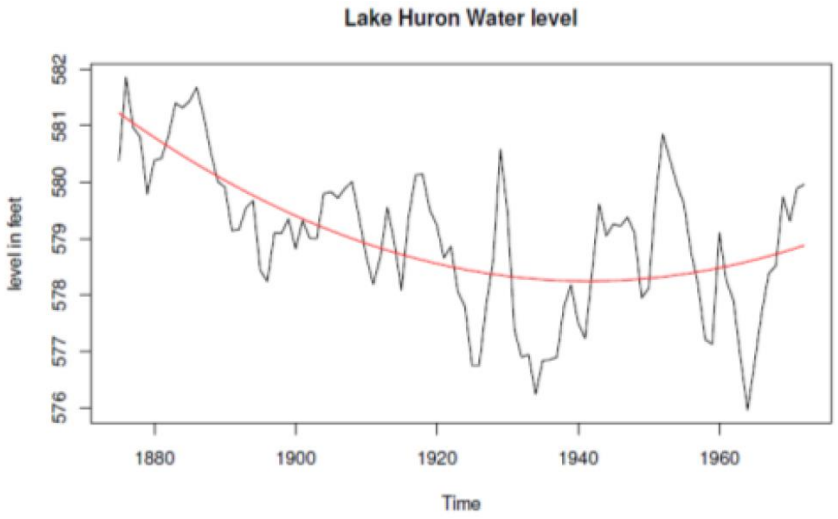


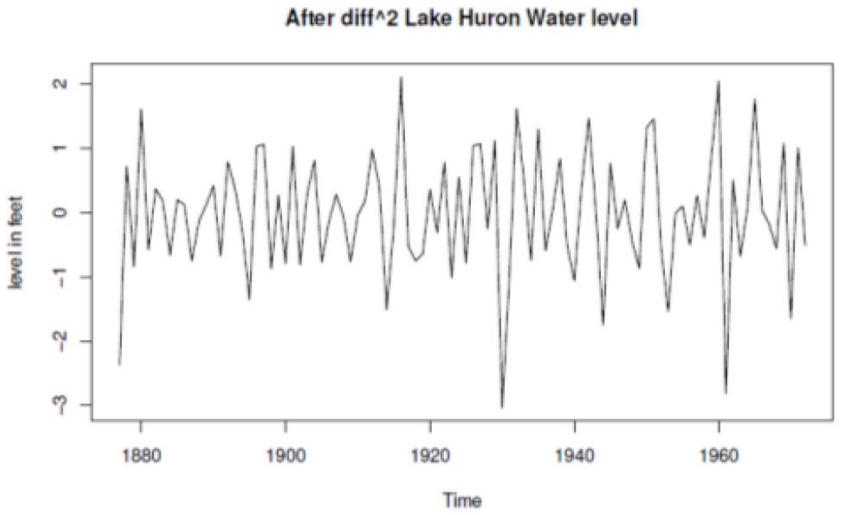


정상성을 만족하지 않는 비정상시계열로 판단!

모형적합절차

비정상일 경우 차분을 통해 정상화



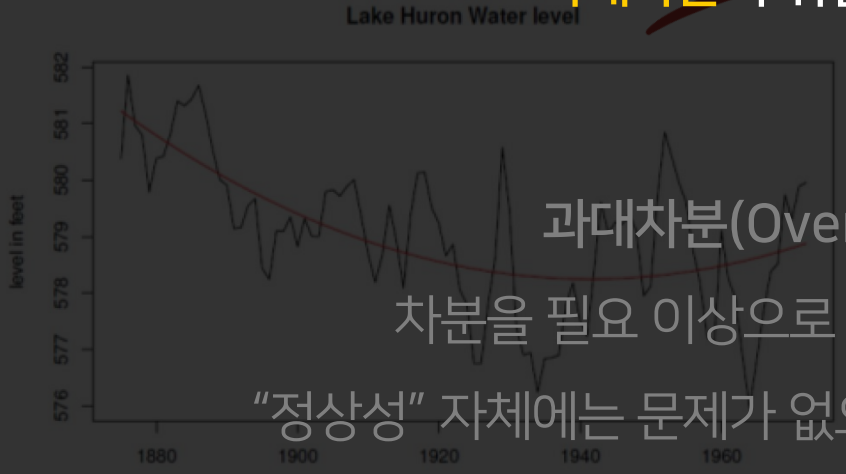


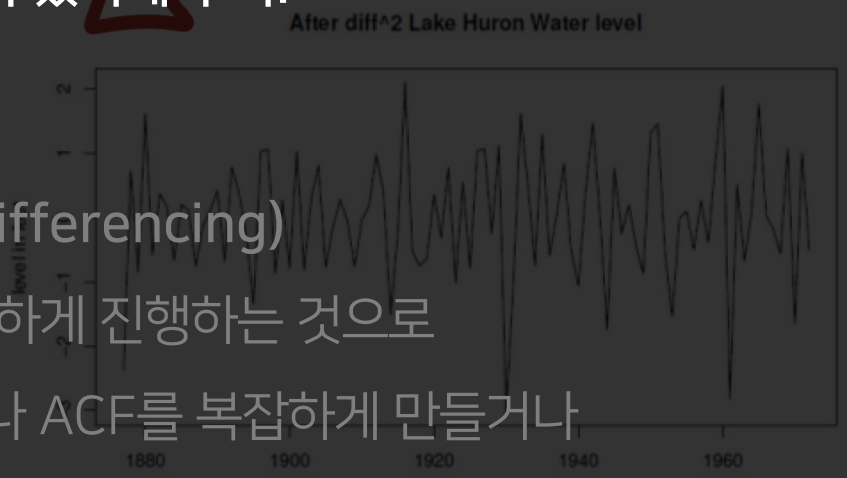
모형적합절차

비정상일 경우 차분을 통해 정상화

차분의 차수 d가 1, 2를 넘어가게 된다면

과대차분의 위험이 있기에 주의!





과대차분(Overdifferencing)

차분을 필요 이상으로 과하게 진행하는 것으로

"정상성" 자체에는 문제가 없으나 ACF를 복잡하게 만들거나

분산이 커지며, 시계열 자료의 불필요한 상관관계를

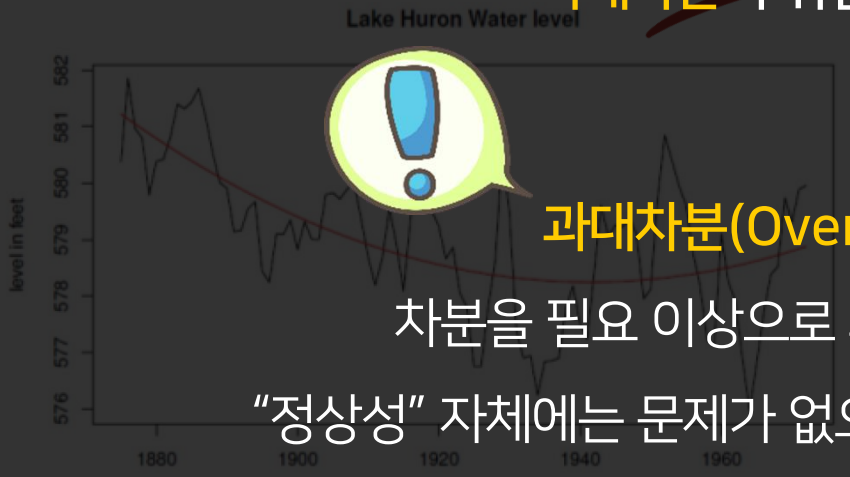
만들고 모형 적합 과정을 복잡하게 만드는 문제가 발생

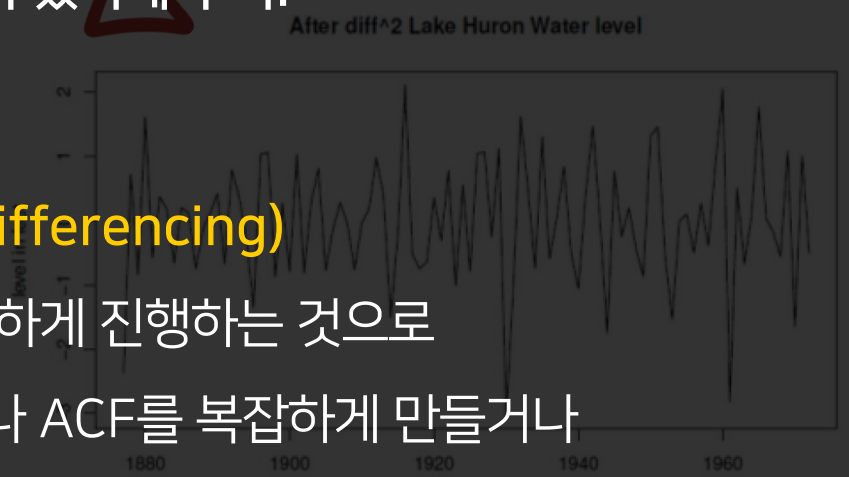
모형적합절차

비정상일 경우 차분을 통해 정상화

차분의 차수 d가 1, 2를 넘어가게 된다면

과대차분의 위험이 있기에 주의!





과대차분(Overdifferencing)

차분을 필요 이상으로 과하게 진행하는 것으로

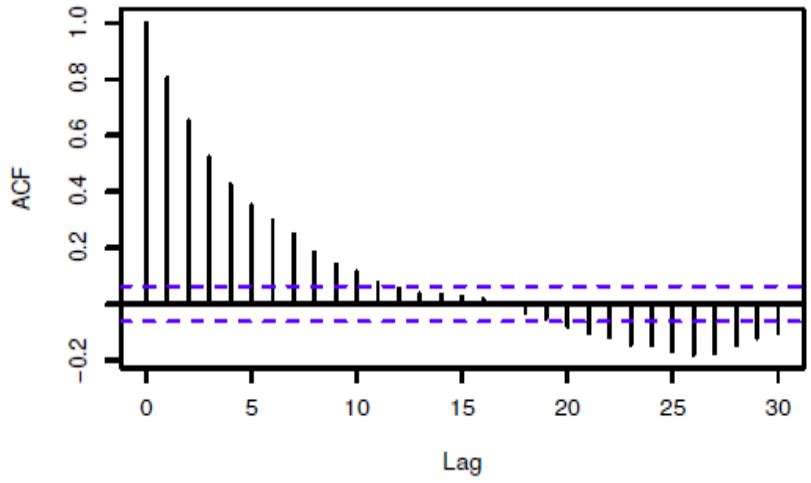
"정상성" 자체에는 문제가 없으나 ACF를 복잡하게 만들거나

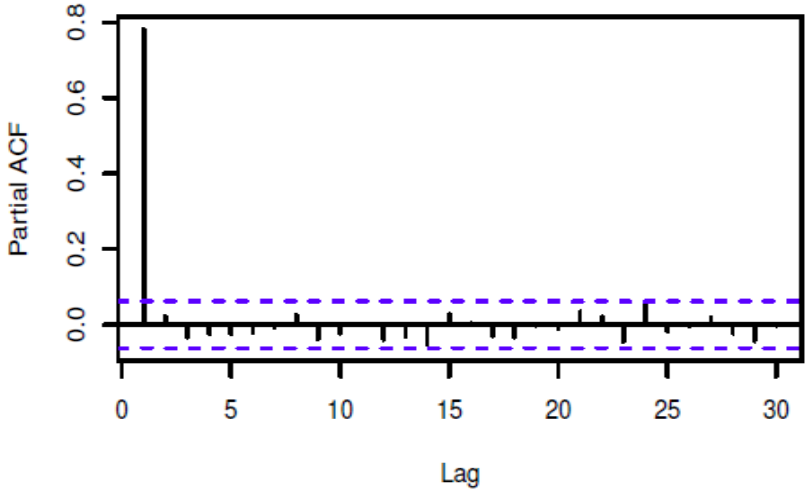
분산이 커지며, 시계열 자료의 불필요한 상관관계를

만들고 모형 적합 과정을 복잡하게 만드는 문제가 발생

모형적합절차

모수 추정 및 모형진단



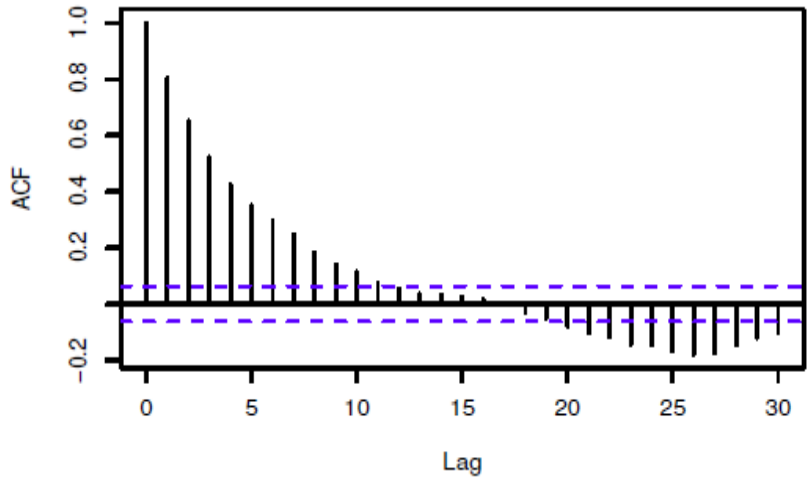


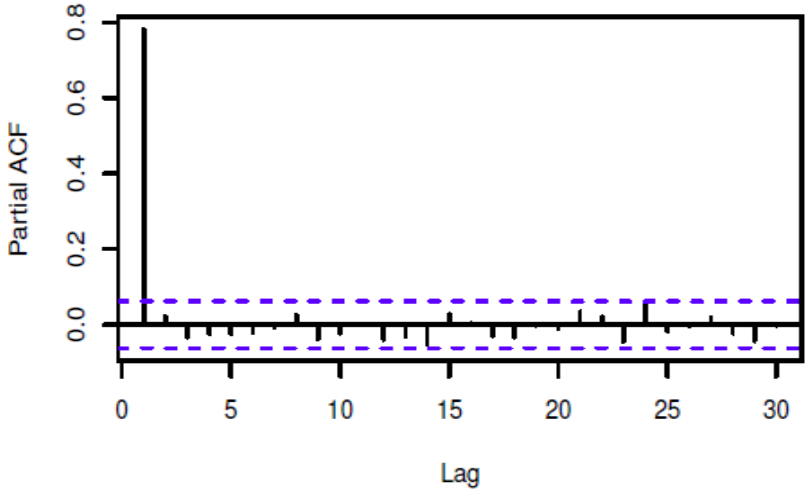
차분 이후의 ACF가 지수적으로 감소하며, PACF는 lag 2부터는 절단되어

AR(1) 모형이 적합하다고 판단할 수 있음!

