The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 333-339

On the Development of Probability Matching Priors for Non-regular Pareto Distribution

Woo Dong Lee[1], Sang Gil Kang[2], Jang Sik Cho[3]

Abstract

In this paper, we develop the probability matching priors for the parameters of non-regular Pareto distribution. We prove the propriety of joint posterior distribution induced by probability matching priors. Through the simulation study, we show that the proposed probability matching prior matches the coverage probabilities in a frequentist sense. A real data example is given.

Keywords : Pareto Distribution, Noninformative Prior, Probability Matching Prior, Non-regular Case

1. 서론

파레토분포는 감소위험률 (decreasing failure rate)를 가지며 양의 방향으로 치우쳐진 형태를 띄는 분포이다. 또한 긴 꼬리 확률을 가지며 사회, 경제분야와 관련한 자료뿐만 아니라 신뢰성분야의 자료를 분석하기에 적합한 분포라고 알려져 있다 (Arnold와 Press, 1983).

사회경제학분야의 자료들은 흔히 오른쪽으로 꼬리가 긴 분포를 하는 경우가 많다. 이러한 자료를 통계적으로 분석하기 위하여 파레토분포나 로그정규분포등을 이용하여 왔다. 특히 경제학자들은 위쪽 꼬리부분에 대해 관심이 많은 경우가 있는데, 로그정규분포가 위쪽 꼬리부분에 대해서 잘 적합시키지 못하는 단점이 있는 반면 파레토 분포의 경우는 꼬리부분에서도 상대적으로 잘 적합시키기 때문에 이 분야에 많이 이용되었다.

Arnold 와 Press (1983, 1989a,1989b), Geisser (1984, 1985), Lwin (1972), Nigm 과 Hamdy (1987) 그리고 Tiwari, Yang과 Zalkikar (1996)은 이 분포에 대한 베이지안 추론에 대해 연구한 바 있다. 특히, Arnold 와 Press (1989b)는 공액사전분포 (conjugate prior)와 수정된 Lwin 사전분포를 이용하여 베이지안 추론에 대해 연구하였다.

그러나 분포에 대한 사전정보가 거의 없는 경우, 공액사전분포의 사용은 주관적일 수 있다. 이러한 점을 보완하기 위하여 무정보적 사전분포를 이용하는데, 대표적인 사전분포로는 제프리스

1) Associate Professor, Faculty of Information and Science, Daegu Haany University, Kyungpook, Korea
E-mail : wdlee@kyungsan.ac.kr
2) Assistant Professor, Department of Applied Statistics, Sangji University, Wonju, Korea
E-mail : sangkg@mail.sangji.ac.kr
3) Associate Professor, Department of Statistical Information Science, Kyungsung University, Pusan, Korea
E-mail : jscho@star.ks.ac.kr

(Jeffreys')사전분포, 기준사전분포 (reference prior) 그리고 확률대응 사전분포 (probability matching prior)등이 있다. 제프리스 사전분포는 피셔정보행렬의 행렬식의 제곱근에 비례하는데, 많은 통계분석에서 유용하게 이용되어왔다. 그러나 제프리스 사전분포는 관심모수 (parameter of interest)와 장애모수 (nuisance parameter)가 있는 통계모형에서는 불일치성 (inconsistent), 확률대응성 등의 문제점이 있다.

제프리스 사전분포에 대한 대안으로 Tibshirani (1989)는 장애모수를 가지는 모형에 대해 1차 확률대응사전분포를 개발하였다. 그 후, Datta와 Ghosh (1995a, 1995b, 1996), Murkerjee 와 Ghosh (1997), Murkerjee 와 Dey (1993) 등은 확률대응사전분포를 개발하고 여러 통계모형에 적용한 사례가 있다. 앞서 언급된 연구는 대부분 정칙인 분포 (regular class) 에 대한 결과이다. 한편, Ghosal (1999)은 비정칙인 분포에 대해 확률대응사전분포를 개발하는 방법을 제시하였다.

이 논문에서는 비정칙인 파레토분포에 대해 확률대응사전분포를 개발한다. 그리고 개발된 사전분포들에 의해 유도된 사후분포에 대한 적절성을 조사할 것이다. 개발된 사전분포들을 이용하여 실제 자료를 분석하는 예를 보인다. 모의실험을 통하여 사전분포들의 확률대응성을 조사할 것이다.

