2005 Proceedings of the Spring Conference, Korean Statistical Society

# 확률밀도함수가 표현되지 않는 경우 수치적 최우추정법 - 웨이크비 분포 적용

박 정 수[1)]

요 약

확률밀도함수가 명확히 표현되지 않고 오직 백분위함수로만 표현되는 분포에서 최우추정치를 구하는 수치적 최적화 알고리즘에 대해서 연구하였다. 이 최우추정 알고리즘을 수문학 등에서 사용되는 5-모수의 웨이크비 분포에 적용하였으며, 몬테카를로 시뮬레이션을 통하여 L-적률추정법과 그 성능을 비교하였다.

## 1. 서론

통계학에서 확률분포의 모수를 추정하는 방법으로 최우추정치가 가장 널리 알려져 있고, 이론상으로도 점근최적 방법이다. 그런데 만약 확률밀도함수(pdf)가 명확히 표현되지 않는 경우는 확률밀도함수의 곱인 우도함수 또한 명확히 표현되지 않으므로 최우추정치를 구하는데 어려움이 생긴다. 예를들어 백분위함수가 다음과 같이 매우 간단한 1-모수 분포의 경우에도 (Skew 로지스틱 분포),

$$x_F = Q(F) = \alpha \ln(F) - (1-\alpha)\ln(1-F),$$

pdf를 명확히 표현할 수 없다. 그러면 이런 경우에 최우추정치(MLE)는 어떻게 구할 것인가?

본 연구에서는 이러한 문제의 해답을 구하는 수치적 최우추정 알고리즘을 개발하고 그것을 웨이크비 분포 (Wakeby distribution)에 적용하였다. 웨이크비 분포는 수문학 분야에서 중요 관심사이지만 pdf가 명확히 표현되지 않아 지금까지 최우추정치가 사용되지 못했고 L-적률추정치가 사용되어 왔다.

웨이크비 분포는 일일 강수량 또는 년 중 최대 일일 강수량 (극값)의 확률분포로서 수문학, 토목공학, 수자원연구에 자주 쓰이는 분포이다. 백분위함수는 두 개의 일반화 파레토 분포의 합으로 구성되는 5-모수 분포이며, 다음과 같이 표현된다.

$$Q(F) = \xi + \frac{\alpha}{\beta}\{1-(1-F)^{\beta}\} - \frac{\gamma}{\delta}[1-(1-F)^{-\delta}]. \quad \text{----- (1.1)}$$

$\theta = (\xi, \alpha, \beta, \gamma, \delta)$ 라는 다섯 개의 모수를 가지는 특별한 분포로서 정의역은 다음과 같다.

$$\xi \le x < \infty \quad \text{if } \delta \ge 0 \text{ and } \gamma > 0,$$
$$\xi \le x \le \xi + \alpha/\beta - \gamma/\delta \text{ if } \gamma = 0 \text{ or } \delta < 0. \quad \text{---------- (1.2)}$$

이 분포는 특별한 경우로서, 일반화 극단분포(GEV), 일반화 파레토분포( $\gamma = 0$ 또는 $\alpha = 0$ 일 때), 로그-정규분포, 로그-감마분포를 포함하는 매우 폭넓은 분포이다.

이 분포는 pdf나 누적분포함수가 명확히 표현되지 않으며 다만 백분위 함수만 명백히 표현된다. 모수 추정방법으로 지금까지는 백분위수에 대한 최소제곱법 (Karian and Dudewicz, 2000)이나 L-적률 추정법 (Hosking, 1990)이 사용되어 왔다. 최우추정법은 계산상의 어려움 때문에 지금껏 이용되지 못했지만, 대표본일 때는 최우추정치가 L-적률 추정법보다 좋을 것으로 기대된다. 본 또한 소표본일 때 L-적률 추정법과 최우추정법의 성능을 비교하였다.

1) 전남대학교 자연대 통계학과 교수, jspark@chonnam.ac.kr

## 2. 수치적 최우추정 알고리즘

오직 백분위함수 만을 이용하여 최우추정치를 계산하는 알고리즘의 개략적인 단계를 보면 다음과 같다.