모형적합절차

모수 추정 및 모형진단





이때까지의 모형적합절차를 모두 일컬어

ARIMA(1,1,0)이 적합하다고 표현!

3

SARIMA

SARIMA의 흐름

ex) 주기가 s=12인 시계열 데이터

	Month 1	Month 2	⋯	Month 12
Year 1	Y_1	Y_2	⋯	Y_{12}
Year 2	Y_{13}	Y_{14}	⋯	Y_{24}
⋯	⋮	⋮	⋯	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	⋯	$Y_{12+12(r-1)}$

확률적인 계절성분을 고려하는 모델

SARIMA의 흐름

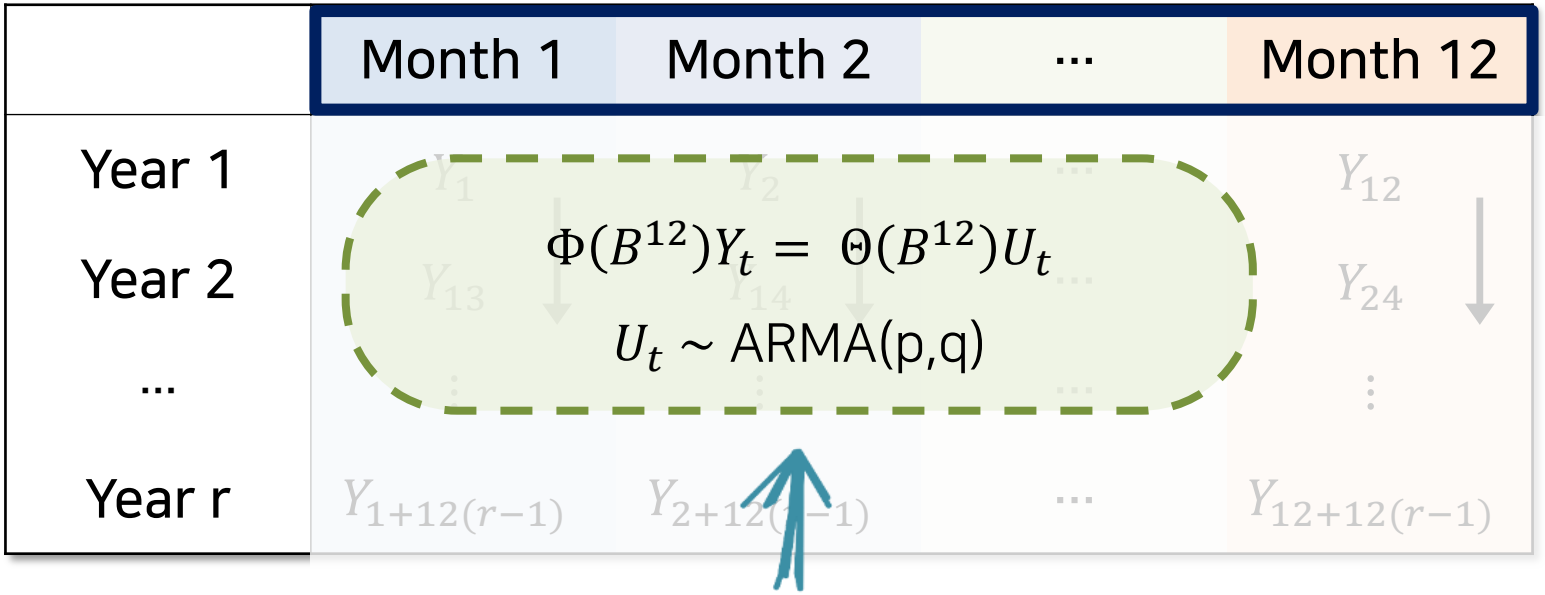
ex) 주기가 s=12인 시계열 데이터

	Month 1	Month 2	⋯	Month 12
Year 1	Y_1	Y_2	⋯	Y_{12}
Year 2	Y_{13}	Y_{14}	⋯	Y_{24}
⋯	⋮	⋮	⋯	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	⋯	$Y_{12+12(r-1)}$

column을 고정했을 때 시계열들이 ARMA(P,Q)를 따른다고 가정

SARIMA의 흐름

ex) 주기가 s=12인 시계열 데이터



한 주기 내에도 Correlation이 존재 가능하기 때문에

오차항이 ARMA(p,q)를 따르는 시계열이라고 가정

SARIMA의 흐름

ex) 주기가 s=12인 시계열 데이터

	Month 1	Month 2	⋯	Month 12
Year 1	Y_1	Y_2	⋯	Y_{12}

STEP 1과 STEP 2를 합치면

Seasonal ARMA model : SARMA(p,q) x (P,Q)

$$\Phi(B^{12})Y_t = \Theta(B^{12})\phi^{-1}(B)\theta(B)Z_t$$

$$\phi(\mathrm{B})\Phi(\mathrm{B}^{12})\mathrm{Y_t} = \theta(B)\Theta(B^{12})Z_t$$

$$Z_t \sim WN(0,\sigma^2)$$

차분을 추가하면 SARIMA!

SARIMA

전통적 분해법 ⇒ 결정적 계절성분을 가정

SARIMA ⇒ 확률적 계절성분을 가정

SARIMA 종류

1. 순수 SARIMA ⇒ 계절적 성분만 고려
2. 승법 SARIMA ⇒ 비계절적 & 계절적 성분 모두 고려

추세와 계절성이 모두 존재할 때 사용 가능

순수 SARIMA

주기가 s인 순수 SARIMA

$$\Phi(B^s)(1-B^s)^D X_t = \Theta(B^s)Z_t$$

$$\Phi(\mathrm{B}^s) = (1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$$

$$\Theta(\mathrm{B}^s) = (1+\Theta_1 \mathrm{B}^s + \Theta_2 \mathrm{B}^{2s} + \cdots + \Theta_\mathrm{Q} \mathrm{B}^{Qs})$$

$$\mathrm{Z_t} \sim WN(0, \sigma^2)$$

순수 SARIMA

주기가 s인 순수 SARIMA

$$\Phi(B^s)(1-B^s)^D X_t = \Theta(B^s)Z_t$$

$$\Phi(\mathrm{B}^s) = (1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$$

$$\Theta(\mathrm{B}^s) = (1+\Theta_1 \mathrm{B}^s + \Theta_2 \mathrm{B}^{2s} + \cdots + \Theta_\mathrm{Q} \mathrm{B}^{Qs})$$

$$\mathrm{Z_t} \sim WN(0, \sigma^2)$$

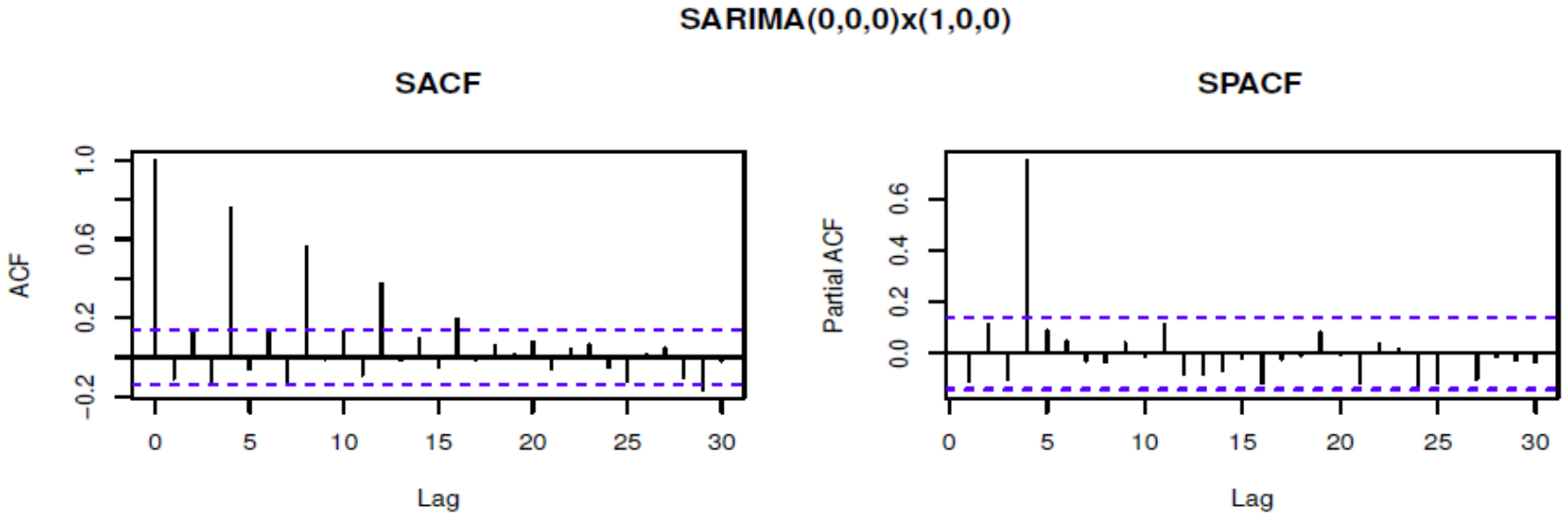
계절성만을 고려하는 모델

P개의 과거 관측치와 D번의 계절차분, Q개의 과거 오차항으로

현재 관측치를 설명하는 모형

순수 SARIMA

$$(1 - .8B^4)X_t = Z_t.$$

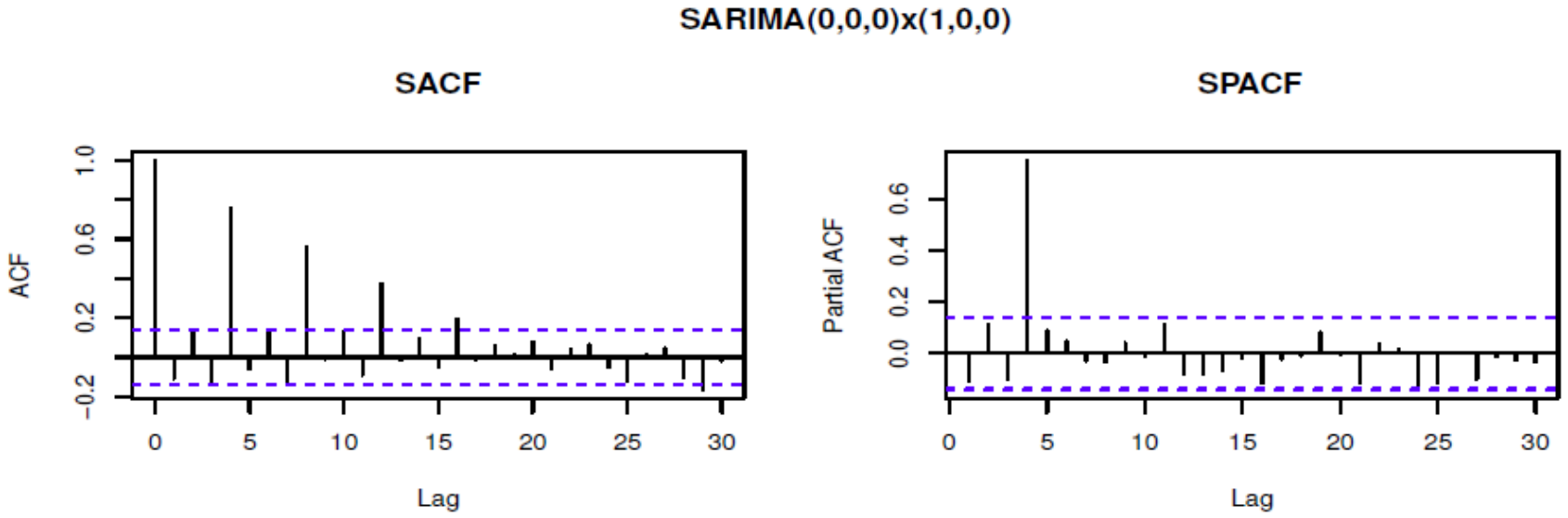


순수 SARIMA모형의 ACF/PACF는

s에 해당되는 계절 주기에서만 0이 아니고, 다른 시차에서는 0임

순수 SARIMA

$$(1 - .8B^4)X_t = Z_t.$$



모든 시차가 아닌 주기에 해당하는 시차를 집중해서 관측해야 함!

