한국통계학회논문집
8권 기념호, 2001
pp. 47-64

베이지안 통계학의 과거 · 현재 · 미래

김용대[1], 김혜중[2], 오만숙[3], 오현숙[4], 정윤식[5]

1. 머리말

본 논문은 한국통계학회 창립 30주년을 기념하기 위하여 베이지안 통계학의 역사, 현재 및 앞으로의 연구방향을 소개하면서 우리나라 베이지안 통계학의 현황과 미래를 조명하여 논의하고자 한다. 특히 베이지안 통계학의 소개는 한국 통계학회 창립이후 처음 시도되는 것이라 더욱 의미가 있으리라 할 것이다.

우선 "베이지안 통계학"이란 관심 있는 모든 것(모수, 결측치, 예측치 등)들에 대한 불확실성(uncertainty)을 확률분포로써 나타낸다는 가정에서 출발한다. 따라서, 전통적인 통계학의 기본 가정인 "고정된 미지"라는 관점과는 근본적으로 출발점이 다르다. 따라서, 베이지안 추론의 근간은 사후분포 또는 사후밀도함수이다. 사후밀도함수는 우도함수와 사전밀도함수의 곱에 비례하는 함수로서 우도함수에 압축된 표본정보와 사전밀도함수에 압축된 사전정보를 베이즈 정리에 의하여 합성한 것이다. 따라서 베이지안 패러다임은 개념적으로 간단하고, 직관적, 확률적 타당성을 지닌다고 볼 수 있다.

그러나 베이지안 접근 방법이 통계학의 워낙 넓은 분야에서 이루어져 있으므로 이를 체계적으로 열거하여 소개하는 것은 쉬운 일이 아니다. 특히 베이지안 통계학은 국내에서의 연구보다는 외국에서의 연구가 보다 활발히 전개되어 그에 대한 영향을 받아오고 있으므로 우선 국제적인 학문의 흐름을 소개하고자 한다. 우선 1970년대부터의 JASA(Journal of American Statistical Association)와 AS(Annals of Statistics)를 중심으로 그 흐름을 간략히 살펴보고자 한다. 그 다음으로 한국 통계학자들의 연구현황을 알아보기 위하여 한국통계학회에서 발간하는 통계학연구, 응용통계연구 및 한국통계학회 논문집을 중심으로 분석한다. 마지막으로, 현재 활발히 연구되어지는 분야들 중, 사전분포함수, 베이지안 계산, 베이지안 검정 및 모형선택, 베이지안 다변량연구 및 비모수적 베이지안 생존자료 분석에 대한 연구 동향을 소개한다.

2. 베이지안 통계학의 연구 현황

2.1 국제적인 현황

베이지안 활동이 급진전한 추세는 실제 숫자들을 통해서 부분적으로 나타내어질 수 있다.

1) 이화여자대학교 통계학과
2) 동국대학교 통계학과
3) 이화여자대학교 통계학과
4) 경원대학교 응용통계학과
5) 부산대학교 통계학과

JASA 논문집을 중심으로 살펴보면 1973년에서 1980년까지는 매년 6편 정도가 게재되었으며(총 55편), 1981년부터 1989년까지는 매년 평균 7편(총 68편)으로 약간 증가하였다. 그러나 1990년 Gelfand and Smith가 Gibbs sampler를 소개한 후 논문 수는 매우 큰 폭으로 증가하였다. 1990년에 10편을 시작으로 1991년에는 20편이 게재되었으며 그 이후로 매년 15편 이상씩을 발표하고 있다. 더욱이 1996, 1997, 1998년도에 베이지안 논문이 수적으로나 비율적으로 현격한 증가가 있어왔다. (표 2.1 참조). 이와 같은 현상은 MCMC방법들이 복잡한 통계적 모형들에 다양하게 응용되었으며, 더욱이 비모수적 베이지안 접근법들도 폭넓게 사용된 것이다. 또한 이론적인 측면을 강조하는 AS에서도 1990년도 이후에 베이지안 논문 편수가 급증함을 볼 수 있다. 이는 MCMC의 수렴성에 대한 연구와 비모수적인 베이지안 접근 방법에 대한 이론적 연구들이 기존의 베이지안 연구에 첨부되어 연구 영역이 확장되어 베이지안 논문 편수가 증가한 것이다. 이러한 현상들은 베이지안 통계학에 관한 책들의 수에 대하여서도 보여질 수 있다. 1990년 전에 발행된 베이지안 다변량 통계 관련 참고 서적은 Box와 Tiao (1973)의 *Bayesian Inference in Statistical Analysis*, Press (1982)의 *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods and Inference* 그리고 Anderson (1984)의 *An Introduction to Multivariate Statistical Analysis*가 고작 이었으나, 1990년 이후 이에 대한 참고서적이 현저히 증가하고 있다 이를 수치적으로 표현하면 처음 200년 동안(1769 - 1969) 베이지안 통계학에 관한 책은 오직 15권 정도였다. 또한, 그 다음 20년간 (1970-1989) 제작된 베이지안 통계학 책들은 30권 정도이다. 그 후 10년(1990-1999)간 베이지안 통계학 책들은 대체로 60권이다.

또한 그간에 있던 베이지안들의 연구모임도 1990년대에 접어들면서 확대되었다. 이들 중에서 현재 활동이 활발한 것은 1992년 창립된 Intenational Society for Bayesian Analysis (ISBA)와 같은 해 S. J. Press교수의 주도로 북미주지역의 베이지안들을 위해 결성된 미국통계협회의 베이지안 통계 연구회(Section on Bayesian Statistical Science; SBSS)가 있다. 이 외에도 베이지안들의 연구발표모임으로는 A. Zellner교수가 1970년에 시작한 Seminar on Bayesian Inference in Statistics and Econometrics (SBIES)가 매년 개최되고 있고, Bernardo교수의 주도로 스페인에서 3년 또는 4년마다 개최되는 "Valencia Conference", E. Jane교수의 업적을 기리기 위해 1981년부터 세계 각지에서 매년 번갈아 열리는 "Max Ent Workshop", 그리고 W. Edward교수에 의해 구성된 Bayesian Method in Decision Theory 연구모임 등이 있어 베이지안들 간에 활발한 연구교류가 이루어지고 있다.

표 2.1. JASA 와 AS에서의 베이지안 논문 편수 비율

년	JASA			Annals of Statistics		
	총편수	Bayesian	비율(%)	총편수	Bayesian	비율(%)
1973-1980	1147	55	4.8	920	69	7.5
1981-1989	1181	68	5.6	1030	78	7.6
1990-1995	833	102	12.3	680	62	9.1
1996	166	31	18.7	143	17	11.9
1997	161	28	17.4	106	9	8.5
1998	119	21	17.7	101	11	10.9
1999	107	13	12.1	88	9	10.2
2000	143	20	14	75	8	10.7
2001	58	18	13.8	9	0	0

최근 베이지안들의 연구동향을 조사 비교하기 위해 1990년 이후 개최된 제4차~제6차 Valencia Conference에서 발표된 베이지안통계 논문들을 AMS subject classification를 참조하여 연구분야별로 분류한 것을 표2.2에 나타내었다. 표2.2은 단변량 과 다변량통계로 구분하였는데 이는 조사한 논문에서 다룬 모집단의 변수나 관측치들이 단변량 인지 다변량인 지에 기준을 두었다. 이 표에 따라 베이지안들의 연구동향을 다음과 같이 정리할 수 있다:

(i) Proceeding에 수록된 논문 167편중에 베이지안 다변량통계 논문이 차지하는 비율은 49.7% (83편)이며, 년도별로 볼 때 이 비율은 점점 증가하여 1998년 제6차 발표회에서는 발표논문 47편 중 58.3% (28편)가 다변량통계 관련 논문임을 보여주고 있다.

