Journal of the Korean Institute of Industrial Engineers
Vol. 39, No. 2, pp. 109-118, April 2013.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2013.39.2.109

<Original Research Paper>

비정규 오차를 고려한 자기회귀모형의 추정법 및 예측성능에 관한 연구

임보미 · 박정술 · 김준석 · 김성식 · 백준걸†

고려대학교 산업경영공학과

A Study of Estimation Method for Auto-Regressive Model with Non-Normal Error and Its Prediction Accuracy

Bo Mi Lim · Cheong-Sool Park · Jun Seok Kim · Sung-Shick Kim · Jun-Geol Baek

School of Industrial Management Engineering, Korea University

We propose a method for estimating coefficients of AR (autoregressive) model which named MLPAR (Maximum Likelihood of Pearson system for Auto-Regressive model). In the present method for estimating coefficients of AR model, there is an assumption that residual or error term of the model follows the normal distribution. In common cases, we can observe that the error of AR model does not follow the normal distribution. So the normal assumption will cause decreasing prediction accuracy of AR model. In the paper, we propose the MLPAR which does not assume the normal distribution of error term. The MLPAR estimates coefficients of auto-regressive model and distribution moments of residual by using pearson distribution system and maximum likelihood estimation. Comparing proposed method to auto-regressive model, results are shown to verify improved performance of the MLPAR in terms of prediction accuracy.

Keywords: Time Series Analysis, Auto-Regressive Model, Pearson Distribution System, Maximum Likelihood Estimation, Non-Normal Data

1. 서 론

시계열 데이터는 일반적인 데이터에 시간의 차원이 추가된 형태로 시간에 따라 관측되는 데이터이며, 경제, 환경, 사회 등 다양한 분야에서 수집된다(Last *et al*., 2001). 시계열 데이터를 분석하는 목적은 과거 값을 토대로 미래 값의 예측이다. 이러한 분석은 경제 분야에서 주식, 환율 등의 예측으로 수익을 창출하고, 환경 분야에서 강우량과 하천 유출수량 등의 예측으로 수해 및 자연재해를 예방한다. 또한 사회현상을 비롯하여 보안이나 수요와 공급 등에 이르기까지 합리적인 의사결정을 돕는다. 그러므로 시계열 데이터의 분석에서 예측 오차를 줄여 정확성을 높이고자 하는 연구들은 다양한 분야에서 요구되는 중요한 연구주제이다.

시계열 데이터의 분석 방법론은 지수평활법(Exponential Smoothing Method), 분해법(Decomposition Method), 박스-젠킨스 모형(Box-Jenkins Model) 등이 있다(Bowerman *et al*., 2005). 지수평활법은 이동 평균 방법을 보완하여 최근 자료에 큰 가중치를 부여하는 방법이며, 분해법은 패턴을 부분패턴으로 분해함으로써 예측뿐만 아니라 계절성, 순환성, 추세성과 같은 시계열 데이터의 특성들에 대한 정보와 영향을 분석하는 방법이

본 연구는 지식경제부 및 정보통신산업진흥원의 정보통신연구기반구축사업의 연구결과로 수행되었음(NIPA-2012-(B1100-1101-0002)).
이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2012-0008332).
† 연락저자 : 백준걸 교수, 136-701 서울특별시 성북구 안암동 5가 1번지 고려대학교 산업경영공학과, Tel : 02-929-5888, Fax : 02-3290-3396, E-mail : jungeol@korea.ac.kr
2013년 1월 31일 접수; 2013년 2월 27일 수정본 접수; 2013년 3월 6일 게재 확정.

다. 그러나 평활법은 불규칙 변동을 제거한 후 시계열 데이터의 추정이 가능한 방법이며, 분해법은 특정 패턴이 시계열에 내제되어 있을 때 가능한 방법이다. 이에 반에 박스-젠킨스 모형은 다양한 시계열 데이터에 적용이 가능하며, 정확성과 예측성이 높아 가장 많이 이용되는 방법이다(Bruce *et al*., 2005).

박스-젠킨스 모형은 시계열 데이터의 자기상관관계(Auto-correlation)를 이용한 다중회귀기법(Multi-linear Regression)으로 모형과 모형의 계수를 추정하는 방법이다(Pankratz, 2008). 이러한 박스-젠킨스 모형은 모형 식별, 계수 추정, 모형 검증 세 단계 절차로 구성되어 있다. 먼저 자기상관계수와 부분자기상관계수를 이용해 박스-젠킨스 모형은 시계열 데이터가 어떤 모형에 적합한지를 판단한다. 적합한 모형이 선택되면 모형의 계수 값을 추정하는 것이 두 번째 단계이다. 마지막으로 오차항의 자기상관계수와 부분자기상관계수에 의해 그 모형이 적합한지를 판단한다. 박스-젠킨스 모형은 정상성(Stationarity)과 가역성(Invertibility)에 따라 크게 자기회귀모형(AR : Auto-regressive Model)과 이동평균모형(MA : Moving Average Model) 두 종류로 나뉜다. 정상성은 시계열의 확률적 성질들이 시간의 흐름에 따라 변하지 않음을 뜻하며, 가역성은 계수가 -1에서 1사이에 있으면 관측되지 않는 오차항의 값을 데이터로부터 추정할 수 있음을 뜻한다. 이에 자기회귀모형은 정상성 특징은 있지만 가역성 특징은 없으며, 이동평균 모형은 가역성 특징은 있지만 정상성 특징은 없다. 그러므로 자기회귀모형과 이동평균 모형을 적용하거나 응용할 때는 다른 방식으로 접근해야 하며, 본 연구에서는 우선적으로 자기회귀모형에서의 문제점을 해결하고자 한다.

자기회귀모형은 Yule-Walker 방정식을 이용한 다중선형회귀모형(Multiple Linear Regression Model)이다(Endo and Randall, 2007). 이 모형은 계수를 추정하기 전에 알려지지 않은 오차항의 분포가 평균 0, 분산 σ^2인 정규분포를 따른다는 가정이 있다(Akaike, 1969). 그러나 인위적인 데이터가 아닌 실제 시계열 데이터에서 오차항의 분포가 정규분포인 가정을 성립하는 경우는 거의 없다(Endo and Randall, 2007). <Figure 1>는 자기상관계수와 부분자기상관계수를 확인하여 자기회귀모형의 차수 1에 적합한 시계열 데이터를 자기회귀모형에 적용한 후 계산한 오차항의 분포이며, 실험에서 사용되는 데이터이다. 그렇지만 <Figure 1>에 제시된 오차항의 분포가 정규분포로 가정한 곡선에 잘 표현되지 않는다. 또한 경제 분야의 주가, 환율 등 금융 시계열 데이터에서 오차항의 첨도가 정규분포의 첨도보다 크고, 왜도가 0이 아닌 음이나 양의 값을 갖는 특성들이 빈번하게 나타난다(Rydberg, 2000).