2. 확률대응사전분포

2개의 모수를 가지는 비 정칙 파레토분포의 확률밀도 함수는 다음과 같다.

$$f(x;\alpha,\beta)=\alpha\beta^{\alpha}x^{-(\alpha+1)},\ x>\beta,\alpha,\beta>0 \tag{1}$$

여기에서 α는 형태모수 (shape parameter)이고 $\frac{1}{\beta}$는 정도 모수 (precision parameter)이다. 식 (1)에서 보는 것과 같이 x가 β에 의존하므로 이 분포는 비 정칙인 분포이다.

Ghosal (1999)에 의하면 관심모수가 α인 경우, 이 모수에 대해 점근적으로 $O(n^{-1})$로 일치하는 확률대응사전분포 $\pi_M^{\alpha}(\alpha,\beta)$는 다음의 미분 방정식을 만족한다.

$$\frac{1}{\lambda(\alpha,\beta)}\ \frac{\partial}{\partial\alpha}\log\pi_M^{\alpha}(\alpha,\beta)=-\ \frac{\partial}{\partial\alpha}\left\{\frac{1}{\lambda(\alpha,\beta)}\right\}$$

여기에서

$$\lambda(\alpha,\beta)=\sqrt{-2E\left[\frac{\partial^2}{2\partial\alpha^2}\log f(X;\alpha,\beta)\right]}$$

이다. 식 (1)로 부터 $\lambda(\alpha,\beta)=1/\alpha$가 됨을 알 수 있다. 그러므로 관심모수가 α인 경우, 위의 미분 방정식을 만족하는 확률대응사전분포는 다음과 같다.

$$\pi_M^{\alpha}(\alpha,\beta)\propto K(\beta)\frac{1}{\alpha} \tag{2}$$

여기에서 $K(\beta)$는 β에 대한 임의의 양의 함수이다.

이제 β가 관심모수인 경우에 frequentist의 확률과 점근적으로 $O(n^{-2})$로 일치하는 확률대응사전분포 $\pi_M^{\beta}(\alpha,\beta)$는 다음의 미분 방정식을 만족하는 함수이다.

$$\frac{1}{c(\alpha,\beta)}\frac{\partial}{\partial\beta}\log\pi_M^{\beta}(\alpha,\beta)+\frac{2A_{11}(\alpha,\beta)}{c(\alpha,\beta)\lambda^2(\alpha,\beta)}\frac{\partial}{\partial\alpha}\log\pi_M^{\beta}(\alpha,\beta)$$
$$=-\frac{\partial}{\partial\beta}\left\{\frac{1}{c(\alpha,\beta)}\right\}-2\frac{\partial}{\partial\alpha}\left\{\frac{A_{11}(\alpha,\beta)}{c(\alpha,\beta)\lambda^2(\alpha,\beta)}\right\}$$

여기에서

$$A_{11}(\alpha,\beta)=\frac{1}{2}E\left[\frac{\partial^2}{\partial\alpha\partial\beta}\log f(X;\alpha,\beta)\right],$$

그리고

$$c(\alpha,\beta)=E\left[\frac{\partial}{\partial\beta}\log f(X;\alpha,\beta)\right]$$

이다.

주어진 확률밀도함수로부터 $A_{11}(\alpha,\beta)=\frac{1}{2\beta}$, $c(\alpha,\beta)=\frac{\alpha}{\beta}$가 됨을 알 수 있다. 그러므로, 위의 미분 방정식은

$$\frac{\beta}{\alpha}\frac{\partial}{\partial\beta}\log\pi_M^{\beta}(\alpha,\beta)+\alpha\frac{\partial}{\partial\alpha}\log\pi_M^{\beta}(\alpha,\beta)=-\frac{1}{\alpha}-1$$

으로 쓸 수 있다. 이 미분방정식을 만족하는 해는

$$\pi_M^{\beta}(\alpha,\beta)=\frac{e^{\frac{1}{\alpha}}h(e^{\frac{1}{\alpha}}\beta)}{\alpha}$$