표2.2. Proceedings of Valencia International Meeting (제4차~제6차)에 수록된 베이지안통계 논문들의 연구분야별 분류

년 도 / 구 분 / 연구분야	1991(제4차)		1994(제5차)		1998(제6차)		합 계
	단변량 통계	다변량 통계	단변량 통계	다변량 통계	단변량 통계	다변량 통계	
Exposition, Historical	5	2					7
Probability Theory	1					1	2
Decision Theory	4	1	1	6	2		14
Sampling Theory		1				1	2
Distribution Theory		1	1	1	1		4
Parametric Inference	10	2	9	9	6	1	37
Nonparametric Inference	3	1	4	2	2	2	14
Regression and Correlation	4	7	4	3	2	6	26
Experimental Design		1				1	2
Sequential Methods	3		3		1	1	8
Inference from Stochastic Process	5	1	1	4	2	7	20
Engineering Statistics	2		1	1	1	1	6
Applications	1	4	3	8	2	7	25
합 계	38	21	27	34	19	28	167

(ii) 다변량통계의 연구분야별 분류의 순위를 보면 Applications(전체 논문의 11.3%), Regression and Correlation (9.5%), Parametric Inference (7.1%), Inference from stochastic process (7.1%)등의 순으로 나타나고 있다. 또한, 년도별로 볼 때 Applications와 Inference from stochastic process의 논문편수 성장이 두드러지게 나타난다. 이들의 논문을 분석하면 전자에서는 패턴인식, 후자에서는 MCMC방법론에 대한 연구가 대부분이다. 그리고, 이 두 분야는 최근에 들어 다변량통계의 논문수가 증가하는 경향이 주목된다.

(iii) 다변량통계에서 Probability theory, Sampling theory와 Experimental design분야에 대한 베이지안들의 관심은 적은 것으로 나타나고, 최근에 베이지안들의 관심 분야로 부상한 Nonparameteric inference에 대한 연구는 꾸준히 진행되고 있음을 보여준다.

2.2. 국내연구 현황

국내 베이지안 통계학의 연구 동향을 살펴보기 위하여 한국통계학회에서 발간되는 세 가지 논문집인, 통계학연구(JKSS), 응용통계연구 및 한국통계학회 논문집들을 조사하였다. 이들을 베이지안 추론에 관한 논문 편수와 논문집에 게재된 총 논문 편수를 연도별로 정리하여 표2.3에 나타내었다. 이 표에 의하면 세 가지 논문집에 실린 베이지안 논문 편수는 전체 1593편중 133편으로 약 8.4%를 차지한다. 이를 논문집 별로 나누어 보면, 통계학 연구(JKSS)에서는 전체(618편)의 9%(56편), 응용통계연구에서는 6.5%, 그리고 한국통계학회 논문집에서는 9%를 차지하고 있어서, 방법론에 비해 응용분야에 대한 연구가 활발하지 않은 것으로 나타났다. 또한 각 논문집들의 논문 수가 90년대 중반이후로 두 배 이상 증가됨을 볼 수 있다. 이는 베이지안 통계학자의 증가이기도 하고, 90년에 깁스 샘플러가 소개된 후 어렵게 느껴지던 계산 방법이 보다 쉽게 접근할 수 있다는 이점에 의하여 보다 복잡한 모형에서의 해석에서 많은 통계학자들이 베이지안 방법을 선호하기 때문이다.

표 2.3. 한국통계학회 발간 논문집에서의 베이지안 논문 비율

년	통계학연구(JKSS)			응용통계연구			한국통계학회 논문집		
	총편수	베이지안	비율(%)	총편수	베이지안	비율(%)	총편수	베이지안	비율(%)
1973-1986	174	9	5.2	0	0	0	0	0	0
1987-1993	113	9	8	124	8	6.5	0	0	0
1994	37	3	8.1	41	3	7.3	24	1	4.2
1995	42	1	2.4	29	0	0	102	9	8.8
1996	60	4	6.7	27	1	3.7	59	4	6.8
1997	43	4	9.3	31	2	6.6	62	7	11.3
1998	38	5	13.2	36	3	8.3	88	5	5.7
1999	39	8	20.5	57	3	5.3	93	11	11.8
2000	38	9	23.7	47	4	8.5	87	9	10.3
2001	34	4	11.8	39	4	10.3	29	3	10.3
합	618	56	9	431	28	6.5	544	49	9

또한 최근 국내의 베이지안 다변량통계 분야의 연구 동향을 살펴보기 위해 1996년부터 2000년까지 5년 간 한국통계학회가 발간하고 있는 각 학회지에 발표된 논문들을 조사하였다. 이들 중에서 베이지안 다변량 통계분야의 논문편수와 학회지에 게재된 총 논문 편수를 년도별로 구하여 표2.4에 나타내었다. 이 표에 의하면 베이지안 통계논문이 JKSS에서는 전체의 8.2%, 응용통계연구

에는 2.0%, 그리고 한국통계학회 논문집에는 4.2%를 차지하고 있어서, 방법론에 비해 응용분야에 대한 연구가 활발하지 않은 것으로 나타났다.

20세기 후반부터 급진전된 베이지안 통계계산법, 로버스터 베이지안 분석법, 그리고 비모수적방법에 대한 연구는 그동안 큰 발전이 없었던 다변량 베이지안통계에 대한 연구에 활로를 열어주었다. 이를 계기로 국내에서도 일반화회귀모형, 분류분석, 패턴인식등에 사용되는 여러 가지 새로운 통계적모형 및 분석법들이 연구되고 있다. 그러나, 연구결과가 직접 응용분야에 사용되는 예는 상당히 적다. 이는 개발된 모형 및 방법론을 응용당사자가 직접 이해하고, 복잡한 알고리즘을 이용해 만든 컴퓨터프로그램으로 실 자료를 분석하는 자체가 대단히 어려운 작업이기 때문이다. 따라서 베이지안 다변량통계의 응용분야에서는 산학협동 이나 학연협동연구등를 통해 연구결과의 실용성을 높일 수 있는 방향으로 연구를 진행할 필요가 있다. 다행히 이러한 문제점을 인식한 몇몇 산업체 (주로 Risk Management 또는 CRM 관련 회사)에서 현재 베이지안들과의 공동연구를 진행하고 있어, 앞으로 응용분야의 발전에 희망을 주고있다.

표2.4 최근 5년간 한국 통계학회지별 베이지안 다변량 통계분야논문 건수

학회지명	96년	97년	98년	99년	00년	합계
JKSS	2(48)	4(43)	2(38)	4(39)	5(40)	17(208)
응용통계연구	0(27)	1(31)	2(20+)	1(52)	0(47)	4(177+)
한국통계학회논문집	4(91)	3(94)	3(84)	6(93)	3(87)	19(449)

※ 괄호 안의 숫자는 년도별 총 게재 논문편수

1990년까지 국내 베이지안 통계학자들의 수는 매우 적은 수준이었으나, 그 이후로 많은 통계학자들이 베이지안 방법론에 많은 관심을 보여 왔다. 더욱이 외국에서 연구하고 온 베이지안 학자들의 증가로 인하여 베이지안 통계학자들 모임의 필요성이 대두되어 1996년 1월에 첫 번째 모임을 충남대에서 갖게 되었다. 그 후 비정기적으로 춘·추계 학회 학술 논문 발표회에서 베이지안 section을 구성하여 베이지안 통계학자들간의 연구 교환을 활발히 하였다. 그러나 많은 베이지안 통계학자들의 욕구를 충족하기에는 비정기적인 모임으로는 부족하여 1999년 2월 가칭 베이지안 통계 연구회를 동국대에서 발족하였으며, 2000년 6월 한국통계학회의 정식연구회로 가입하여 현재 약 40여명의 회원으로 매년 2월과 8월에 정기 논문 발표회를 갖고 있다.