자기회귀모형은 계수를 추정하기 전에 알 수 없는 오차항의 분포를 정규분포로 가정하기 때문에 모형의 계수를 정확하게 추정하지 못하고, 예측성을 감소시킨다. 그러므로 본 연구에서는 이 문제점을 해결하고자 피어슨 분포 시스템(Pearson Distribution System)을 이용한다. 피어슨 분포 시스템은 네 개의 적률(Moment)을 이용해 정규분포 외에도 다양한 분포를 포

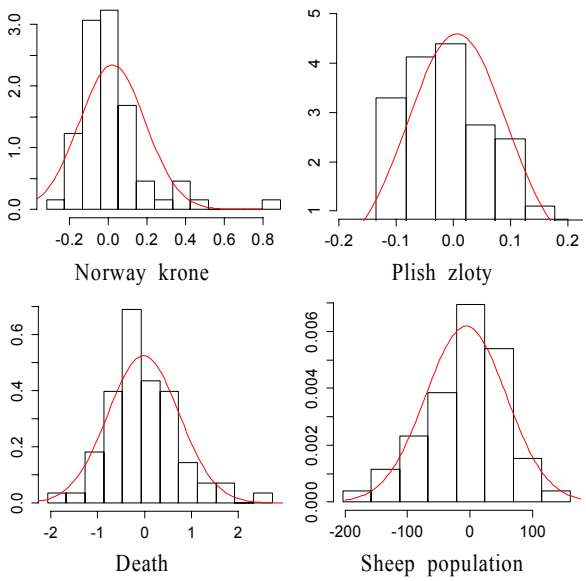

Figure 1. Distribution of error term

현한다(Nagahara, 2000). 네 개의 적률은 1차 적률 평균, 2차 적률 분산, 3차 적률 왜도, 4차 적률 첨도로 형상모수이다. 피어슨 분포시스템은 네 개의 적률로 오차항의 분포를 추정하기 때문에 기존 자기회귀모형에서 오차항의 분포를 정규분포로 가정 하에 평균과 분산 두 모수를 추정한 것보다 더 정확하게 분포를 표현한다.

피어슨 분포의 네 개의 적률(평균, 분산, 왜도, 첨도)과 자기회귀모형의 계수를 추정하기 위해 본 연구에서는 최우추정법(MLE : Maximum Likelihood Estimation)을 이용한다. 최우추정법은 통계학에서 주된 모수 추정 법으로 표본의 수가 크면 다른 추정법과 비교하였을 때 정확도가 높은 방법이다(Walpole *et al*., 2006). 또한 최소제곱법(LSE : Least Squared Estimation)은 정규분포를 가정 하에 분산을 추정하는 방법이지만 최우추정법은 분포 가정 없이 모수를 추정하는 방법이다.

본 연구에서 제안하는 MLPAR(Maximum Likelihood of Pearson system for Auto-Regressive model) 방법은 자기회귀모형의 계수를 추정함에 있어 오차항의 분포가 정규분포를 따른다는 가정으로 인해 정확성과 예측성을 감소시키는 문제점을 해결하고자 한다. 제안하는 방법은 자기회귀모형의 계수에 초기 값을 설정한 후 계수 값을 바꾸면서 피어슨 분포 타입별로 음의 로그 우도 값(Negative Log Likelihood)을 계산하고, 가장 작은 값에서의 자기회귀모형 계수와 오차항 분포 모수를 추정한다. 제안한 방법의 성능을 평가하기 위해 본 연구는 시계열 데이터의 자기상관계수(AC : Autocorrelation Coefficient)와 부분자기상관계수(PAC : Partial Autocorrelation Coefficient)를 확인하여 자기회귀모형에 적합한 데이터이지만 오차항의 분포가 정규분포를 따르지 않은 데이터로 실험을 했다. 결과적으로 제안한 방법(MLPAR)이 기존 방법(AR)보다 적률의 특징을 반영하여, 정확성과 예측성이 증가하였다.

2. 본론

2.1 자기회귀모형(Auto-Regressive Model)

자기회귀모형은 현 시차의 값이 전 시차들 간의 관계, 상수, 현 시차의 오차의 합으로 표현되며 시간과의 관계를 추론한다(Yule *et al*., 1927).

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + c + \epsilon_t \tag{1}$$

박스-젠킨스 모형 중 자기회귀모형에 가장 적합한 데이터인지 확인해보기 위해서 모형의 식별, 계수의 추정, 모형의 검토의 3단계 절차가 필요하다. 먼저 주어진 시계열 데이터의 자기상관계수(AC : Autocorrelation Coefficient)와 부분자기상관계수(PAC : Partial Autocorrelation Coefficient)를 구한다. 자기상관계수는 한 시차의 관측치와 k 시차 떨어진 관측치 간의 상호관계 연관성이 있는지를 측정하는 척도로 식 (2)와 같으며, 부분자기상관계수는 한 시차 이외의 모든 시차 관측치에 의한 영향력을 제외한 상태에서 특정한 두 관측치가 상호관계 연관성이 있는지를 측정하는 척도로 식 (3)과 같다. 박스-젠킨스 모형은 이 두 계수를 이용하여 모형과 모형의 차수를 결정한다.

$$p_k = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)\,Var(X_{t+k})}} = \frac{\gamma_k}{\gamma_0} \tag{2}$$

$$\phi_{kk} = Corr(X_t, X_{t+k} | X_{t+1}, \cdots, X_{t+k-1}) \tag{3}$$

두 계수의 값을 그래프로 그리면 자기회귀모형의 자기상관계수는 시차가 증가함에 따라 싸인 곡선이나 지수 곡선으로 0을 향해 점차적으로 감소하는 형태이다(Box *et al*., 1994). 또한 자기회귀모형의 부분자기상관계수는 시차 p까지만 유의한 값이며 나머지 시차는 신뢰구간 내에 있는 형태로 모형의 차수를 결정한다(Box *et al*., 1994).

자기회귀모형의 계수는 오차항의 분포가 정규분포를 따른다는 가정 하에 최우추정법 또는 최소자승법으로 추정된다. 마지막으로 모형이 통계적으로 유의한지 확인하기 위해 잔차의 자기상관계수와 부분자기상관계수의 그래프를 확인한다. 두 그래프는 시차 0을 제외한 나머지 시차에서 신뢰구간 내에 있어야 한다. 만약 신뢰구간에 벗어난 차수가 있으면 추정한 모형은 적합하지 않으며, 다른 모형이 더 적합할 수 있으므로 다시 식별되어야 한다.

