한은 통계직렬 스터디: 설명편

2021-15115 권이태

November 24, 2024

Contents

2024년

Problem. 두 개의 시계열 x_t와 y_t가 불안정(non-stationary)한 시계열인 경우 두 변수 간의 인과관계를 추정할 때 발생할 수 있는 문제점이 무엇인지 설명하고, 해결책을 간단히 제시하시오.

Solution. 먼저 아래를 정의하자.

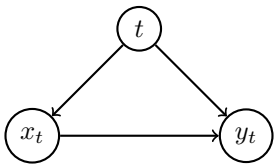
Definition 1. *어떠한 단변량 확률과정 X_t가 안정(weak-stationary)한 확률과정이라는 것은,*

1. $\mathbb{E}[X_t] = \mu$ *for all* t
2. $Cov(X_t, X_{t-j}) = \gamma_j < \infty$ *for all* t*, given* j*.*

임을 의미한다.

한편 이는 확률과정 X_t의 평균과 공분산 구조가 시간 t에 의해 의존하지 않으며, 어떠한 결정적(deterministic)인 추세가 존재하지 않음을 의미한다. 반대로 x_t, y_t가 불안정하다는 것은 곧 x_t, y_t에 시간 t의 흐름에 따른 추세가 존재하거나, 분산이 불안정함을 의미한다.

먼저 추세가 존재할 때의 문제를 들여다보자. 예를 들어, 시간이 흐름에 따라 경제의 산출량은 증가하고, 기술 수준은 진보한다. 이때 기술 수준의 증가가 산출량의 증가에 어떠한 영향을 미치는지에 대한 인과관계를 분석하려 한다. 그렇다면 우리의 모형은 아래와 같이 그릴 수 있다.



즉 우리는 $x_t \to y_t$처럼 x_t가 y_t에 미치는 영향만 보고자 하나, x_t와 y_t가 모두 시간 t의 흐름에 따라 추세를 가지고 변화하면서 t가 교락요인으로 작용한다. 따라서 단순히 x_t와 y_t에 대한 OLS

$$y_t = \alpha + \beta x_t + u_t$$

를 적합하면, x_t와 u_t에 내생성이 존재하므로 정확한 인과효과 β를 추정할 수 없다. 다르게 말하면, 우리가 회귀분석을 적합하여 얻은 $\hat{\beta}$는 $x_t \to y_t$에 의한 인과관계만을 포함하지 아니하고, $x_t \leftarrow t \to y_t$로 인한 상관관계까지 포함한다. 따라서 이 경우 인과관계의 명확한 파악을 위해서는 t에 의한 교락효과를 통제해야 한다.
둘째로는 시계열에 단위근이 존재하여 불안정한 상황을 고려해볼 수 있다. x_t, y_t가 모두 random walk를 따르는 서로 독립인 $I(1)$이라고 하자.

Definition 2. *어떤 시계열이 $I(k)$를 따른다는 것, 혹은 $X_t \sim I(k)$라는 것은*

$$\tilde{X}_t := (I - L)^k X_t$$

이 stationary임을 의미한다. 만약 stationary인 시계열이라면, $X_t \sim I(0)$과 같이 쓸 수 있다. 이때 L은 lag operator이며, 아래와 같이 정의된다.

$$LX_t := X_{t-1}$$

만약 x_t, y_t가 random walk를 따른다면

$$y_t = y_{t-1} + \zeta_t \quad \zeta_t \sim_{i.i.d} (0, \sigma_\zeta^2)$$
$$x_t = x_{t-1} + \xi_t \quad \xi_t \sim_{i.i.d} (0, \sigma_\xi^2)$$

이므로, 이들을 차분하면 IID 과정은 white noise로 stationary임에 따라

$$(I-L)y_t = y_t - y_{t-1} = \zeta_t \sim I(0)$$
$$(I-L)x_t = x_t - x_{t-1} = \xi_t \sim I(0)$$

이다. 즉 y_t, x_t는 $I(1)$이 된다. 이 경우 서로 독립이므로, 회귀식

$$y_t = \alpha + \beta x_t + u_t$$

에서 얻는 $\hat{\beta}$는 0으로 수렴하기를 기대한다. 즉, 일치추정량이기를 기대한다. 그러나 적절한 조건 하에서 $T \to \infty$라면 두 독립인 Wiener process $Z(t)$와 $\Xi(t)$에 대하여,

$$\hat{\beta} \Rightarrow \frac{\sigma_\zeta}{\sigma_\xi} \times \frac{\int_0^1 Z(t)\Xi(t)dt - \int_0^1 Z(t)dt \int_0^1 \Xi(t)dt}{\int_0^1 \Xi(t)^2 dt - \left(\int_0^1 \Xi(t)dt\right)^2}$$

임이 알려져 있으며, 이는 $O_p(1)$ 수준이다. 즉 인과효과 $\beta = 0$를 효율적으로 추정하지 못한다. 유사한 이유로 t통계량은 $O_p(\sqrt{T})$ 수준으로 주어지기에, 실제로는 $\beta = 0$이고 표본의 크기 T가 증가하더라도 제1종의 오류율이 0으로 수렴하지 않는다. 이는 실제로는 인과관계, 더 나아가서는 상관관계가 없음에도 인과관계가 유의한 것처럼 만들므로, spurious regression이라는 현상으로 불리기도 한다.
이러한 상황을 종합하면, non-stationary한 두 시계열 사이의 단순 회귀분석은 정확한 인과관계를 포착하지도 못할 뿐더러, 회귀계수 추정량을 매우 불안정하게 만들 수 있다. 이러한 문제점을 해결하기 위해서 가장 많이 사용되는 방법은 공적분 관계를 이용하는 것이다.

Definition 3. *어떤 두 시계열 y_t, x_t가 공적분 관계라는 것은, 벡터 $(c_1, c_2)^T \in \mathbb{R}^2$가 있어*

$$c_1 y_t + c_2 x_t \sim I(0)$$

임을 의미한다. 이 경우 (c_1, c_2)를 공적분 벡터라 부르며,

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} \sim C(c_1, c_2)$$

처럼 쓰기도 한다.

만약 공적분 관계가 존재한다면, $c_1 \neq 0$임을 가정할 때

$$y_t = -\frac{c_2}{c_1} x_t + u_t$$

에서 u_t가 stationary이다. 이처럼 residual이 stationary가 되는 경우에는 spurious regression과 같은 병적인 상황이 발생하지 않고 회귀계수에 대한 일치추정량을 얻을 수 있음이 알려져 있다. 따라서 모형 혹은 자료로부터 적절한 공적분 관계를 찾고, 그를 통해 회귀분석을 사용해 인과관계를 추정해도 될지 고려해보는 것이 가장 간단한 해결책 중 하나이다.

만약 공적분 관계가 관찰되지 아니하고 y_t와 x_t가 동일한 차수 k를 가진다면, 차분을 수행하는 것 역시 선택지가 될 수 있다. 모형

$$y_t = \alpha + \beta x_t + u_t$$

에서 전 시기의 모형

$$y_{t-1} = \alpha + \beta x_{t-1} + u_{t-1}$$

을 빼면

$$\Delta y_t = \beta \Delta x_t + \Delta u_t$$

가 되며, y_t, x_t, u_t가 $I(1)$이었던 경우 $\Delta y_t, \Delta x_t, \Delta u_t$는 stationary가 되므로 이 모형은 유효해진다. 따라서 이에 대한 회귀분석을 수행하면 β에 대한 일치추정량을 얻어줄 수 있다. 단 이는 y_t, x_t, u_t가 모두 같은 차수를 공유하는 경우에만 사용 가능하다.

Problem. 두 시계열의 공적분관계를 테스트하는 방법을 간략하게 설명하고, 이때 유의해야 할 점을 서술하시오.

Solution. 공적분 관계의 검정은 특정한 모형을 두고 이루어진다. 예를 들어 아래의 모형을 고려하자.

$$y_t = \beta x_t + u_t$$
$$u_t = \rho u_{t-1} + \epsilon_t$$

공적분 관계가 존재한다면 u_t가 stationary여야 하며, $u_t = \rho u_{t-1} + \epsilon_t$로 모형화하였으므로 $|\rho| \geq 1$이면 unit root가 존재하여 non-stationary이며, $|\rho| < 1$일 때 u_t가 stationary가 된다.

Definition 4. *어떠한 시계열 u_t가 $AR(1)$을 따를 때, 즉*

$$u_t = \rho u_{t-1} + \epsilon_t$$

*이고 ϵ_t이 white noise일 때, $\rho = 1$이면 이 시계열이 **단위근**(unit root)를 가진다고 한다. 이 시계열이 stationary할 조건은 $|\rho| < 1$와 동일하다. 이는 단기적인 쇼크가 장기적으로 남아있지 않을 조건과 동일하다.*

따라서 공적분 관계가 존재하는지 검정하는 것은 u_t가 stationary인지 검정하는 것과 같으며, 이는 단위근 검정

$$H_0 : \rho = 1$$
$$H_1 : \rho < 1$$

을 수행하는 것과도 동일하다. 따라서 아래의 과정을 따라 공적분 관계를 검정한다.

1. 먼저 $y_t = \beta x_t + u_t$를 OLS로 적합하여 추정된 잔차 $\hat{u}_t = y_t - \hat{\beta} x_t$를 얻는다.

2. $\hat{u}_t$에 대한 단위근 검정을 수행하여, 해당 시계열이 정상시계열이면 공적분 관계가 있다고 판단한다.

이때 공적분 관계가 존재한다는 귀무가설 하에서는 그 존재로 인하여 OLS가 super-consistent하므로 $\hat{u}_t$를 적절히 얻을 수 있고, 이에 따라 그를 통한 단위근 검정 역시 유효하다.
이제 단위근 검정의 방법을 확인하자. 공적분 관계의 테스트에서는 단위근 검정에서의 적절한 모형과 검정 방법 선택에 유의해야 한다. 단위근 검정 방법에는 Dickey-Fuller test, Augmented Dickey-Fuller test, Phillips-Perron test, KPSS test 등이 있다. 한편 모형에는 크게 세 종류가 존재한다. Dickey-Fuller test에서의 상황을 통해 세 종류의 모형을 비교하여 보자.

1. Unit root

$$\textbf{true DGP}: \begin{cases} y_t = \beta x_t + u_t \\ u_t = u_{t-1} + \epsilon_t \end{cases} \quad \textbf{estimated}: \begin{cases} y_t = \beta x_t + u_t \\ u_t = \rho u_{t-1} + \epsilon_t \end{cases}$$

여기에서는 오차항 u_t가 실제로는 random walk를 따르는 non-stationary process임을 가정하되, 해당 모형을 $AR(1)$ 모형으로써 추정할 때 상수항이 없는 모형 $u_t = \rho u_{t-1} + \epsilon_t$를 이용한다.

2. Unit root with constant

$$\textbf{true DGP}: \begin{cases} y_t = \beta x_t + u_t \\ u_t = \alpha + u_{t-1} + \epsilon_t \end{cases} \quad \textbf{estimated}: \begin{cases} y_t = \beta x_t + u_t \\ u_t = \alpha + \rho u_{t-1} + \epsilon_t \end{cases}$$

여기에서는 모형 추정의 과정에서 상수항 α의 존재 역시 고려한다. 즉 u_t가 평균이 0이 아닐 수 있는 상황을 고려한다.

3. Unit root with constant and deterministic time trend

$$\textbf{true DGP}: \begin{cases} y_t = \beta x_t + u_t \\ u_t = \alpha + \rho u_{t-1} + \delta t + \epsilon_t \ (\rho = 1 \text{ or } \delta \neq 0) \end{cases} \quad \textbf{estimated}: \begin{cases} y_t = \beta x_t + u_t \\ u_t = \alpha + \rho u_{t-1} + \delta t + \epsilon_t \end{cases}$$

여기에서는 $\rho = 1$과 $\delta = 1$을 joint하게 검정하여 trend와 unit root가 모두 없어 u_t가 stationary인지 검정한다.

즉 이 세 모형은 각각 u_t가 평균과 추세가 모두 없는 경우, 평균은 있고 추세가 없는 경우, 둘 모두 있는 경우를 모형화한다. 이들 모형을 어떻게 세우냐에 따라 검정의 결과가 달라질 수 있기에, 경제학적 모형 및 데이터의 형태를 바탕으로 적절한 모형을 세우고 해당 모형 하에서 단위근 검정을 수행해야만 한다. 만약 잔차에 추세가 있으면 셋째 모형을, 추세는 없으나 잔차 평균이 0이 아니면 둘째 모형을, 둘 다 없는 경우에는 첫째 모형을 적용하면 된다. ADF test나 PP test 역시 동일하게 세 개의 상황에서 각각 검정통계량의 분포와 기각역이 달라지며, 이에 따라 검정의 결과 역시 달라질 수 있다. 단 그 결과는 여기에 담기에는 너무 복잡하여 크게 다루지는 않는다. PP test의 경우 DF test와 모형 자체는 완전히 동일하되 검정통계량이 다를 뿐이며, ADF test의 경우에는 상수항과 추세항의 포함 여부를 달리 하여

$$\begin{aligned} y_t &= \beta x_t + u_t \\ u_t &= \alpha + \rho u_{t-1} + \delta t + \xi_1 \Delta u_{t-1} + \cdots + \xi_{p-1} \Delta u_{t-p+1} + \epsilon_t \end{aligned}$$

에 기반한 세 종류의 모형이 존재하며, joint하게 $\delta = 0$과 $\rho = 1$을 검정하거나 $\rho = 1$을 검정한다. 즉 우리가 데이터, 혹은 경제학적 모형과 직관으로부터 어떠한 true DGP를 가정하고 어떠한 모형으로써 추정하냐에 따라 잔차의 정상성 검정 방법과 검정 결과가 달라진다. 따라서 공적분 검정의 과정에서는 모형과 검정 방법의 선택을 사려깊게 수행해야만 한다.

Problem. $I(1)$인 x_t, y_t 두 시계열은 $y_t = 5x_t + \epsilon_t$로 공적분 관계가 성립한다. 이때 ϵ_t는 $I(0)$이다. 또한 두 변수 간에 다음과 같은 모델이 추정되었다. 아래의 ARDL(Autoregressive distributed lag) 모델로부터 오차수정모형(error correction model)을 도출하고 θ의 값을 찾으시오.

$$y_t = 4 + 0.5y_{t-1} + 2x_t + \theta x_{t-1} + u_t, \quad u_t \overset{i.i.d}{\sim} N(0, \sigma)$$

Solution.

Definition 5. *integrated time series* $(y_t, x_t)^T \in \mathbb{R}^2$*에 대한 Autoregressive distributed lag(ARDL) 모형은 아래와 같이 표현된다.*

$$y_t = \gamma + \sum_{i=1}^{p} \phi_j y_{t-i} + \sum_{j=0}^{q} \theta_j x_{t-j} + u_t$$

이러한 모형을 ARDL(p, q)와 같이 쓰기도 한다.