이 된다. 여기에서 $h(\cdot)$는 미분가능한 함수이다. $h(e^{\frac{1}{\alpha}}\beta)$를 $(e^{\frac{1}{\alpha}}\beta)^{-1}$로 잡는다면 확률대응 사전분포는

$$\pi_M^{\beta}(\alpha,\beta)=\frac{1}{\alpha\beta} \tag{3}$$

이 된다. 이 사전분포는 α가 관심모수인 경우에 $K(\beta)=\frac{1}{\beta}$로 취하는 경우와 같다. 이 확률대응 사전분포는 관심모수에 관계없이

$$\pi_M(\alpha,\beta)=\pi_M^{\alpha}(\alpha,\beta)=\pi_M^{\beta}(\alpha,\beta)=\frac{1}{\alpha\beta} \tag{4}$$

로 두기로 하자.

3. 사후분포

이 절에서는 위에서 유도된 확률대응사전분포에 대한 적절성을 조사하려고 한다. 먼저 X_1, $X_2,\cdots,X_n$을 확률밀도함수 (1)을 가지는 모집단으로부터 추출된 n개의 확률표본이라고 하자. 이 표본에 대한 우도함수는 다음과 같다.

$$L(\alpha,\beta)=\alpha^n\beta^{n\alpha}\prod_{i=1}^{n}x_i^{-(\alpha+1)}I(w>\beta),$$

여기에서 $I(\cdot)$는 지시함수 (indicator function)이며, $w=\min(x_1,x_2,\cdots,x_n)$이다.

앞절에서 개발된 확률대응사전분포 (4)에 의한 사후분포의 적절성에 대해 고려해 보자.

정리 1. $n>1$이면, 사전분포 $\pi_M(\alpha,\beta)\propto\frac{1}{\alpha\beta}$를 이용한 사후분포는 적절 분포이며, 다음과 같은 확률밀도함수를 가진다.

$$\pi(\alpha,\beta\mid \mathrm{x})=\frac{n\left[\sum_{i=1}^{n}\log(\frac{x_i}{w})\right]^{n-1}\alpha^{n-1}\beta^{n\alpha-1}\prod_{i=1}^{n}x_i^{-\alpha}I(w>\beta>0)I(\alpha>0)}{\Gamma(n-1)} \tag{5}$$

증명. $\pi(\alpha,\beta)$와 우도함수를 결합한 사후분포는

$$\pi(\alpha,\beta\mid \mathrm{x})\propto\alpha^{n-1}\beta^{n\alpha-1}\prod_{i=1}^{n}x_i^{-(\alpha+1)}I(w>\beta>0)I(\alpha>0)$$

이다. 위의 결합사후분포를 β에 대해 적분하여 정리하면,

$$\int_0^\infty\int_0^w\pi(\alpha,\beta\mid \mathrm{x})d\beta d\alpha = \frac{\prod_{i=1}^{n}x_i^{-1}}{n}\int_0^\infty\alpha^{n-2}\exp\left\{-\alpha\sum_{i=1}^{n}\log(\frac{x_i}{w})\right\}d\alpha$$
$$= \frac{\prod_{i=1}^{n}x_i^{-1}}{n}\frac{\Gamma(n-1)}{\left[\sum_{i=1}^{n}\log(\frac{x_i}{w})\right]^{n-1}}$$

이 된다. 두 번째 식은 $n>1$인 경우 감마함수형태로 계산된다.

α와 β에 대한 결합사후확률분포를 이용하여 각각의 주변사후확률분포와 분포함수 등을 고려해 보자. 먼저 α에 대한 주변사후확률분포와 분포함수는 다음과 같다.

정리 2. $n>1$인 경우 α에 대한 주변사후확률밀도함수와 분포함수는 다음과 같다. 먼저, 주변사후확률밀도함수는

$$\pi(\alpha\mid \mathrm{x})=\frac{\left[\sum_{i=1}^{n}\log(\frac{x_i}{w})\right]^{n-1}\alpha^{n-2}\exp\left\{-\alpha\sum_{i=1}^{n}\log(\frac{x_i}{w})\right\}}{\Gamma(n-1)},\ 0<\alpha<\infty \tag{6}$$

이다. 즉, 모수 $n-1$과 $\sum_{i=1}^{n}\log(\frac{x_i}{w})$를 가지는 감마분포 (Gamma distribution)이다. 그리고, 분포함수는

$$F_\alpha(\alpha_0\mid \mathrm{x})=IG(\alpha_0;n-1,\sum_{i=1}^{n}\log(\frac{x_i}{w})), \tag{7}$$

이다. 여기에서 $IG(x_0;a,b)=\int_0^{x_0}\frac{b^a y^{a-1}\exp\{-by\}}{\Gamma(a)}dy$ 이다.