3. 베이지안 통계학의 연구동향

3.1. 사전분포함수

Bayes(1763)이래, 특히 Fisher(1922)이래 베이지안 추론의 장점에 대한 논쟁이 매우 활발하였다. 이러한 논쟁의 초점은 사전분포함수들의 선택의 임의성이었다. 그러나 오늘날 베이지안 방법들이 통계학의 이론, 응용분야 모두에서 폭발적으로 대중화되어 가고 있다. 이는 사전정보가 거의 없다 하더라도 신뢰할 수 있는 추론을 유도 할 수 있는 것은 무 정보적 사전 분포를 이용할 수 있기

때문이다. 따라서 지난 수년간 매우 많은 범위의 무정보적 사전분포가 제안되고 연구되어 왔다 (Berger and Yang, 1996). Laplace(1812)는 전 모수공간에서 무정보 사전분포로써 flat prior를 이용할 것을 제안하였다. 그러나 제안된 사전분포는 일대일 재모수화(one-to-one reparametrization)에서 불변을 유지하지 못하다는 비평을 받고 있다. 따라서 Jeffreys(1961)는 일대일 재모수화에서 불변을 유지하는 사전 분포를 제안하였다. 이 분포는 피셔 정보 행렬(Fisher information matrix)의 행렬식에 양의 제곱근으로 유도된다. 이러한 불변성에 불구하고, Jeffreys 사전분포는 nuisance 모수이 있는 경우에 비판을 받고 있다. 예를 들면, Bernardo(1979)는 Jeffreys 사전분포가 모형이 평균 μ와 분산 σ^2을 갖는 정규분포를 따를 때 $\frac{\mu}{\sigma}$의 추론에서 marginalization paradox(Dawid, Stone and Zidek, 1973)을 나타냄을 보였다. 두 번째 예로써 Berger and Bernardo(1992)는 Jeffreys 사전 분포가 균형 일원 ANOVA 모형에서 샘플 크기에 비례하면서 cell의 수가 무한이 커질 때 오차 분산의 불일치 추정치를 이끌고 있음을 보였다. 그러므로 Jeffreys 사전분포는 Neyman -Scott(1948) 현상을 피하지 못한다.

3.1.1. Reference Priors

Bernardo(1979)는 reference prior로 알려진 무정보 사전분포를 제안하였다. 이러한 사전 분포는 적당한 entropy distance(missing information)를 최대화 하므로써 얻어진다. 이때 nuisance 모수가 없는 경우는 Jeffreys 사전분포와 같다. 이러한 개념을 바탕으로 Berger and Bernardo(1989, 1992)는 reference prior를 찾는 알고리즘을 개발하였다. 그 후로 Ye and Berger(1991)는 exponential regression 모형에, Berger and Bernardo(1992)는 multinomial 모형, Chang and Eaves(1990)은 그룹모형에서 reference prior를 연구하였으며, Berger and Yang(1994)는 AR(1)모형, Sun and Ye(1994, 1995)는 정규평균의 곱에 관심을 갖고 연구하였다. 또한 Wolfinger and Kass(1996)는 variance components 모형에서 reference prior를 구하였다.

3.1.2. Matching Priors

무정보 사전 분포를 발전시키는 약간 다른 형태는 베이지안 credible set의 사후 coverage 확률과 그에 대응되는 프리컨티스트 coverage 확률을 조화시키는데 근거하고 있다.

사후 quantiles에 근거한 matching prior는 Welch and Peers(1963)이 처음 제안하였으며, 여기서 그들은 nuisance 모수가 존재하지 않는 경우이며 단변량 모수 θ에 국한하였다. 그 후로 nuisance 모수가 존재한 경우에 관심있는 모수에 대한 matching prior를 구하는 연구가 Peers(1965), Stein(1985)등에 의하여 진행되었다. 즉, $X = (X_1, \ldots, X_n)$가 $f(x;\theta)$를 확률밀도 함수로 갖는다 하자. 여기서 $\theta = (\theta_1, \ldots, \theta_p)^t$이고 θ_1이 관심있는 모수라 하자. $\theta_1^{(1-\alpha)}(\pi, x)$를 $(1-\alpha)$ 사후 quantile 라 하자. 즉, $P^{\pi}[\theta_1 \leq \theta_1^{(1-\alpha)}(\pi, x) \mid X] = 1-\alpha$ 이다. 이때

$$P_{\theta}[\theta_1 \leq \theta_1^{(1-\alpha)}(\pi, X)] = 1-\alpha+o(n^{-u})$$

을 만족하는 사전분포 $\pi(\theta)$를 matching prior라 한다. 여기서 $u=1/2$이면 이러한 matching prior를 first order probability matching prior라 하고 $u=1$이면 second order probability

matching prior라 한다. 이러한 사전분포함수를 구하는 것들은 적당한 미분방정식을 푸는 것이다. 특히 Tibshirani(1989)는 관심있는 모수가 nuisance 모수에 직교인 경우 그러한 matching prior의 완전한 해를 제시하였다. 이러한 연구들은 Datta and Ghosh(1995), Sun and Ye(1996), Mukerjee and Ghosh(1996), Garavn and Ghosh(1998)등에 의하여 계속되었다.

또한 Mukerjee and Dey(1993)는 nuisance 모수가 단변량인 경우에 first order probability matching prior들의 집합중에서 $u=1$ 만족시키는 matching prior를 만족시키는 사전분포를 구하는 방법을 제시하였다. 즉, second order probability matching prior를 구하였다. 이러한 결과들을 Mukerjee and Ghosh(1996)는 여러개의 nuisance 모수들이 있는 경우로 확장하였다. 이러한 연구들이 90년대들어와 왕성하게 될 수 있었던 것은 Bickel and Ghosh(1990)와 Ghosh and Mukerjee(1992)의 연구가 커다란 밑거름이 되었다.

3.1.3. Power priors

베이지안 추론에서 사전 분포함수의 유도는 가장 중요한 역할을 한다. 무 정보적이고 비 적절한 사전 분포들이 어떤 특정한 문제에 대하여 구체화하기 쉽고, 유용하다 할지라도 베이즈 요인(Bayes factor), 사후 모델 확률(posterior model probability)을 계산하기 위하여 적절한 사전분포가 필요하다는 것이 잘 알려져 있는 것처럼 무 정보적이고 비 적절한 사전 분포들은 모형선택 또는 모형비교와 같은 많은 응용분야에서 사용되는데 많은 문제점을 안고 있다. 추가적으로 베이즈 요인들은 모호(vague) 사전 분포들의 초모수 (hyperpapameter)들의 선택에 일반적으로 매우 민감하다고 알려져 있다. 그러므로 우리들은 정보적 사전분포의 유도를 피하기 위하여 모형선택에서 모호 사전분포들을 단순히 구체화시킬 수 없다. 또한 무정보 사전 분포는 사후 추정치에서의 불안전성과 깁스 샘플러에 대한 수렴성의 문제점을 야기 시킬 수 있다. 더구나, 무정보 사전 분포는 어떤 특별한 응용에서 가질 수 있는 실제 사전 정보의 사용을 이룰 수 없다. 그러므로, 이러한 상황에서는 정보적 사전 분포들이 본질적이다. 이러한 정보적 사전 분포는 연구자가 현재 연구에서와 같이 같은 반응 값과 설명변수를 측정할 수 있는 과거연구에 접근 할 수 있는 응용분야에서 이용될 수 있다. 예를 들면, 많은 암연구들과 AIDS clinical trials등에서 현재 연구는 가끔 과거 연구에서 사용되었던 treatment 들과 유사하거나, 약간 수정된 treatments들을 사용한다. 이러한 유사한 연구들에서 발생한 자료들을 ***Historical data***라 한다. 베이지안 관점에서 유사한 연구를 수행한 과거 자료들이 현재 연구의 결과들을 설명하는데 매우 많은 도움을 줄 수 있다. 이러한 관점에서 historical data에서 사전분포를 구할 수 있으며, 이를 ***Power prior***라 하며 다음과 같이 정의한다. power prior는 historical data에 근거한 우도함수에 a_0 승하여 정의하며, 여기서 $0 \leq a_0 \leq 1$ 는 모수이고, 현재연구에 historical data가 미치는 영향의 정도를 나타낸다. Chen, Manatunga and Williams(1998)가 인간 쌍둥이 자료에서 유전성 추정치 연구를 위한 power prior를 처음 시도하였다. 그 후로 Chen, Ibrahim and Yiannoutsos(1999)는 로지스틱 회귀모형에서 변수선택을 위하여 power prior를 연구하였다. 일반화 선형모형(GLM)에서는 Ibrahim, Chen and Ryan(2000), Chen, Ibrahim, Shao and Weiss(1999)등이 연구하였다. 또한 여러 형태의 생존자료 분석에서 Ibrahim and Chen(1998), Ibrahim, Chen and MacEachern(2000), Chen, Ibrahim and Sinha(1999), Chen, Dey and Sinha(1999)에 의하여 연구되었다.