Definition 6. *(Engle and Granger approach)*
y_t, x_t*가* $I(1)$*을 따르고 stationary* ϵ_t*에 대해 공적분 관계* $y_t - \beta_1 x_t = \epsilon_t$*를 가진다면, 오차수정모형(error correction model; ECM)은 아래와 같은 형태를 가진다.*

$$A(L)\Delta y_t = \gamma + B(L)\Delta x_t + \alpha\epsilon_{t-1} + v_t$$

주어진 모형에서 Δy_t를 만들기 위하여 양변에서 y_{t-1}을 빼면,

$$\Delta y_t = 4 - 0.5y_{t-1} + 2x_t + \theta x_{t-1} + u_t$$

을 얻고, 아래와 같이 변형할 수 있다.

$$\begin{aligned}\Delta y_t &= 4 - 0.5y_{t-1} + 2x_t + \theta x_{t-1} + u_t \\ &= 4 - 0.5(y_{t-1} - 5x_{t-1}) + 2(x_t - x_{t-1}) + (\theta - 0.5)x_{t-1} + u_t \\ &= 4 + 2\Delta x_t - 0.5\epsilon_{t-1} + (\theta - 0.5)x_{t-1} + u_t\end{aligned}$$

이때 좌변은 $y_t \sim I(1)$임에 따라 stationary이고, 우변은 $(\theta - 0.5)x_{t-1}$ 항을 제외하면 모두 stationary이며 x_{t-1}은 non-stationary이다. 따라서 $\theta = 0.5$여야 이 항이 사라지고 양변이 모두 stationary가 될 수 있으며, 이에 따른 오차수정모형은

$$\Delta y_t = 4 + 2\Delta x_t - 0.5\epsilon_{t-1} + u_t$$

으로 주어진다.

Note. 이 모형은 ϵ_{t-1}을 이용하여 오차를 수정한다는 점에서 오차수정모형이라 불린다. 특히 공적분 관계에 대한 식 $y_t = 5x_t + \epsilon_t$는 장기적으로 y_t와 x_t 사이의 선형관계를 묘사하는 반면, ECM은 단기적으로 그들의 변동이 어떠한 연관을 맺는지를 의미한다. 즉 ECM은 장기적인 공적분 관계, 혹은 제약 하에서 설명변수의 단기적인 변동에 반응변수가 어떻게 반응하는지를 묘사한다. 이때 수정되는 오차항을 보면, 고정된 계수 -0.5와 $t-1$기에서의 장기균형으로부터 벗어난 정도인 ϵ_{t-1}의 곱으로 구성되어 있다. 계수가 음수라는 것은 균형으로부터 양으로 벗어나 $\epsilon_{t-1} > 0$이 되면 Δy_t가 음이 되도록 error correction이 진행됨으로써 y_t가 장기적으로 균형 수준으로 회복될 수 있도록 한다. 즉 error correction은 차분된 시계열로 구성된 회귀모형에서 양변의 균형을 맞추는 correction 역할을 하는 동시에, 경제적으로도 충격으로 인한 단기적 불균형을 장기균형으로 correction하기도 한다.

Note. 이 문제에서 주로 다루는 주제들은 정상성과 공적분에 대한 논의 중 **Engle & Granger Approach**에 해당합니다. `https://warwick.ac.uk/fac/soc/economics/staff/gboero/personal/hand2_cointeg.pdf`이 그나마 명료하게 정리된 글 같습니다.

Note. 추후 시간이 되면 여기에 등장한 사실들에 대한 증명들과 unit root test들에 대한 검정통계량과 기각역도 추가하겠습니다.

2023년

Problem. 다음을 간략히 기술하시오: 스피어만의 순위상관계수(Spearman rank correlation coefficient)

Solution. 두 표본 $x_1, x_2, \cdots, x_n$과 $y_1, y_2, \cdots, y_n$에 대하여 스피어만의 순위상관계수는 아래와 같이 정의된다.

Definition 7. *두 표본 $x_1, x_2, \cdots, x_n$과 $y_1, y_2, \cdots, y_n$의 순위상관계수는*

$$r_s = \frac{\sum_{i=1}^{n}(s_i - \bar{s})(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2 \sum_{i=1}^{n}(t_i - \bar{t})^2}}$$

으로 정의된다. 이때 $\{s_1, \cdots, s_n\}$은 $\{x_1, \cdots, x_n\}$의 순위변환이며, $\{t_1, \cdots, t_n\}$은 $\{y_1, \cdots, y_n\}$의 순위변환이다.

스피어만의 순위상관계수는 순위변환된 표본들의 피어슨 상관계수와도 동일하며, 항상 -1과 1 사이의 값을 가진다. 스피어만의 순위상관계수가 1에 가까운 값을 가질수록 x_i와 y_i 사이에는 양의 상관관계가, -1에 가까운 값을 가질수록 음의 상관관계가 존재한다고 말할 수 있다. 선형적인 관계를 주로 포착하는 피어슨의 상관계수와는 달리 스피어만 순위상관계수는 비선형적인 상관계수 역시 잘 포착할 수 있으며, 순위변환을 하므로 이상점과 영향점에 강건하게 상관관계를 파악할 수 있다는 장점을 가지고 있다. 그러나 순위변환 과정에서 자료가 가진 정보가 소실되며, 대응되는 모집단의 모수가 없어 해석이 어렵다는 단점도 있다.

Problem. 다음을 간략히 기술하시오: 인자분석에서의 공통성(communality)

Solution.

Definition 8. *인자분석(Factor Analysis)에서의 직교인자모형(orthogonal factor model)은 아래와 같은 형태를 가진다.*

$$X_i = \mu + L \cdot f_i + \epsilon_i$$

이때 unit $i = 1, 2, \cdots, n$에 대해 $X_i \in \mathbb{R}^p$는 관측가능한 자료이자 random vector이며, $\mu \in \mathbb{R}^p$는 그 평균이다. 우리는 이 X_i가 평균에 더하여 관측되지 않은 인자(factor) f_i들의 변동에 의해 설명된다고 믿으며, 그들의 가중치를 factor loading이라 부르고 모아서 L로 표기한다. 만약 m개의 인자가 있다면, $L \in \mathbb{R}^{p \times m}$, $f_i \in \mathbb{R}^m$이다. ϵ_i는 f_i와는 무관하게 X_i에만 영향을 주는 오차항이다. 직교인자모형에서는 특히 아래를 가정한다.

- $\mathbb{E}[f] = 0, Cov(f) = I$: *orthogonal factors*
- $\mathbb{E}[\epsilon] = 0, Cov(\epsilon) = \Psi = diag(\psi_i)$
- $Cov(\epsilon, f) = 0$

즉 이 모형에서는 관측값 X_i가 어떠한 내재적인 인자들 f_i의 선형결합으로 좌우된다고 가정하며, 적절한 identifying assumption을 통해 모수 L과 Φ를 추정하고 각 개체의 f_i를 예측하는 것을 목표로 한다.

여기에서 X_i의 공분산 행렬은

$$\mathrm{Var}(X) = \mathrm{Var}(\mu + Lf_i + \epsilon_i) = LL^T + \Psi$$

으로 주어지고, X_{ij}, 즉 i번째 unit의 j번째 성질은

$$\mathrm{Var}(X_{ij}) = \sum_{k=1}^{m} l_{jk}^2 + \psi_j$$

으로 표현되며, 여기에서 앞부분인 $l_{j1}^2 + \cdots + l_{jm}^2$을 공통성(communality), 뒷부분인 ψ_j를 specific variance 라 부른다. 전체 분산 중 공통성이 차지하는 비중이 높을수록 관측값의 변동이 factor f_i에 의해 잘 설명됨을 의미한다.

Problem. 다음을 간략히 기술하시오: ARCH(Autoregressive Conditional Heteroskedasticity) 모형

Solution.

Definition 9. *ARCH(autoregressive conditional heteroskedasticity) 모형은 자기회귀적으로 시계열의 조건부 분산을 묘사한다. 이는 아래의 모형처럼 쓸 수 있다.*

$$\begin{aligned} &y_t = \sigma_t \xi_t,\ \{\xi_t\} \sim_{i.i.d} N(0,1) \\ &\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_q y_{t-q}^2 \\ &\alpha_0 > 0,\ \alpha_i \geq 0 \end{aligned}$$

이때 이를 차수가 q인 ARCH 모형, 혹은 $ARCH(q)$처럼 쓴다.

ARCH 모형에서는 이전 q기의 $y_{t-1}, \cdots, y_{t-q}$가 평균인 0으로부터 멀어지는 경우 시점 t에서의 조건부 분산 σ_t^2가 커지게 모형화함으로써, 금융시계열 등에서 분산이 집중화되는 현상을 묘사할 수 있다.

Problem. 다음과 같은 $AR(2)$ 시계열에 대해 답하시오.

$$\dot{X}_t = 0.8\dot{X}_{t-1} - 0.1\dot{X}_{t-2} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma_\epsilon^2), \quad \dot{X}_t = X_t - \mathbb{E}[X_t]$$

위 시계열이 정상(stationary) 시계열인지 판단하고, 정상 시계열이라면 MA 모형으로 나타내시오.

Solution. 이 시계열은 정상시계열이다. AR 모형을 따르는 시계열에 대하여 정상성을 판정하는 가장 쉬운 방법 중 하나는 그 특성방정식의 근을 파악하는 것이다.

Definition 10.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

위의 $AR(p)$ 모형을 따르는 시계열 y_t의 특성방정식(characteristic equation)은

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

으로 정의된다. 이 특성방정식의 근인 특성근(characteristic root)이 모두 $|z| > 1$으로 단위원 밖에 있다면, 이 시계열은 정상시계열이다.

평균이 0이고 $AR(2)$ 모형을 따르는 시계열 $\dot{X}_t$의 특성방정식은

$$\phi(z) = 1 - 0.8z + 0.1z^2$$

으로 주어지며, 그 근은

$$z = \frac{0.8 \pm \sqrt{0.64 - 0.4}}{0.2} = 4 \pm \sqrt{6}$$

으로 모두 그 절댓값이 1보다 크다. 따라서 $\dot{X}_t$는 정상시계열이다.
한편 lag operator를 이용하여 주어진 $AR(2)$ 모형을 표현하면,

$$(1 - 0.8L + 0.1L^2)\dot{X}_t = \epsilon_t$$

이다. 이를 $MA(\infty)$ 모형으로 표현하면,

$$\dot{X}_t = (1 - 0.8L + 0.1L^2)^{-1}\epsilon_t = \sum_{i=0}^{\infty} a_i \epsilon_{t-i}$$

처럼 쓸 수 있다. 이때 a_i는

$$\frac{1}{1 - 0.8z + 0.1z^2} = \sum_{i=0}^{\infty} a_i z^i$$

를 만족하게 하는 상수 a_i이다. 한편 아래처럼 좌변을 분해할 수 있다.

$$\begin{aligned}
\frac{1}{1 - 0.8z + 0.1z^2} &= \frac{10}{(z - (4 + \sqrt{6}))(z - (4 - \sqrt{6}))} \\
&= \frac{1}{(1 - \frac{z}{4+\sqrt{6}})(1 - \frac{z}{4-\sqrt{6}})} \\
&= \left(\sum_{j=0}^{\infty} \frac{1}{(4+\sqrt{6})^j} z^j \right) \left(\sum_{k=0}^{\infty} \frac{1}{(4-\sqrt{6})^k} z^k \right) \\
&= \sum_{i=0}^{\infty} \left(\sum_{j+k=i} \frac{1}{(4+\sqrt{6})^j (4-\sqrt{6})^k} \right) z^i
\end{aligned}$$

따라서 테일러 급수의 수렴성에 의하여,

$$a_i = \sum_{j+k=i} \frac{1}{(4+\sqrt{6})^j (4-\sqrt{6})^k}$$

으로 주어진다. 이때 j와 k는 0 이상의 정수여야 한다. 따라서 이 시계열을 MA 모형으로 표현하면,

$$\dot{X}_t = \sum_{i=0}^{\infty} \left(\sum_{j+k=i} \frac{1}{(4+\sqrt{6})^j (4-\sqrt{6})^k} \right) \epsilon_{t-i}$$

Problem. 다음과 같은 $AR(2)$ 시계열에 대해 답하시오.

$$\dot{X}_t = 0.8\dot{X}_{t-1} - 0.1\dot{X}_{t-2} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma_\epsilon^2), \quad \dot{X}_t = X_t - \mathbb{E}[X_t]$$

자기상관함수(ACF; autocorrelation function) ρ_1, ρ_2의 값을 구하시오.

Solution.

Definition 11. *정상시계열 y_t의 자기상관함수(ACF; autocorrelation function) $\rho(h)$는 아래와 같이 정의된다.* $(h = 0, 1, 2, \cdots)$

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{Cov(y_{t+h}, y_t)}{Var(y_t)}$$

이때 $\gamma(h)$는 자기공분산함수(autocovariance function)

$$\gamma(h) = Cov(y_{t+h}, y_t)$$

이다.

먼저 아래의 식들을 얻을 수 있다.

$$\begin{aligned}\gamma(0) = \mathrm{Var}(\dot{X}_t) &= \mathrm{Cov}\left(\dot{X}_t, \dot{X}_t\right) \\ &= \mathrm{Cov}\left(0.8\dot{X}_{t-1} - 0.1\dot{X}_{t-2} + \epsilon_t, 0.8\dot{X}_{t-1} - 0.1\dot{X}_{t-2} + \epsilon_t\right) \\ &= 0.65\gamma(0) - 0.32\gamma(1) + \sigma_\epsilon^2\end{aligned}$$

$$\begin{aligned}\gamma(1) &= \mathrm{Cov}\left(\dot{X}_t, \dot{X}_{t-1}\right) \\ &= \mathrm{Cov}\left(0.8\dot{X}_{t-1} - 0.1\dot{X}_{t-2} + \epsilon_t, \dot{X}_{t-1}\right) \\ &= 0.8\gamma(0) - 0.1\gamma(1)\end{aligned}$$

이들에 대한 연립방정식을 풀면

$$\gamma(0) = \frac{1100}{641}\sigma_\epsilon^2, \quad \gamma(1) = \frac{800}{641}\sigma_\epsilon^2$$

을 얻는다. 따라서 $\rho_1 = \frac{\gamma(1)}{\gamma(0)} = \frac{8}{11}$이다.
한편

$$\begin{aligned}\gamma(2) &= \mathrm{Cov}(\dot{X}_t, \dot{X}_{t-2}) \\ &= \mathrm{Cov}(0.8\dot{X}_{t-1} - 0.1\dot{X}_{t-2} + \epsilon_t, \dot{X}_{t-2}) \\ &= 0.8\gamma(1) - 0.1\gamma(0)\end{aligned}$$

이므로,

$$\gamma(2) = \frac{640}{641}\sigma_\epsilon^2 - \frac{110}{641}\sigma_\epsilon^2 = \frac{530}{641}\sigma_\epsilon^2$$

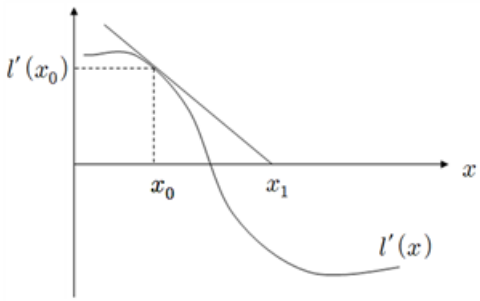
이고

$$\rho_2 = \frac{\gamma(2)}{\gamma(0)} = \frac{53}{110}$$

이다.