정리 3. $n>1$인 경우 β에 대한 주변사후확률밀도함수와 분포함수는 $0<\beta<w$인 경우,

$$\pi(\beta \mid \mathrm{x}) = \frac{n(n-1)\left[\sum_{i=1}^{n} \log(\frac{x_i}{w})\right]^{n-1}}{\beta\left[\sum_{i=1}^{n} \log(\frac{x_i}{\beta})\right]^{n}}, 0<\beta<w \tag{8}$$

이며, 분포함수는 $0<\beta_0<w$에 대하여,

$$F_{\beta}(\beta_0 \mid \mathrm{x}) = \left[\frac{\sum_{i=1}^{n} \log(\frac{x_i}{w})}{\sum_{i=1}^{n} \log(\frac{x_i}{\beta_0})}\right]^{n-1}, 0<\beta_0<w \tag{9}$$

이 된다.

4. 모의실험 및 예제

이 절에서는 위에서 제안된 사전분포를 이용하여 확률대응성에 대한 모의실험과 Arnold 와 Press (1989)의 논문에서 분석된 적이 있는 미국 1년 급여자료를 이용하여 분석하는 실제 예를 보일 것이다.

먼저, 제안된 확률대응사전분포에 대한 확률대응성을 조사하기 위하여 난수를 이용한 모의 실험을 실시하였다. 그 결과는 아래의 표 1과 같다. 표 1을 작성하기 위한 모수값은 $\alpha=3, \beta=3$인 경우로 가정하였다. 참고로, 모수의 값을 변화시켜보았으나 결과는 거의 유사하였다.

표 1을 계산하기 위한 절차는 다음과 같다. α와 β에 대한 주변사후확률분포함수 (7)과 (9)의 0.05와 0.95에 해당하는 분위수를 구하였다. α의 주변사후분포에서 확률 p에 해당하는 분위수를 $\alpha^{\pi}(p;\ \mathrm{x})$과 β에 대한 분위수를 $\beta^{\pi}(p;\ \mathrm{x})$라고 두자. 컴퓨터에서 난수를 10,000번 반복 생성하여 $p=0.05, 0.95$인 경우, 10,000번 중 $\{\alpha<\alpha^{\pi}(p;\ \mathrm{x})\}$, $\{\beta<\beta^{\pi}(p;\ \mathrm{x})\}$를 만족하는 사건에 대한 비율을 구하였다.

표 1 보듯이 제안된 확률대응 사전분포는 표본수가 작은 경우인 5나 10인 경우도 확률대응성을 잘 만족한다는 사실을 알 수 있었다.

n	$\pi_M^{\alpha}(\alpha, \beta)$		$\pi_M^{\beta}(\alpha, \beta)$	
5	0.0488	0.9505	0.0474	0.9532
10	0.0504	0.9482	0.0471	0.9466
15	0.0504	0.9535	0.0540	0.9491
20	0.0489	0.9520	0.0506	0.9509
25	0.0528	0.9468	0.0486	0.9529
30	0.0517	0.9506	0.0515	0.9507

표 1. 추정된 포함확률

예제. Arnold 와 Press (1989)는 30명의 미국 산업체 근로자의 1년 임금을 확률표본으로 이 자

료가 파레토 분포라고 가정하고 종속인 공액사전분포와 수정된 Lwin 사전분포를 이용하여 베이지안 분석을하였다. 30 개의 자료는 112, 154, 119, 108, 112, 156, 123, 103, 115, 107, 125, 119, 128, 132, 107, 151, 103, 104, 116, 140, 108, 105, 158, 104, 119, 111, 101, 157, 112, 115 (단위: ×US $). 이 자료를 확률대응사전분포에 적용하여 분석하려고 한다.