승법 SARIMA

$$\phi(\mathrm{B})\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)Z_t$$

$$\Phi(\mathrm{B}^s) = (1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$$

$$\Theta(\mathrm{B}^s) = \left(1+\Theta_1 \mathrm{B}^s + \Theta_2 \mathrm{B}^{2s} + \cdots + \Theta_{\mathrm{Q}} \mathrm{B}^{Qs}\right)$$

$$\phi(\mathrm{B}) = \left(1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)$$

$$\theta(B) = \left(1+\theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)$$

계절 이외의 요소의 연관성도 고려하는 모형

순수 SARIMA는 오차가 백색잡음이지만,

승법 SARIMA는 오차가 ARMA (혹은 ARIMA)를 따름

승법 SARIMA

$$\phi(\mathrm{B})\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)Z_t$$

$$\Phi(\mathrm{B}^s) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$$

$$\Theta(\mathrm{B}^s) = (1 + \Theta_1 \mathrm{B}^s + \Theta_2 \mathrm{B}^{2s} + \cdots + \Theta_\mathrm{Q} \mathrm{B}^{Qs})$$

모수의 절약 측면에서 순수 SARIMA보다

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

승법 SARIMA를 더 자주 사용!

$$\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)$$

계절 이외의 요소의 연관성도 고려하는 모형

순수 SARIMA는 오차가 백색잡음이지만,

승법 SARIMA는 오차가 ARMA (혹은 ARIMA)를 따름

4

이분산 시계열모형

이분산 시계열모형

이분산 시계열모형의 필요성

지금까지 살펴본 시계열 모형

1. 시간의 따른 분산의 변화가 없다고 가정 ⇒ 등분산성을 가정
2. 주로 평균 부분의 움직임에 관심을 갖는 모형

4 이분산 시계열모형

이분산 시계열모형의 필요성

지금까지 살펴본 시계열 모형

1. 시간의 따른 분산의 변화가 없다고 가정 ⇒ 등분산성을 가정
2. 주로 평균 부분의 움직임에 관심을 갖는 모형

But! 금융 관련 시계열의 경우 분산이 과거자료에 의존

↓

시간에 따른 이분산성을 분석하기 위한 모형 필요

이분산 시계열모형

변동성

통계학

시간에 따른 이분산성

경제학

변동성 (Volatilty)

조건부 분산으로 측정가능

이분산 시계열모형

변동성

통계학

시간에 따른 이분산성

경제학

변동성 (Volatilty)

조건부 분산으로 측정가능

이분산 시계열모형은 조건부 분산을 시간의 함수로 표현하는 시계열 모형

4 이분산 시계열모형

변동성

통계학 | 조건부 분산이란? | 경제학

시간에 따른 이분산성 $Var(r_t|\mathcal{F}_{t-1})$ 변동성 (Volatilty)

Where $\mathcal{F}_{t-1}$ is σ-field generated by historical information

과거의 정보의 영향에 의한 분산

조건부 분산으로 측정가능

이분산 시계열모형은 조건부 분산을

시간의 함수로 표현하는 시계열 모형

4 이분산 시계열모형

변동성

통계학 — 시간에 따른 이분산성

조건부 분산이란? $Var(r_t|\mathcal{F}_{t-1})$

경제학 — 변동성 (Volatilty)

Where $\mathcal{F}_{t-1}$ is σ-field generated by historical information

과거의 정보의 영향에 의한 분산

조건부 분산으로 측정가능

조건부 등분산성

$Var(r_t|\mathcal{F}_{t-1})$=constant

과거의 영향에 무관하게 분산이 일정!

VS

조건부 이분산성

$Var(r_t|\mathcal{F}_{t-1}) \neq$ constant

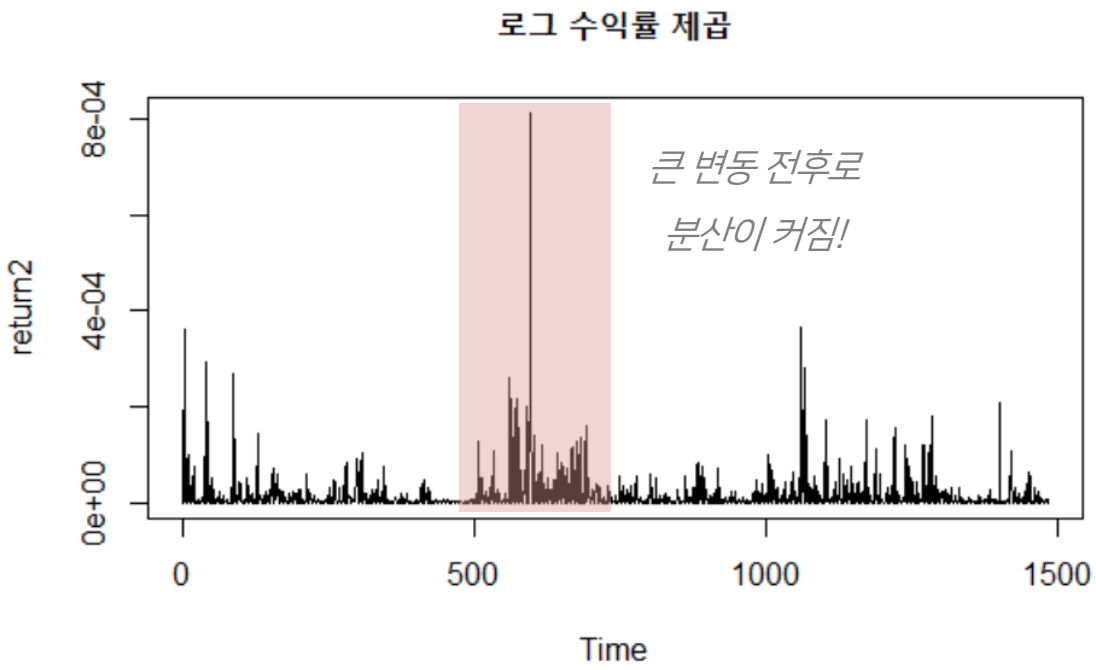
과거의 영향을 받아 분산이 변화!

이분산 시계열모형은 조건부 분산을

시간의 함수로 표현하는 시계열 모형

이분산 시계열모형

변동성 집중

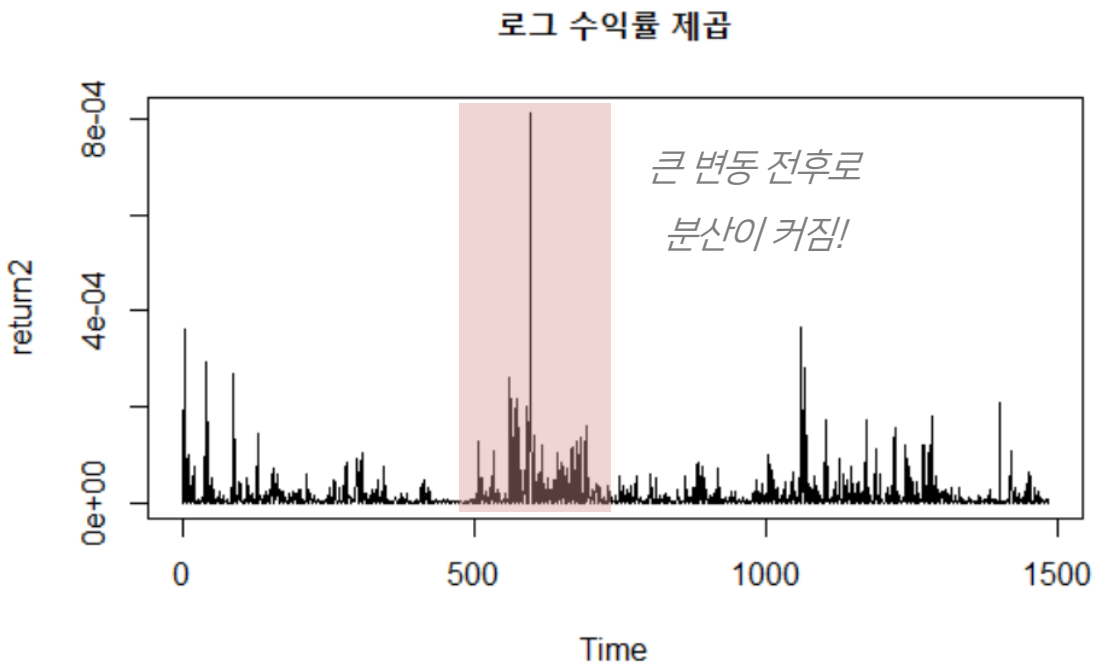


분산(σ_t^2)에 자기상관이 존재

조건부 이분산성의 증거!

이분산 시계열모형

변동성 집중



변동성 집중은 분산 안정화 변환(VST)으로 상쇄 불가

σ_t^2에 대한 모형 필요

이분산 시계열모형

ARCH 모형

AutoRegressive Conditional Heteroskedasticity

ARCH(p)

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i {\varepsilon^2}_{t-i}$$

$$= \alpha_0 + \alpha_1 {\varepsilon^2}_{t-1} + \alpha_2 {\varepsilon^2}_{t-2} + \cdots \alpha_p {\varepsilon^2}_{t-p}$$

$$\alpha_0 > 0, \alpha_j \geq 0, j = 1,2, \ldots, p$$

t시점의 오차의 변동성(조건부 분산) σ_t^2을

과거시점들의 오차항으로 설명하는 비선형 모형

ARCH 모형

ARCH모형의 비선형성

ARCH(1)

$$\varepsilon_t^2 = \sigma_t^2 Z_t^2 = (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) Z_t^2$$

$$= \alpha_0 Z_t^2 + \alpha_1 Z_t^2 \{(\alpha_0 + \alpha_1 \varepsilon_{t-2}^2) Z_{t-1}^2\}$$

$$= \alpha_0 Z_t^2 + \alpha_0 \alpha_1 Z_t^2 Z_{t-1}^2 + \alpha_1^2 \varepsilon_{t-2}^2 Z_t^2 Z_{t-1}^2$$

$$= \alpha_0 \sum_{j=0}^{n} (\alpha_{1C}^{j} Z_t^2 Z_{t-1}^2 \cdots Z_{t-j}^2) + \alpha_1^{n+1} \varepsilon_{t-n-1}^2 Z_t^2 Z_{t-1}^2 \cdots Z_{t-j}^2$$

오차항의 곱으로 이루어진 형태라는 점에서 비선형적 모형

4 이분산 시계열모형

ARCH 모형

추정과 검정

추정 (최대우도추정법)MLE

검정 $H_0 : \alpha_1 = \cdots = \alpha_p = 0$

Ljung- Box Q 검정

ε_t^2이 AR(p)모형을 따름을 이용하여 ε_t^2의 자기상관계수가 유의한지 여부를 판단하는 검정 방법

Engle's ARCH test

LM(Largrange Multiplier)에 기반하여, ε_t가 $i.i.d\ (0, \sigma^2)$을 따른다는 귀무가설 하에서 검정통계량 $F = \frac{(SST-SSR)/p}{SSR/(T-p-p-1)}$이 점근적으로 $\chi^2(p)$를 따른다는 사실을 활용하여 검정

오차항의 정규성 검정

QQ-plot, Jarque-Bera test

정규성 검정에 대한 자세한 내용은 회귀분석팀 2주차 클린업 참고!