2022년

Problem. 다음은 최대우도추정량(MLE)을 찾기 위한 Newton-Raphson 알고리즘의 의사코드(pseudocode)이다. 빈칸 A, B, C를 채우시오.



Solution.

1. 로그우도함수의 1차 미분함수를 $l'(x)$로 정의한다.
2. 로그우도함수의 2차 미분함수를 $l''(x)$로 정의한다.
3. 초기추측값을 x_0에 저장하고
   반복임계치를 e에 최대반복 횟수를 N에 저장한다.
4. 반복 카운터 $i=1$로 지정한다.
5. If $l''(x_0)=0$ then "에러" 메시지를 출력하고 종료한다.
6. $x_1=$ A
7. $i=$ B
8. If $i \geq N$ then "수렴하지 않음" 메시지를 출력하고 종료한다.
9. If $|l'(x_1)|>e$ then $x_0=$ C 그리고 go to (5).
10. x_1을 반환하고 종료한다.

Definition 12. *나이브한 뉴턴-랩슨 방법의 각 iteration에서, $x^{(j)}$는 $x^{(j+1)}$로 업데이트되며, 이는 $(x^{(j)}, f(x^{(j)}))$에서의 접선이 x축과 만나는 점이다. 접선의 기울기가*

$$f'(x^{(j)}) = \frac{f(x^{(j)})}{x^{(j)} - x^{(j+1)}}$$

이므로, 업데이트는

$$x^{(j+1)} = x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})}$$

으로 이루어진다.

따라서 A는

$$x_0 - \frac{l'(x_0)}{l''(x_0)}$$

, B는 $i+1$, C는 x_1이다.

Problem. 부트스트랩(Bootstrap) 방법에 대해 간략히 기술하시오.

Solution. 통계적 추론 하에서 모집단의 분포족을 모르는 경우, 추정량의 표본분포를 알 수 없으므로 추정량의 편향과 표준오차, 신뢰구간을 구하거나 가설검정을 하기 어렵다. 따라서 붓스트랩은 표본으로부터 얻은 경험적 분포함수로써 원래 모집단의 분포함수를 근사하여 해당 표본에서 표본과 같은 크기의 부트스트랩 표본을 정해진 반복수 B만큼 추출하고, 각각으로부터 얻은 B개의 붓스트랩 추정량으로써 표본분포를 근사하여 통계적 추론에 이용한다.

붓스트랩의 기본 가정은 아래와 같다.

$$X_1, X_2, \cdots, X_n \sim_{i.i.d} F$$

$$\hat{\theta}_n = T(X_n) = T(X_1, \cdots, X_n) : \text{ 표본으로부터 얻은 모수의 추정량}$$

그러나 우리는 원래 함수 F에 대해 모집단 가정을 가하지 않는 이상 알 수 없다. 비모수적 상황에서는 추론에 가해지는 가정이 최소화되는 것을 선호하기 때문에, F에 무관하게 표본분포를 알 수 있는 방법이 필요하다. 혹은 F나 T가 너무 복잡한 경우에도 표본분포를 계산하기 어렵다. 부트스트랩에서는 F에서 표본을 새로 뽑을 수 없으므로, 우리의 표본 $X_1, \cdots, X_n$으로부터 모집단의 분포 F를 추정하여 그 추정된 분포 F_n을 가상의 모집단으로 두고, 여기에서 표본을 다시 추출하여 표본분포를 근사하는 것을 목표로 한다. 이때 F_n은 **경험적 분포함수**(eCDF)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

이다. 이는 사실 관측값 X_i에 질량 $1/n$이 배정된 이상분포와 동일하므로, F_n에서 표본을 추출하는 것은 $X_1, \cdots, X_n$에서 n개의 값을 다시금 단순복원추출하여 크기 n인 i번째 **부트스트랩 표본**

$$\mathbf{X}_n^{*(i)}$$

을 얻을 수 있다. 또한 이들 n개로부터 부트스트랩 표본에서의 통계량 $\hat{\theta}_n^{*(i)}$ 역시 $i = 1, 2, \cdots, B$에 대해 계산할 수 있다. 이렇게 B개의 $\hat{\theta}_n^{*(i)}$으로부터 얻어낸 경험적 분포는 $\hat{\theta}_n$의 **부트스트랩 표본분포**가 된다. 이를 도식화하면 아래와 같다. 이 부트스트랩 표본분포를 $\hat{\theta}_n$의 표본분포의 근사치로 사용하여, 점근적으로 옳은 통계적 추론을 수행할 수 있다.

$$\text{Population} : F$$

$$\Downarrow \text{ Choose } n \text{ observations to build sample with sample size } n \Downarrow$$

$$\text{Sample} : \mathbf{X}_n = (x_1, x_2, \cdots, x_n)$$

$$\Downarrow \text{ Under } F_n \text{ instead of } F, \text{ make B bootstrap samples with sample size } n \Downarrow$$

$$\text{Error} : \frac{1}{\sqrt{n}}$$

$$\begin{gathered}\text{Bootstrap sample } 1 : \mathbf{X}_n^{*(1)} = (x_1^{*(1)}, \cdots, x_1^{*(n)}),\ \ \hat{\theta_n}^{*(1)} = T(\mathbf{X}_n^{*(1)}) \\ \vdots \\ \text{Bootstrap sample } B : \mathbf{X}_n^{*(B)} = (x_1^{*(B)}, \cdots, x_1^{*(B)}),\ \ \hat{\theta_n}^{*(B)} = T(\mathbf{X}_n^{*(B)})\end{gathered}$$

$$\Downarrow \text{ Estimate boostrap sample distribution using } (\hat{\theta_n}^{*(1)}, \cdots, \hat{\theta_n}^{*(B)}) \Downarrow$$

$$\text{Error} : \frac{1}{\sqrt{B}}$$

$$\text{Estimate: } G^*(x) = \frac{1}{B}\sum_{i=1}^{B} I(\hat{\theta_n}^{*(i)} \le x)$$

Problem. 다음은 차원축소 등에 이용되는 주성분분석(principal component analysis; PCA)에 대한 기술이다. 빈칸 A, B, C를 채우시오.

> 양정치(positive definite) 공분산행렬 Σ에 대해 확률변수 X가 다변량 정규분포 $N_n(\mu,\Sigma)$를 따른다고 하자($\sigma_i^2=\Sigma_{ii}$).
> 주성분분석은 공분산행렬 Σ의 ___A___분해를 통하여 Σ를 $\Sigma=\Gamma\Lambda\Gamma$로 나타낸다. Γ의 열벡터 $v_1,\cdots,v_n$을 고유벡터, 이에 대응하는 Λ의 대각성분 $\lambda_1,\cdots,\lambda_n$을 고윳값이라고 한다. 여기서 고윳값 및 고유벡터는 $\lambda_1\geq\cdots\geq\lambda_n>0$으로 구성한다. 확률벡터 Y를 $Y=\Gamma(X-\mu)$로 정의하자. 그러면 Y는 ___B___의 분포를 따르는데, 이때 Y를 주성분 벡터라고 한다.
> 확률변수 X와 Y의 총변동(total variation, TV)은 다음의 이유로 동일하다.
>
> $$TV(X)=\sum_{i=1}^{n}\sigma_i^2=\underline{\quad C\quad}=\sum_{i=1}^{n}\lambda_i=TV(Y)$$

Solution.

Definition 13. *다변량 자료의 공분산 구조는 많은 경우 자료의 개수(n)보다 훨씬 적은 개수의 선형결합들에 의해 기술될 수 있다.* ***주성분분석****(PCA)는 n개의 변수로 구성된 시스템에서 나타나는 총 변동, 혹은 그들의 공분산 구조를 상당 부분 설명할 수 있는 변수들의 선형결합인* ***주성분****(principal component)를 골라내는 것을 목표로 한다.*
확률벡터 X가 $X\sim N_n(\mu,\Sigma)$를 따른다고 하면, Σ는 대칭인 양정치행렬이므로, 직교대각화(고유값분해)가 항상 가능하다. 따라서

$$\Sigma=\Gamma^T\Lambda\Gamma$$

로 쓰면,

$$\Gamma^T=\begin{bmatrix}| & | & \cdots & |\\ v_1 & v_2 & \cdots & v_n\\ | & | & \cdots & |\end{bmatrix},\quad \Lambda=\begin{pmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{pmatrix}$$

이고 Γ는 직교행렬이며, $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n>0$이도록 적을 수 있다. 그렇다면 단순한 선행대수에 의하여

$$X=\langle X,v_1\rangle v_1+\cdots+\langle X,v_n\rangle v_n$$

을 얻으며, $\langle X,v_i\rangle$이 i번째 주성분이 된다. 한편 i번째 주성분은 아래와 같은 성질을 가진다.

Proposition 1. *1번째 principal component는 $PC_1=v_1^TX=v_{1n}X_1+\cdots+v_{1n}X_n$이며,*

$$v_1=\text{argmax}_b\{Var(b^TX)|\|b\|=1\}$$

이다. 즉 PC_1은 X의 선형결합 중 그 분산이 가장 큰 성분이다.

Proposition 2. *r번째 principal component는 $PC_r=v_r^TX=v_{rn}X_1+\cdots+v_{rn}X_n$이며, 1-$(r-1)$개의 principal component가 주어졌을 때*

$$v_r=\text{argmax}_b\{Var(b^TX)|\|b\|=1, Cov(b^TX,a_j^TX)=0\ for\ j=1,2,\cdots,r-1\}$$

이다. 즉 PC_r은 X의 변동 중 $PC_1,\cdots,PC_{r-1}$으로 설명되지 않는 변동을 가장 잘 설명하는 성분이다.

Proposition 3. *단순 계산을 통하여, 아래를 얻는다.*

$$Var(PC_r)=\lambda_r,\quad Cov(PC_r,PC_s)=0\quad(r\neq s)$$

한편 주성분들을 모은 벡터인 **주성분 벡터**

$$Y = \begin{pmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_n \end{pmatrix} = \begin{pmatrix} v_1^T X \\ v_2^T X \\ \vdots \\ v_n^T X \end{pmatrix} = \Gamma X$$

는 $X \sim N_n(\mu, \Sigma)$임에 따라

$$Y \sim N_n(\Gamma\mu, \Gamma\Sigma\Gamma^T) = N_n(\Gamma\mu, \Lambda)$$

이며, 문제에서처럼 μ를 뺀 버전으로 정의하는 경우

$$Y \sim N_n(0, \Lambda)$$

가 된다. 이는 **Proposition 3**의 결과와도 일치한다. 한편 PCA에서 변수들의 총변동은 $\mathrm{TV}(X) = \mathrm{Tr}(\Sigma)$으로 정의되며, trace는 similarity transformation에 invariant하므로 $\mathrm{Tr}(\Lambda) = \sum_{i=1}^n \lambda_i$와도 같다. 만약 $m < n$이 있어

$$\sum_{i=1}^{n} \lambda_i \approx \lambda_1 + \lambda_2 + \cdots + \lambda_m$$

이라면, $PC_1, \cdots, PC_m$이 X의 대부분의 정보를 가지고 있다고 생각할 수 있다.
이제 문제를 풀면, A는 고윳값, B는 $N_n(0, \Lambda)$, 그리고 C는 $\mathrm{tr}(\Sigma) = \mathrm{tr}(\Gamma^T\Lambda\Gamma) = \mathrm{tr}(\Lambda\Gamma\Gamma^T) = \mathrm{tr}(\Lambda)$이다.

Problem. 다음의 입력값과 출력값을 갖는 K-means 알고리즘의 의사코드를 5줄 내외로 간략히 작성하시오.

> 입력값: k = 군집 수
> D = $n(> k)$개의 샘플을 포함하는 데이터
> 출력값: k개의 군집

Solution.

Definition 14. *K-means 알고리즘은 군집화를 수행함에 있어 n개의 데이터로부터 구성된 k개의 군집*

$$C_1, \cdots, C_k, \quad (\cup_{i=1}^k C_k = D, C_i \cap C_j = \emptyset \text{ if } i \neq j)$$

에 대하여

$$\sum_{i=1}^{k} W(C_i)$$

를 최소화하는 $C_1, \cdots, C_k$를 찾고자 하는 알고리즘이다. 이때 $W(C_i)$는 군집 C_i 내에서의 변동을 의미한다. 즉 군집화는 동질한 개체들을 모은 군집을 여러 개 형성하는 방법이다. K-means 알고리즘에서는 먼저 관측치들을 $C_1, \cdots, C_k$에 랜덤하게 할당한 뒤, 군집이 바뀌지 않을 때까지

1. *각 군집마다 군집의 중앙을 계산하고*
2. *모든 관측치를 가장 가까운 중심의 군집에 할당하는*

과정을 반복하여 진행한다. 이는 $\sum_{i=1}^k W(C_k)$를 점진적으로 감소시켜나가며 군집화를 수행한다.

따라서 의사코드는 아래와 같다.

Algorithm 1 K-means Clustering Algorithm

Require: $k \geq 1$, $D = (x_1, x_2, \cdots, x_n)$ $(n > k)$
  Initialize: n개의 관측값 $x_1, \cdots, x_n$을 k개의 집합으로 임의로 할당하여 $C_1^{(0)}, C_2^{(0)}, \cdots, C_k^{(0)}$을 만든다.
  $j \leftarrow -1$
  while $j == -1$ or $C_i^{(j)} \neq C_i^{(j-1)}$ for any $i = 1, 2, \cdots, k$ **do**
    $j \leftarrow j + 1$
    $i = 1, 2, \cdots, k$에 대하여 $C_i^{(j)}$의 중심인 $m_i^{(j)} = \dfrac{1}{|C_i^{(j)}|} \displaystyle\sum_{x_l \in C_i^{(j)}} x_l$을 계산한다.
    $d_l^{(j+1)} = \underset{\{1,2,\cdots,k\}}{\operatorname{argmin}} \|m_i^{(j)} - x_l\|$을 얻고, x_l을 $C_{d_l^{(j+1)}}$에 포함시켜 $C_1^{(j+1)}, C_2^{(j+1)}, \cdots, C_k^{(j+1)}$을 만든다.
  end while
  Output: $C_1^{(j)}, C_2^{(j)}, \cdots, C_k^{(j)}$

Problem. 최단입점보간(nearest neighbor interpolation) 방법에 대해 간략히 기술하고 동 방법의 단점을 제시하시오.

Solution.

Definition 15. *최단입점보간은 보간법의 일종으로, 데이터가 존재할 수 있는 공간 중 일부 점에서만 그 값이 주어져 있을 때, 값이 존재하지 않는 다른 점에서의 값을 해당 점과 가장 가까운 값을 가진 점의 값으로써 보간하는 방법이다. 즉 데이터가*

$$\{(x_1, y_1), \cdots, (x_n, y_n)\}$$

으로 주어져 있을 때, $\{x_1, \cdots, x_n\}$에 포함되지 않는 새로운 점 x^*에서의 값 y^*는

$$y^* = y_{\operatorname{argmin}_{1 \leq i \leq n} \|x^* - x_i\|}$$

으로 보간한다.

최단입점보간의 단점에는 크게 아래의 것들이 있다.