2.2 피어슨 분포 시스템(Pearson Distribution System)

피어슨 분포 시스템은 네 적률을 이용해 정규분포 외에도 감마분포, 베타분포, T 분포 등 다양한 분포를 포현한다(Nagahara, 2000). 네 개의 적률은 1차 적률 평균, 2차 적률 분산, 3차 적률 왜도, 4차 적률 첨도로 형상모수이다. Pearson(1895, 1916)은 확률밀도함수 p에 대한 미분 방정식으로 식 (4)와 같이 피어슨 분포 시스템을 정의했다.

$$-\frac{p'}{p} = \frac{b_0 + b_1 x}{c_0 + c_1 x + c_2 x^2} \tag{4}$$

피어슨 분포 시스템의 타입(Ⅰ, Ⅱ, Ⅲ, Ⅳ, Ⅴ, Ⅵ, Ⅶ)은 네 개의 적률(평균, 분산, 왜도, 첨도)에 의해서 결정된다(Parrish, 1983). 또한 네 적률은 적률의 순서대로 타입 결정의 영향력이 크다. 3차 적률인 왜도는 중심축을 기준으로 분포가 좌우로 얼마나 대칭적인지를 나타내는 통계 값으로 비대칭도라고도 하며, 4차 적률인 첨도는 분포 모양이 중간위치에서 뾰족한 정도를 나타내는 통계 값이다. 피어슨 분포 시스템의 타입은 분석할 데이터의 네 개의 적률을 구한 후 1차 적률을 0, 2차 적률을 1로 표준화한 값으로 결정된다. 이는 식 (5)와 같이 3차 적률의 제곱인 β_1과 4차 적률에 3을 더한 β_2의 식으로 정의된다(Nagahara, 2004).

$$k = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)} \tag{5}$$

따라서 피어슨 분포 시스템은 <Figure 2>과 같이 β_1과 β_2의 축으로 타입을 구분할 수 있다. k < 0, 0 < k < 1, 1 < k일 때 피어슨 분포 시스템은 타입 Ⅰ, Ⅳ, Ⅵ를 선택하며, 이 세 타입이 <Figure 2>에서 B, E, G 영역에 해당하는 주요 타입이다(Elderton and Johnson, 1969). <Figure 2>에서 타입 Ⅰ(B)과 Ⅵ(G)의 영역 사이인 타입 Ⅲ(D)은 k = ±∞일 때 선택되며, 타입 Ⅳ(E)와 Ⅵ(G)의 영역 사이인 타입 Ⅴ(F)는 k가 1일 때 선택된다. 또한, k = 0일 때는 타입 Ⅶ(H), Ⅱ(C), 정규분포(A)의 영역으로 β_2의 값에 따라 타입이 결정된다.

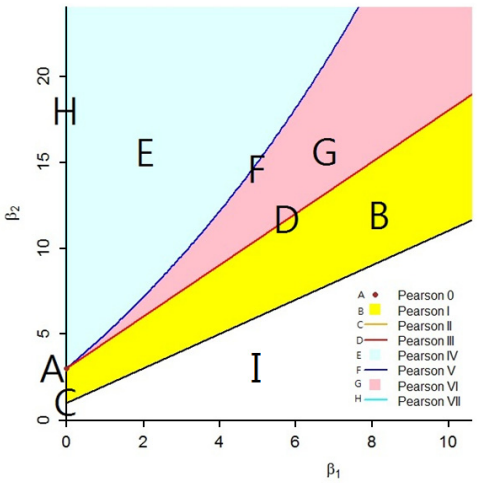

Figure 2. Diagram of pearson distribution system

피어슨 분포 시스템에서 각 타입의 확률밀도함수는 <Table 1>과 같으며, 이를 이용해 데이터의 분포를 표현한다(Nagahara, 2000). <Table 1>을 이용해 피어슨 분포 시스템은 오차항의 분포를 더 유사하게 표현할 수 있다.

기존 자기회귀모형은 오차항의 분포를 평균 0, 분산 σ^2인 정규분포인 가정으로 인해 실제 오차항의 분포를 잘 표현하지 못함을 서론에 있는 <Figure 1>에서 확인했다. <Figure 1>에서 제시한 데이터를 피어슨 분포 시스템으로 추정하면 오차항의 분포는 <Figure 3>의 점이 있는 곡선으로 표현된다. 즉 피어슨 분포 시스템이 정규분포보다 분포를 더 정확하게 추정한다.

Table 1. PDF of pearson distribution system

Type	PDF(Probability Density Function)
I (B)	$\frac{(x-a)^{p-1}}{b^p B(p,\ q)}[1-\frac{(x-a)}{b}]^{q-1}$
II(C)	$\frac{(x-a)^{p-1}}{b^p B(p,\ q)}[1-\frac{(x-a)}{b}]^{p-1}$
III(D)	$\frac{(\lvert x-\gamma\rvert)^{\alpha-1}}{\beta^{\alpha}\gamma(\alpha)}\exp[-\frac{(\lvert x-\gamma\rvert)}{\beta}]$
IV(E)	$\frac{\gamma(b+b\delta i)\gamma(b-b\delta i)\tau^{2b-1}}{\gamma(b)\gamma(b-\frac{1}{2})\pi^{\frac{1}{2}}}\frac{\exp[2b\delta\arctan(\frac{x-\mu}{\tau})]}{[(x-\mu)^2+\tau^2]^b}$
V(F)	$\frac{\lambda^v}{\gamma(v)(\lvert x-a\rvert)^{v+1}}exp[-\frac{\lambda}{\lvert x-a\rvert}]$
VI(G)	$\frac{a^m(x-a)^{\beta-1}}{B(\beta,\ m)(\alpha+x-a)^{m+\beta}}$
VII(H)	$\frac{\gamma(b)}{\gamma(b-\frac{1}{2})\sqrt{\pi}}\frac{1}{\tau[1+(\frac{x-\mu}{\tau})^2]^b}$

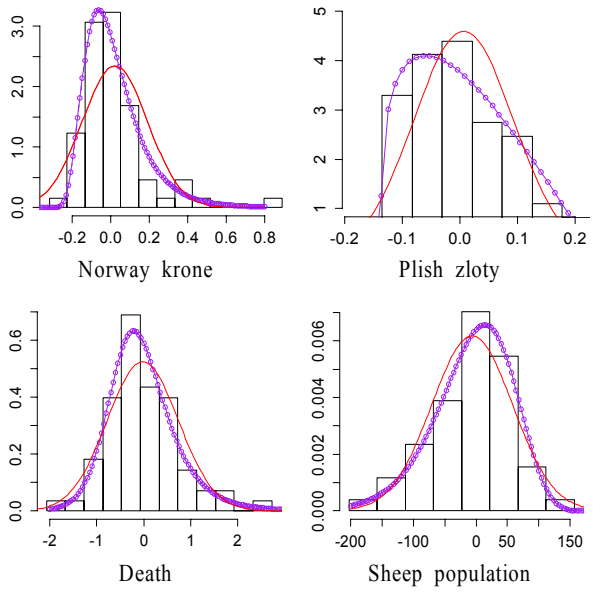

Figure 3. Distribution of error term by pearson distribution system

<Figure 1>에서 제시한 데이터를 피어슨 분포 시스템으로 추정하면 오차항의 분포는 <Figure 3>의 점이 있는 곡선으로 표현된다. 즉 피어슨 분포 시스템이 정규분포보다 분포를 더 정확하게 추정한다.