그 결과, 제곱손실오차를 가정하는 경우, 사후분포 (6)과 (8)을 이용한 사후분포의 평균인 베이즈 추정량을 구하면 $\tilde{\alpha}=E(\alpha \mid \mathrm{x})=5.7206$, $\tilde{\beta}=E(\beta \mid \mathrm{x})=100.3943$이며, 사후분포의 표준편차는 $\sqrt{Var(\alpha \mid \mathrm{x})}=1.0623$, $\sqrt{Var(\beta \mid \mathrm{x})}=0.624$이었다.

5. 결론

비 정칙인 파레토 분포에 대한 사전분포로서는 Lwin 사전분포 (1972)가 있다. 그러나 Lwin의 공액사전분포에서는 초모수들이 제약조건을 가지게 되므로 Arnold와 Press (1983)에 의하여 수정된 Lwin 사전분포가 제안되었다. Arnold와 Press (1989)에 의해 제안된 주관적인 공액사전분포나 수정된 Lwin 사전분포를 사용한 자료 분석에서는 반드시 초모수들 (hyperparameters) 을 결정해야한다. 그들은 실험자의 정보를 이용하여 초모수들을 결정하는 방법을 제안하였다. 그러나 실험자의 정보가 없다면 그들의 방법을 이용하여 자료 분석을 하기는 쉽지 않다.

본 논문에서는 비 정칙인 파레토 분포를 따르는 자료를 분석하기 위해서 무정보적 사전분포인 확률대응사전분포를 개발하였다. 모의실험을 통하여 제안된 사전분포에 대한 확률대응성을 조사하였으며, 모의실험 결과 표 1에서 보듯이 확률대응성이 잘 만족됨을 알 수 있었다. 그러므로 모수에 대한 정보가 거의 없거나 사용하지 못하는 경우가 있다면 위에서 제안된 확률대응사전분포를 사용하여 자료를 분석하는 것이 바람직 할 것이다.

참고문헌

[1] Arnold, B. C. and Press, S. J. (1983). Bayesian Inference for Pareto Populations, Journal of Econometrics, 21, 287-306.

[2] Arnold, B. C. and Press, S. J. (1989a). Bayesian Inference and Decision Techniques, eds. P. Goel and A. Zellner, Amsterdam: North-Holland, pp. 157-173.

[3] Arnold, B. C. and Press, S. J. (1989b). Bayesian Estimation and Prediction for Pareto Data, Journal of the American Statistical Association, 84, 1079-1084.

[4] Datta, G. S. and Ghosh, J. K. (1995a). On Priors Providing Frequentist Validity for Bayesian Inference, Biometrika, 82, 37-45.

[5] Datta, G. S. and Ghosh, M. (1995b). Some Remarks on Noninformative Priors, Journal of the American Statistical Association, 90, 1357-1363.

[6] Datta, G. S. and Ghosh, M. (1996). On the Invariance of Noninformative Priors, The Annal of Statistics, 24, 141-159.

[7] Geisser, S. (1984). Prediction Pareto and Exponential Observables, Canadian Journal of Statistics, 12, 143-152.

[8] Geisser, S. (1985). Interval Prediction for Pareto and Exponential Observables, Journal of

Econometrics, 29, 173-185.

[9] Ghosal, S. (1999). Probability Matching Priors for Non-Regular Cases, Biometrica, 86, 4, 956-964.

[10] Lwin, T. (1972). Estimation of the Tail of the Paretian Law, Scandinavian Actuarial Journal, 55, 170-178.

[11] Mukerjee, R. and Dey, D. K. (1993). Frequentist validity of Posterior Quantiles in the Presence of a Nuisance Parameter : Higher Order Asymptotics, Biometrika, 80, 499-505.

[12] Mukerjee, R. and Ghosh, M. (1997). Second Order Probability Matching Priors, Biometrika, 84, 970-975.

[13] Nigm, A. M. and Hamdy, H. L. (1987). Bayesian Prediction Bounds for the Pareto Lifetime Model, Communications in Statistics, 16, 1761-1772.

[14] Tibshirani, R. (1989). Noninformative Priors for One Parameter of Many, Biometrika, 76, 604-608.

[15] Tiwari, R. C. Yang, Y. and Zalkikar, J. N. (1996). Bayes Estimation for the Pareto Failure-Model Using Gibbs Sampling, IEEE Trans. Reliability, R-45, 471-476.

[2003년 3월 접수, 2003년 7월 채택]