- 보간된 함수가 step function의 형태로, 연속하거나 미분가능하지 않다. 보간된 함수가 매끄러워야 하는 경우, 이 방법은 선호될 수 없다.
- x_i의 차원이 증가하면, 차원의 저주로 전체 공간에서 n개의 점 $x_1, \cdots, x_n$은 매우 드물게 존재한다. 이는 x^*와 x_i 간의 간격을 늘림으로써 보간이 제대로 이루어지지 못하게 한다.
- 보간된 값들이 유한 개의 값만을 가질 수 있다.
- 동점이 존재하며 해당 동점에서의 값이 상이한 경우, 선택 방법에 따라 보간값이 매우 달라질 수 있다.

Problem. $Y_1, \cdots, Y_n$은 베르누이분포(Bernoulli distribution)를 따르는 i.i.d.(independent and identically distributed) 확률변수이다.

$$Y_i \sim \text{Bernoulli}(p)$$

성공확률 p에 대한 사전분포함수로 다음과 같은 베타분포(beta distribution)를 고려하라.

$$p \sim \text{Beta}(\alpha, \beta)$$

베타분포의 확률밀도함수는 다음과 같다.

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1 - p)^{\beta - 1}$$

성공확률 p에 대한 사후분포함수 $\pi(p|y_1, \cdots, y_n)$을 구하시오.

Solution.

$$\begin{aligned}\pi(p|y_1, \cdots, y_n) &\propto \pi(p) \times L(p; y_1, \cdots, y_n) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \times \binom{n}{y_1 + \cdots + y_n} p^{y_1+\cdots+y_n}(1-p)^{n-(y_1+\cdots+y_n)} \\ &\propto p^{\alpha+\sum_{i=1}^n y_i - 1}(1-p)^{\beta+n-\sum_{i=1}^n y_i - 1}\end{aligned}$$

이므로, 사후분포는 $\text{Beta}(\alpha + \sum_{i=1}^n y_i, \beta + n - \sum_{i=1}^n y_i)$이다. 따라서 분포함수는

$$\pi(p|y_1, \cdots, y_n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum_{i=1}^n y_i)\Gamma(\beta+n-\sum_{i=1}^n y_i)} p^{\alpha+\sum_{i=1}^n y_i - 1}(1-p)^{\beta+n-\sum_{i=1}^n y_i - 1}$$

Problem. 위 문제에서, 성공확률의 사후평균 $\mathbb{E}_{p|y_1,\cdots,y_n}(p)$을 구하시오.

Solution.

$$\mathbb{E}_{p|y_1,\cdots,y_n}(p) = \frac{\alpha + \sum_{i=1}^n y_i}{\alpha + \beta + n}$$

임이 베타분포의 평균에서 잘 알려져 있다.

Problem. 손실함수 $L(a, p)$에 대하여 사후기대손실(expected posterior risk) $\rho(a)$는 베이즈추정량 a의 함수로 다음과 같이 정의된다.

$$\rho(a) = \mathbb{E}_{p|y_1,\cdots,y_n}[L(a, p)]$$

손실함수가 제곱오차인 경우, 즉,

$$L(a, p) = (a-p)^2$$

의 경우, 사후기대손실을 최소화하는 베이즈추정량 a는 사후평균과 같음을 보이시오. 즉, $a = \mathbb{E}_{p|y_1,\cdots,y_n}(p)$.

Solution. 아래처럼 $\rho(a)$를 쓸 수 있다.

$$\begin{aligned}\rho(a) &= \mathbb{E}_{p|y_1,\cdots,y_n}[L(a, p)] \\ &= \mathbb{E}_{p|y_1,\cdots,y_n}[(a-p)^2] \\ &= \mathbb{E}_{p|y_1,\cdots,y_n}[(a - \mathbb{E}_{p|y_1,\cdots,y_n}(p))^2 + (\mathbb{E}_{p|y_1,\cdots,y_n}(p) - p)^2] \\ &+ 2\mathbb{E}_{p|y_1,\cdots,y_n}[(a - \mathbb{E}_{p|y_1,\cdots,y_n}(p))(\mathbb{E}_{p|y_1,\cdots,y_n}(p) - p))] \\ &= (a - \mathbb{E}_{p|y_1,\cdots,y_n}(p))^2 + \text{Var}_{p|y_1,\cdots,y_n}(p) \\ &+ 2(a - \mathbb{E}_{p|y_1,\cdots,y_n}(p))\mathbb{E}[\mathbb{E}_{p|y_1,\cdots,y_n}(p) - p] \\ &= \text{Var}_{p|y_1,\cdots,y_n}(p) + (a - \mathbb{E}_{p|y_1,\cdots,y_n}(p))^2 \geq \text{Var}_{p|y_1,\cdots,y_n}(p)\end{aligned}$$

따라서 $\rho(a)$가 최소화되는 것은 $(a - \mathbb{E}_{p|y_1,\cdots,y_n}(p))^2 = 0$일 때, 즉 $a = \mathbb{E}_{p|y_1,\cdots,y_n}(p)$일 때이다.

Problem. 다음과 같은 AR(1) 모형에 대하여 자기상관함수(auto-correlation function; ACF) ρ_k (k는 시차,

이하 동일)를 구하시오. (X_t는 정상시계열로 가정)

$$X_t = \delta + \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim_{i.i.d.} N(0, \sigma^2)$$

Solution. 자기공분산함수 γ_k는 아래의 점화식을 가진다. $k > 1$일 때,

$$\begin{aligned}\gamma_k &= \mathrm{Cov}(X_t, X_{t+k}) \\ &= \mathrm{Cov}(X_t, \delta + \phi X_{t+k-1} + \epsilon_{t+k}) \\ &= \phi\gamma_{k-1}\end{aligned}$$

이때 $\epsilon_{t+k} \perp\!\!\!\perp X_t$임을 이용한다. 양변을 γ_0으로 나누면,

$$\rho_k = \phi\rho_{k-1}$$

을 얻는다. $\rho_0 = 1$임은 자명하므로,

$$\rho_k = \phi^{|k|}$$

을 얻게 된다.

Problem. 다음과 같은 MA(1) 모형에 대하여 자기상관함수 ρ_k를 구하시오.

$$X_t = \epsilon_t - \theta\epsilon_{t-1}, \quad \epsilon_t \sim_{i.i.d.} N(0, \sigma^2)$$

Solution. 자기공분산함수 γ_k에 대해 먼저 논의하자. $k > 1$에 대하여,

$$\begin{aligned}\gamma_k &= \mathrm{Cov}(X_t, X_{t+k}) \\ &= \mathrm{Cov}(X_t, \epsilon_{t+k} - \theta\epsilon_{t+k-1}) \\ &= 0\end{aligned}$$

이다. $k = 1$인 경우에는,

$$\begin{aligned}\gamma_1 &= \mathrm{Cov}(X_t, X_{t+1}) \\ &= \mathrm{Cov}(\epsilon_t - \theta\epsilon_{t-1}, \epsilon_{t+1} - \theta\epsilon_t) \\ &= -\theta\mathrm{Var}(\epsilon_t) = -\sigma^2\theta\end{aligned}$$

이고 $k = 0$인 경우에는

$$\begin{aligned}\gamma_0 &= \mathrm{Cov}(X_t, X_t) \\ &= \mathrm{Cov}(\epsilon_t - \theta\epsilon_{t-1}, \epsilon_t - \theta\epsilon_{t-1}) \\ &= (1 + \theta^2)\sigma^2\end{aligned}$$

이다. 따라서 이로부터

$$\rho_0 = 1, \quad \rho_1 = \rho_{-1} = -\frac{\theta}{1 + \theta^2}, \quad \rho_k = 0 \quad \text{if } |k| \geq 2$$

을 얻는다.

Problem. 다음과 같은 ARMA(1,1) 모형에 대하여 자기상관함수 ρ_k를 구하시오. (X_t는 정상시계열로 가정)

$$X_t = \delta + \phi X_{t-1} + \epsilon_t - \theta\epsilon_{t-1}, \quad \epsilon_t \sim_{i.i.d.} N(0, \sigma^2)$$

Solution. 자기공분산함수 γ_k에 대한 논의부터 시작하자. $k > 1$에 대하여,

$$\begin{aligned}
\gamma_k &= \mathrm{Cov}(X_t, X_{t+k}) \\
&= \mathrm{Cov}(X_t, \delta + \phi X_{t+k-1} + \epsilon_{t+k} - \theta\epsilon_{t+k-1}) \\
&= \mathrm{Cov}(X_t, \phi X_{t+k-1}) \\
&= \phi\gamma_{k-1}
\end{aligned}$$

이다. 따라서 γ_0으로 양변을 나누면,

$$\rho_k = \phi\rho_{k-1}$$

을 얻는다. 이는 ρ_1만 구하면 됨을 의미한다.

$$\begin{aligned}
\gamma_1 &= \mathrm{Cov}(X_t, X_{t+1}) \\
&= \mathrm{Cov}(X_t, \delta + \phi X_t + \epsilon_{t+1} - \theta\epsilon_t) \\
&= \phi\gamma_0 - \theta\mathrm{Cov}(X_t, \epsilon_t) \\
&= \phi\gamma_0 - \theta\mathrm{Cov}(\delta + \phi X_{t-1} + \epsilon_t - \theta\epsilon_{t-1}, \epsilon_t) \\
&= \phi\gamma_0 - \theta\sigma^2
\end{aligned}$$

이고,

$$\begin{aligned}
\gamma_0 &= \mathrm{Cov}(X_t, X_t) \\
&= \mathrm{Cov}(\delta + \phi X_{t-1} + \epsilon_t - \theta\epsilon_{t-1}, \delta + \phi X_{t-1} + \epsilon_t - \theta\epsilon_{t-1}) \\
&= \phi^2\gamma_0 + (1+\theta^2)\sigma^2 - 2\phi\theta\mathrm{Cov}(X_{t-1}, \epsilon_{t-1}) \\
&= \phi^2\gamma_0 + (1+\theta^2)\sigma^2 - 2\phi\theta\sigma^2
\end{aligned}$$

이므로

$$\gamma_0 = \frac{\sigma^2}{1-\phi^2}(1 - 2\phi\theta + \theta^2), \quad \gamma_1 = \frac{\sigma^2}{1-\phi^2}(\phi - \theta)(1 - \phi\theta)$$

을 얻는다. 따라서 모두를 γ_0으로 나누면,

$$\begin{aligned}
\gamma_0 &= 1 \\
\gamma_1 = \gamma_{-1} &= \frac{(\phi-\theta)(1-\phi\theta)}{1 - 2\phi\theta + \theta^2} \\
\gamma_k &= \phi^{|k|-1}\frac{(\phi-\theta)(1-\phi\theta)}{1 - 2\phi\theta + \theta^2}
\end{aligned}$$

을 얻을 수 있다.

Problem. 위와 같은 결과를 바탕으로 정상패턴이 확인된 ARMA모형을 자기상관함수 및 부분자기상관함수 (partial auto-correlation function; PACF)를 이용하여 식별하는 방법 및 한계점을 설명하고, 모형 선택의 기준이 되는 통계량의 예와 이를 이용한 모형선택 방법을 설명하시오.

Solution.

Definition 16. *부분자기상관함수는 아래와 같이 정의된다. $k \geq 1$에 대하여,*

$$\phi_{kk} = Corr(X_{k+1} - \text{proj}_{\text{span}\{X_2,\cdots,X_k\}} X_{k+1}, X_1 - \text{proj}_{\text{span}\{X_2,\cdots,X_k\}} X_1)$$

Proposition 4. *정상 ARMA 모형에서, 아래의 표가 성립한다.*

모형	ACF	PACF
AR(p) 모형	지수적으로 감소소멸하는 형태	p시차 초과로는 0으로의 절단 형태
MA(q) 모형	q시차 초과로는 0으로의 절단 형태	지수적으로 감소소멸하는 형태
ARMA(p,q) 모형	시차 $q-p$ 초과로는 지수적으로 감소소멸	시차 $p-q$ 초과로는 지수적으로 감소소멸

따라서 ACF, PACF 함수의 개형을 통하여 이 시계열이 어떠한 p와 q를 가지는 ARMA모형인지 대략 파악할 수 있다.

Definition 17. *정상시계열의 표본자기공분산함수(sample autocovariance function)은*

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(X_t - \bar{X})(X_{t+|h|} - \bar{X}), \quad |h| < n$$

으로 정의한다. 이때 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$이다. 이로부터 표본자기상관함수(sample autocorrelation function; SACF)

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

을 얻을 수 있다. 또한 표본부분자기상관함수(sample partial autocorrelation function; SPACF) $\hat{\phi}_{hh}$는

$$\begin{pmatrix} \hat{\rho}(0) & \hat{\rho}(1) & \cdots & \hat{\rho}(h-1) \\ \hat{\rho}(1) & \hat{\rho}(0) & \cdots & \hat{\rho}(h-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(h-1) & \hat{\rho}(h-2) & \cdots & \hat{\rho}(0) \end{pmatrix} \begin{pmatrix} \hat{\phi}_{h1} \\ \hat{\phi}_{h2} \\ \vdots \\ \hat{\phi}_{hh} \end{pmatrix} = \begin{pmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ \vdots \\ \hat{\rho}(h) \end{pmatrix}$$

를 풀어 얻어진다. 이들은 각각 자기공분산함수, 자기상관함수, 부분자기상관함수의 추정량이다.

따라서 ACF, PACF를 이용한 ARMA모형의 식별은 먼저 주어진 시계열로부터 SACF와 SPACF를 계산한 뒤, 그들의 개형이 언제부터 절단되는지, 혹은 언제부터 지수적으로 감소하는지 확인함으로써 ARMA 모형의 시차 p, q를 결정하는 방식으로 진행된다. 그러나 이러한 식별 방법은 표본으로부터 얻은 SACF와 SPACF가 언제부터 지수적으로 감소하는지 정확히 판단하기 어려워 인간의 휴리스틱에 일부 기댄다는 문제점이 있다.

Definition 18. *ARMA 모형에서, AIC(Akaike's information criterion)은 아래와 같이 정의된다.*

$$AIC(p,q) = n\log(\hat{\sigma}^2) + 2(p+q)$$

또한, AICc(corrected AIC)는 여기에 correction term이 붙어

$$AICc(p,q) = n\log(\hat{\sigma}^2) + n + \frac{n}{n-(p+q+1)}2(p+q)$$

로 정의된다.

Definition 19. *ARMA 모형에서, BIC(Bayesian information criterion)은 아래와 같이 정의된다.*

$$AIC(p,q) = n\log(\hat{\sigma}^2) + (p+q)\log(n)$$

Definition 20. *ARMA 모형에서, HQC(Hannan-Quinn information criterion)는 아래와 같이 정의된다.*

$$HQC(p,q) = n\log(\hat{\sigma}^2) + 2(p+q)\log(\log(n))$$

모형 적합에서, 이러한 통계량들이 작을수록 모형이 잘 적합되었음을 의미한다. 따라서 이들 적당한 시차 p, q에 대하여 이러한 정보기준들을 계산하고, 그리드 서치를 통하여 이들이 최소화되는 p, q을 구하면 모형선택을 SACF, SPACF에 대한 휴리스틱한 선택 없이 수행해줄 수 있다.