GARCH 모형

GARCH(Generalized AutoRegressive Conditional Hetroskedasticity)

ARCH(p)

$$\sigma_t^2 = \alpha_0 + \alpha_1 {\varepsilon^2}_{t-1} + \alpha_2 {\varepsilon^2}_{t-2} + \cdots \alpha_p {\varepsilon^2}_{t-p}$$

이분산 시계열모형

GARCH 모형

GARCH(Generalized AutoRegressive Conditional Hetroskedasticity)

ARCH(p)

$$\sigma_t^2 = \alpha_0 + \alpha_1 {\varepsilon^2}_{t-1} + \alpha_2 {\varepsilon^2}_{t-2} + \cdots \alpha_p {\varepsilon^2}_{t-p}$$

p가 커지면 추정량의 정확도 떨어짐

GARCH 모형

GARCH(Generalized AutoRegressive Conditional Hetroskedasticity)

ARCH(p)

$$\sigma_t^2 = \alpha_0 + \alpha_1 {\varepsilon^2}_{t-1} + \alpha_2 {\varepsilon^2}_{t-2} + \cdots \alpha_p {\varepsilon^2}_{t-p}$$

p가 커지면 추정량의 정확도 떨어짐

ARCH 모형의 정상성 조건 만족시키지 못할 가능성 ↑

GARCH 모형

GARCH(Generalized AutoRegressive Conditional Hetroskedasticity)

ARCH(p)

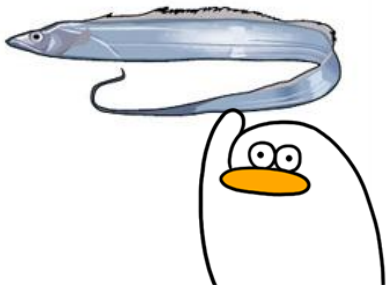
$$\sigma_t^2 = \alpha_0 + \alpha_1 {\varepsilon^2}_{t-1} + \alpha_2 {\varepsilon^2}_{t-2} + \cdots \alpha_p {\varepsilon^2}_{t-p}$$

p가 커지면 추정량의 정확도 떨어짐

ARCH 모형의 정상성 조건 만족시키지 못할 가능성 ↑

일반화된 GARCH 모형 도입

팔딱팔딱 살아 숨쉬는 GARCH가 왔어요~



GARCH 모형

Generalized AutoRegressive Conditional Hetroskedasticity

GARCH(p,q)

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

$$= \alpha_0 + \alpha_1 \varepsilon^2{}_{t-1} + \alpha_2 \varepsilon^2{}_{t-2} + \cdots \alpha_p \varepsilon^2{}_{t-p} + \beta_1 \sigma^2{}_{t-q} + \cdots + \beta_q \sigma^2{}_{t-q}$$

$$\alpha_0 > 0, \alpha_j \geq 0, \beta_j \geq 0 \quad \& \quad \sum \alpha_j + \sum \beta_j < 1$$

t시점의 오차항의 변동성을 p시점 이전의 오차항의 제곱과

q시점 이전의 변동성으로 설명한 모형

GARCH 모형

모수의 절약

GARCH(1,1)

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

$$\sigma_t^2 = \frac{\alpha_0}{(1 - B\beta)} + \frac{\alpha_1}{(1 - B\beta)} \varepsilon_{t-1}^2$$

$$= \frac{\alpha_0}{(1 - B\beta)} + \alpha_1 (1 + B\beta + B^2\beta^2 + \cdots) \varepsilon_{t-1}^2$$

$$\sigma_t^2 = \frac{\alpha_0}{(1 - B\beta)} + (\alpha_1 \varepsilon_{t-1}^2 + \alpha_1 \beta \varepsilon_{t-2}^2 + \alpha_1 \beta^2 \varepsilon_{t-3}^2 + \cdots)$$

$$= \frac{\alpha_0}{1 - \beta} + \sum_{j=1}^{\infty} \alpha_1 \beta^{j-1} \varepsilon_{t-j}^2$$

GARCH 모형

모수의 절약

GARCH(1,1)

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

$$\sigma_t^2 = \frac{\alpha_0}{(1 - B\beta)} + \frac{\alpha_1}{(1 - B\beta)} \varepsilon_{t-1}^2$$

$$= \frac{\alpha_0}{(1 - B\beta)} + \alpha_1 (1 + B\beta + B^2\beta^2 + \cdots) \varepsilon_{t-1}^2$$

$$\sigma_t^2 = \frac{\alpha_0}{(1 - B\beta)} + (\alpha_1 \varepsilon_{t-1}^2 + \alpha_1 \beta \varepsilon_{t-2}^2 + \alpha_1 \beta^2 \varepsilon_{t-3}^2 + \cdots)$$

$$= \frac{\alpha_0}{1 - \beta} + \sum_{j=1}^{\infty} \alpha_1 \beta^{j-1} \varepsilon_{t-j}^2$$

$ARMA(\infty)$ 로 표현 가능

모형의 쌍대성과 유사!

이분산 시계열모형

GARCH 모형

모수의 절약

GARCH(1,1)

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

GARCH Zoo

Leverage effect, 비정상성, 장기 상관관계 등 문제점 多

$$\sigma_t^2 = \frac{\alpha_0}{(1-B\beta)} + \frac{\alpha_1}{(1-B\beta)} \varepsilon_{t-1}^2$$

$$= \frac{\alpha_0}{(1-B\beta)} + \alpha_1 (1 + B\beta + B^2\beta^2 + \cdots)\varepsilon_{t-1}^2$$

$$\sigma_t^2 = \frac{\alpha_0}{(1-B\beta)} + (\alpha_1 \varepsilon_{t-1}^2 + \alpha_1 \beta \varepsilon_{t-2}^2 + \alpha_1 \beta^2 \varepsilon_{t-3}^2 + \cdots)$$

$$= \frac{\alpha_0}{1-\beta} + \sum_{j=1}^{\infty} \alpha_1 \beta^{j-1} \varepsilon_{t-j}^2$$

$ARMA(\infty)$ 로 표현 가능

모형의 쌍대성과 유사!

이분산 시계열모형

GARCH 모형

모수의 절약

GARCH(1,1)

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

GARCH Zoo

$$\sigma_t^2 = \frac{\alpha_0}{(1-B\beta)} + \frac{\alpha_1}{(1-B\beta)} \varepsilon_{t-1}^2$$

Leverage effect, 비정상성, 장기 상관관계 등 문제점 多

▼

$$= \frac{\alpha_0}{(1-B\beta)} + \alpha_1 (1 + B\beta + B^2\beta^2 + \cdots)\varepsilon_{t-1}^2$$

수많은 Upgrade GARCH 모형 존재

$$\sigma_t^2 = \frac{\alpha_0}{(1-B\beta)} + (\alpha_1 \varepsilon_{t-1}^2 + \alpha_1 \beta \varepsilon_{t-2}^2 + \alpha_1 \beta^2 \varepsilon_{t-3}^2 + \cdots)$$

ex) AGARCH, MGARCH, TGARCH, NGARCH, EGARCH, IGARCH, GJR-GARCH…

-> 데이터에 적합한 GARCH 선택!

$$= \frac{\alpha_0}{1-\beta} + \sum_{j=1}^{\infty} \alpha_1 \beta^{j-1} \varepsilon_{t-j}^2$$

$ARMA(\infty)$ 로 표현 가능

모형의 쌍대성과 유사!

5

ARFIMA

ARFIMA 모형의 필요성

장기기억 보존

차수가 정수인 차분을 시행할 경우

과거의 관측치를 빼 주어 장기 기억을 분석하는 것은 불가능

ARFIMA 모형의 필요성

장기기억 보존

차수가 정수인 차분을 시행할 경우

과거의 관측치를 빼 주어 장기 기억을 분석하는 것은 불가능

실수 차원의 차분을 통해 메모리를 최대한 보존하는 시계열

ARFIMA(Autoregressive Fractionally Integrated Moving Average)

ex) d=0.1, 0.4, 0.5, ···

ARFIMA

ARFIMA 모형의 필요성

장기기억 보존

- 장기기억을 가진 시계열은 ACF가 훨씬 천천히 감소함
- 차분을 여러 번해도 추세가 온전히 제거되지 않을 수 있음
- 차분의 횟수가 늘어나면 추정해야 할 모수 또한 증가

ARFIMA 모형의 필요성

장기기억 보존

- 장기기억을 가진 시계열은 ACF가 훨씬 천천히 감소함
- 차분을 여러 번해도 추세가 온전히 제거되지 않을 수 있음
- 차분의 횟수가 늘어나면 추정해야 할 모수 또한 증가

ARFIMA

ARFIMA 모형

차수 d의 특성

d차 차분을 이항급수 식으로 표현

$$(1-B)^d = 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots$$

ARFIMA 모형

차수 d의 특성

차수 d에 차례대로 0.2, 0.4, 0.6, 0.8, 1을 대입

$$(1-B)^{0.2}X_t = X_t - 0.2X_{t-1} - 0.08X_{t-2} - 0.048X_{t-3} - 0.0336X_{t-4} - \cdots$$

$$(1-B)^{0.4}X_t = X_t - 0.4X_{t-1} - 0.12X_{t-2} - 0.064X_{t-3} - 0.0416X_{t-4} - \cdots$$

$$(1-B)^{0.6}X_t = X_t - 0.6X_{t-1} - 0.12X_{t-2} - 0.056X_{t-3} - 0.0336X_{t-4} - \cdots$$

$$(1-B)^{0.8}X_t = X_t - 0.8X_{t-1} - 0.08X_{t-2} - 0.032X_{t-3} - 0.0176X_{t-4} - \cdots$$

$$(1-B)^{1.0}X_t = X_t - 1.0X_{t-1} - 0X_{t-2} - 0X_{t-3} - 0X_{t-4} - \cdots$$

ARFIMA 모형

차수 d의 특성

차수 d에 따른 특성

d가 0에 가까워질수록 보존되는 데이터의 양은 증가하지만,

non-stationary한 특성은 커짐

d가 1에 가까워질수록 보존되는 데이터의 양은 감소하지만,

stationary한 특성은 커짐

$$(1-B)^{0.2}X_t = X_t - 0.2X_{t-1} - 0.08X_{t-2} - 0.048X_{t-3} - 0.0336X_{t-4} - \cdots$$

$$(1-B)^{0.4}X_t = X_t - 0.4X_{t-1} - 0.12X_{t-2} - 0.064X_{t-3} - 0.0416X_{t-4} - \cdots$$

$$(1-B)^{0.6}X_t = X_t - 0.6X_{t-1} - 0.12X_{t-2} - 0.056X_{t-3} - 0.0336X_{t-4} - \cdots$$

$$(1-B)^{0.8}X_t = X_t - 0.8X_{t-1} - 0.08X_{t-2} - 0.032X_{t-3} - 0.0176X_{t-4} - \cdots$$

$$(1-B)^{1.0}X_t = X_t - 1.0X_{t-1} - 0X_{t-2} - 0X_{t-3} - 0X_{t-4} - \cdots$$

ARFIMA 모형

차수 d의 특성

차수 d에 따른 특성

d가 0에 가까워질수록 보존되는 데이터의 양은 증가하지만,

non-stationary한 특성은 커짐

d가 1에 가까워질수록 보존되는 데이터의 양은 감소하지만,

stationary한 특성은 커짐

$$(1-B)^{0.2}X_t = X_t - 0.2X_{t-1} - 0.12X_{t-2} - 0.064X_{t-3} - 0.0416X_{t-4} - \cdots$$

$$(1-B)^{0.4}X_t = X_t - 0.4X_{t-1} - 0.12X_{t-2} - 0.064X_{t-3} - 0.0416X_{t-4} - \cdots$$

$$(1-B)^{0.6}X_t = X_t - 0.6X_{t-1} - 0.12X_{t-2} - 0.056X_{t-3} - 0.0336X_{t-4} - \cdots$$

$$(1-B)^{0.8}X_t = X_t - 0.8X_{t-1} - 0.08X_{t-2} - 0.032X_{t-3} - 0.0176X_{t-4} - \cdots$$

$$(1-B)^{1.0}X_t = X_t - 1.0X_{t-1} - 0X_{t-2} - 0X_{t-3} - 0X_{t-4}$$

d값이 작아질수록 더 먼 과거의 데이터가 반영되고,

1에 가까워질수록 보존되는 과거 정보의 양이 감소

ARFIMA 모형

ARFIMA(p,d,q)

$$\phi(B)(1-B)^d Y_t = \theta(B)Z_t\,,\ Z_T \sim WN(0,\sigma^2),\ 0 < d < \frac{1}{2}$$

$$\phi(B) = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)$$

$$\theta(B) = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)$$

정상성 조건을 만족하기 위해 $|d| < 0.5$ 를 만족해야 함

d가 0.5보다 커지게 되면, 분산이 ∞로 발산!

ARFIMA

ARFIMA 모형

장기기억이란?

정상성을 만족하는 확률과정 $\{Z_t\}$의 자기상관함수 $\rho(k)$가

ARFIMA(p,d,q)

$0 < d < \frac{1}{2}$인 어떤 실수에 대하여 $\rho(k) \sim Ck^{2d-1}, k \to \infty\ (C > 0)$을 만족

$$\phi(B)(1-B)^d Y_t = \theta(B)Z_t\ ,\ Z_T \sim WN(0,\sigma^2),\ 0 < d < \frac{1}{2}$$

$$\phi(B) = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)$$

$$\theta(B) = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)$$

정상성 조건을 만족하기 위해 $|d| < 0.5$ 를 만족해야 함

d가 0.5보다 커지게 되면, 분산이 ∞로 발산!