그렇지만 피어슨 분포 시스템이 모든 비정규 분포를 분석하는 것은 아니다. 분석하지 못하는 부분은 <Figure 2>에서 I 영역이며, I 영역에는 확률밀도함수를 구현하지 못하였거나 아직 연구가 되어 있지 않은 분포가 포함되어 있다(Nagahara, 2004). 그러므로 본 연구에서는 오차항의 분포를 추정함에 있어 시스템에 구현되어 있는 타입 7개에 최우추정법을 이용해 가장 적합한 타입을 결정한다.

2.3 MLPAR(Maximum Likelihood of Pearson system for Auto-Regressive model)

최우추정법은 점 추정의 한 방법으로써 알지 못하는 모집단의 모수를 추정함에 있어 우도함수(Likelihood Function)가 최대인 모수를 결정하는 방법이다(Aldrich, 1997). 우도란 어떤 가설이 진실이고, 시행의 결과가 주어졌을 때 그 결과가 나올 확률이다. MLPAR에서 이 가설이 자기회귀모형의 계수와 오차항의 적률이라고 하면 최우추정법에 의한 추정치는 가장 적절한 계수와 적률이다.

$$L(\beta|x_1,\ x_2,\cdots,\ x_n)=\prod_{i=1}^{n}f(x_i|\beta) \tag{6}$$

$$\widehat{\theta_{MLE}}\subseteq \arg max_{\beta\in\Theta}\hat{L}(\beta|x_1,x_2,\cdots,\ x_n) \tag{7}$$

최우추정법은 로그함수를 이용해 수치적 언더플로우(Underflow) 현상을 해결한다. 데이터의 수가 많고, 작은 확률 값들이 곱해지면 우도 값이 0에 근접하게 된다. 이를 해결하기 위해 로그우도함수(Log Likelihood Function)가 이용된다. 로그우도함수는 우도에 로그를 취해 각 확률을 곱에서 합으로 바꾼다. 그러므로 본 연구에서의 목적식은 식 (8)처럼 로그우도함수를 이용한다.

$$\ln L(\beta|x_1,x_2,\cdots,x_n)=\sum_{i=1}^{n}\ln f(x_i|\beta) \tag{8}$$

MLPAR 수도 코드(Pseudo Code)는 <Table 2>와 같다. MLPAR은 자기회귀모형에 적합한 데이터 X로 전 시차의 수(p), 전 시차와 현 시차간의 관계상수(ϕ), 상수(c), 오차항의 적률(평균, 분산, 왜도, 첨도)을 추정한다.

MLPAR은 자기회귀모형의 계수와 오차항의 분포 모수를 추정하기 위한 알고리즘이다. 이를 구하기 위해 먼저 자기회귀모형의 차수(p)는 부분자기상관계수를 이용한 함수에 의해 결정된다(step 1). 차수가 정해지면 MLPAR은 β(c, $\phi_{1,}\cdots$, ϕ_p)을 0으로 초기화한 후(step 2) β를 자기회귀모형에 대입해 오차항(ε)을 계산한다(step 3). 구해진 오차항의 네 개 적률 값은 <Table

1>에 제시한 피어슨 분포 타입 별 확률밀도 함수에 오차항의 분포를 가장 잘 표현할 수 있는 적률 값으로 계산된다(step 4). 그러면 구해진 적률의 성질을 가진 오차항의 분포 타입이 결정된다(step 5). 기존 자기회귀모형은 계수를 추정함에 있어 오차항의 분포를 정규분포로 가정한 것에 반면에 이 시점에서 MLPAR은 피어슨 분포 타입으로 표현해 자기회귀모형의 계수를 추정한다.

Table 2. MLPAR pseudo code

MLPAR(X)

Input : X Data
Output : coefficients of auto-regressive model and distribution moments of ε

Constant : c = 0
Best number of the past variables : p*
Constant relating X_t to X_{t-i} : $\phi_i = 0$, i = 1, ⋯, p*
White noise : $\varepsilon_t = 0$, t = 2, ⋯, T
Distribution moments of ε
: pa = (mean, sigma, skewness. kurtosis)
Density distribution of ε : $f(\beta)$
Coefficients of auto-regressive model : β Current β_n
Best negative log likelihood : min
Iteration number : n =1

step 1) p* = determine the best using PACF (Partial Autocorrelation Coefficient Function)
step 2) Initialize $\beta = c(c, \phi_1, \phi_2, \cdots, \phi_p) = 0$
step 3) Calculate ε variable

$$\epsilon_t = X_t - \sum_{i=1}^{p} \phi_i X_{t-i} - c$$

step 4) pa = Calculate the empirical moments of ε as moments are altered in order to assure that the whole sample lies in support of the distribution (Reference to table 1)
step 5) $f(\beta)$ = Determine density distribution of ε by pa
step 6) Use the newton method with density distribution of ε

$$q(\beta) = f(\beta_n) + \nabla f(\beta_n)(\beta - \beta_n) + \frac{1}{2}\nabla^2 f(\beta_n)(\beta - \beta_n)^2$$

step 7) Calculate minimization of function($q(\beta)$)
step 8) Calculate sum of negative log likelihood

$$\min = \sum_{i=1}^{n} \ln f(x_i|\beta)$$

step 9) IF $\nabla f(\beta_c) > \lambda$(1e-10) then
β values change as much as step size($\nabla^2 f(\beta_c)$)
n = n+1
Go to step 3
end IF
step 10) return β and pa

자기회귀모형 계수를 추정하기 위해 최적화 방법 중 뉴턴법이 이용된다(step 6). step 7은 step 6에 있는 뉴턴 식을 미분하여 식 (9)와 같은 식을 만들어 식 (9)에서 0이 되는 β값을 구한다(Gay, 1983).

$$\frac{dq}{d\beta} = \nabla f(\beta_n) + \nabla^2 f(\beta_n)\beta - \nabla^2 f(\beta_n)\beta_n = 0 \qquad (9)$$

MLPAR은 구해진 β값으로 오차항의 확률밀도함수에 대입해 음의 로그우도 값을 구한다(step 8). 마지막으로 변화도(Gradient)가 λ보다 크면 헤시안(Hessian)을 이용해 자기회귀모형의 계수(β)를 변화시킨 후 step 3부터 step 9까지 과정을 반복한다. 반대로 변화도가 λ보다 작으면 step 10으로 이동해 MLPAR은 최종적으로 자기회귀모형의 계수(c, ϕ_1, ⋯, ϕ_p)와 오차항의 분포 적률(평균, 분산, 왜도, 첨도)을 도출한다.

기존 자기회귀모형은 오차항의 분포를 가정한 후 계수를 추정하지만, 제시하는 MLPAR은 계수 값에 따른 오차항의 분포를 계산하면서 계수를 추정한다. 결과적으로 MLPAR은 오차항의 분포 가정이 없어 기존 추정방법보다 적률의 특징을 반영하여 정확성과 예측성이 증가하는 장점을 지닌다.