**단계 1.** 주어진 한 관측치 $x$ 에 대해서 $x=Q(F)$ 을 풀어서 F(x)를 구한다. 이때 Newton-Raphson 형의 반복적 수치계산을 통하여 F(x)를 구한다.

**단계 2.** 단계1에서 구한 F(x)에 대해 다음과 같은 식을 통하여 확률밀도함수 f(x)를 구한다.

$$f(x) = \frac{1}{dQ(F)/dF}\Big|_{F=F(x)}.$$

**단계 3.** 단계1과 2를 모든 x에 대해 반복하여 실행한 뒤, 음의 로그우도함수( $\lambda$)를 계산한다.

$$\lambda_{\underline{\theta}} = -\sum_{i=1}^{n} \log f_{\underline{\theta}}(x_i) .$$

**단계 4.** $\lambda$ 를 최소로 하는 모수를 (1.2)의 제약조건 하에서의 비선형 최적화 알고리즘을 이용하여 수치적으로 구한다. 이때 필요한 각 모수에 대한 $\lambda$의 미분벡터도 수치미분으로 구해서 이용한다.

**단계 5.** 초기치 여러 개를 주어서, 각각 도달한 최저치 중에서 가장 작은 $\lambda$ 값을 갖는 모수를 최종적인 최우추정치(global optimizer)로 간주한다.

위의 단계2에서 구해지는 pdf는 정의역 (1.2) 하에서 다음과 같은 형태가 된다.

$$f_{\underline{\theta}}(x) = \frac{(1-p_x)^{\delta+1}}{\alpha(1-p_x)^{\beta+\delta}+\gamma}. \quad \text{--- (2.1)}$$

여기서 $p_x$ 는 위의 단계1에서 x 에 대응하여 구해지는 F(x) 를 뜻하며, 이는 다섯 개의 모수 $\underline{\theta}=(\xi, \alpha, \beta, \gamma, \delta)$ 의 함수이다. 이제 음의 로그우도함수는

$$\lambda = -\sum_{i=1}^{n}[(\delta+1)\ln(1-p_{x_i}) - \ln\{\alpha(1-p_{x_i})^{\beta+\delta}+\gamma\}] \quad \text{---- (2.2)}$$

이 된다.

이제 식(2.2)를 최소로 하는 모수들을 구하기 위해 수치적 최적화 알고리즘을 이용하여야 한다. 그런데 (1.2)와 같은 모수에 대한 비선형 제약조건이 있는데다가 최적점을 찾아가는 과정에서 이 제약조건 밖으로 나가면 식(2.2)가 정의되지 않기 때문에, 일반적인 최적화 알고리즘이 아닌 feasible sequential quadratic programming을 사용하였다(Nocedal and Wright, 1999). 이를 위해 FFSQP 라는 포트란 프로그램을 적용하였다(Lawrence and Tits, 2001). 또한 단계1에서 $x=Q(F)$ 를 푸는데 필요한 Newton-Raphson 알고리즘은 Hosking(2000)이 작성하고 홈페이지에 올려둔 포트란 프로그램들을 사용하였다.

## 3. 웨이크비 분포에서 추정방법의 비교

웨이크비 분포에 대해서는 지금까지 주로 L-적률추정법이 적용되어 왔다. 이 절에서는 L-적률추정법(Method of L-moment estimation; L-ME)과 최우추정법의 성능을 몬테카를로 시뮬레이션을 통하여 비교하였다. L-적률은 순서통계량의 선형결합으로 정의되며, 모집단의 L-적률과 표본의 L-적률을 등치시킨 연립방정식의 해로써 L-적률추정치가 구해진다. 이 추정치는 일반적으로 모수의 수가 많은 경우 (3개 이상), 적률추정치보다 더 좋은 추정을 하며 특히 소표본에서 최우추정치보다 더 좋다고 알려져 있다. L-적률추정법에 관한 자세한 내용은

Hosking(1990)을 보기 바란다.

웨이크비 분포에 대한 기존의 연구(Landwehr, Matalas and Wallis, 1979, 1980) 및 본 연구에서는 크게 6개로 구분한 다음(표 3.1)과 같은 모수들에 대해 시뮬레이션 하였다. 여기서 $\xi$ 는 location-equivariant 하므로 항상 0 으로 두었고 $\alpha/\beta$ 는 scale-equivariant 하므로 $\alpha=\beta$ 로 놓고 난수를 발생시켰다. 또한 아래와 같은 설정 하에서 이루어졌다.

표본의 크기: 50, 150, 300, 500
백분위수 추정에 이용된 백분위: 90, 95, 98, 99.5
반복횟수: 1500.