ARFIMA 모형

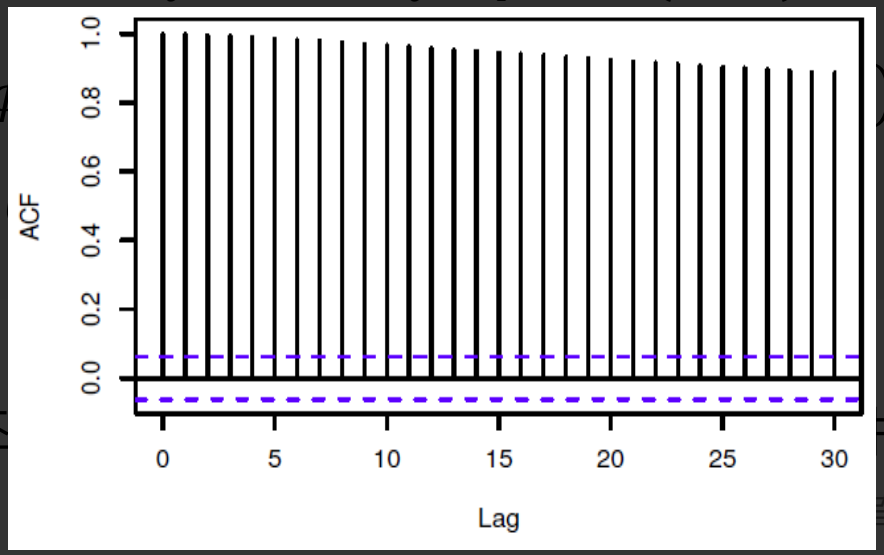
장기기억이란?

정상성을 만족하는 확률과정 $\{Z_t\}$의 자기상관함수 $\rho(k)$가

ARFIMA(p,d,q)

$0 < d < \frac{1}{2}$인 어떤 실수에 대하여 $\rho(k) \sim Ck^{2d-1}, k \to \infty\ (C > 0)$을 만족

$$\phi(B)(1-B)^d Y_t = \theta(B)Z_t\ ,\ Z_T \sim WN(0,\sigma^2),\ 0 < d < \frac{1}{2}$$



정상성 조 : 함

발산!

(반박불가)

ACF의 합이 무한대로 발산하며, ACF가 빠르게 0으로 수렴하지 않음

6

ARMAX

ARMAX

ARMAX

기존 ARMA모형에 외부요인(eXogenous)을 추가한 모형

$$\phi(B)Y_t = \theta(B)Z_t + \beta_0 + \beta_1 X_t$$

외부요인

1. X_t와 Y_t의 관측 값의 수는 일치해야 함
2. 외부요인은 연속형/범주형 변수 모두 가능

ARMAX

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} - \sum_{i=1}^{q} \theta_i Z_{t-i} + Z_t + \beta_0 + \beta_1 X_t$$

치환

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

모형의 식을 잔차를 가진 회귀식으로 변환 후

회귀식을 추정한 뒤 나온 잔차에 ARMA 모형을 적합!

ARMAX

ARIMAX(⇒ ARMAX에 차분을 포함)

$$\phi(B)(1-B)^d Y_t = \theta(B)Z_t + \beta^T \underline{X}$$

SARIMAX(⇒ ARIMAX에 계절성을 고려)

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)Z_t + \beta^T \underline{X}$$

7

VAR

VAR(Vector Auto Regressive)

현재의 관측값을 자신 및 다른 변수의 과거 관측값으로 설명

$$VAR(1) = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

$$X_t = c_1 + \phi_{11} X_{t-1} + \phi_{12} Y_{t-1} + \varepsilon_1$$

$$Y_t = c_2 + \phi_{21} X_{t-1} + \phi_{22} Y_{t-1} + \varepsilon_2$$

$$\varepsilon_i \sim WN(0, \sigma^2)$$

1. 여러 변수들 간의 의존성 고려 가능
2. 변수간 상호작용 고려 가능

VAR을 사용하는 이유

충격반응분석(Impulse Response)

각 변수가 예상치 못한 충격 이후 어떻게 변화하는지 분석

분산분해(Variance Decomposition)

각각의 변수들의 변동이 전체 변동에 기여한 부분이 어느 정도인지 확인

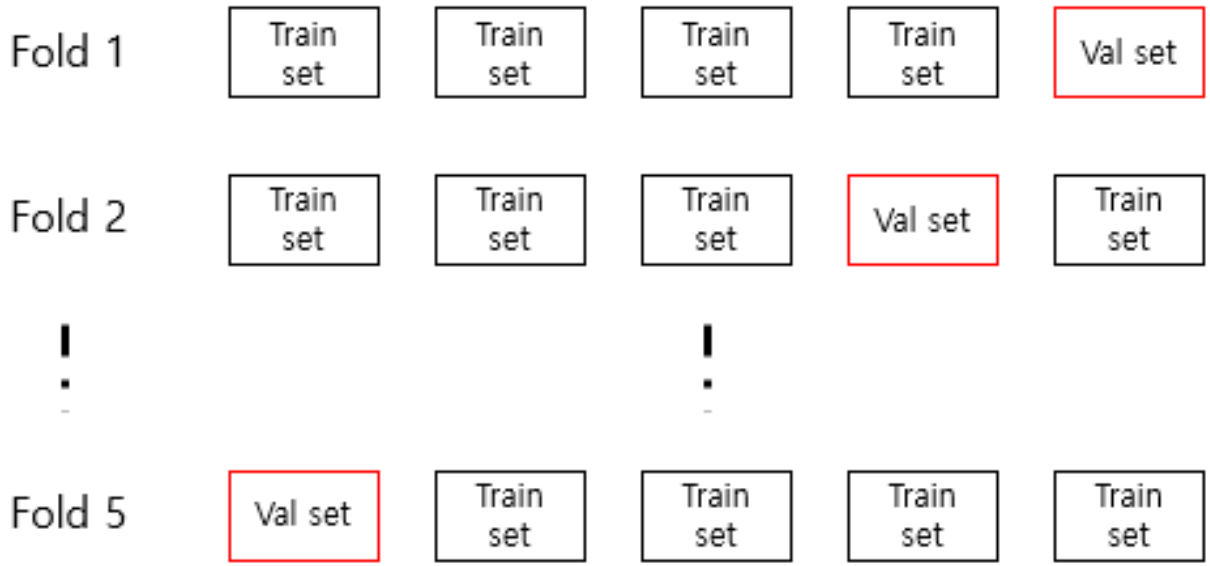
8

Time-Series CV

Time-Series CV

교차검증(Cross Validation)

과적합을 방지하기 위한 방법



데이터마이닝팀 1주차 클린업 참고!

Train set을 다시 Train / Validation set으로 나누어 모델 평가

교차검증(Cross Validation)

과적합을 방지하기 위한 방법

Fold 1	Train set	Train set	Train set	Train set	Val set
Fold 2	Train set	Train set	Train set	Val set	Train set
⋮			⋮		
Fold 5	Val set	Train set	Train set	Train set	Train set

Train set을 다시 Train / Validation set으로 나누어 모델 평가

Time-Series CV의 필요성

시계열은 시간 순서가 중요한 자료

무작위로 섞어서 교차검증 시 **자료의 손실** 발생

"시간"이라는 속성을 보존하는 차별화된 교차검증방식 필요!

교차검증(Cross Validation)

과적합을 방지하기 위한 방법

Fold 1	Train set	Train set	Train set	Train set	Val set
Fold 2	Train set	Train set	Train set	Val set	Train set
⋮			⋮		
Fold 5	Val set	Train set	Train set	Train set	Train set

*데이터마이닝팀 1주차 클린업 참고!

Train set을 다시 Train / Validation set으로 나누어 모델 평가

Time-Series CV의 필요성

시계열은 시간 순서가 중요한 자료

무작위로 섞어서 교차검증 시 **자료의 손실** 발생

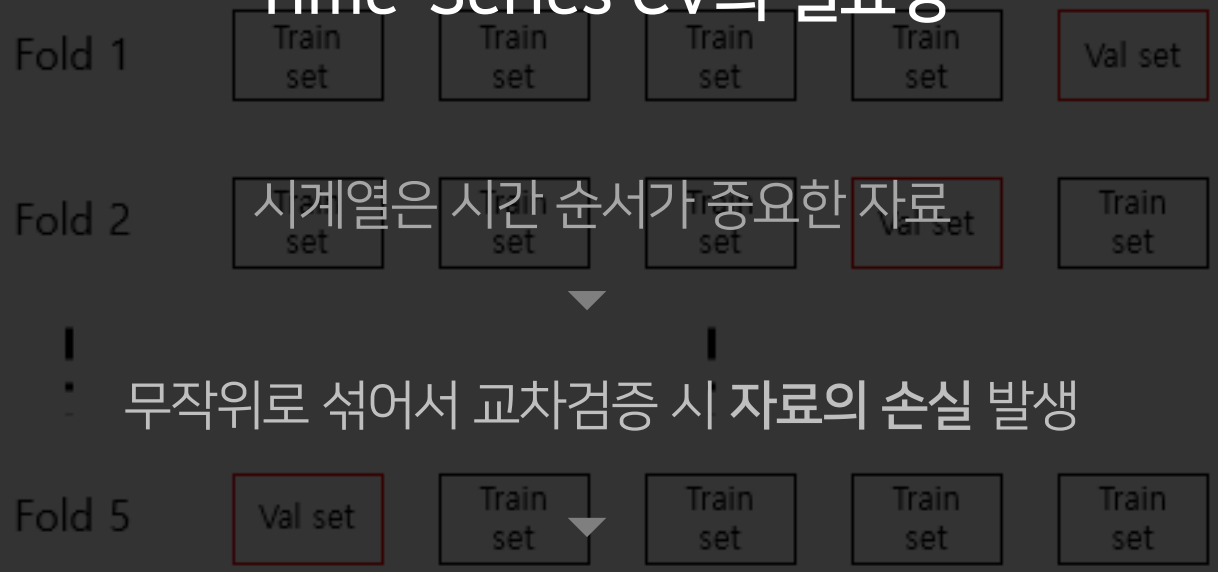
“시간”이라는 속성을 보존하는 차별화된 교차검증방식 필요!

8 Time-Series CV

교차검증(Cross Validation)

과적합을 방지하기 위한 방법

Time-Series CV의 필요성



시계열은 시간 순서가 중요한 자료

무작위로 섞어서 교차검증 시 **자료의 손실** 발생

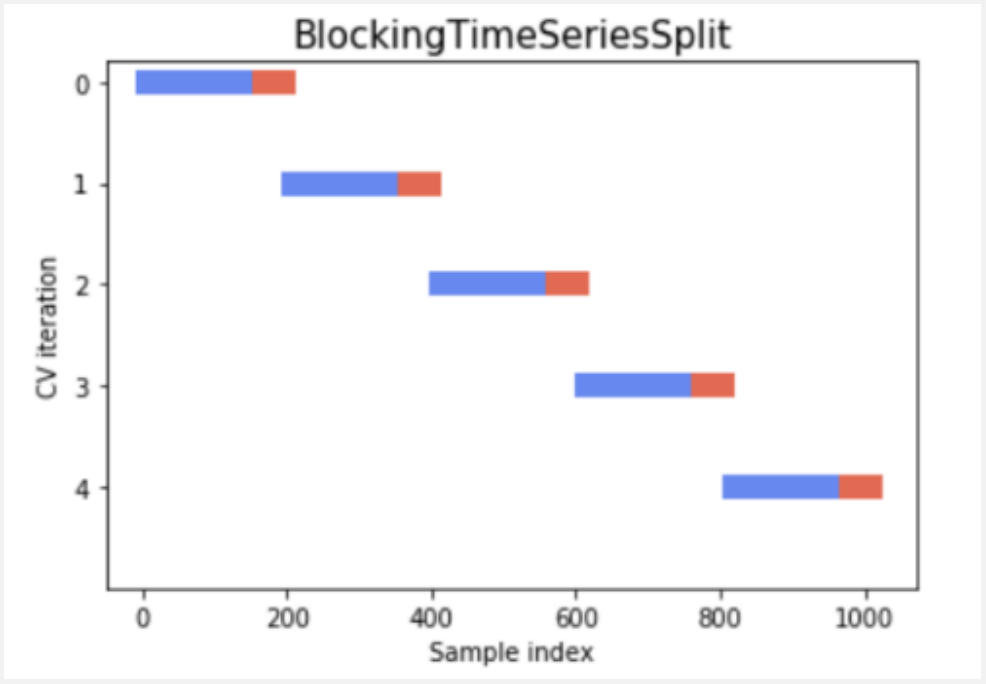
"시간"이라는 속성을 보존하는 차별화된 교차검증방식 필요!