3. 2. 베이지안 계산

베이지안 추론의 근간은 사후분포 또는 사후밀도함수이다. 사후밀도함수는 우도함수와 사전밀도함수의 곱에 비례하는 함수로서 우도함수에 압축된 표본정보와 사전밀도함수에 압축된 사전정보를 베이즈 정리에 의하여 합성한 것이다. 따라서 베이지안 패러다임은 개념적으로 간단하고 직관적, 확률적 타당성을 지닌다고 볼 수 있다. 그러나 베이지안의 실제적용은 단순하지 않은 경우가 종종 발생하는데 이는 근간이 되는 사후밀도함수가 수리적으로 주어지지 않고 단지 그 함수형태만 알 수 있는 경우가 많기 때문이다. 따라서 사후밀도함수와 나아가서는 사후추론을 위한 사후통계량을 구하는 베이지안 계산기법이 요구된다. 이 장에서는 이러한 베이지안 계산기법을 알아보기로 한다.

첫 번째 방법으로는, 표본의 크기가 큰 경우에 사후분포에 대한 정규근사를 이용하는 것이다. 표본의 크기가 크면 사전정보에 비하여 표본정보가 지배적이어서 사후밀도함수를 모수 θ의 MLE를 평균으로 그리고 헤시안 행렬의 역함수에 부호를 바꾼 것을 분산으로 하는 정규밀도함수로 근사시킬 수 있다. 이는 매우 간단하나 표본의 크기가 작으면 근사가 정확하지 않은 단점이 있다. 두 번째로는 모수 θ의 함수 $u(\theta)$의 사후기대치에 대한 근사식을 구하는 것으로 앞의 정규근사에 보정치를 추가하여 정확도를 높이는 방법으로 Lindley(1980)에 의해 제안되었고 Tierney and Kadane(1984)은 Lindley의 근사식을 개선하여 정확도가 더 높은 근사식을 제안하였다.

위의 두 방법들은 수리적 근사를 이용한 것으로 표본의 크기가 충분히 크지 않을 경우에는 상당히 오차가 클 수 있다. Naylor and Smith (1982)는 이점을 지적하고 Gaussian quadrature 방법을 이용하여 수치적으로 사후기대치를 추정할 것을 제안하였다. Gaussian quadrature 방법은 사후기대치 계산에 필요한 적분의 차원이 작은 경우에는 효율이 좋으나 차원이 높아지면 효율이 매우 떨어지는 단점이 있다.

다차원 적분을 효율적으로 수행하기 위한 수치적 기법으로 몬테칼로 기법이 1970년 후반부터 통계학자들의 관심을 끌기 시작하였는데 이는 컴퓨터의 기능 향상과 개인용 컴퓨터의 보급과도 관련이 있어 보인다. 몬테칼로 기법 중 주표본기법(importance sampling)은 초기에 Kloek and van Dijk(1978), Van Dijk and Kloek (1980, 1983, 1984), Stewart (1983), Geweke (1988) 등에 의하여 연구되었다. 주표본기법은 알고리즘이 단순하고 다차원 적분에도 효율이 뛰어난 장점이 있으나 단점으로는 기법의 효율과 정확도가 표본생성함수에 크게 의존한다는 것이다. 좋은 표본생성함수의 선택이 사후밀도함수가 다차원 공간에서 복잡한 형태를 가질 경우 실제적으로 매우 어려울 수 있기 때문이다.

1990년 Gelfand and Smith (1990)은 그동안 물리학계에서 사용되어지던 마코브 체인 몬테칼로 기법 (Markov chain Monte Carlo, MCMC)이 베이지안 추론의 여러 분야에서 매우 유용하게 사용되어질 수 있음을 보여주었다. MCMC 기법은 주표본기법에 비하여 효율은 다소 떨어지나 전문가가 아니라도 사용이 용이하고 그 적용범위가 매우 넓다는 장점을 가지고 있어 90년대 들어서 폭발적인 관심을 끌었으며 베이지안 패러다임의 적용 범위를 크게 확장시켰다는데 그 의의가 있다. 특히 모형이 복잡한 다차원 문제에도 적용이 용이하며, 필요에 따라 잠재변수의 사용으로 문제를 단순화시킬 수 있고 결측치의 처리 또한 용이하여 Gelfand and Smith (1990) 이후 MCMC 기법의 개발과 응용에 관한 논문이 계속 출간되고 있다. MCMC 기법 중 특히 깁스표본기법은 다차원 모수의 사후표본 생성이 각 원소 모수의 조건부 확률분포로부터의 난수생성으로 이루지

므로 차원의 저주를 상당 부분 피해갈 수 있으며 잠재변수를 용이하게 처리할 수 있는 등의 장점이 있어 널리 사용되고 있다. 그러나 깁스표본기법은 각 조건부 사후분포가 편리한 형태로 주어져야 한다는 조건이 필요한데 이를 위하여 보통 짝사전분포의 사용이 요구된다는 단점이 있다. 이를 해결하기 위한 방법으로 Metropolis-Hastings 기법(Hastings, 1970)이 있다.

Metropolis-Hastings 기법은 사후분포 대신 주표본기법에서 처럼 표본생성분포로부터 난수를 생성하되 생성된 난수를 랜덤하게 채택함으로써 사후분포와 표본생성분포의 차이를 보정하는 기법으로, 다차원 난수 생성 혹은 깁스표본기법과 혼합하여 사용되어지고 있다. 최근들어 모형선택에 사용되어질 수 있는 MCMC기법이 관심을 끌고 있는데 Green(1995)의 역점프 MCMC (Reversible jump MCMC) 기법은 차원이 서로 다른 모형들 사이에 마코브 체인을 형성함으로써 각 모형의 사후확률을 추정할 수 있는 장점이 있어 관심을 끌고 있다. 이상의 MCMC 기법은 중요한 장점들이 있어 널리 사용되고 있으나 단점으로는 수렴성의 확인이 어렵고 타 몬테칼로 기법에 비하여 효율이 떨어지는 것이다. 이러한 수렴성은 Cowles and Carlin(1996), Brooks and Roberts(1998), Mergersen, Robert and Guiheunneuc-Jouyaux(1998)등에 의하여 연구되었다. 특히 최근에 Robert(1998, Chap. 2)는 MCMC방법의 수렴성 확인을 위한 여러 가지 방법들을 소개한다. 베이지안 계산에서 최근의 추세는 수리적 근사 보다는 수치적 기법을 선호하는데 이는 모형이 예전에 비하여 복잡해짐에 따라 표본에 비하여 모수가 많은 경우가 빈번하고, 또한 컴퓨터의 발달과 보급으로 많은 통계학자들이 수치적 기법에 접근이 용이하다는 면이 작용하는 듯하다. 수치적 기법 중 MCMC 기법이 널리 사용되고 있는 추세인데 이는 언급한 바와 같이 MCMC 기법이 적용이 용이하고 응용범위가 광범위한 것으로 사료된다.