3. 실험 및 결과 분석

3.1 실험 데이터

이 절에서는 본 연구에서 제안한 알고리즘의 성능을 평가하기 위해 기존방법과 MLPAR의 실험을 실시하였다. 실험은 실제 데이터로 datamarket.com에 있는 다양한 종류의 시계열 데이터를 이용하였으며, 데이터는 <Table 3>과 같다. <Table 3>의 시계열 데이터는 자기상관계수와 부분자기상관계수를 확인하여 박스-젠킨스 모형 중 자기회귀모형에 적합한 데이터임을 확인했다. 또한 자기회귀모형의 계수를 추정한 후 계산된 오차항의 분포는 정규분포를 따르지 않았다. 이에 대한 검사는 앤더슨-달링(Anderson-Darling) 검정방법을 이용했다. 앤더슨-달링 검정은 데이터가 특정 분포를 얼마나 잘 따르는지

Table 3. Experiment data

Dataset	Instances		Feature
	Train	Test	
Norway krone	2003. 1 ~2008. 11	2008. 12 ~2010. 07	Monthly Mean
Sheep population	1867~1925	1926 ~1939	Yearly Mean
Communications	1998. 12 ~2011. 04	2011. 05 ~2012. 07	Monthly Mean
Death	1915 ~1983	1984 ~2004	Yearly Mean
Plish zloty	2001. 05 ~2007. 05	2007. 06 ~2008. 11	Monthly Mean
Southern Oscillation	1964. 12 ~1970. 09	1970. 10 ~1972. 05	Monthly Mean

측정하는데 일반적으로 많이 이용되는 적합도 검정 방법이다 (Anderson and Darling, 1952). <Table 3>에서 제시한 시계열 데이터에 엔더슨-달링 검정을 한 결과, 오차항의 분포가 정규분포를 따르지 않음을 확인하였다.

3.2 실험 설계

실험은 <Table 3>의 데이터를 MLPAR과 자기회귀모형에 적용하여 모형의 정확성과 예측성을 평가하였다. 먼저 모형의 정확성에서는 평균제곱오차(MSE : Mean Squared Error)와 AIC (Akaike Information Criterion) 척도를 이용해 평가하였다. 평균제곱오차는 학습용 데이터(Train Data)로 자기회귀모형을 만든 후 전 데이터의 참 값을 대입해 계산한 값과 실제 값인 학습용 데이터 간의 오차로 계산했다. AIC(Akaike Information Criterion) 척도는 모형의 정확성을 평가하는 지표이며, 식 (10)과 같다.

$$AIC = 2k - 2\ln(L) \tag{10}$$

AIC 척도는 값이 작을수록 데이터가 모형에 적합함을 나타내며, 식 (10)에서 k는 모형의 미지수 개수이다. 기존 방법과 제시한 MLPAR에서 k는 같으므로 $-2\ln(L)$ 즉 로그우도에 의해 AIC 값이 결정된다.

모형의 예측성에서는 평균제곱오차(MSE)를 이용해 두 방법을 평가하였다. 첫 번째 방법에서 평균 제곱오차는 학습용 데이터로 추정한 모형으로 전 데이터의 참 값을 대입해 예측한 값과 실제 값인 검증용 데이터(Test Data) 간의 오차로 계산하였다. 그렇지만 시간이 흐르게 되면 시계열 형태는 변화한다. 한 모형을 만든 후 계속 같은 모형으로 예측을 수행하는 것은 점진적으로 오차율을 높일 수 있다. 또한 많은 데이터가 연속적으로 수집되는 경우에는 데이터를 다 반영할 수 없다. 그러므로 두 번째 방법에서 평균제곱오차는 Rolling Horizon 방법을 이용해 계산하였다. Rolling-Horizon 방법이란 타임 윈도우를 시간의 흐름에 따라 이동하면서 새로 입력되는 데이터를 모형에 적용하고, 가장 오래된 데이터를 모형에서 제거하는 방식이다(Sethi *et al.*, 1991). 윈도우 사이즈는 <Table 3>에서 제시한 데이터들의 학습용 데이터로 두었고, 예측은 윈도우을 한 시차씩 옮겨 반복하였다.

3.3 실험 결과

(1) 모형의 정확성

먼저 모형의 정확성을 확인하기 위해서 <Table 3>에서 제시한 학습용 데이터로 자기회귀모형의 계수를 추정하여 모형을 만든 후 전 데이터의 참 값을 대입해 계산한 값과 실제 값인 학습용 데이터 간의 오차를 계산했다. 이와 같은 실험으로 <Table 4>는 자기회귀모형 계수를 추정함에 있어 기존 방식(AR)과 MLPAR에 적용해 평균제곱오차를 구한 결과이며, 감소율은 AR의 평균제곱오차를 1로 두었을 때 MLPAR의 평균 제곱오차 감소비율이다. <Table 4>을 보면 6개의 데이터 중 Communications와 Death 데이터를 제외한 4개의 데이터에서 AR보다 MLPAR이 더 작은 값의 평균제곱오차를 보였다.

Table 4. Result in accuracy of model(MSE)

Dataset	AR	MLPAR	Reduction ratio
Norway krone	2.02854	**1.74332***	16.36%
Sheep population	309383	**276826***	11.76%
Communications	**48.6646***	48.6712	-0.01%
Death	**39.3038***	40.5741	-3.23%
Plish zloty	0.5230	**0.5217***	0.25%
Southern Oscillation	37.4629	**37.4246***	0.10%

기존 방식과 달리 MLPAR은 오차항의 분포를 정규분포로 가정하지 않았다. <Figure 4>에서 β_1(왜도^2)이 0, β_2(첨도+3)이 3일 때 정규분포(A)이며, <Table 4>에 제시한 데이터에서 구해진 오차항의 분포는 다른 타입에 해당된다. 또한 <Table 4>에서 모형의 정확성이 높아지지 않은 Communications와 Death 데이터의 이유를 오차항의 분포에서 알 수 있다. 두 데이터의 오차항의 분포는 왜도에 비해 첨도가 큰 타입 Ⅳ(E)이다. 피어슨 분포 시스템에서 분포를 결정할 때 적률 중 4차 적률인 첨도가 영향력이 가장 작다. 그래서 오차항의 분포를 표현함에 있어 첨도보다 정규분포에서 분산의 수치를 줄이는 방법이 데이터를 중간위치에 몰리게 해 더 정확성을 높였다. 그러므로 Communications와 Death 데이터의 MSE가 기존 방법(AR)에 비해 감소하지 않은 결과를 도출한 것이다.