모수의 추정치에 대한 RMSE를 보면 $\xi$, $\alpha$, $\beta$ 에 대해서는 전체적으로 MLE 가 작고 $\gamma$, $\delta$ 에 대해서는 표본의 크기가 150이나 300 까지는 L-ME 가 작고 그 보다 큰 경우는 MLE가 작은 경향을 보인다. 결론적으로 이 분포에서는 모수보다는 백분위수의, 특히 98백분위수 이상에서의 정확한 추정이 더 중요한 점을 감안하면 표본의 크기가 300 정도까지는 L-ME를 사용하고 그 이상에서는 MLE의 사용이 권장된다.

<표3.1: 시뮬레이션에 사용되는 6개의 웨이크비 분포의 모수 설정>

| 분포 | $\alpha=\beta$ | $\gamma$ | $\delta$ |
|---|---|---|---|
| WA-1 | 16.0 | 0.8 | 0.2 |
| WA-2 | 7.5 | 0.6 | 0.12 |
| WA-3 | 1.0 | 0.6 | 0.12 |
| WA-4 | 16.0 | 0.4 | 0.04 |
| WA-5 | 1.0 | 0.4 | 0.04 |
| WA-6 | 2.5 | 0.2 | 0.02 |

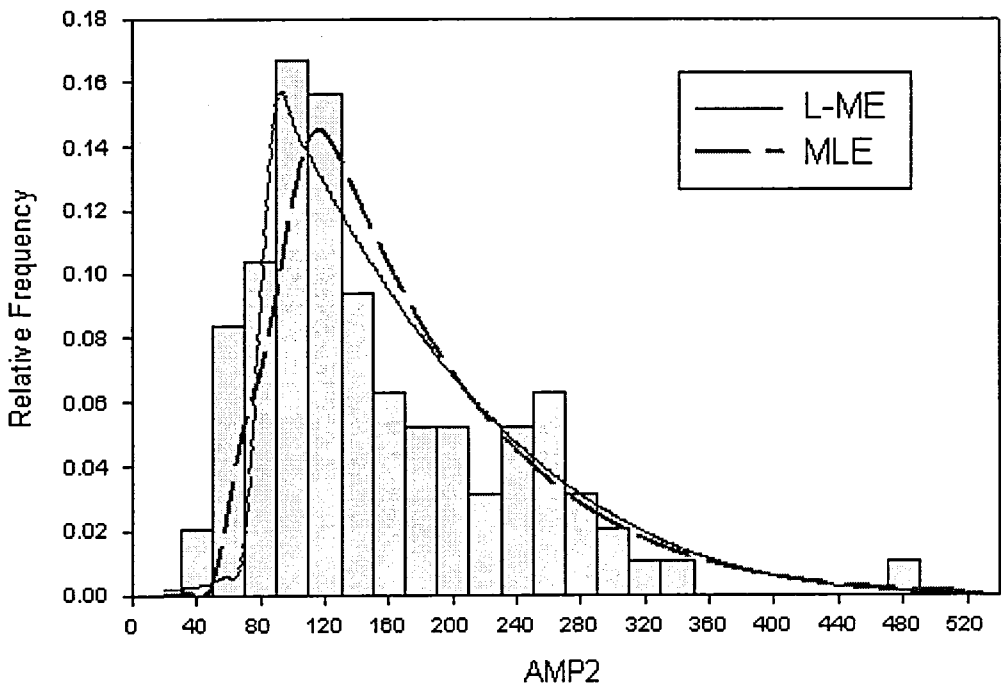
## 4. 토의 및 결론

위에서 기술한 수치적 MLE 알고리즘의 한 문제는 한번 우도함수를 계산하는데 L-적률방법에 비하여 시간이 오래 걸린다는 점이다. 예를 들어 자료의 수가 200개인 경우, 열 번의 초기치를 시도한다고 했을 때 MLE를 구하는데 걸리는 총 계산 시간은 PC에서 5초 정도이다. 그런데 L-적률 방법은 0.1초도 안 걸린다. 따라서 MLE 계산 시간을 단축시키기 위하여, 단계1에서 누적분포함수의 값 F(x)를 구하기 위해, Newton-Raphson 알고리즘의 2차형으로서 매우 빠른 Halley의 방법 (Huh, 1986)이 이용되었는데 이는 Hosking(2000)에 의해 구현되었다. (여기서 Huh(1986)의 연구와 다른 점은 백분위수가 주어졌을 때 F(x)를 구한다는 것이다.) 이 경우 Newton-Raphson 알고리즘보다 2배정도 빨라졌다.