2021년

Problem. 확률과정 Z_t가 아래와 같은 MA(2) 모형을 따른다고 할 때 다음 물음에 답하시오.

$$Z_t = \epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2}, \quad \epsilon_t \sim W.N.(0,1)$$

Z_t의 자기공분산 함수를 구하시오.

Solution.

$$\begin{aligned}\gamma(0) &= \mathrm{Cov}(Z_t, Z_t) \\ &= \mathrm{Cov}(\epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2}, \epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2}) \\ &= (1^2 + 2.4^2 + 0.8^2)\mathrm{Cov}(\epsilon_t, \epsilon_t) \\ &= 7.4\end{aligned}$$

$$\begin{aligned}\gamma(1) &= \mathrm{Cov}(Z_t, Z_{t-1}) \\ &= \mathrm{Cov}(\epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2}, \epsilon_{t-1} + 2.4\epsilon_{t-2} + 0.8\epsilon_{t-3}) \\ &= 2.4 + 0.8 \times 2.4 = 4.32\end{aligned}$$

$$\begin{aligned}\gamma(2) &= \mathrm{Cov}(Z_t, Z_{t-2}) \\ &= \mathrm{Cov}(\epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2}, \epsilon_{t-2} + 2.4\epsilon_{t-3} + 0.8\epsilon_{t-4}) \\ &= 0.8\end{aligned}$$

이며 이에 따라 $\gamma(-1) = \gamma(1) = 4.32, \gamma(-2) = \gamma(2) = 0.8$이다. 또한 $k \geq 3$인 경우

$$\begin{aligned}\gamma(k) &= \mathrm{Cov}(Z_t, Z_{t-k}) \\ &= \mathrm{Cov}(\epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2}, \epsilon_{t-k} + 2.4\epsilon_{t-k-1} + 0.8\epsilon_{t-k-2}) = 0\end{aligned}$$

이 $t-k-2 < t-k-1 < t-k < t-2 < t-1 < t$에서 성립한다. 따라서 Z_t의 자기공분산함수는

$$\gamma(h) = \begin{cases} 7.4 & h = 0 \\ 4.32 & |h| = 1 \\ 0.8 & |h| = 2 \\ 0 & |h| \geq 3 \end{cases}$$

이다.

Problem. 위 문제에서, 확률과정 Z_t가 가역성(invertibility)을 만족하지 않음을 보이시오.

Solution.

Definition 21. *어떠한 확률과정 Z_t가 ARMA 모형을 따를 때, 만약*

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

인 π_j가 존재한다면 이 시계열은 가역적인 시계열이다. 만약

$$\phi(L) Z_t = \theta(L) \epsilon_t$$

와 같이 ARMA 모형이 표현된다면, 이 시계열이 가역적일 필요충분조건은 $\theta(z)$의 근 중 $|z| \leq 1$인 것이 없는 것이다.

우리의 경우 $\theta(z) = 1 + 2.4z + 0.8z^2$이며, 그 근은

$$z = \frac{-2.4 \pm \sqrt{2.4^2 - 4 \times 1 \times 0.8}}{2 \times 0.8} = -\frac{1}{2}, -\frac{5}{2}$$

이다. 따라서 절대값이 1보다 작은 근 $-\frac{1}{2}$이 존재함에 따라, 이 확률과정은 가역성을 만족하지 못한다.

Problem. 확률과정 Z_t와 동일한 자기공분산함수를 가지면서 가역성 조건을 만족하는 새로운 MA(2) 과정을 찾으시오.

Solution. 원하는 시계열을 W_t라고 하고,

$$W_t = \epsilon_t + a\epsilon_{t-1} + b\epsilon_{t-2}, \quad \epsilon_t \sim W.N.(0, \sigma^2)$$

라고 하자. 그렇다면 이 시계열의 자기공분산함수는

$$\tilde{\gamma}(h) = \begin{cases} (1 + a^2 + b^2)\sigma^2 & h = 0 \\ (a + ab)\sigma^2 & |h| = 1 \\ b\sigma^2 & |h| = 2 \\ 0 & |h| \geq 3 \end{cases}$$

이다. 이것이 Z_t의 자기공분산함수와 동일해야 하므로,

$$\begin{aligned} (1 + a^2 + b^2)\sigma^2 &= 7.4 \\ (a + ab)\sigma^2 &= 4.32 \\ b\sigma^2 &= 0.8 \end{aligned}$$

이 풀고자 하는 문제이다. 셋째 식을 둘째 식에 넣으면, $a(\sigma^2 + 0.8) = 4.32$이다. 그러면

$$\sigma^2 + a^2\sigma^2 + b^2\sigma^2 = \sigma^2 + \frac{108^2\sigma^2}{25^2(\sigma^2 + 0.8)^2} + \frac{16}{25\sigma^2} = \frac{37}{5}$$

이고

$$625\sigma^4(\sigma^2 + 0.8)^2 + 108^2\sigma^4 + 400(\sigma^2 + 0.8)^2 = 4625\sigma^2(\sigma^2 + 0.8)^2$$

이다. $\sigma^2 = w$로 치환하면

$$625w^4 + 1000w^3 + 400w^2 + 11664w^2 + 400w^2 + 640w + 256 = 4625w^3 + 7400w^2 + 2960w$$

이고 정리하면

$$625w^4 - 3625w^3 + 5064w^2 - 2320w + 256 = 0$$

이다. 양변을 w^2으로 나누면,

$$625w^2 - 3625w - \frac{2320}{w} + \frac{256}{w^2} = -5064$$

이다. 그러면

$$k = 25w + \frac{16}{w}$$

으로 두면,

$$k^2 = 625w^2 + 800 + \frac{256}{w^2}$$

이므로, 주어진 식은

$$k^2 - 145k + 4264 = 0$$

이다. 그런데 $w = 1$일 때 $k = 41$이 근임을 Z_t로부터 안다. 따라서

$$(k - 41)(k - 104) = 0$$

로 인수분해하여 $k = 41, 104$를 근으로 얻는다. 먼저 $k = 41$일 때 $w = 1$ 외의 다른 근을 찾아 보자.

$$25w + \frac{16}{w} = 41$$

은 정리하면

$$25w^2 - 41w + 16 = (w - 1)(25w - 16) = 0$$

이고, 원하는 근은 $w = \sigma^2 = \frac{16}{25}$이다. 다시 대입하면,

$$b = 0.8 \times \frac{25}{16} = 1.25$$

이고,

$$a = 4.32 \times \frac{1}{\frac{16}{25} + 0.8} = 3$$

이다. 그러나 여기에 대응되는 $1 + 3z + 1.25z^2$는 절대값이 1보다 작은 근 -0.4를 가지므로, 가역시계열은 될 수 없다.

그 다음으로는 $k = 104$에 대응되는 w를 찾으면,

$$25w + \frac{16}{w} = 104$$

에서

$$25w^2 - 104w + 16 = (25w - 4)(w - 4) = 0$$

이므로 $w = 4$과 $w = \frac{4}{25}$를 근으로 얻는다. 먼저 $w = 4$인 경우에는 $b = 0.2, a = 0.9$를 얻으며,

$$W_t = \epsilon_t + 0.9\epsilon_{t-1} + 0.2\epsilon_{t-2}, \quad \epsilon_t \sim W.N.(0, 4)$$

는 $\theta(z)$의 근이

$$z = \frac{-0.9 \pm \sqrt{0.9^2 - 0.8}}{2 \times 0.2} = -\frac{5}{2}, -2$$

으로 모두 절대값이 1보다 크기에 가역시계열이다. $w = \frac{4}{25}$인 경우에는 $b = 5$, $a = 4.5$를 얻을 수 있으며,

$$W_t = \epsilon_t + 4.5\epsilon_{t-1} + 5\epsilon_t, \quad \epsilon_t \sim W.N.(0, 0.16)$$

은 $\theta(z)$의 근이

$$z = \frac{-4.5 \pm \sqrt{20.25 - 20}}{2 \times 5} = -\frac{1}{2}, -\frac{2}{5}$$

로 절대값이 1보다 작은 근이 있기에 가역시계열이 아니다.
따라서 이를 종합하면, 주어진 Z_t와 동일한 자기공분산함수를 공유하는 가역시계열은

$$W_t = \epsilon_t + 0.9\epsilon_{t-1} + 0.2\epsilon_{t-2}, \quad \epsilon_t \sim W.N.(0, 4)$$

가 유일하다.
Note. 더 쉬운 답안도 있다. 예를 들어

$$Z_t = (1 - aB)(1 - bB)\epsilon_t$$

과 동일한 자기상관함수를 가지는 MA 시계열을 찾으려면,

$$W_t = (a - B)(b - B)\epsilon_t$$

와 같이 뒤섞으면 된다. 위의 경우

$$Z_t = \epsilon_t + 2.4\epsilon_{t-1} + 0.8\epsilon_{t-2} = (1 - 0.4B)(1 - 2B)\epsilon_t$$

이므로, 우리가 찾는 건 정상성을 위해 $(1 - 2B)$ 부분을 뒤집은 시계열이므로,

$$W_t = (1 - 0.4B)(2 - B)\epsilon_t = (1 - 0.4B)(1 - 0.5B)(2\epsilon_t) = (2\epsilon_t) - 0.9(2\epsilon_{t-1}) + 0.2(2\epsilon_{t-2})$$

이 답안이 된다. 스케일만 조정해주자.

2019년

Problem. 다음과 같은 정상 AR(1) 모형을 고려하라.

$$X_t = \phi_0 + \phi_1 X_{t-1} + \epsilon_t, \quad \{\epsilon_t\} \sim_{i.i.d.} N(0, \sigma^2),\ t = 1, 2, \cdots, n$$

적률추정법과 $X_1 = x_1$이 주어졌을 때의 조건부최대가능도추정법을 이용해 ϕ_0, ϕ_1을 추정하시오.

Solution.

Definition 22. *정상 AR(p) 모형*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t, \quad \epsilon_t \sim_{i.i.d.} (0, \sigma^2)$$

에 대하여,

$$\mathbb{E}[(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}) X_{t-j}] = \mathbb{E}[\epsilon_t X_{t-j}] = 0$$

이 $j = 1, 2, \cdots, p$에 대해 성립한다. 따라서 자기공분산함수를 통해 이를 표현하면

$$\gamma(j) = \phi_1 \gamma(j-1) + \cdots + \phi_p \gamma(j-p)$$

를 얻고, 이를 Yule-Walker 방정식이라 부른다.

Definition 23. *Yule-Walker 방정식을 행렬로 표현하면,*

$$\begin{pmatrix} \rho(0) & \rho(1) & \cdots & \rho(p-1) \\ \rho(1) & \rho(0) & \cdots & \rho(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \cdots & \rho(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(p) \end{pmatrix}$$

을 얻는다. 따라서 여기의 자기상관함수 $\rho(j)$를 표본자기상관함수 $\hat{\rho}(j)$로 대체하여 ϕ_j의 추정량

$$\hat{\phi}_j = \frac{\left| \begin{pmatrix} \hat{\rho}(0) & \hat{\rho}(1) & \cdots & \hat{\rho}(1) & \cdots & \hat{\rho}(p-1) \\ \hat{\rho}(1) & \hat{\rho}(0) & \cdots & \hat{\rho}(2) & \cdots & \hat{\rho}(p-2) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \hat{\rho}(p-1) & \hat{\rho}(p-2) & \cdots & \hat{\rho}(p) & \cdots & \hat{\rho}(0) \end{pmatrix} \right|}{\left| \begin{pmatrix} \hat{\rho}(0) & \hat{\rho}(1) & \cdots & \hat{\rho}(j) & \cdots & \hat{\rho}(p-1) \\ \hat{\rho}(1) & \hat{\rho}(0) & \cdots & \hat{\rho}(j-1) & \cdots & \hat{\rho}(p-2) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \hat{\rho}(p-1) & \hat{\rho}(p-2) & \cdots & \hat{\rho}(p-j) & \cdots & \hat{\rho}(0) \end{pmatrix} \right|}$$

을 얻을 수 있으며, 이를 ϕ_j의 적률추정량 혹은 Yule-Walker 추정량이라 부른다.

우리 모형의 경우

$$\mathbb{E}[(X_t - \phi_0 - \phi_1 X_{t-1})X_{t-1}] = 0$$

임을 이용하면 $\gamma(1) = \phi_0\mathbb{E}[X_{t-1}] + \phi_1\gamma(0)$ 이고, 평균을 이용하면

$$\mathbb{E}[X_t] = \mathbb{E}[\phi_0 + \phi_1 X_{t-1} + \epsilon_t] = \phi_0 + \phi_1\mathbb{E}[X_{t-1}]$$

이다. X_t는 정상이기에,

$$\begin{cases} \phi_0\mathbb{E}[X_t] + \phi_1\gamma(0) & = \gamma(1) \\ \phi_0 + \phi_1\mathbb{E}[X_t] & = \mathbb{E}[X_t] \end{cases}$$

을 풀면

$$\phi_0 = \frac{\mathbb{E}[X_t](\gamma(1) - \gamma(0))}{\mathbb{E}[X_t]^2 - \gamma(0)}$$
$$\phi_1 = \frac{\mathbb{E}[X_t]^2 - \gamma(1)}{\mathbb{E}[X_t]^2 - \gamma(0)}$$

을 얻는다. 따라서 각각의 적률추정량은 표본자기공분산함수와 표본평균

$$\hat{\mathbb{E}}[X_t] = \bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$
$$\hat{\gamma}(0) = \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2$$
$$\hat{\gamma}(1) = \frac{1}{n}\sum_{t=2}^{n}(X_t - \bar{X})(X_{t-1} - \bar{X})$$

을 이용하여

$$\hat{\phi}_0 = \frac{\bar{X}(\hat{\gamma}(1) - \hat{\gamma}(0))}{\bar{X}^2 - \hat{\gamma}(0)}$$
$$\hat{\phi}_1 = \frac{\bar{X}^2 - \hat{\gamma}(1)}{\bar{X}^2 - \hat{\gamma}(0)}$$

으로 주어진다.