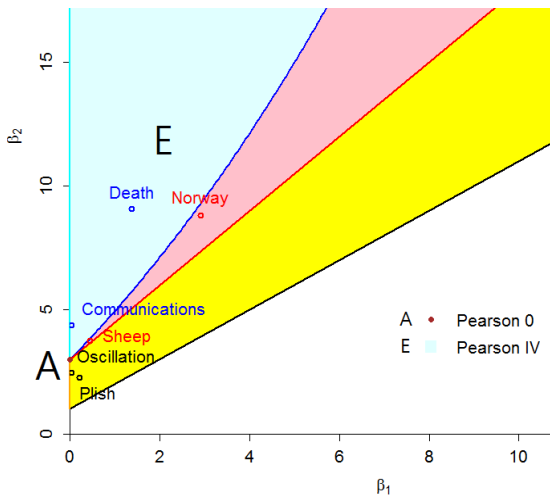

Figure 4. β_1 and β_2 of error term

MLPAR을 적용 후 계산된 오차항의 분포는 <Figure 5>와 같다. 점이 있는 곡선은 피어슨 분포시스템으로 분포를 표현한 것이다.

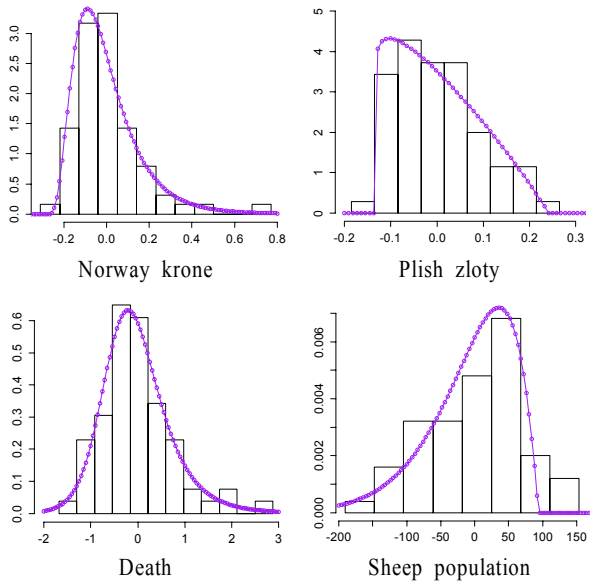

Figure 5. Distribution of error term by pearson distribution system (MLPAR)

<Table 4>에서 제시한 데이터 중 대표적으로 Norway krone 데이터의 그래프를 확인하고자 한다. <Figure 6>은 Norway krone 데이터를 기존 방법(AR)으로 추정한 결과이다. <Figure 6>에서 점이 있는 선은 실제 데이터, 세모가 있는 선은 AR로 추정한 것이며, 실제 데이터보다 AR로 추정한 결과가 큰 값인 양의 오차항(Positive error)이면 점 수직선, 작은 값인 음의 오차항(Negative error)이면 수직선이다. 수직선과 점 수직선으로 표시된 것의 수는 비슷하지만, 음의 오차항 중 차이가 적은 수직선이 많다. 이는 <Figure 1>에서 오차항의 분포가 X축 -1에서 0.5 사이에 밀도가 높은 것을 예상할 수 있다.

AR과 MLPAR을 비교하기 위해 두 결과를 함께 그린 그래프가 <Figure 7>이다. <Figure 7>을 보면 실제 데이터인 점이 있는 선에 AR로 추정한 세모가 있는 선보다 MLPAR로 추정한 네모가 있는 선의 오차가 더 작다. 즉 평균 제곱오차(MSE)는 AR에서 2.02854와 MLPAR에서 1.74332로 기존의 값을 1로 두었을 때 MLPAR에서 16.36% 감소함을 보였다.

두 번째로, 모형의 정확성을 평가하기 위해 모형의 AIC를 측정하였다. 결과는 <Table 5>와 같다.

Table 5. Result in accuracy of model(AIC)

Dataset	AR	MLPAR	Reduction ratio
Norway krone	-53.7004	**-78.1591***	31.29%
Sheep population	693.7204	**659.7684***	5.15%
Communications	264.1162	**249.0894***	6.03%
Death	165.8190	**157.4213***	5.33%
Plish zloty	-135.3040	**-149.826***	9.69%
Southern Oscillation	162.2950	**161.7643***	0.33%

제시한 MLPAR 방법은 음의 로그우도 값을 최소화하는 방법이므로 모든 데이터에서 MLPAR의 AIC 값이 작다. 이 결과는 MLPAR이 AR보다 더 적합한 모형임을 나타낸다.

(2) 모형의 예측성

모형의 예측성을 검증하기 위해 MLPAR과 기존 방법에서의 학습용 데이터로 구축한 모형에 전 데이터의 참 값을 대입해 계산된 값과 실제 미래 값인 검증용 데이터 간의 평균제곱오차를 계산했다. 결과는 <Table 6>와 같이 Communications 데이터를 제외한 데이터에서 AR보다 MLPAR이 더 작은 값의 평균제곱오차(MSE)을 보였다.

<Table 6>에서 제시한 데이터 중 대표적으로 Norway krone 데이터의 그래프를 확인하고자 한다. <Figure 8>의 수직선 이후 점 있는 선이 검증용 데이터, 점선의 테두리 부분이 이 부분