3. 3. 베이지안 가설 검정(베이지안 모형 선택)

베이지안 가설검정의 기본 원리는 설정된 가설, H_i ($i=0,1$)의 사후확률(posterior probability of H_i)을 구한 후, 서로 비교하여 값이 큰 쪽을 선택하는 것이다. 귀무가설과 대립가설을 각각 $H_0: \theta \in \Theta_0$, $H_1: \theta \in \Theta_1$이라 하고 π_0, π_1을 각각 Θ_0과 Θ_1의 사전확률이라고 하자. 주어진 자료 x에 대하여 가설 H_i의 사후확률은 $P(H_i|x) = f(x|\Theta_i)\pi_i / m(x)$ 이며, $m(x)$는 주변확률밀도함수로서 두 사후확률의 비교시에는 상쇄되어지므로 쉽게 얻어지는 경우외에는 계산에서 고려하지 않아도 된다. 다음, 가설의 형태에 따라 사후확률을 구하는 방법을 살펴보면, $\Theta_i = \{\theta_i\}$와 같이 한 포인트인 경우에는

$$P(H_i|x) = f(x|\theta_i)\pi_i / m(x)$$

가 되지만 $\Theta_i = \{\theta \le \theta_0\}$ 처럼 어떤 구간이나 집합을 나타내는 경우에는

$$P(H_i|x) = \int_{\Theta_i} dF^{\pi(\theta|x)}(\theta) = \int_{\theta \le \theta_0} f(x|\theta)\pi_i g_i(\theta)\, d\theta / m(x) \qquad (1)$$

이며, 여기서 $g_i(\theta)$는 Θ_i상에서의 확률밀도함수로서 Θ_i의 사전확률 π_i의 분포형태를 나타내준다. 이러한 원리는 두 개 이상의 여러 개의 가설검정에도 적용되어 각 가설의 사후확률을 구하여 서로 비교하면 되므로 확장성이 좋다고 할 수 있다. 이것이 베이지안 가설검정이라는 용어보다는 흔히 베이지안 모형 선택 또는 베이지안 모형 비교라고 불리우는 이유 중 하나라고 여겨진다.

베이즈 팩터(Bayes factor)는 대립가설 H_1에 대한 귀무가설 H_0의 사후확률의 오즈와 사전확률의 오즈의 비로서, $\alpha_i\ (i=0,1)$ 를 H_i의 사후확률이라고 할 때, H_0의 베이즈 팩터는

$$B=\frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}$$

와 같이 정의된다. 이는 H_0의 H_1에 대한 상대비율이 자료를 관측한 후 몇 배가 되었느냐에 대한 척도로서 $P(H_0|x)=(1+\frac{\pi_1}{B\pi_0})^{-1}=1-P(H_1|x)$ 의 관계가 성립된다. 또한, 베이즈 팩터, B는 H_0의 H_1에 대한 가중우도함수의 비(weighted likelihood ratio)로 해석될 수 있는데, 그 이유는

$$B=\frac{\int_{\Theta_0} f(x|\theta)g_0(\theta)\,d\theta}{\int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta}$$

을 만족하기 때문이다. 단, 여기서 Θ_i, ($i=0,1$)가 한 포인트일 때에는 $g_i(\theta)=1$ 이고, 구간을 나타낼 때에는 위에서 언급했던 바와 같이 구간내의 분포함수 $g_i(\theta)$가 된다. 즉, 베이즈 팩터는 단순히 두 가설의 사후확률의 크기만 비교할 뿐만 아니라 상대적 크기를 비교하므로써 더욱 심도있는 정보를 제시해 준다. 그러나 무정보사전확률 등과 같이 사전확률 π_i가 부적합(improper)할 때에는 베이즈 팩터의 사용에 문제가 발생한다. 즉, 베이즈 팩터 B가 임의의 상수의 비를 포함하게 되는 것이다. 이러한 문제점을 보완한 것으로 표본에서 최소시험자료(minimal training sample)를 추출하여 부적합한(improper) 사전확률함수를 적합한(proper) 확률함수로 전환시키는 방법이 Spiegelhalter and Smith(1982) 등에 의해 제안되었고, 최근 여기서 한걸음 더 발전시켜, 베이즈 팩터들의 평균적인 개념을 도입하여, 추출되는 시험자료의 편의성을 고려한 Intrinsic Bayes Factor(IBF)가 Berger and Pericchi(1996)에 의해 제시되었다. 또한, 계층적 모형이나 합성모형의 경우, 초기단계의 모형설정에서 고전적 가설검정은 결과에 따라 채택 또는 기각을 함으로써 오차를 무시하지만 베이지안 방법에서는 베이즈 팩터를 이용하여 가중치를 줌으로써 유연성을 보인다.

다른 베이지안 분석과 마찬가지로, 베이지안 가설검정에서도 모형의 사전확률(prior probability)의 영향을 받게 되는데, 각 모형의 사전확률에 따라 P-값을 이용한 고전적 가설검정의 결과와 불일치할 수 있다는 사실이 Lindley(1957)에 의해 최초로 발표되었으며 "Lindley's paradox"라고 불리워지고 있다. 이로부터 베이지안과 고전적 가설검정간의 결과의 불일치성에 대한 연구가 활발히 이루어졌다. Pratt(1965), DeGroot(1973), Dempster(1973), Dickey(1977), Zellner and Siow(1980), Hill(1982)은 한쪽검정(one-sided test)에서는 P-값이 무정보 사전확률(vague prior probability)에 대한 H_0의 사후확률과 근사적으로 일치한다는 것을 제시했다. 다음의 예를 보기로 하자.

[예 1]*(Berger(1985))* 확률변수 X의 분포는 정규분포, $X\sim N(\theta,\sigma^2)$를 따르고 분산 σ^2은 알고 있다고 가정하자. 평균 θ에 대한 사전확률은 무정보사전확률분포, 즉, $\pi(\theta)=1$이라고 하면 θ의 사후확률분포는 $\theta\sim N(x,\sigma^2)$이다. 따라서 가설 $H_0:\theta\leq\theta_0$ vs $H_1:\theta>\theta_0$을 검정하고자 할 때, H_0의 사후확률은 $P(H_0|x)=P(\theta\leq\theta_0)=\Phi((\theta_0-x)/\sigma)$로서 구해진다.

한편, 고전적 가설검정에서의 P-값은 $\theta=\theta_0$일 때, 확률변수 X가 관측값 x보다 클 확률이므로,

$P-$ 값 $= P(X \geq x) = 1 - \Phi((x-\theta_0)/\sigma)$이다. 그러므로 정규분포의 대칭성에 의해, $P(H_0|x)$는 P-값과 같다.

Casella and Berger(1987)는 한쪽검정에서 H_0의 사후확률의 infimum을 구하여 P-값과 비교하여 이 둘이 근사적으로 일치하거나 P-값이 더 큼을 보임으로써 베이지안 검정결과와 고전적 검정결과의 일치성을 확인하였다. 여기서 H_0의 사후확률의 infimum은 타당한 사전확률분포들의 군집하에서 구한 것으로 특정 사전확률을 사용했을 경우보다 객관적 해석을 유도할 수 있다. 그러나, 한쪽검정이라고 해서 베이지안과 고전적 가설검정 결과가 항상 일치하는 것은 아니며, 예를 들어 $H_0: \theta = 0$ vs $H_1: \theta > 0$과 같이 귀무가설이 한 개의 값인 경우에는 P-값과 H_0의 사후확률이 상당히 다를 수 있다. 특히, 귀무가설이 한 개의 값이고 양쪽검정(two-sided test)의 경우 불일치가 심화되는데, 즉, 귀무가설이 한 개의 값(point null hypothesis)인 경우에 고전적 검정결과와 베이지안 검정결과의 불일치가 발생한다고 볼 수 있다.