한편 x_1이 주어지는 경우, 조건부 가능도함수는

$$\begin{aligned} L(\phi_0, \phi_1, \sigma^2; x_2, \cdots, x_n | X_1 = x_1) &= f_{X_2,\cdots,X_n|X_1}(x_2, x_3, \cdots, x_n | X_1 = x_1; \phi_0, \phi_1) \\ &= f_{X_2|X_1}(x_2|X_1 = x_1; \phi_0, \phi_1) \times f_{X_3|X_2,X_1}(x_3|X_2 = x_2, X_1 = x_1; \phi_0, \phi_1) \\ &\times \cdots \times f_{X_n|X_{n-1},\cdots,X_1}(x_n|X_{n-1} = x_{n-1}, \cdots, X_2 = x_2, X_1 = x_1; \phi_0, \phi_1) \\ &= \prod_{t=2}^{n} f_{X_t|X_{t-1}}(x_t|X_{t-1} = x_{t-1}; \phi_0, \phi_1) \\ &= \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_t - \phi_0 - \phi_1 x_{t-1})^2}{2\sigma^2}\right) \end{aligned}$$

이며 조건부 로그가능도함수는

$$l(\phi_0, \phi_1, \sigma^2; x_2, \cdots, x_n | X_1 = x_1) = -\frac{n-1}{2}\log(2\pi) - \frac{n-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(x_t - \phi_0 - \phi_1 x_{t-1})^2$$

으로 주어진다. ϕ_0, ϕ_1으로 이들을 각각 미분하면

$$\frac{\partial l}{\partial \phi_0} = \frac{1}{2\sigma^2}\sum_{t=2}^{n} 2(x_t - \phi_0 - \phi_1 x_{t-1})$$
$$\frac{\partial l}{\partial \phi_1} = \frac{1}{2\sigma^2}\sum_{t=2}^{n} 2x_{t-1}(x_t - \phi_0 - \phi_1 x_{t-1})$$

을 얻고, 일계조건으로부터

$$\phi_0(n-1) + \phi_1\sum_{t=2}^{n} x_{t-1} = \sum_{t=2}^{n} x_t$$
$$\phi_0\sum_{t=2}^{n} x_{t-1} + \phi_1\sum_{t=2}^{n} x_{t-1}^2 = \sum_{t=2}^{n} x_t x_{t-1}$$

을 얻고 조건부최대가능도추정량으로

$$\hat{\phi}_0 = \frac{\sum_{t=2}^{n} x_t \sum_{t=2}^{n} x_{t-1}^2 - \left(\sum_{t=2}^{n} x_{t-1}\right)^2}{(n-1)\sum_{t=2}^{n} x_{t-1}^2 - \left(\sum_{t=2}^{n} x_{t-1}\right)^2}$$
$$\hat{\phi}_1 = \frac{(n-1)\sum_{t=2}^{n} x_t x_{t-1} - \left(\sum_{t=2}^{n} x_{t-1}\right)\left(\sum_{t=2}^{n} x_t\right)}{(n-1)\sum_{t=2}^{n} x_{t-1}^2 - \left(\sum_{t=2}^{n} x_{t-1}\right)^2}$$

을 얻는다.

2013년

Problem. 2개 그룹으로부터 다음과 같은 확률표본 $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$를 얻었다고 할 때, 다음 물음에 답하시오.

$$\text{그룹A} : \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$
$$\text{그룹B} : \begin{pmatrix} -3 \\ 5 \end{pmatrix}, \begin{pmatrix} -2 \\ 7 \end{pmatrix}, \begin{pmatrix} -1 \\ 6 \end{pmatrix}$$

그룹A와 그룹B를 좌표평면에 표시하고, 그룹별 평균 벡터와 공분산 행렬을 구하시오.

Solution. 평균벡터는 단순히 평균을 구하여

$$\hat{\mu}_A = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$
$$\hat{\mu}_B = \begin{pmatrix} -2 \\ 6 \end{pmatrix}$$

을 얻는다. 공분산 행렬은 표본공분산으로써

$$S_A = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$
$$S_B = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

에서 동일하게 나타난다. 좌표평면에의 표현은 아래 문제에서 함께 확인한다.

Problem. Fisher의 판별함수를 구하고 $\begin{pmatrix} -1 \\ 5 \end{pmatrix}$는 어느 그룹으로 분류되는지 판별하시오.

Definition 24. *공분산이 같고 평균만 다른 두 다변량 분포 $A \sim (\mu_A, \Sigma), B \sim (\mu_B, \Sigma)$에 대하여, Fisher의 판별함수, 혹은 separating hypterplane은*

$$a^T X = b$$

형태이다. 이때 a는

$$a = \text{argmax}_a \frac{(a^T \mu_B - a^T \mu_A)}{a^T \Sigma a}$$

을 만족하는 벡터이며, 간단히 쓰면 $a = \Sigma^{-1}(\mu_B - \mu_A)$이다. b는 $b = \frac{1}{2}(\mu_A + \mu_B)$로 설정한다. 만약 어떠한 새로운 표본 X가 $a^T X > b$를 만족한다면 이는 B에서 비롯된 것으로, 반대라면 A에서 비롯된 것으로 판정할 수 있다.

한편 표본 버전으로 하는 경우,

$$\hat{a} = S_{\text{pooled}}^{-1}(\bar{x}_B - \bar{x}_A)$$
$$\hat{b} = \frac{\bar{x}_A + \bar{x}_B}{2}$$

를 사용하여 자료에 대한 판별함수를

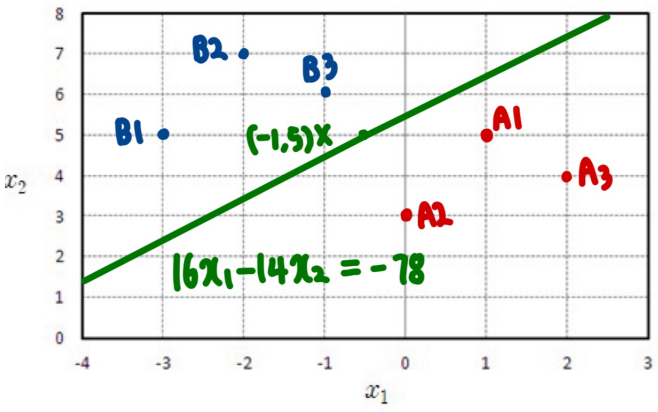
$$\phi(\mathbf{x}) = \begin{cases} A & \text{if } (\bar{x}_B - \bar{x}_A)^T S_{\text{pooled}}^{-1}\mathbf{x} < \frac{1}{2}(\bar{x}_B - \bar{x}_A)^T S_{\text{pooled}}^{-1}(\bar{x}_B + \bar{x}_A) \\ B & \text{if } (\bar{x}_B - \bar{x}_A)^T S_{\text{pooled}}^{-1}\mathbf{x} > \frac{1}{2}(\bar{x}_B - \bar{x}_A)^T S_{\text{pooled}}^{-1}(\bar{x}_B + \bar{x}_A) \end{cases}$$

처럼 설정할 수 있다. 우리의 경우

$$\bar{x}_A = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$
$$\bar{x}_B = \begin{pmatrix} -2 \\ 6 \end{pmatrix}$$
$$S_{\text{pooled}} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

이므로, 판별함수는

$$\phi(x_1, x_2) = \begin{cases} A & \text{if } \frac{16}{3}x_1 - \frac{14}{3}x_2 < -26 \\ B & \text{if } \frac{16}{3}x_1 - \frac{14}{3}x_2 > -26 \end{cases}$$



으로 위 그림처럼 주어진다. 한편 $\begin{pmatrix} -1 \\ 5 \end{pmatrix}$는 판별함수에서 $\frac{16}{3} - \frac{70}{3} = -\frac{54}{3} = -18$로 -26보다 크기 때문에, B 그룹으로 분류된다.

Problem. A제과점에서는 1년 동안 새로 출시된 9개 빵에 대해 제과점 내부 평가점수와 고객의 선호도간 관계를 조사하고자 한다. 아래 자료를 바탕으로 평가점수와 고객 선호도 간 양의 연관성이 있다고 할 수 있는지를 유의수준 10%에서 켄달의 독립성 검정과 스피어만 상관계수를 각각 이용해 검정하시오.

	1	2	3	4	5	6	7	8	9
내부 평가점수	22.2	23.0	21.0	26.7	22.4	22.0	25.4	22.6	29.0
선호도	5.2	6.2	5.0	9.6	7.2	8.0	9.8	5.6	7.6

Definition 25. *두 쌍 (X_i, Y_i)와 (X_j, Y_j)에 대하여, 만약 $X_i < X_j$, $Y_i < Y_j$이거나 $X_i > X_j$, $Y_i > Y_j$라면 이 두 쌍을 조화롭다(concordant)고 부르고, 반대로 $X_i < X_j$, $Y_i > Y_j$거나 $X_i > X_j$, $Y_i < Y_j$라면 이 두 쌍을 조화롭지 못하다(discordnat)고 부른다. 이를 이용하여 조화로우면 1, 조화롭지 않으면 -1의 값을 가지는*

$$U_{ij} = \begin{cases} +1, & (X_i - X_j)(Y_i - Y_j) > 0 \\ -1, & (X_i - X_j)(Y_i - Y_j) < 0 \\ 0, & (X_i - X_j)(Y_i - Y_j) = 0 \end{cases}$$

을 $1 \le i < j \le n$에 대해 정의할 수 있다. 그렇다면 이변량 자료

$$\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$$

의 켄달의 타우는

$$\tau_n = \frac{1}{\binom{n}{2}} \sum_{i<j} U_{ij}$$

이다. 이는 $\tau = 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1$의 추정량이기도 하다.

Proposition 5. *만약 (X_i, Y_i)가 이변량 정규분포를 따르는 랜덤표본이라면, 피어슨 상관계수 ρ에 대하여*

$$\tau = \frac{2}{\pi} \arcsin(\rho)$$

가 성립한다.

Proposition 6. *만약 (X_i, Y_i)가 이변량 정규분포를 따르는 랜덤표본이라면, n이 클 때*

$$\tau_n \dot{\sim} N\left(0, \frac{4n+10}{9n(n-1)}\right)$$

이다.

켄달의 타우를 이용하면 비모수적 순열검정을 통해 구하는 표본의 독립성 검정을 할 수 있다. 이때 귀무가설과 대립가설은

$$H_0 : \tau = 0 \text{ (내부 평가점수와 선호도는 독립적이다.) vs. } H_1 : \tau > 0$$

으로 주어진다. 비모수적 상황이므로 순열검정을 통해 검정하려 했으나, 순열검정 시에는 내부 평가점수는 고정하고 선호도만 뒤섞는 총 9!개의 총 뒤섞기가 존재한다. 따라서 이를 직접 수행하는 것은 불가능하다. 다행히도 켄달통계량의 분포표가 뒤에 주어져 있다. 여기에서 $K = 16$이고, $n = 9$이면

$$P_0(K \ge 16) = 0.060$$

으로 유의수준 0.1보다 작다. 따라서 유의수준 0.1에서 귀무가설을 기각할 수 있다.
한편 스피어만의 상관계수를 계산하는 경우, 그 값은 계산을 통해

$$r_s = 0.6$$

으로 얻어지며, 분포표에서 $r_s(0.1, 9) = 0.4667$보다 크므로 귀무가설

$$H_0 : r_s = 0 \text{ (내부 평가점수와 선호도는 독립적이다.)}$$

을 기각할 수 있다. (이때, 대립가설은 $H_1 : r_s > 0$이다.)

2012년

Problem. A공장에서 생산되는 7개의 전구를 무작위로 추출하여 그 수명을 측정한 결과가 다음과 같다. 전구의 평균 수명이 35시간보다 길다고 할 수 있는지에 대해 아래 3가지 검정을 시행하시오. (단, 유의수준 5%)

25	16	44	62	36	58	38

(표본평균 $\overline{X}=39.857$, 표본표준편차 $S=16.557$)

보기: t-검정, 부호검정, 정규분포 근사를 이용한 검정

Solution. 먼저 t검정을 이용할 때에는, 귀무가설 하에서

$$T = \frac{\bar{X} - 35}{S/\sqrt{7}} \sim t(6)$$

임을 이용할 수 있다. 검정통계량의 값이

$$t = \frac{39.857 - 35}{16.557/\sqrt{7}} = 0.776$$

으로 자유도가 6인 t분포의 95퍼센트 퀀타일보다 훨씬 작다. 따라서 귀무가설을 기각할 수 없으며, 평균 수명이 35시간보다 길다고 말할 충분한 근거가 없다.

Definition 26. *중앙값이 μ라는 귀무가설에 대한 부호검정의 검정통계량은 아래와 같이 정의된다.*

$$S = \sum_{i=1}^{n} I(X_i > \mu)$$

만약 분포의 실제 중앙값이 μ보다 크다면 S는 큰 경향성을 가질 것이며, μ보다 작다면 작은 값을 가지게 될 것이다. 따라서 해당 값이 얼마나 극단적인지를 통해 중앙값에 대한 검정을 수행할 수 있다.

우리의 상황에서는 S가 크게 나오는 경우 귀무가설을 기각하여 평균 수명이 35시간보다 길다고 말할 수 있다. 이때 우리가 얻은 데이터로부터, $S = 5$이며, p값은 $P_0(S \geq 5) = 1 - P(S < 4) = 0.2266$으로 유의수준 0.05보다 크다. 따라서 귀무가설을 기각할 만한 충분한 근거가 없다.
정규분포 근사를 이용하는 경우, t-검정과 유사하되 그 분포가 $t(6)$가 아닌 정규분포를 따른다고 근사할 수 있다. 앞서 확인하였듯 그 값은 0.776으로 1.96보다 작기에, 평균 수명이 35시간보다 길다고 말할 충분한 근거가 없다.

2011년

문제 없음

2010년

Problem.

□ 서로 독립인 확률표본 $X_1,\ldots,X_n$은 정규분포 $N(\theta,\sigma^2)$을 따르고, θ의 사전분포(prior distribution)는 정규분포 $N(\theta_0,\sigma_0^2)$인 확률밀도함수 $p(\theta)$를 가진다. 이 때 $X_i=x_i$ $(i=1,\ldots,n)$로 주어졌을 때 θ의 사후분포(posterior distribution)의 확률밀도함수를 $f(\theta|x_1,\ldots,x_n)$이라고 하자 (단, σ^2, θ_0, σ_0^2은 알려져 있다). 다음 중 **옳은** 것은?

A. $f(\theta|x_1,\cdots,x_n)=p(\theta)\times L(\theta|x_1,\ldots,x_n)$
(단, $L(\theta|x_1,\ldots,x_n)$는 우도함수(likelihood function))

B. θ의 사후분포는 평균 $\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\cdot\theta_0+\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\cdot\bar{x}$ 를 갖는다. (단, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$)

C. θ의 사후분포는 분산 $\frac{\sigma_0^2/n}{\sigma_0^2+\sigma^2/n}\cdot\sigma^2$ 을 갖는다.

D. 표본크기가 클수록 θ의 베이즈 추정량(제곱오차 손실함수 사용)은 사전분포의 평균 θ_0에 근접한다.

E. θ에 대한 95% 확률의 고밀도(HDR; highest density region) 신용구간(credible region)은 전통적 분석의 95% 신뢰구간과 같다.

Solution.

Note 1. ***(베이즈 추론의 세 가지 요소)***

1. *사전분포(prior distribution): 자료를 관측하기 전 모수 θ의 분포*
2. *모수모형 하에서의 가능도: $X|\theta\sim f(x|\theta)$*
3. *사후분포(posterior distribution): 자료를 관측한 뒤 가능도로써 업데이트한 모수 θ의 분포*

Theorem 1. ***(베이즈 규칙)***
θ의 사전밀도함수 $\pi(\theta)$와 $X|\theta$의 가능도함수 $f(x|\theta)$가 주어졌을 때, θ의 사후분포는

$$\pi(\theta|x)=\frac{\pi(\theta)\cdot f(x|\theta)}{\int_\Theta \pi(\theta)\cdot f(x|\theta)d\theta}\propto\pi(\theta)f(x|\theta)$$

처럼 주어진다. 이때 Θ는 모수공간이다.