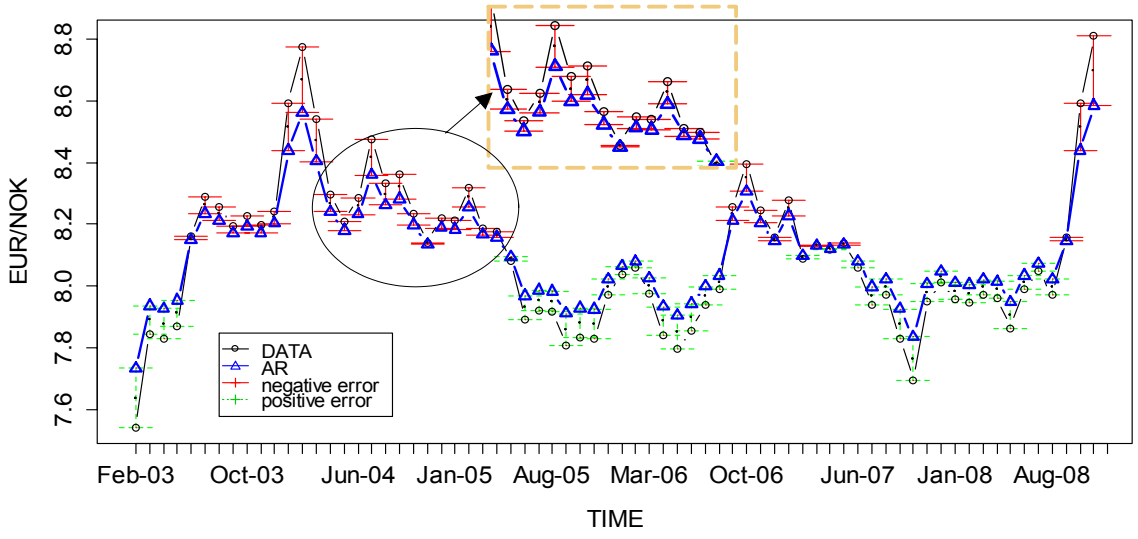

Figure 6. Data resulting from AR model and data

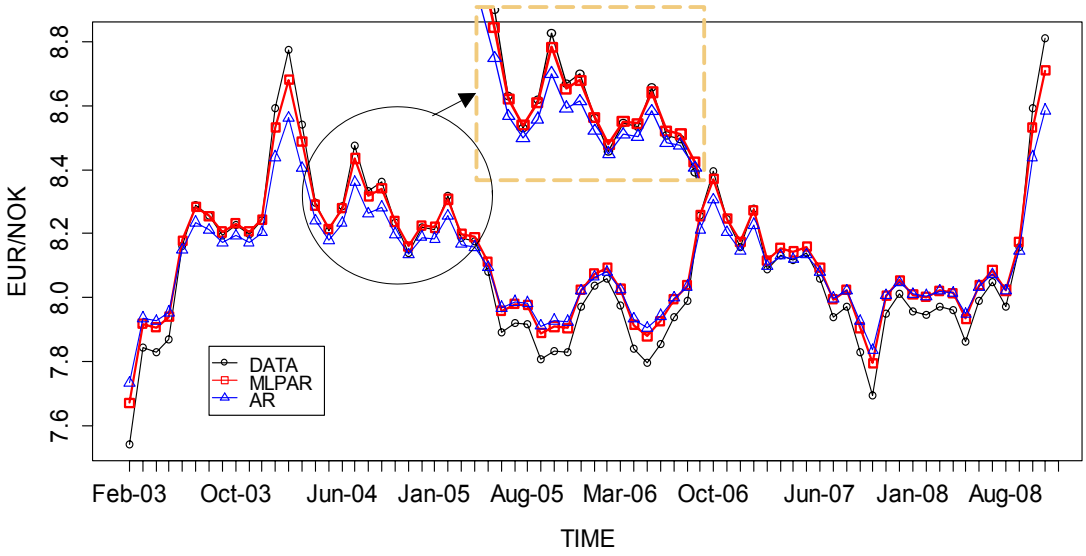

Figure 7. Data resulting from LPAR model and data(Train Data)

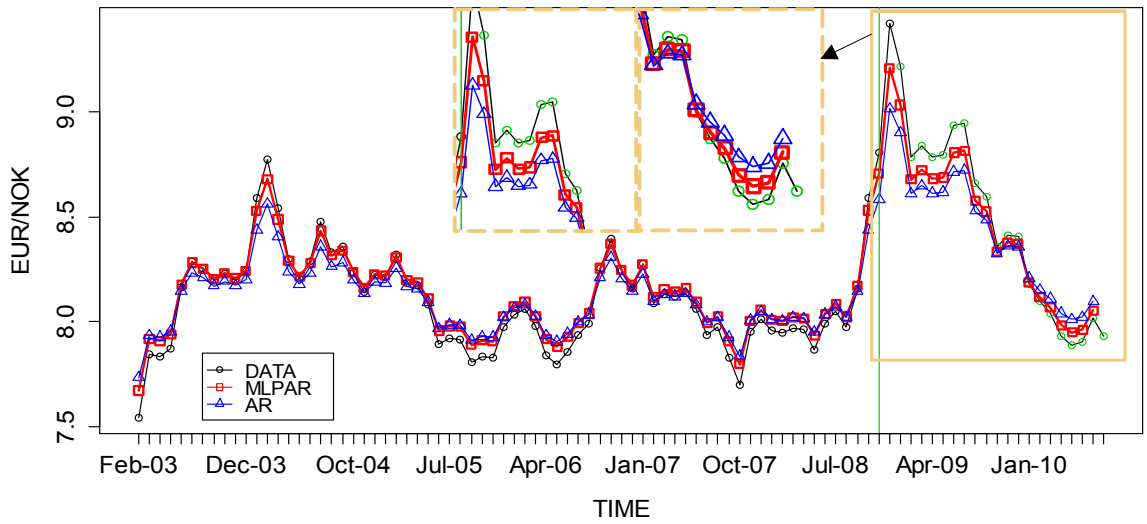

Figure 8. Data resulting from MLPAR model and data(Test Data)

을 확대한 것이다. <Figure 8>을 보면 세모 있는 선의 기존 방법(AR)보다 네모 있는 선의 MLPAR에서 오차가 더 작다. 즉 평균제곱오차(MSE)는 기존 방법(AR)에서 0.4567과 MLPAR에서 0.3447로 기존의 값을 1로 두었을 때 MLPAR에서 32.49% 감소함을 보였다.

Table 6. Result in prediction of model(MSE)

Dataset	AR	MLPAR	Reduction ratio
Norway krone	0.4567	**0.3447***	32.49%
Sheep population	77234.34	**76585.21***	0.85%
Communications	**0.9098***	0.9164	-0.72%
Death	13.1315	**12.1718***	0.79%
Plish zloty	0.1875	**0.1873***	0.10%
Southern Oscillation	15.8746	**15.5020***	2.40%

Table 7. Result in prediction of model(MSE) by rolling horizon method

Dataset	AR		MLPAR		Reduction ratio
	Mean	SD	Mean	SD	
Norway krone	0.3532	0.238	**0.3197***	0.246	10.48%
Sheep population	300.56	110.6	**86.095***	86.74	291%
Communications	**0.2715***	0.241	0.2764	0.241	-0.17%
Death	**0.7082***	0.437	0.7181	0.443	-1.38%
Plish zloty	0.0801	0.044	**0.0735***	0.043	8.98%
Southern Oscillation	1.2868	0.396	**1.2384***	0.344	3.91%

두 번째로 Rolling-Horizon 방법을 이용하였다. 결과는 <Table 7>과 같이 표준편차는 기존 방법(AR)과 MLPAR이 비슷하였지만, 평균은 6개 중 타입 Ⅳ에 속하는 Communications와 Death 데이터를 제외한 4개 데이터에서 감소함을 확인할 수 있었다.

4. 결 론

본 연구에서는 추정하기에 앞서 오차항의 분포를 정규분포로 가정 하에 자기회귀모형의 계수를 추정하는 기존 방법의 문제점을 해결하기 위해 새로운 방식(Maximum Likelihood of Pearson system for Auto-Regressive model)을 제안하였다. 자기회귀모형에 적합한 데이터를 이용해 기존 방법보다 MLPAR이 적률의 특성을 반영해 모형의 정확성과 예측성을 높였고, 더 적합한 것을 확인하였다.

그렇지만 MLPAR 추정이 항상 최적의 점을 선택하는 것은 아니다. 최우추정법에서 이용하는 최적화는 함수가 볼록함수이면 전역 최적화이다(Gay, 1990). 하지만 함수가 다봉분포(Multi-modal)라면 지역적인 최적화일 수 있다. 그러므로 자기회귀모형의 계수를 추정함에 있어 계수에 초기 값을 얼마로 설정하는 것이 가장 좋은지 고려해야 한다.