[예 2]*(Berger and Delampady(1987))* 평균이 θ이고 분산이 σ^2(known)인 정규분포로부터 $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$을 검정하기 위해 n 개의 표본을 추출하였다고 하자. 표본평균의 분포는 $\overline{X} \sim N(\theta, \sigma^2/n)$ 이고, $T = \sqrt{n}(\overline{X} - \theta_0)/\sigma$ 라고 할 때, P-값은 관측값 $T = t$에 대하여

$$p-value = 2[1 - \Phi(|t|)]$$

이다. 다음 θ에 대한 사전확률분포를 $\theta \sim N(\theta_0, \sigma^2)$ 라고 하자 (정확하게는 $\{\theta | \theta \neq \theta_0\}$에서의 확률분포로서 위에서 설명한 식 (1)에서 $g_1(\theta)$를 의미한다). 베이즈 팩터와 H_0의 사후확률은

$$B = \sqrt{1+n}\, \exp\{-\frac{1}{2} t^2/(1+n^{-1})\}, \quad P(H_0|x) = [1 + \frac{(1-\pi_0)}{\pi_0} \frac{1}{B}]^{-1}$$

이다. H_0과 H_1에 동일한 사전확률 $\pi_0 = 1/2$ 을 부여하고 P-값과 H_0의 사후확률을 비교해 보면 ([표 1] 참조), $P(H_0|x)$가 P-값보다 5배에서 50배 정도까지 크다는 것을 알 수 있다. 예를 들어, $t = 1.96$의 값에 대하여 P-값은 0.05로서 항상 일정하지만(따라서 유의수준 5%에서 H_0을 기각하게 됨), $P(H_0|x)$의 값은 n의 값에 따라 약 1/3부터 거의 1에 가까운 값까지 갖게된다. 이것은 단지 $t = 1.96$이라는 정보는 베이지안 검정에는 아무 역할도 할 수 없음을 의미할 뿐만 아니라 한쪽검정에서처럼 P-값을 H_0의 사후확률로 여길 수 없음을 제시해 준다.

[표 1] P-값과 H_0의 사후확률

		n						
t	p-value	1	5	10	20	50	100	1000
1.645	0.10	0.42	0.44	0.47	0.56	0.65	0.72	0.89
1.960	0.05	0.35	0.33	0.37	0.42	0.52	0.60	0.80
2.576	0.01	0.21	0.13	0.14	0.16	0.22	0.27	0.53
3.291	0.001	0.086	0.026	0.024	0.026	0.034	0.045	0.124

그러나, 위의 예는 특정 사전확률분포에 대한 경우이므로 이를 일반적으로 규명하기 위해 Berger and Sellke(1987), Berger and Delampady(1987), Delampady and Berger(1990)는 Dickey(1977)가 접근한 방법과 유사하게 여러 개의 사전확률분포들의 군집을 고려하여 H_0의 사후확률의 infimum을 P-값과 비교한 결과 H_0의 사후확률이 P-값보다 큰 경우가 많음을 보였다. 주로 고려되어진 사전확률의 군집들은

$\Gamma_A = \{all\ distributions\}$,

$\Gamma_s = \{all\ distributions\ symmetric\}$,

$\Gamma_{US} = \{all\ distributions\ with\ unimodal\ \text{and}\ symmetric\}$,

$\Gamma_{NOR} = \{all\ Normal\ distributions\}$

이다. [예 2]에서 다룬 사전확률분포 $\theta \sim N(\theta_0, \sigma^2)$를 모든 정규분포군으로 확장했을 때, 즉, $\theta \in \Gamma_{NOR}$에 대하여 H_0의 사후확률의 최소값을 구해보면 (Berger and Sellke(1987)),

$$\underline{P}(H_0|x, \Gamma_{NOR}) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{\exp(t^2/2)}{\sqrt{e}\,t}\right]^{-1}$$

이며, $\pi_0 = 1/2$ 에 대하여 P-값과 비교한 결과는 [표 2]와 같다. P-값과 $\underline{P}(H_0|x, \Gamma_{NOR})$의 차이가 큼을 알 수 있다. 이는 한쪽검정(복합 대 복합 검정)에서 P-값과 H_0의 사후확률(무정보 사전확률에 대하여)이 근사적으로 동일하다는 것과는 사뭇 다른 결과로서 H_0이 한 개의 값인 경우에는, 무정보사전확률에 대해서도 H_0의 사후확률과 P-값을 비슷한 의미로 받아들일 수 없을 뿐 아니라 P-값에 의한 판별의 결과를 믿을 수 없음을 의미한다.

[표 2] P-값과 $\underline{P}(H_0|x, \Gamma_{NOR})$

p-value	t	$\underline{P}(H_0\|x, \Gamma_{NOR})$
0.10	1.645	0.412
0.05	1.960	0.321
0.01	2.576	0.133
0.001	3.291	0.0235

최근 베이지안과 고전적 방법이 일치될 수 있는 분석방법에 대한 연구가 계속되고 있다 (Good(1992)). Berger, Brown and Wolpert(1994)는 단순 대 단순 검정(testing of simple versus simple hypoteses)에서 conditional frequentist 방법이 베이지안 방법과 오류율에서 일치함을 보였으며 Berger, Boukai and Wang(1997)은 이를 단순 대 복합검정(testing of simple versus composit hypotheses)으로의 확장을 시도하였다.

3.4. 베이지안 다변량통계

다변량 해석은 19세기 후반 F. Galton이 상관개념을 처음으로 도입한 후 다양한 분야에서 자료

분석의 기법으로 개발되었다. 교육학과 심리학 분야에서 개발된 C. Spearman의 인자분석, 유전학적 현상을 설명하기 위해 R. A. Fisher가 개발한 여러 다변량 해석기법, 천문학분야에서 개발된 Gauss의 회귀모형과 경제학분야에서 이를 일반화 시켜 개발한 여러 종류의 다변량 회귀모형들, 농학분야에서 개발한 다변량 실험계획모형; 교육학, 심리학 및 사회학분야에서 개발한 잠재구조모형과 다차원 척도모형, 화학, 경제학, 공학분야에서 개발한 조절모형; 재정학분야에서 사용되는 다변량 안정분포(stable distribution)이론, 그리고 인류학, 분류학 및 마케팅분야에서 사용되는 분류 및 판별 분석 등의 기법들이 다변량 분포이론들과 어우러져 응용성이 높은 통계학의 한 분야로 자리 메김하고 있다.

다변량 해석의 추론 및 의사결정기법에는 크게 표본이론에 근거한 전통적인 방법과 베이지안 방법의 두 가지 방법으로 나눌 수 있다. 베이지안 방법이란 서로 상관을 가진 확률변수들의 벡터 또는 행렬의 분포 및 표본분포이론을 바탕으로, 분석하고자 하는 현상에서 수집한 다변량 자료정보와 이에 대한 개인적 또는 주관적인 정보를 베이즈 정리를 통해 결합시켜 그 현상을 추론하는 기법들의 집합을 말한다.

다변량 해석에서 위 두 가지 방법의 우열에 관한 뚜렷한 결론은 없으나, 이들을 이용하여 추론 및 의사결정을 함에 있어 각 추론법이 주관적 그리고 기술적인 어려움을 안고 있다는 것은 이미 밝혀진 사실이다. 예를 들면, 다변량 해석에서 전통적 방법을 사용할 경우 표본의 크기와 제 1종 및 제 2종 오류간의 승산 (trade-off)에 관해 분석자의 주관적인 결정이 필요하며, 제 1종 및 제 2종 오류의 개념을 사용하지 않는 베이지안 방법에서는 주관적인 사전확률분포 설정 문제를 지니고 있다. 한편, 전통적인 방법에서 최적의 분석법이 없는 다변량 회귀분석과 이분산 정규판별분석 등을 베이지안 방법으로 간단하게 분석할 수 있는데 반하여, 복잡한 사후확률분포(zonal polynomials 분포)를 가진 주성분모형이나 정준분석모형 등은 전통적 방법이 간단한 분석법을 제공한다. 그러므로, 이 두 방법이 현재 다변량 해석의 기술적인 측면에서 서로 보완적인 관계를 가진다고 볼 수 있다. 20세기 후반부터 급진전된 베이지안 통계계산법, 로버스트 베이지안 분석법, 그리고 비모수적 방법에 대한 연구는 그 동안 큰 발전이 없었던 다변량 베이지안 통계에 대한 연구에 활로를 열어주었다.