Train set을 다시 Train / Validation set으로 나누어 모델 평가

Time-Series CV

Blocked Time-Series CV

a.k.a Rolling Window CV



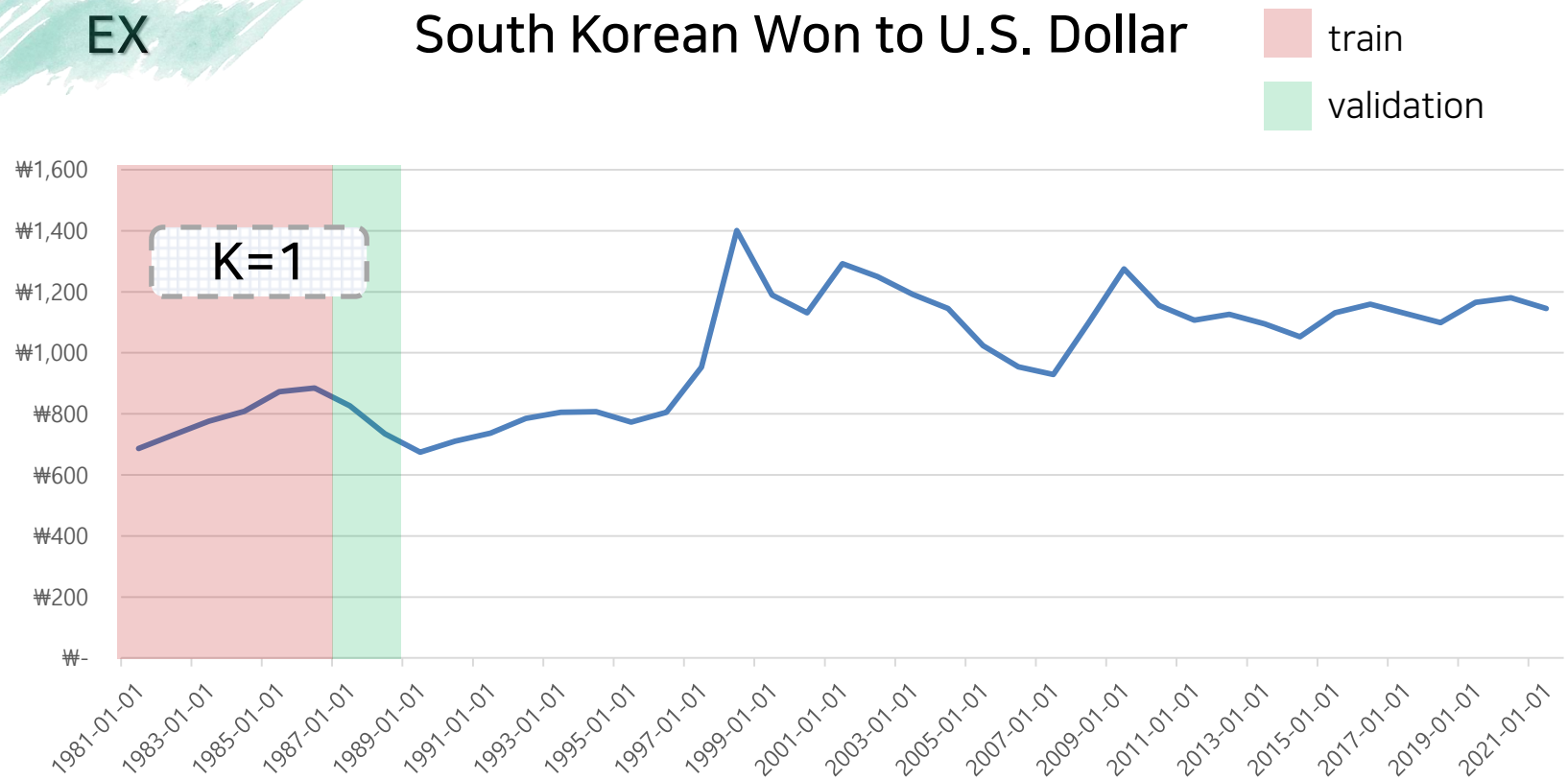
동일한 사이즈의 window 내에서 train/valid set 분리

Time-Series CV

Blocked Time-Series CV

a.k.a Rolling Window CV

EX

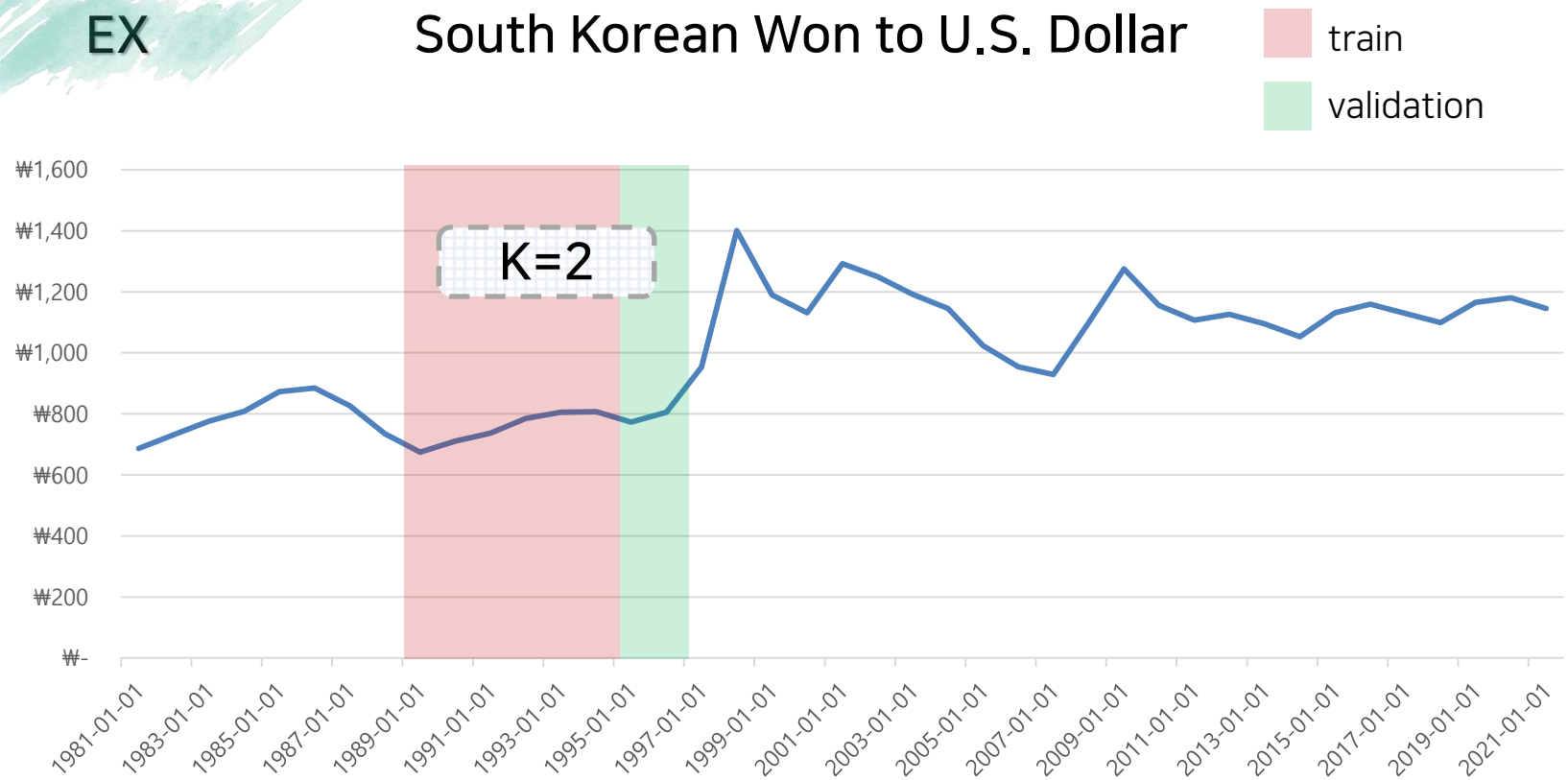


Time-Series CV

Blocked Time-Series CV

a.k.a Rolling Window CV

EX

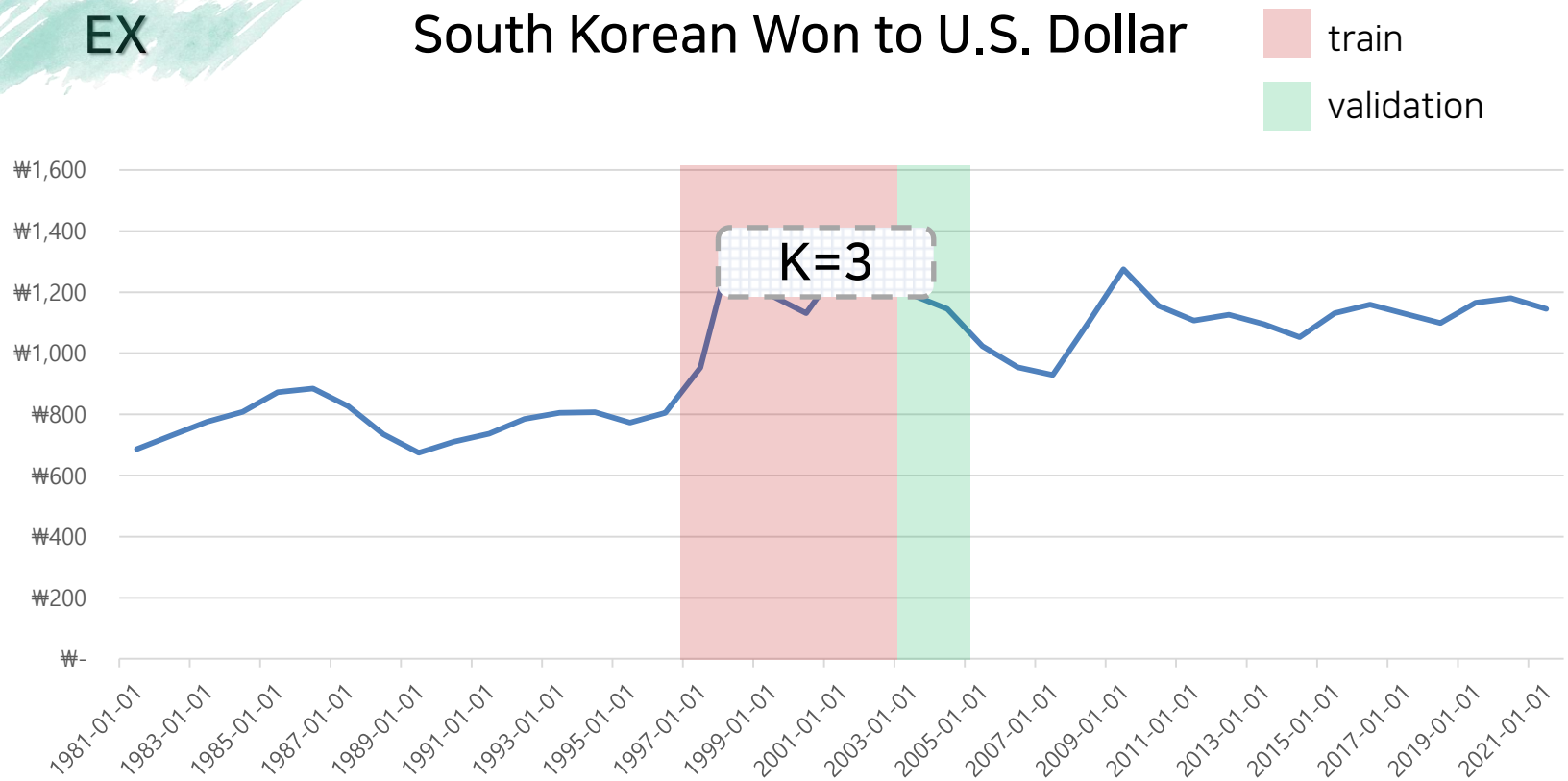


Time-Series CV

Blocked Time-Series CV

a.k.a Rolling Window CV

EX

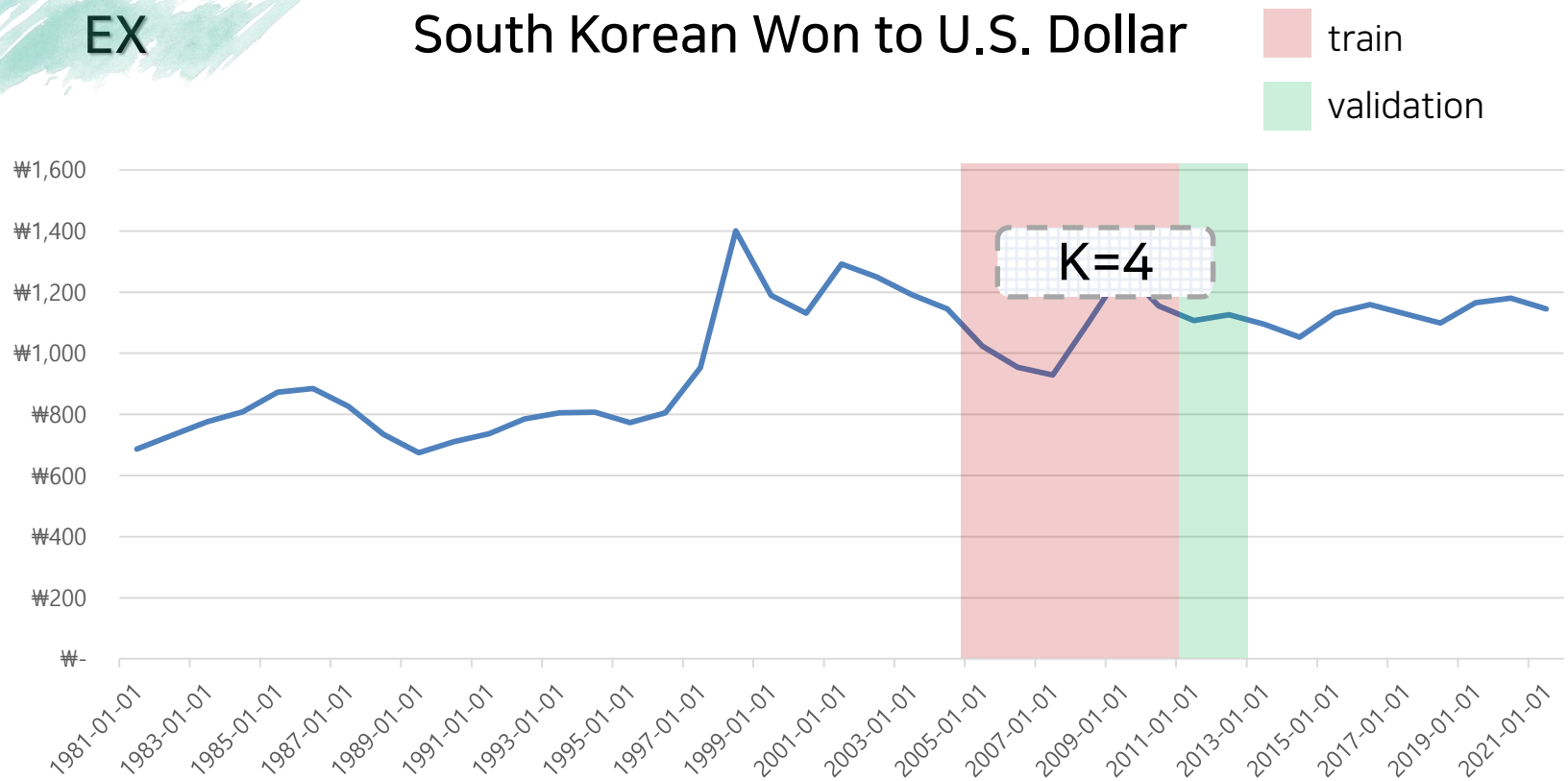


Time-Series CV

Blocked Time-Series CV

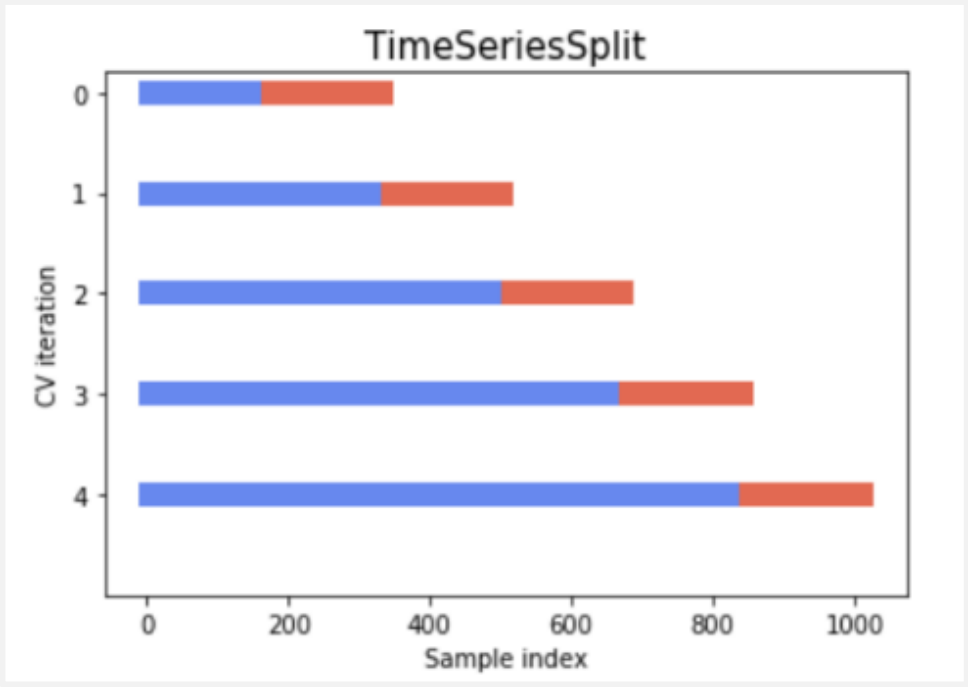
a.k.a Rolling Window CV

EX



Time-Series CV

a.k.a Expanding Window CV



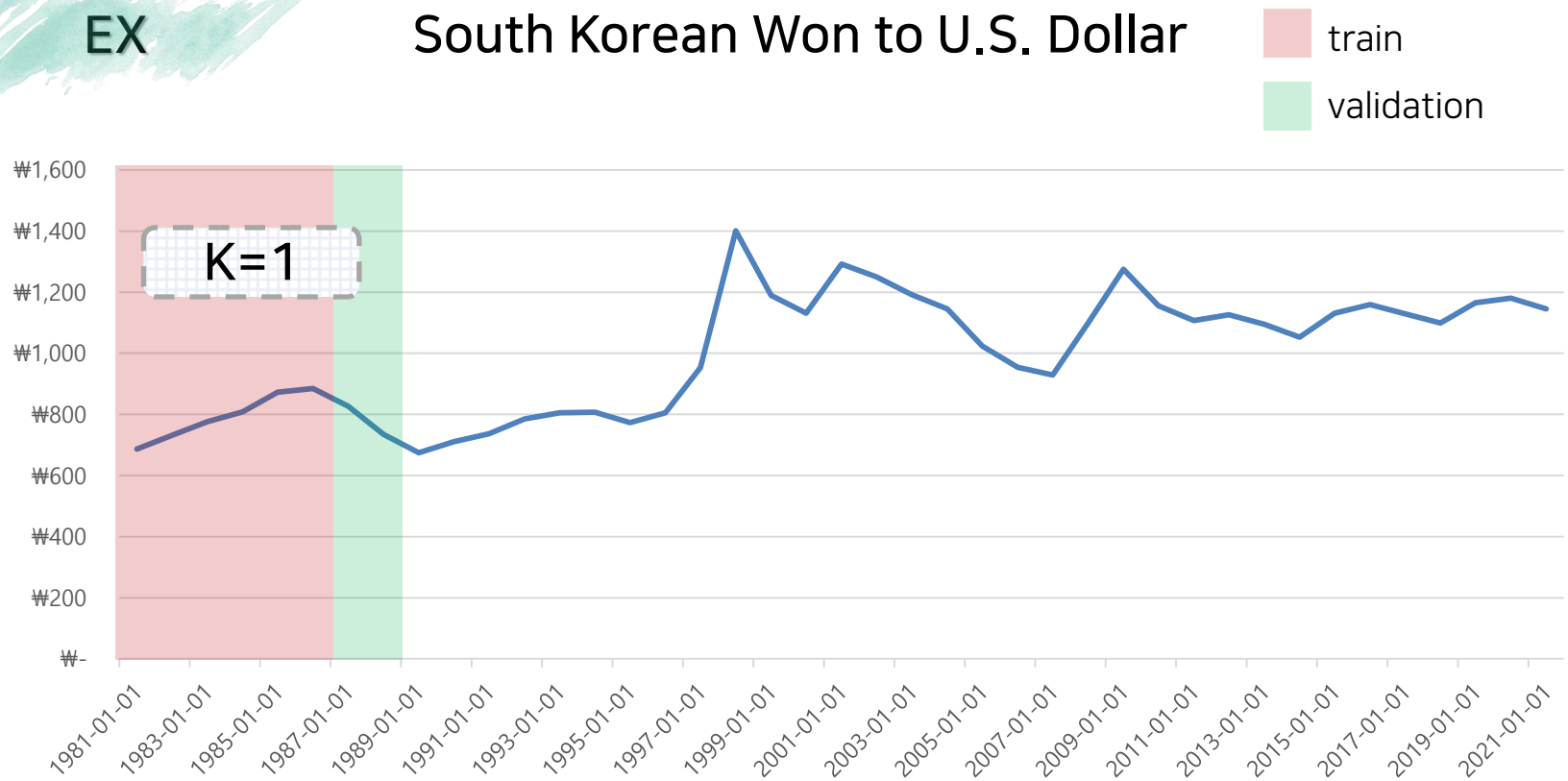
누적하며 이동하는 window 내에서 train/valid set 분리

Time-Series CV

Time-Series CV

a.k.a Expanding Window CV

EX