Definition 27. *베이즈 추정량은 사후분포를 한 점으로써 요약할 수 있는 값이다.*
(1) ***사후평균***

$$E[\theta|x] = \int \theta\pi(\theta|x)d\theta = \hat{\theta}^B$$

(2) ***최대사후밀도 추정량****(Maximum a posteriori; MAP)*

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \pi(\theta|x) = \arg\max_{\theta} \pi(\theta) \cdot L(\theta; x)$$

만약 $\pi(\theta)$가 상수라면, $\hat{\theta}^{MAP} = \hat{\theta}^{MLE}$이다. 즉 MAP는 사전분포를 고려하는 generalized MLE이다.
(3) ***사후분포의 중앙값***

$$\hat{\theta}^{med} = \arg\min_{a} \int |\theta - a|\pi(\theta|x)d\theta$$

Note 2. *베이즈 추정량은 추정량이 실제 값을 잘못 추정했을 때의* ***손실함수(loss function)****을 적절히 두었을 때 그 기댓값인* ***베이즈 위험****을 최소로 하는 값과도 같다. 즉 θ가 실제 모수일 때 $\hat{\theta}$로 잘못 추정함에 따른 손실함수를 $L(\theta, \hat{\theta})$라고 둘 때, 베이즈 기대손실은*

$$R(\pi, \hat{\theta}) = E_{\theta|x}[L(\theta, \hat{\theta})] = \int_{\Theta} L(\theta, \hat{\theta})\pi(\theta|x)d\theta$$

처럼 주어진다. 특정 손실함수를 가정할 때의 베이즈 추정량 $\hat{\theta}^{\pi}$은

$$\hat{\theta}^{\pi} = \arg\min_{\hat{\theta}\in\Theta} \int_{\Theta} L(\theta, \hat{\theta})\pi(\theta|x)d\theta$$

으로 주어진다. (우리는 위에서 $\hat{\theta}$ 대신 action의 준말로서 a를 사용하였다.)
사후평균은 손실함수를 제곱손실

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

으로 둘 때의 베이즈 추정량이며, 최대사후밀도추정량은 손실함수를 0-1 손실

$$L(\theta, \hat{\theta}) = \begin{cases} 1 & \hat{\theta} = \theta \\ 0 & \hat{\theta} \neq \theta \end{cases}$$

로 둘 때의 베이즈 추정량이다. 마지막으로 사후분포의 중앙값은 절대오차손실

$$L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

하에서의 베이즈 추정량이다.

Definition 28. ***(신용집합; credible set)***
$\alpha \in (0, 1)$에 대하여, 모수공간의 집합 $C \subseteq \Theta$가

$$\pi(\theta \in C|x) \geq 1 - \alpha$$

라면 이 C를 θ의 $100(1-\alpha)\%$ 신용집합이라고 부른다.

Definition 29. ***(동일꼬리 신용집합; equal tail credible set)***
$\Theta = \mathbb{R}$이며 $100(1-\alpha)\%$ 신용집합 C가 어떤 실수 a, b에 대하여 (a, b)의 형태라고 하자. 만약

$$\pi(\theta < a|x) = \pi(\theta > b|x) = \frac{\alpha}{2}$$

로 신용집합의 양 꼬리에 대한 포함확률이 $\alpha/2$로 동일하다면, 이 C를 ***동일꼬리 신용집합****이라 부른다.*

Definition 30. ***(최고사후밀도 신용집합; highest posterior density credible set)***

신용집합 $C \subseteq \Theta$가

$$C_k = \{\theta \in \Theta : \pi(\theta|x) \geq k\}$$

인 형태면서 $P(\theta \in C_k|x) \geq 1-\alpha$라면, 이 C_k를 θ의 $100(1-\alpha)\%$ HPD 신용집합이라고 부른다.

Note 3. *$100(1-\alpha)\%$ 신용집합들을 고려할 때, 그 중 가장 체적이 가장 작은 신용집합은 HPD 신용집합이다. 왜 그런지 적당한 경우에 대하여 생각해보자. 신용집합 C가 있다고 할 때, 이것은 $100(1-\alpha)\%$ 신용집합이어야 하므로*

$$\int_C \pi(\theta|x)d\theta \ = 1-\alpha$$

라고 하자. 그렇다면 적절한 $100(1-\alpha)\%$ HPD 신용집합 C_k를 고려했을 때에도

$$\int_{C_k} \pi(\theta|x)d\theta = 1-\alpha$$

이다. 그렇다면 둘을 서로 빼

$$\int_{C-C_k} \pi(\theta|x)d\theta = \int_{C_k-C} \pi(\theta|x)d\theta$$

를 얻을 수 있을 것이다. 그런데 $\theta \in C-C_k$에 있다는 것은 곧 $\pi(\theta|x) < k$라는 것이기에

$$\int_{C-C_k} \pi(\theta|x)d\theta \leq kVol(C-C_k)$$

이다. 반면 $\theta \in C_k - C$라면 $\pi(\theta|x) \geq k$여야 하므로

$$\int_{C_k-C} \pi(\theta|x)d\theta \geq kVol(C_k-C)$$

이다. 둘이 서로 같아야 하므로

$$kVol(C-C_k) \geq kVol(C_k-C)$$

의 관계가 성립하고, k가 양수이므로 지워준다면 $C-C_k$의 체적이 C_k-C의 체적보다 크거나 같다. 따라서

$$Vol(C) = Vol(C-C_k) + Vol(C \cap C_k) \geq Vol(C_k-C) + Vol(C \cap C_k) = Vol(C_k)$$

임을 얻는다.

이제 문제를 풀어 보자. 베이즈 정리에 의하여,

$$f(\theta|x_1,\cdots,x_n) = \frac{p(\theta) \times L(\theta|x_1,\cdots,x_n)}{\int_\Theta p(\theta) \times L(\theta|x_1,\cdots,x_n)d\theta}$$

이며, 일반적으로는 $f(\theta|x_1,\cdots,x_n) \propto p(\theta) \times L(\theta|x_1,\cdots,x_n)$만이 성립한다. 한편 우리의 경우

$$\begin{aligned}
f(\theta|x_1,\cdots,x_n) &\propto p(\theta) \times L(\theta|x_1,\cdots,x_n) \\
&\propto \exp(-(\theta-\theta_0)^2/2\sigma_0^2) \times \prod_{i=1}^n \exp(-(x_i-\theta)^2/2\sigma^2) \\
&\propto \exp\left(-\frac{\theta^2-2\theta\theta_0}{2\sigma_0^2} - \frac{n\theta^2-2\theta\sum_{i=1}^n x_i}{2\sigma^2}\right) \\
&\propto \exp((\theta-\theta_1)^2/2\tau^2)
\end{aligned}$$

이며, 이때

$$\theta_1 = \frac{n\sigma^{-2}\bar{x} + \sigma_0^{-2}\theta_0}{n\sigma^{-2} + \sigma_0^{-2}}, \quad \tau^2 = (n\sigma^{-2} + \sigma_0^{-2})^{-1}$$

이 성립함이 잘 알려져 있다. 문제에 나온 형태대로 정리하면, θ의 사후분포는 평균으로

$$\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\theta_0$$

을 가지며, 분산은

$$\frac{\sigma_0^2/n}{\sigma_0^2 + \sigma^2/n} \cdot \sigma^2$$

으로 B는 틀리고 C는 맞게 된다. 한편 표본크기가 커짐에 따라

$$\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n} \xrightarrow{p} 1, \quad \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \xrightarrow{p} 0$$

이므로 θ의 제곱오차 손실함수를 사용한 베이즈 추정량, 즉 사후평균은 $\bar{x}$에 근접한다.
마지막으로 E를 보자. θ에 대한 95퍼센트 HDR 신용집합은, 사후분포가 종상의 정규분포를 따름에 따라 사후분포 상에서 2.5퍼센타일부터 97.5퍼센타일까지에 해당하는

$$(\theta_1 - 1.96\tau, \theta_1 + 1.96\tau)$$

로 주어진다. 한편 전통적 분석의 95퍼센트 신뢰구간은, 사전분포에 대한 지식이 없으므로

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

로 주어진다. 즉 둘은 일반적으로 $\theta_1 \neq \bar{x}$, $\tau \neq \sigma/\sqrt{n}$임에 따라 다르다.

Problem. 표본공분산행렬 $\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$을 이용하여 주성분분석에 필요한 고유값과 고유벡터를 구하였다. 이때 첫 번째 주성분 계산에 사용된 고유값과 고유벡터를 각각 λ_1, e_1, 두 번째 주성분 계산에 사용된 고유값을 λ_2라고 한다. e_1의 둘째 원소가 $1/\sqrt{2}$라고 할 때, 각각을 구하시오.

Solution.
주어진 행렬의 고유값은 0.7과 1.3임을 쉽게 알 수 있다. 이때 그 중 더 큰 것이 첫 번째 주성분에 해당하므로, $\lambda_1 = 1.3, \lambda_2 = 0.7$이다. 또한 $\lambda_1 = 1.3$에 상응하는 고유벡터 중 둘째 원소가 $1/\sqrt{2}$인 것은

$$e_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

이다.

Problem. A헬스클럽에서는 왼손과 오른손의 악력을 비교하기 위해 8명의 회원을 선발하여 오른손과 왼손의 악력을 측정하였다. 아래의 측정결과를 보고 오른손의 악력이 왼손의 악력보다 더 세다고 할 수 있는지를 검정하려고 한다. 이를 위해 이용할 수 있는 모수적 검정방법과 비모수적 검정방법을 각각 1개씩 제시하고 그 중 하나를 택하여 검정통계량을 구하시오.

악력(握力) 측정 결과

회원번호	A102	A205	B100	A287	B059	B125	A031	B203
오른손	16	18	35	42	13	11	27	36
왼손	20	12	28	37	15	16	30	35

Solution. 모수적으로는 t검정, 비모수적으로는 부호검정을 사용할 수 있다. 여기에서는 비모수적 검정방법인 부호검정을 택한다. 이때 부호검정의 통계량은 회원 중 오른손의 악력이 왼손의 악력보다 더 큰 사람의 수인 4이다.

2009년

Problem.

( ) □ 계절조정에 관한 다음 설명 중 **옳은** 것은?

A. 계절성분은 대부분 일정한 1년 이내의 주기를 가지고 규칙적으로 반복되는 변동을 의미하며, 계절조정이란 계절성분을 추정하여 원계열로부터 얻은 추세·순환 계열에 이를 추가하는 절차이다.

B. 계절조정계열에는 자연재해, 파업 등으로 인한 불규칙 성분이 제거되어 있다.

C. 승법모형은 시계열의 수준에 따라 계절성분의 진폭이 달라질 경우에 사용할 수 있으며 변환을 통하여 가법모형으로의 전환이 가능하다.

D. 승법모형을 가정할 때 계절성분과 불규칙성분이 불안정할 경우 전년동기비를 이용하면 타 방법에 비해 추세와 순환성분을 비교적 정확하게 반영할 수 있다.

E. 불규칙성분의 추정계열의 시계열을 그려보았을 때, 지속적으로 하락한다면 시계열의 분해가 잘 된 것으로 판단할 수 있다.

Solution.

A. 그렇지 않다. 계절조정은 이를 추정한 뒤 추세 순환 계열에서 이를 제거하는 방법이다.

B. 아니다. 불규칙 성분은 계절조정에 포함되지 않는다.

C. 그렇다.

D. 아니다. 계절성분이 불안정한 경우 계절성분이 일정하다고 가정하는 전년동기비를 통한 계절조정이 불완전해 오류를 만들 수 있으므로, 오히려 부적절하다.

E. 아니다. 불규칙성분이 어떠한 시계열적 경향성을 가지고 있다는 것은 해당 성분이 불규칙하지 않다는 것이므로, 분해가 제대로 이루어지지 않았음을 암시한다. 이 경우 추세성분의 분해가 제대로 이루어지지 않았으리라 생각된다.

Problem. X_1, X_2, X_3의 표본공분산행렬을 이용하여 표본주성분을 구하는 과정에서 다음과 같은 고유값과 고유벡터를 얻었다.

$$\lambda_1 = 5.83, e_1 = (0.38, -0.92, 0)^T$$
$$\lambda_2 = 2.00, e_2 = (0, 0, 1)^T$$
$$\lambda_3 = 0.17, e_3 = (0.92, -0.38, 0)^T$$

이때, 첫 번째 주성분을 쓰고 첫 번째 주성분의 분산을 구하시오.

Solution. 첫 번째 주성분은 $0.38X_1 - 0.92X_2$이며, 그 분산은 5.83이다. 이에 대한 설명은 앞 참고.

Problem. 시계열 $\tilde{x}_t = \phi\tilde{x}_{t-1} + a_t$의 자기상관함수(ACF)를 구하고 ACF의 대략적인 형태를 $\phi > 0$와 $\phi < 0$의 경우로 나누어 약술하시오. 단 $|\phi| < 1$이며 a_t는 백색잡음이며, $\tilde{x}_t = x_t - \mu = x_t - \mathbb{E}[X_t]$이다.

Solution. $\gamma(h)$를 자기공분산함수로 하자. 그렇다면 $h \geq 1$일 때

$$\begin{aligned}\gamma(h) &= \mathrm{Cov}(\tilde{x}_t, \tilde{x}_{t+h}) \\ &= \mathrm{Cov}(\tilde{x}_t, \phi\tilde{x}_{t+h-1} + a_{t+h}) \\ &= \mathrm{Cov}(\tilde{x}_t, \phi\tilde{x}_{t+h-1}) \\ &= \phi\gamma(h-1)\end{aligned}$$

이다. 여기에서 양변을 $\gamma(0)$으로 나누면, ACF를 $\rho(h)$로 쓸 때

$$\rho(0) = 1, \quad \rho(h) = \phi\rho(h-1) = \phi^h$$

를 $h \geq 1$에 대해 얻는다. 따라서 자기상관함수는

$$\rho(h) = \phi^{|h|}$$

로 주어진다. $\phi > 0$이면 이는 h가 양수이며 증가함에 따라 지수적으로 감소하는 모습을 보이며, $\phi < 0$이면 진동하면서 절대값은 지수적으로 감소하는 모습을 보인다.

Problem. 외과수술에서 새로운 방법(A)이 개발되어 기존의 방법(B)과 비교하고자 한다. A, B 두 방법을 각각 7명의 환자에게 실시하여 회복일수를 관측한 결과가 아래와 같다. A가 B보다 더 효과적이라고 할 수 있는지 t-검정, 윌콕슨 순위합 검정을 실시하여 확인하고 두 검정의 결과를 비교하고 어느 것이 나은지, 그리고 그 이유는 무엇인지 설명하라.

A	9	11	12	16	17	19	49
B	15	18	20	21	23	25	67

Solution. 먼저 t-검정을 실시하면, Welch의 t검정을 수행하기에는 무리가 있는 상황이므로, 등분산을 가정하고 pooled t-검정을 실시할 수 있다.

$$S_p^2 = \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{7 + 7 - 2} = 254.67$$

이므로,

$$t = \frac{\bar{x} - \bar{y}}{S_p\sqrt{\frac{1}{7} + \frac{1}{7}}} = \frac{19 - 27}{15.96 \times 0.535} = -0.937$$

이다. 이는 유의수준 5퍼센트에서 자유도가 14인 t분포의 95퍼센타일인 기각역에 미치지 못하므로, A가 B보다 더 낫다는 충분한 근거가 없다.