3.5. 비모수적 생존자료 분석

본 절에서는 베이지안 생존분석의 역사를 되 짚어보고 향후 연구분야에 대해서 고찰하였다. 베이지안 생존분석을 논하기에 앞서서, 생존분석이 다른 통계분석과 다른 특징을 살펴보면 다음과 같다. 첫째, 생존분석에 쓰이는 대부분의 모형이 비모수/준모수 모형이다. 그 이유는, 생존분석의 자료는 많은 경우 중도절단 자료를 포함하며, 이런 경우 분포의 꼬리부분에 대한 정보가 거의 없고, 따라서 모수적 모형의 타당성에 대한 검증이 매우 어렵다. 둘째, 생존분석에서는 정보의 양이 자료의 개수에 따라 증가한다는 일반적인 통계학의 원리보다는 주어진 자료 하에서 정보의 양이 시간에 따라 증가한다는 개념이 사용된다. 이러한 새로운 개념은 생존시간 자료를 counting process모형으로 분석할 수 있는 토대를 제공한다. 셋째로는, 생존분석의 자료는 중도절단자료를 포함한다. 즉, 완전한 자료를 관측하지 못하고 부분적으로만 자료를 관측하게 된다. 비슷한 문제로는 결측치, 편의된 자료 (biased sampling) 등이 있다. 따라서, 생존분석을 위한 베이지안 방법의 개발을 위하여는 (i) 비모수적 사전분포의 개발, (ii) 자료가 counting process형태로 주어진 경우의 베이지안 추론 개발과 (iii) 중도절단 자료로부터 사후분포의 계산방법의 개발 등이 필수적이다.

본 논문에서는 이러한 문제들을 베이지안 통계학자들이 어떻게 해결하여 왔으며, 어떠한 문제들이 아직 미지로 남아 있는지에 대해서 고찰한다.

Ferguson (1973)이 비모수 사전분포로써 Dirichlet process를 제안한 후 많은 통계학자에 의하여 중도절단 자료에서의 사후분포의 유도가 시도되었다. Susarla and van Ryzin (1976)에 의하여 우절단 자료에서 생존함수의 베이즈추정량이 계산되었다. Doksum (1974)은 Dirichlet process 사전분포를 일반화한 neutral to right process 사전분포를 계발하였으며, Ferguson and Phadia (1978)에 의하여 우절단 자료에서 neutral to right process 사전분포가 공액류가 됨이 보여졌다. 하지만, neutral to right process 사전분포는 그 범위가 너무 커서 자료분석에 필요한 실제적인 의미는 거의 없다.

베이지안 생존분석은 1980년대의 암흑기를 거쳐서 드디어 Hjort (1990)에 의하여 Diichlet process 사전분포와 우절단자료로부터 사후분포를 유도하는 문제가 완전히 풀리게 된다. Hjort (1990)는 누적분포함수대신 누적위험함수를 이용하였고, 이를 위하여 beta process 사전분포를 개발하였다. 또한, beta process 사전분포가 우절단자료에서 공액사전분포가 됨을 보인 후, Dirichlet proces를 따르는 누적분포함수의 누적위험함수가 beta process가 됨을 보임으로써 Dirichlet process 사전분포를 사용한 경우의 사후분포가 beta process 가 됨을 증명하였다. Hjort (1990)의 결과 중 흥미로운 사실은 누적위험함수 A의 베이즈추정량이

$$E(A(t)|data) = \int_0^t \frac{c(s)}{Y_n(s)+c(s)}\, dA_0(s) + \int_0^t \frac{Y_n(s)}{c(s)+Y_{n(s)}}\, d\hat{A}(s)$$

로 주어진 다는 것이다. 여기서, $A_0(s)$는 사전분포의 평균, $\hat{A}$은 Aalen-Nelson 추정량으로 비모수적 최대우도 추정량, c(s)는 사전분포의 분산의 역수, $Y_n(s)$는 $s-$시간에 생존하는 개체의 수이다. 즉, 베이즈 추정량이 사전분포의 평균과 최대우도추정량과의 선형결합으로 주어진다.

사후분포의 유도와 함께, 사후분포의 효율적인 계산을 위한 연구가 MCMC알고리즘의 개발 후에 많이 수행되었다. Doss (1994)에 의하여 중도절단 자료에서 Dirichlet process 사전분포를 사용한 경우에 사후분포를 계산하는 효율적인 Gibbs sampling방법이 개발되었으며, Damien et al. (1996)에 의하여 beta process를 이용한 MCMC알고리즘이 개발되었다. 최근에는 사후분포의 이론적 성질에 대한 연구가 활발히 진행되고 있는데, Kim and Lee (2001,a)에 의하여 사후분포의 점근적 편의 (posterior inconsistency)의 가능성과 함께, 사후분포가 점근적으로 불편의(posterior consistency)할 충분조건이 제안되었다. 다행히도, 실제문제에 많이 쓰이는 사전분포들(Dirichlet process, beta process, gamma proces)은 모두 점근적으로 불편의한 사후분포를 갖음이 증명되었다

비례위험모형에서의 베이지안 분석방법은 Kalbfleisch (1978) 와 Hjort (1990)에 의하여 시도 되었는데, 사후분포를 구하는 공식이 너무 복잡하여 큰 의미를 갖지는 못하였다. 그러다가, 최근에 와서 Laud et al. (1998)에 의하여 MCMC알고리즘이 개발되었고, Kim and Lee (2001,b)에 의하여 사후분포의 이론적 성질뿐 아니라, left truncation이 포함된 자료로 확장 되었다.

향후 베이지안 생존분석의 과제로는, 무정보 사전분포의 개발, 다변량 생존분석에서의 베이지안 추론방법의 개발, 임의변량을 포함하는 모형에서의 효율적인 사후분포 계산 알고리즘의 개발 등이 있다.

참고문헌

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley & Sons.

[2] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York

[3] Berger, J. O, Boukai, B. and Wang, Y. (1997). Unified Frequentist and Bayesian Testing of a Precise Hypothesis, *Statistical Science*, 12, 133-160.

[4] Berger, J. O, Brown, L. D. and Wolpert, R. L. (1994). A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing, *Annals of Statistics*, 22, 1787-1807.

[5] Berger, J. O. and Delampady, M. (1987). Testing Precise Hypotheses, *Statistical Science*, 3, 317-352.

[6] Berger, J. O. and Pericchi, L.R.(1996) The intrinsic Bayes factor for the model selection and prediction. *Journal of the American Statistical Association*, 91, 109-122.

[7] Berger, J. O. and Sellke, T. (1987). Testing a Point Null Hypothesis: the Irreconcilability of P-values and Evidence, *Journal of American Statistical Association*, 82, 112-122.

[8] Bernardo, J.M.(1979). Reference posterior distributions for Bayesian inference. *Journal of Royal Statistical Society, Ser. B.* 41, 113-147.

[9] Bickel, P.J. and Ghosh, J.K.(1990) "A decomposition for the likelihood ratio statistic and the bartlett correction - A Bayesian argument", *Annals of Statistics*, 18, 3, 1070-1090.

[10] Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* London: Addison-Wesley Publishing Company.

[11] Brooks, S.P. and Roberts, G.O.(1998). Diagnosing convergence of Markov chain Monte Carlo algorithm. *Statistics and Computing* 8, 319-335.

[12] Casella, G. and Berger, R. (1987). Reconciling Bayesian and Frequentist Evidence in the One-sided Testing Problem, *Journal of American Statistical Association*, 82, 106-111.