위의 알고리즘은 모든 x 에 대해서 F(x)를 구했는데, 이는 시간이 많이 걸리므로 이를 개선하기 위해 일부의 x 에 대해서만 F(x)를 구한 뒤에 내삽법을 적용하여 나머지에 대해 F(x)의 근사값을 구한다. 특히 F(x) 가 단순 비감소함수라는 특성을 이용하면 내삽오차를 줄일 수 있을 것이다. 이때 기준점으로 어떤 x를 몇 개 선택할 것인가가 연구과제이다.

최우추정치에 대한 분산을 추정하기 위해 핏셔의 정보 행렬의 역행렬을 이용할 수 있다. 이

때 핏셔의 정보 행렬를 구하기 위해서는 로그 우도함수에 대한 2차 미분이 필요하게 되는데, pdf가 명백히 표현되지 않는 경우이므로 2차 수치미분을 이용할 수밖에 없다. 수치미분은 본질적으로 오차를 수반하므로 가장 오차가 작은 알고리즘을 이용해야 할 것이다. 그런데 문제는 $\xi$의 조건 (1.2) 때문에 정상성조건(regularity condition)이 만족되지 않아서 점근정규성이나 점근효율성을 보장하지 못한다는 점이다. 또한 핏셔의 정보 행렬이 어떤 경우는 존재하지 않을 수도 있다는 점이다. 따라서 모수들에 대해서 정상성조건을 만족케하는 어떤 제약조건을 줄 수 있는데, 핏셔의 정보 행렬이 수리적으로 명확히 표현되지 않기 때문에 이러한 일 또한 쉽지 않다. 이러한 어려움 때문에 추정치에 대해 붓스트랩에 의한 분산 계산이 바람직해 보인다.




<그림 3.1: 부산의 2일 년 최대강수량의 시계열(1904년-1999년) 자료의 상대도수 히스토그램과 웨이크비 분포의 L-적률추정법(L-ME)과 최우추정법(MLE)을 사용했을 때의 확률밀도함수>

## 참고문헌

맹승진 (2000), 수문 자료의 통계학적 분석 방법, 한국수자원공사 연구 홈페이지, http://www.kowaco.or.kr/~water/water-dic/seawater/waterco16-4.html

박정수, 황영아 (2005), 3-모수 카파분포에서 추정방법들의 비교, 한국통계학회논문집 투고.

허준행 (1997), 수문통계학의 기초(5), **한국수자원학회지**, 제30권 1호, pp. 88-96.

Dudewicz, E.J., and Karian, Z. (1999), Fitting the generalized lambda distribution by the method of percentiles, *Amer. Jour. Math. Management Sci.* **19**, 1-73.

Gilchrist, W.G.(2000), *Statistical Modelling with Quantile Functions*, Chapman & Hall/CRC, Boca Raton.

Gill, P.E., Murray, W., and Wright, M.H. (1981), *Practical Optimization*. Academic Press, New York.

Hosking JRM (1990), L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of The Royal Statistical Society, Series B*, **52(1)**, 105-124.

Hosking, JRM (1999), LMOMENTS: Fortran routines for use with the method of L-moments, Version 3.02, available at http://www.research.ibm.com/people/h/ hosking/lmoments.html

Hosking, J.R.M., and Wallis, J.R., (1997), *Regional Frequency Analysis: An Approach based on L-moments*. Cambridge University Press, Cambridge.

Karian, Z., and Dudewicz, E.J. (2000), *Fitting Statistical Distribution*, CRC Press, Boca Raton, Florida.

IMSL (1987), *User's Manual Math/Library Version 1.0, Vol. 3*, IMSL Inc., Houston (1150 pp)

Landwehr J.M., Matalas N.C., and Wallis J.R. (1979), Estimation of parameters and quantiles of Wakeby distributions. *Water Resources Research*, **15**, 1361-1379.

Lawrence CT, and Tits A. (2001), A computationally efficient feasible sequential quadratic programming algorithm. *SIAM Jour. Optimization*, 11(4), 1092-1118.

Nocedal, J. and Wright, SJ. (1999), *Numerical Optimization*, Springer, New York.

Park JS, Jung HS, Kim RS, Oh JH (2001), Modelling summer extreme rainfall over the Korean peninsula using Wakeby distribution. *International Journal of Climatology, 21*, 1371-1384.

Park JS, Kim RS, and Jung HS (2000), Parameter estimations of Wakeby distribution and its application to Korean extreme rainfall, *Proceeding of The Tenth Japan and Korea Joint Conference of Statistics*, 157-164, Beppu, Japan.

Huh, M. Y. (1986), Computation of percentage points. *Communications in Statistics-Simulation and Computation*, **15**, pp. 1191-1198.