Definition 31. *$X_1, \cdots, X_m \sim_{i.i.d.} F$이며 $Y_1, \cdots, Y_n \sim_{i.i.d.} G$이며 $F(x) = G(x + \Delta)$로 두 분포에 위치모*

수의 차이만이 있을 때, 귀무가설

$$H_0 : \Delta = 0, \quad v.s. \quad H_1 : \Delta \neq 0$$

을 검정하기 위한 부호순위검정의 검정통계량은

$$W = \sum_{i=1}^{n} R_i$$

이다. 이때 R_i는 총 $m+n$개의 관측값 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 중 Y_i의 순위를 의미한다. 만약 W가 너무 크거나 작다면, 귀무가설을 기각할 수 있다.

우리의 경우 W가 너무 크다면 귀무가설을 기각할 수 있다. 검정통계량 W의 값은 $4+7+9+10+11+12+14 = 67$이다. 한편 제시된 표 하에서 0.036이 $P(W \geq 67)$으로 나타났으므로, 유의수준 0.05보다 작아 귀무가설을 기각할 수 있게 된다. 따라서 $t-$검정은 귀무가설을 기각하지 못하는 반면, 윌콕슨 순위합 검정은 귀무가설을 기각할 수 있다. 이는 해당 자료들이 정규분포와는 거리가 먼 49와 67이라는 이상점을 가지고 있는 꼬리가 긴 분포로부터 추출되었기 때문으로 생각된다. 이러한 꼬리가 긴 분포 하에서는 부호순위검정과 같은 비모수적 검정들의 상대효율이 더 좋기에, 비모수적 검정이 선호될 수 있다(검정력 측면). 또한 자료의 개수가 적어 이 자료가 정규분포로부터 추출되었는지 확인하기 어렵기에, 분포에 무관하게 안정한 유의수준을 보장하는 비모수적 검정이 더 안전하다(유의수준 측면).

2008년

Problem. 시계열 $Z_t = 0.3Z_{t-1} + a_t - 0.5a_{t-1}$($a_t$는 백색잡음)에 대하여 $\dfrac{\gamma(k)}{\gamma(k-1)}$을 구하시오. 이때 $\gamma(k)$는 k시차 자기공분산을 의미하여, $k \geq 2$인 경우에 대해서만 구하시오.

Solution. $k \geq 2$인 경우,

$$\begin{aligned}\gamma(k) &= \mathrm{Cov}(Z_t, Z_{t+k}) \\ &= \mathrm{Cov}(Z_t, 0.3Z_{t+k-1} + a_{t+k} - 0.5a_{t+k-1}) \\ &= 0.3\mathrm{Cov}(Z_t, Z_{t+k-1}) = 0.3\gamma(k-1)\end{aligned}$$

이므로 둘의 비는 0.3이다.

Problem. 확률과정 Z_t가 공분산 정상성을 갖기 위한 조건을 기술하시오.

Solution.

Definition 32. *어떠한 다변량 확률과정 Z_t가 공분산 정상성을 갖는 확률과정, 혹은 약한 정상성을 가지는 확률과정이라는 것은,*

1. $\mathbb{E}[Z_t] = \boldsymbol{\mu}$ *for all* t
2. $Cov(Z_t, Z_{t-j}) = \Sigma_j$ *for all* t, *and given* j. Σ_j *is finite.*

임을 의미한다.

Problem. 다음의 확률과정 Z_t가 공분산 정상성을 갖는지 판단하고 그 근거를 제시하시오.

$$Z_t = 0.8Z_{t-1} - 0.15Z_{t-2} + a_t$$

이때 a_t는 백색잡음과정을 따른다.

Solution. 특성방정식 $\phi(z) = 1 - 0.8z + 0.15z^2$의 근 중 크기가 1 미만의 근이 없는 것을 보이는 것으로 충분하다. 근은 각각 $10/3, 2$으로 모두 단위원 밖에 있기에, 이 확률과정은 공분산 정상성을 가진다.

Problem. 다음의 확률과정 Z_t가 공분산 정상성을 갖는지 판단하고 그 근거를 제시하시오.

$$Z_t = \begin{cases} X_t & 2|t \\ Y_t & 2 \nmid t \end{cases}$$

이때 $X_t \sim N(0,1)$, $P(Y_t = 1) = P(Y_t = -1) = 0.5$이고, Z_t는 시간에 따라 독립적이다.

Solution. $\mathbb{E}[Z_t] = \mathbb{E}[X_t I(2|t) + Y_t I(2 \nmid t)] = \mathbb{E}[X_t] I(2|t) + \mathbb{E}[Y_t] I(2 \nmid t) = 0 + 0 = 0$ 으로 평균은 일정하다. 공분산의 경우 제시된 조건에 의하여 $j \geq 1$에 대해 $\mathrm{Cov}(Z_t, Z_{t-j}) = 0$이므로 일정하다. 마지막으로 분산이 유한하며 일정함을 보이면 되는데, t가 짝수인 경우 Z_t의 분산은 X_t의 분산과 같은 1이며, t가 홀수인 경우 Z_t의 분산은 Y_t의 분산과 같은 1이다. 따라서 공분산 정성성을 갖는 확률과정이다.

2007년

Problem.

☐ 다음 설명 중 옳은 것은 ?

A. 불안정(non-stationary) 시계열의 분산은 시간에 상관없이 일정하다.

B. AR과정에서 k시차 표본부분자기상관함수(SPACF) ϕ_k가 유의한지를 판단하는 한계는 일반적으로 $\pm\frac{1}{\sqrt{n}}$ 이다.

C. 불안정 시계열은 로그변환을 통해 안정적 시계열이 된다.

D. ARMA(p, q)의 정상성 조건은 AR(p)과정의 정상성 조건과 동일하다.

E. MA(q)과정은 AR(∞)로 표현할 수 없다.

A. 아니다. 시간에 따라 분산이 증가하는 시계열은 대표적인 불안정 시계열이다.

B. 아니다. AR 과정에서 SPACF이 유의한지를 판단하는 한계는 일반적으로 $\pm\frac{2}{\sqrt{n}}$이다.

C. 아니다. 그러지 않는 불안정 시계열의 예시들이 많이 있다.

D. 그렇다. ARMA(p, q)의 정상성 조건은 AR 부분에만 의존한다.

E. 아니다. MA(q) 모형의 계수들이 invertibility 조건을 만족해야만 한다.

Theorem 2. *AR(p) 모형의 추정에서 PACF는 p시차 초과로는 0으로 절단되는 형태이다. 만약 AR(p) 모형이 참이라면, SPACF는*

$$\sqrt{n}\hat{\phi}_{kk} \xrightarrow{d} N(0,1), \quad k \geq p+1$$

으로 나타난다.
반면 MA(q) 모형의 추정에서 ACF는 q시차 초과로는 0으로 절단되는 형태이다. 만약 MA(q) 모형이 참이라면, SACF는

$$\sqrt{n}\hat{\rho}(k) \xrightarrow{d} N(0, 1+2\sum_{j=1}^{q}\rho(j)^2), \quad k \geq q+1$$

으로 주어진다. 따라서 그들의 유의성 판단은 $z_{\alpha/2} \approx 2$를 그 분산에 곱한 뒤 $\sqrt{n}$으로 나누어 기준을 잡는다.

Problem. 다음 중 원통계에서 계절성이 제거될 수 있는 방법을 모두 고른 것은?
보기: X-11 이동평균, 전년동기비, 3항 이동평균, 계절더미변수 이용

Solution. 정답은 X-11 이동평균(월별자료의 경우), 전년동기비, 3항 이동평균(분기별자료의 경우), 계절더미변수 이용이다.

Definition 33. *X-11 이동평균 방법은 $Y_t = TC_t S_t I_t$ 형태의 곱으로 표현된 승법모형에서 계절조정을 수행하는 방법이다. 이때 TC_t는 추세순환항, S_t는 계절항, I_t는 불규칙항이다. 이는 월별자료를 가정할 때 아래의 네 과정으로 주어진다.(만약 분기별 자료일 경우, 12를 모두 4로 대체하면 된다. 가법모형인 경우 나눗셈과 곱셈을 뺄셈과 덧셈으로 대체하면 된다.)*

1. *영업일수, 요일구성 등의 차이가 있는 경우나 특정한 보정이 필요한 경우 사전적으로 조정하여 사전조정계열을 얻는다.*
2. *2, 3, 4번째 과정에서는 세 번의 반복이 수행된다. 특히 2번째 과정은 세 개의 소과정으로 나뉜다.*
   (a) *첫째 반복에서는, 12개월 중심이동평균(2×12 이동평균)을 원계열에 수행하여 추세순환항의 추정량 $\widehat{TC}_t$를 얻는다. 이는 12개월 이동평균을 실시하며 **13개** 월의 자료의 이동평균을 취했으므로, S_t와 I_t를 제거할 수 있다. 그 다음 계절불규칙항의 추정량*

$$\widehat{S_t I_t} = \frac{Y_t}{\widehat{TC}_t}$$

   를 얻는다.
   (b) *$\widehat{S_t I_t}$에 3×3 이동평균을 적용하여 계절항의 추정량 $\hat{S}_t$과 불규칙항의 추정량 $\hat{I}_t$의 추정량을 얻는다. 이때 3×3 이동평균은 3항 이동평균의 3항 이동평균을 의미한다. 그 다음 moving standard deviation을 $\hat{I}_t$로부터 계산한 뒤, 이를 이용하여 $\widehat{S_t I_t}$의 극단값들을 수정한다. 여기에 다시 3×3 이동평균을 수행하여 다시 $\hat{S}_t$와 $\hat{I}_t$를 얻는다. 이로써 계절조정계열 $\widehat{SA}_t = Y_t - \hat{S}_t$를 얻고, 여기에 13-term Henderson 이동평균을 취하여 추세순환항 $\widehat{TC}_t$를 또다시 추정한다.*
   (c) *그로써 $\widehat{S_t I_t}$를 다시 구하고, (b)의 과정을 한 번 더 반복한다.*
3. *2번과 동일한 과정을 극단값을 조정한 원계열 Y_t에 다시 수행한다.*
4. *2번과 동일한 과정을 3번에서 극단값을 조정한 원계열 Y_t에 다시 수행한다. 이를 통하여 최종적인 $\widehat{TC}_t, \widehat{S}_t, \widehat{I}_t$를 얻는다. 그 후 계절조정이 잘 이루어졌는지 등에 대한 추가적인 검정을 수행할 수 있다.*

이를 통해 $\tilde{S}_t$를 얻을 수 있고, $Y_t/\tilde{S}_t$가 계절조정계열이 된다.

Note 4. *아래 글들을 참조했습니다.*

- *http://www.math.wpi.edu/saspdf/ets/chap21.pdf*
- *https://www.bok.or.kr/portal/bbs/P0000556/view.do?nttId=1989&menuNo=200536&pageIndex=*
- *https://jdemetradocumentation.github.io/JDemetra-documentation/pages/theory/SA_X11.html*

이외에도 전년동기비는

$$\frac{Y_t}{Y_{t-p}} = \frac{TC_t}{TC_{t-p}} \times \frac{S_t}{S_{t-p}} \times \frac{I_t}{I_{t-p}}$$

에 대해 분석함으로써, p가 주기라 할 때 계절성분이 동일하면 $S_t = S_{t-p}$임에 따라 그 영향이 사라지기에 계절성을 제거할 수 있다. 단 이는 승법모형일 때에만 적용이 가능하다. 3항 이동평균은 X-11 이동평균에 비하여 반복수가 적기는 하지만, 분기별 자료인 경우 이를 제거할 수 있다. 계절더미변수 역시 마찬가지로, S_t를 계절에 따라 다른 piecewise constant로 보고 분석할 수 있게 함으로써 이를 제거하는 데 도움을 줄 수 있다.

Problem. 회귀모형 $y_t = \alpha + \beta x_t + \epsilon_t$ $(t = 1, 2, \cdots, n)$가 주어져 있다. 오차항 ϵ_t에 대한 더빈-왓슨 검정통계량을 제시하고, 위 모형에서 $n = 20$이고 해당 통계량의 값이 0.8일 때 유의수준 5%에서 오차항의 자기상관 존재여부와 그 판단 근거를 기술하시오. 또한, 더빈-왓슨 검정의 한계점을 쓰시오.

Solution.
더빈-왓슨 검정통계량은

$$D = \frac{\sum_{t=2}^{n}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\epsilon}_t^2}$$

으로 주어진다. 자료에서 독립변수가 하나이며 해당 통계량의 값이 0.8으로 이는 $n = 20$일 때의 하한 $d_L = 1.201$보다 작다. 따라서 양의 상관관계가 있다고 말할 수 있다. 더빈-왓슨 검정의 경우, 값이 d_L보다 작으면 자기상관이 있다는 귀무가설을 기각할 수 있고, d_U보다 크면 그렇지 않아야 하지만, d가 둘 사이에 들어오는 경우 검정의 결과가 결정되지 않는다는 단점이 있다. 또한 시차 1의 양/음의 자기상관만 검정가능하다는 단점 역시 있다. 즉, 2 이상의 주기를 가지는 오차항의 자기상관을 파악하기는 어렵다.

Problem. $y_t = \alpha + \beta x_t + \epsilon_t$가 ARCH(1) 모형을 따른다고 가정할 때 오차항 ϵ_t의 조건부 분산 $\mathrm{Var}(\epsilon_t|\epsilon_{t-1})$의 형태를 쓰시오.

Solution.
정답은

$$\mathrm{Var}(\epsilon_t|\epsilon_{t-1}) = \alpha_0 + \alpha_1\epsilon_{t-1}^2$$

이다. 단, $\alpha_0 > 0, \alpha_1 \geq 0$이다.

Problem. 시계열 Y_t가 추세 T_t, 순환 C_t, 계절 S_t, 불규칙 I_t 요인의 곱으로 이루어졌다고 가정할 때, 각 구성성분을 분해하는 방법을 기술하시오.

Solution. 절대적인 기준은 없는 것 같다. 가장 간단한 방법에는 아래가 있다.

1. 주기를 p라 할 때, p가 짝수면 $2 \times p$ 이동평균을, 홀수면 p 이동평균을 이용하여 추세순환항의 추정량 $\widehat{T_tC_t}$를 얻는다. 혹은 이동평균이 아닌 다른 회귀모형이나 ARIMA 모형 등을 이용하여 추정할 수도 있다. 즉 평활화를 통하여 추세순환항을 얻는다.
2. 만약 추가적인 순환요소가 있다고 생각되면, 해당 순환요소의 주기를 m이라 한뒤 이에 맞는 추세항을 $\widehat{T_tC_t}$로부터 추가적으로 얻어내어 $\hat{T}_t$와 $\hat{C}_t$를 분리한다.
3. 추세를 제거한 시계열 $Y_t/\hat{T}_t\hat{C}_t$를 얻는다.
4. 추세를 제거한 시계열에서 계절더미변수를 이용한 회귀모형, 혹은 추가적인 모형 기반의 방법으로 $\hat{S}_t$를 구한다.
5. $\hat{I}_t = Y_t/\hat{T}_t\hat{C}_t\hat{S}_t$를 불규칙요인의 추정량으로 얻는다.
6. $\hat{I}_t$에 대한 진단을 통하여 분해가 잘 이루어졌는지 확인한다.