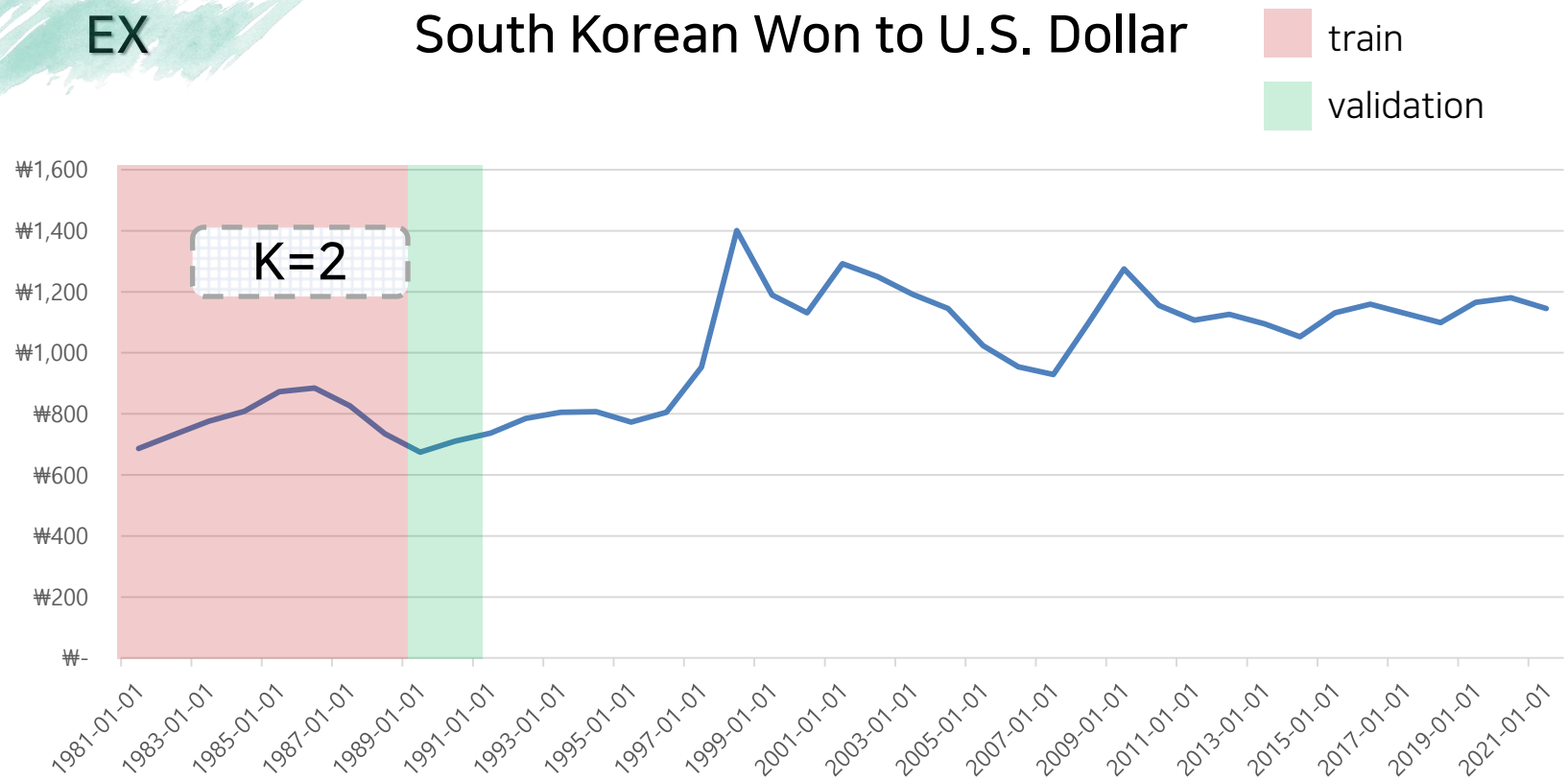


Time-Series CV

Time-Series CV

a.k.a Expanding Window CV

EX

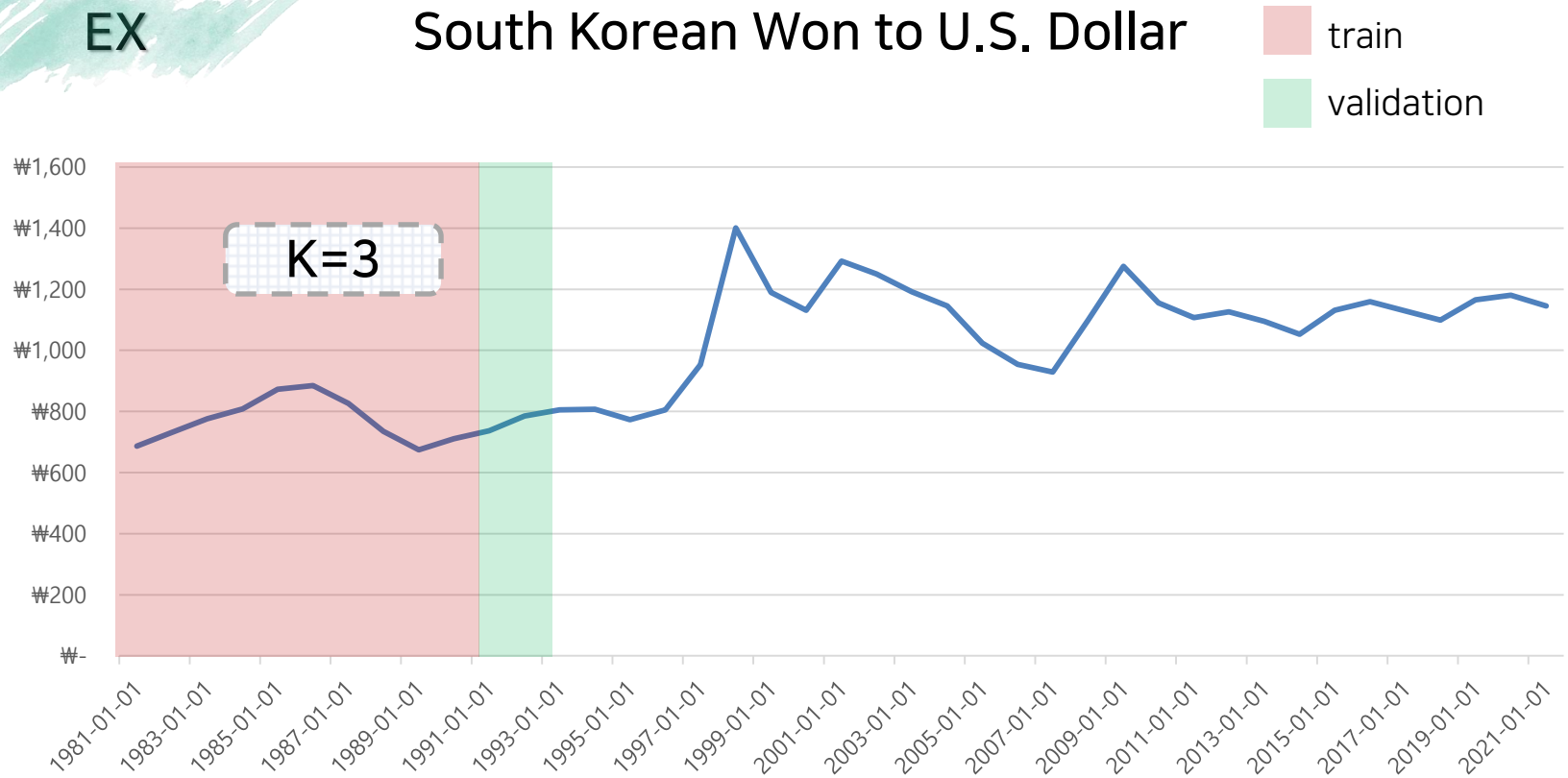


Time-Series CV

Time-Series CV

a.k.a Expanding Window CV

EX

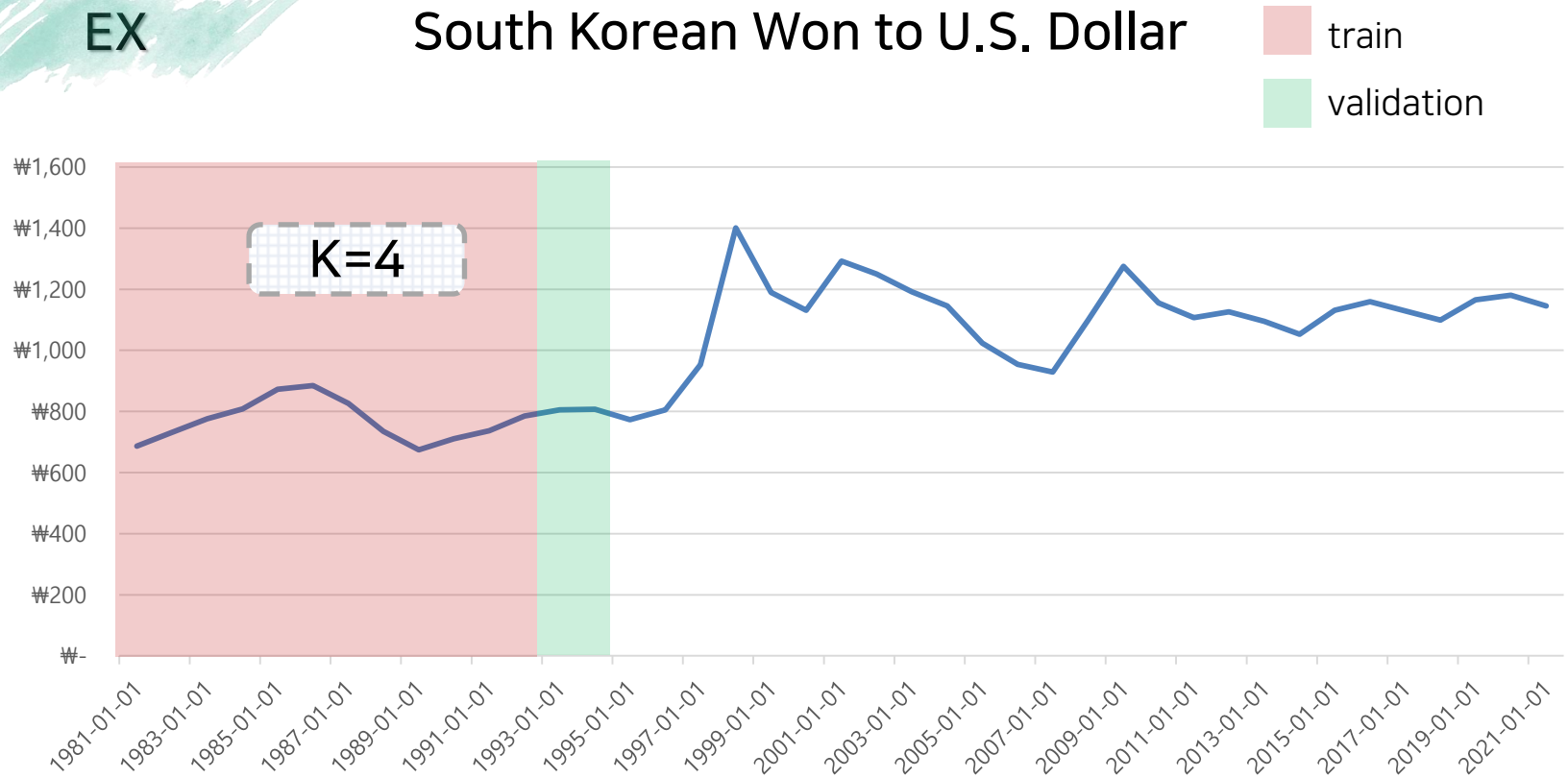


Time-Series CV

Time-Series CV

a.k.a Expanding Window CV

EX




"BRILLIANT"
GQ
THE SUN
METRO
THE TIMES
GOODMAN
BEN AFFLEC
HE TOWN
FROM
ARGO
15
15

9

ARGO

ARGO

Auto Regression with Google search data

PNAS PNAS

CrossMark click for updates

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [c]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with GOogle search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search–based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scal-

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26),

인플루엔자 질병의 발병 수준(ILI activity level)을 예측하기 위해 고안

↓