추후 연구로는 자기회귀모형의 계수를 추정함에 있어 점 추정(Point Estimation)이 아닌 구간 추정(Interval Estimation)을 할 수 있다. 구간 추정은 점 추정이 포함하고 있지 않는 추정오차에 대한 정보를 보강하기에 이점이 있다. 본 연구에서 제안한 방법이 오차항의 분포 가정을 없앴으므로 구간 추정 또한 기존과 다른 방식으로 접근해야 한다.

<Figure 9>를 보면 곡선은 계수 값에 따른 음의 로그우도 함수이다. 기존에 분포를 정규분포로 가정해 95% 신뢰구간을 구할 경우 a 값을 1.96으로 둔 후 함수에 a 값을 빼서 구할 수 있다. 그렇지만 비정규 분포이기 때문에 1.96이 아닌 피어슨 분포로 추정한 타입을 통해 95%인 a 값을 구해야 할 것이다.

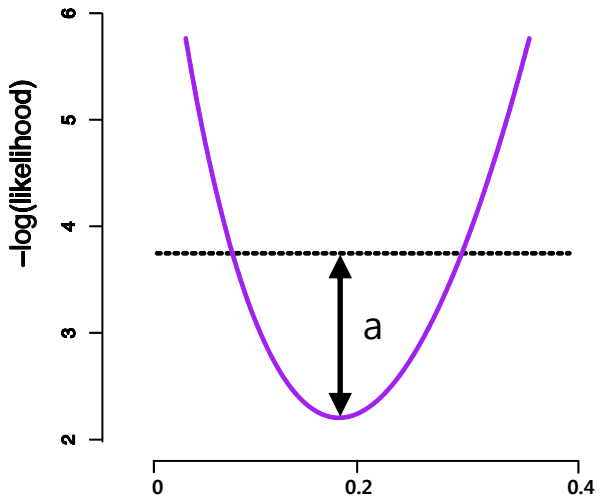

Figure 9. Method calculating confidence interval

<Figure 10>의 점선은 자기회귀모형에서 구간추정을 한 결과이다. 이는 95% 구간 추정을 하였지만 구간에 들어오지 않은 점의 개수 대비 전체 개수를 해본 결과는 81.5%였다. <Figure 10>을 보면 하한은 밀도가 더 높고, 상한은 넓게 분포된 것을 보아 네모가 있는 점선인 비대칭으로 한 구간 추정이 더 정확할 것이다. 이 연구는 본 연구에서 제시하는 평균제곱오차가 아닌 통계적으로 더 나은 지표를 이용하여 합리적인 의사결정에 도움을 줄 수 있을 것으로 예상된다.

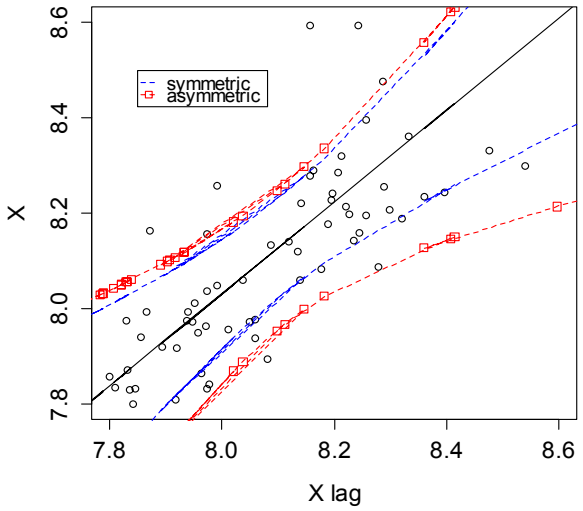

Figure 10. 95% confidence interval estimation

참고문헌

Akaike, H. (1969), Power spectrum estimation through autoregressive model fitting, Annals of t*he institute of statistical mathematics*, **21**, 407-419.

Akaike, H. (1974), A New Look at the Statistical Model Identification, *IEEE Transactions on automatic control*, **19**(6), 716-723.

Aldrich, J. (1997), R. A. Fisher and the making of Maximum Likelihood 1912 ~1922, *Statistical Science*, **12**(3), 162-176.

Bowerman, B. L., O'Connellm R., and Koehler, A. (2005), Forecasting Time Series and Regression, 4th edition, Thomson Brooks/Cole, CA, USA.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994), Time Series Analysis : Forecasting and Control, 3rd edition, Prentice-Hall, NJ, USA.

Elderton, W. P. and Johnson, N. L. (2009), Systems of frequency curves, Cambridge University Press, NY, USA.

Endo, H. and Randall R. B. (2007), Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter, *Mechanical Systems and Signal Processing*, **21**(2), 906-919.

Gay, D. M. (1983), Algorithm 611. Subroutines for Unconstrained Minimization Using a Model.Trust, *ACM Transcations on Mathematical Software*, **9**(4), 503-524.

Gay, D. M. (1990), Usage Summary for Selected Optimization Routines, *Computing Sicience Technical Report*, AT&T Bell Laboratories, NJ, USA.

Last, M., Klein, Y., and Kandel, A. (2001), Knowledge Discovery in Time

Series Databases, *IEEE Transactions on systems*, *Man, and cybernetices-part B : cybernetics*, **31**(1), 160-169.

Nagahara, Y. (2000), Non-Gaussian Filter and Smoother Based on the Pearson Distribution System, *Journal of Time Series Analysis*, **24**(6), 721-738.

Nagahara, Y. (2004), A method of simulating multivariate nonnormal distributions by the Pearson distribution system and estimation, *Computational Statistics and Data Analysis*, **47**(1), 1-29.

Pankratz, A. (2008), Forecasting with Univariate Box-Jenkins Models : Concepts and Cases, John wiley and sons, NY, USA.

Parrish, R. S. (1983), On an integrated approach to member selection and parameter estimation for Pearson distributions, *Computational Statistics and Data Analysis*, **1**, 239-255.

Pearson, K. (1895), Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material, *Philosophical Transactions of the Royal Society of London A*, **186**, 343-414.

Pearson, K. (1916), Mathematical Contributions to the Theory of Evolution. XIX. Second Supplement to a Memoiron Skew Variation, *Philosophical Transactions of the Royal Society of London, Series A, Containing Papersof a Mathematical or Physical Character*, **216**, 429-457.

Rydberg, T. H. (2000), Realistic Statistical Modelling of Financial Data, *International Statistical Review*, **68**(3), 233-258.

Sethi, S. and Sorger G. (1991), A Theory of Rolling Horizon Decision Making, *Annuals of Operations Research*, **29**, 387-416

Walpole, R. E. Myers, R. H. Myers, S. L., and Ye, K. (2006), Probability and Statistics for engineerings and scientists, 9 edition, Prentice Hall, NJ, USA.