ARGO 이전엔 수많은 complex한 모델을 통해 이를 예측하고자 함

ARGO

Auto Regression with Google search data

$$y_t = \mu_y + \sum_{j=1}^{52} \alpha_j y_{t-j} + \sum_{i=1}^{100} \beta_i X_{it} + \epsilon_t$$

외부변수

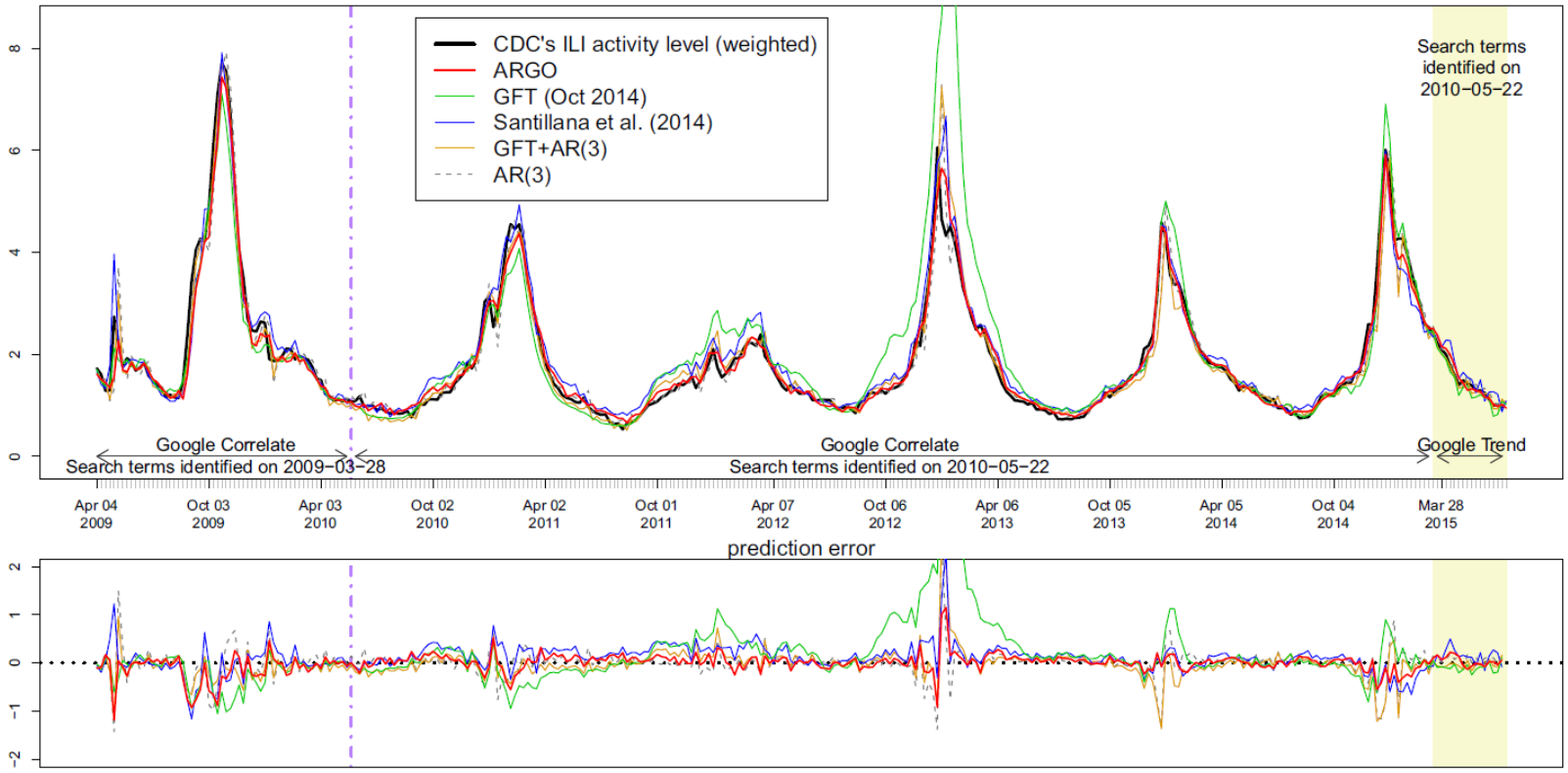
Penalized method

$$argmin \sum_t (y_i - \mu_y - \sum_{j=1}^{52} \alpha_j y_{t-j} - \sum_{i=1}^{100} \beta_i X_{i,t})^2$$
$$+\lambda_\alpha \|\alpha\|_1 + \eta_\alpha \|\alpha\|_2^2 + \lambda_\beta \|\beta\|_1 + \eta_\beta \|\beta\|_2^2$$

기본적인 AR모형에 플루와 관련된 상위 검색어를 외부변수로 사용!

ARGO

Auto Regression with Google search data



다른 복잡한 모델보다 인플루엔자 발병 수준의 경향을 잘 예측!

ARGO

Auto Regression with Google search data

오늘의 교훈

복잡하다 ≠ 성능이 좋다

간단한 모형도 항상 고려하며

분석 목적에 맞게 tuning하는 능력을 기르는 것 필요!

다른 복잡한 모델보다 인플루엔자 발병 수준의 경향을 잘 예측!

Simple is the Best

by 조 "The Professor" 웅빈

김재직 교수님… 좋은 말씀 감사했습니다

- Producer : 건우조 -

THANK YOU