[13] Chen, M.-H., dey, D.K. and Shao, Q.-M. (1999). A new skewed link model for dicthotomous quantal response data, *Journal of the American Statistical Association*, 94.

[14] Chen, M.-H., dey, D.K. and Sinha, D. (2000). Bayesian analysis of multivariate mortality data with large families. *Applied Statistics*, 49, 129-144.

[15] Chen, M-H., Ibrahim, J.G., Shao, Q.-M. and Weiss, R.E. (1999) "Prior elicitation for model selection and estimation in generalized linear mixed models." *Technical Report* MS-01-99-17, Depts. Mathematical Sciences, Worcester Polytechnic Inst..

[16] Chen, M-H., Ibrahim, J.G. and Sinha, D. (1999) "A new Bayesian model for survival data with a surviving fraction." *Journal of American Statistical Association*, 94, 909-919.

[17] Chen, M-H., Ibrahim, J.G. and Yiannoutsos, C. (1999) "Prior elicitation, variable selection and Bayesian compuatation for logistic regression models." *Journal of Royal*

Statistical Society, Ser. B, 61, 223-242.

[18] Chen, M.-H., Manatunga, A.K. and Williams, C.J.(1998). Heritability estimates from human twin databy incorporating historical prior information. *Biometrics,* 54, 1348-1362.

[19] Cowles, M.K. and Carlin, B.P.(1996). Markov chain Monte Carlo convergence disgnostics: A comparative review. *Journal of American Statistical Association* 91, 883-904.

[20] Damien, P., Laud, P.W. and Smith, A.F.M. (1996). Implementation of Bayesian nonparametric inference based on beta processes. *Scandinavian Journal of Statistics*, 23, 27-36.

[21] Datta, G.S. and Ghosh, J.K.(1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82, 37-45.

[22] Datta, G.S. and Ghosh, M.(1996). On the invariance of noninformative priors, *Annals of Statistics*, 24, 141-159.

[23] DeGroot, M. H. (1973). Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio, *Journal of American Statistical Association,* 68, 966-969

[24] Delampady, M. and Berger, J. O. (1990). Lower Bounds on Bayes Factors for the Multinomial Distribution, with Application to Chi-squared Tests of Fit, *Annals of Statistics,* 18, 1295-1316..

[25] Dempster, A. P. (1973). *The Direct Use of Likelihood for Significance Testing,* In Proceedings of the Conference on Foundational Questions in Statistical Inference, University of Aarhus, Aarhus.

[26] Dickey, J. M. (1977). Is the Tail Area Useful as an Approximate Bayes Factor? *Journal of American Statistical Association,* 72, 138-142.

[27] Doksum, K.A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability,* 2, 183-201

[28] Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics*, 22, 1763-1786.

[29] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics,* 1, 209-230.

[30] Ferguson, T.S. and Phadia, E.G. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics*, 7, 163-186.

[31] Gelfand, A.E. and Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, vol. 85, 398-409.

[32] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721-741.

[33] Geweke, J. (1988), Antithetic acceleration of Monte Carlo integration in Bayesian inference, *Journal of Econometrics*, vol. 38, pp. 73-90.

[34] Ghosh, J.K. and Mukerjee, R.(1991). Characterizatin of priors under which Bayesian and

frequentist Bartlett corrections are equivalent in the multiparameter case. *Journal of Multivariate Analysis*, 38, 385-393.

[35] Ghosh, J.K. and Mukerjee, R.(1992) "Bayesian and frequentist Bartlett corrections for likelihood ratio and conditional likelihood ratio tests", *Journal of Royal Statistical Society, Ser. B*, 54, 867-875.

[36] Good, I. J. (1992). The Bayesian/Non-Bayesian Compromise: a Brief Review, *Journal of American Statistical Association*, 87, 597-606.

[37] Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, vol. 82, pp. 711-732.

[38] Hastings, W.K. (1970). Monte Carlo sampling methods using markov chains and their applications, *Biometrika*, vol. 57, pp. 97-109.

[39] Hill, B. (1982). Comment on "Lindley's Paradox", by G. Shafer, *Journal of American Statistical Association*, 77, 344-347.

[40] Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18, 1259-1294.

[41] Ibrahim, J.B. and Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazards models, *Sankhya* Ser. B, 60, 48-64.

[42] Ibrahim, J.B., Chen, M.-H. and Maceachern, S.N. (2000). Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics*, To appear.

[43] Ibrahim,J.B., Chen, M.-H. and Ryan L.M.(1998). Use of historical controls to adjust for covariate in trend tests for binary data. *Journal of the American Statistical Association*, 93, 1282-1293.

[44] Ibrahim, J.B., Chen, M.-H. and Ryan L.M.(2000). Bayesian variable selection for time sries count data, *Statistical Sinica*.(to appear).

[45] Kalbfleisch, J.D. (1978). Non-parametric Bayesian snalysis of survival time data. *Journal of Royal Statistical Society, Ser. B*, 40, 214-221.

[46] Kloek, T. and Van Dijk, H.K.(1978), Bayesian estimates of equation system parameters: An application of integration by Monte Carlo, *Econometrica*, vol. 46, pp. 1-19

[47] Kim, Y. and Lee, J. (2001,a). On posterior consistency of survival models. *Annals of Statistics*,(to appear).

[48] Kim, Y. and Lee, J. (2001,b). Bayesian Analysis of Proportional Hazard Models. Preprint.

[49] Laud, P.W., Damien, P. and Smith, A.F.M. (1998). Bayesian nonparametric and covariate analysis of failure time data. in *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds: Dey, D., Muller, P. and Sinha, D.)

[50] Lindley, D. V. (1957). A Statistical Paradox, *Biometrika*, 44, 187-192.

[51] Mergersen, K.L., Robert, C.P. and Guiheunneuc-Jouyaux, Ch.(1998). MCMC convergence diagnostics: A Review(with discussion). In *Bayesain Statistics 6*, Oxford University Press, 415-440.

[52] 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.

[53] Naylor, J.C. and Smith, A.F.M. (1982), Applications of a method for efficient computation of posterior distributions, *Applied Statistics*, vol. 31, pp. 214-225.

[54] Nicoaou, A.(1993).Bayesian intervals with good frequentist behaior in the presence of nuisance parameters. *Journal of Royal Statistical Society, Ser. B,* 55, 377-30.

[55] Peers, H.W.(1965). On cofidence sets and Bayesian probability points in the case of several parameters. *Journal of Royal Statistical Society, Ser. B,* 27, 9-16.

[56] Pratt, J. W. (1965). Bayesian Interpretation of Standard Inference Statements (with Discussion), *Journal of Royal Statistical Society (Series B)*, 27, 169-203.

[57] Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods and Inference.* Florida: Robert E. Krieger Publishing Company.

[58] Robert,C.P.(1998). *Discretization and MCMC convergence assessment.* New York: Wiley.

[59] Stein, C.(1985). On the coverage probability of confidence sets based on the prior dsitribution. In: *Sequential methods in Statistics*, Banach Center Publication 16, 485-514, Polish Scientific Publishers, Warsaw.

[60] Stewart, L.T. (1983), Bayesian analysis using Monte Carlo integration-A powerful methodology for handling some difficult problems, *Pratical Bayesian Statistics*, A.P. Davis and A.F.M. Smith eds. Harlow, England: Longman.

[61] Susarla and Van Ryzin (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71, 897-902.

[62] Tierney, L. and Kadane, J.B. (1986), Accurate approximations for posterior moments and marginals, *Journal of the American Statistical Association*, vol. 81, pp. 82-86.

[63] Van Dijk, H.K. and Kloek, T. (1980), Further experience in Bayesian analysis using Monte Carlo integration, *Journal of Econometrics*, vol. 14, pp. 307-328.

[64] Van Dijk, H.K. and Kloek, T. (1983), Monte Carlo analysis of skew distributions: An illustrative econometric example, *Pratical Bayesian Statistics*, A.P. Davis and A.F.M. Smith eds. Harlow, England: Longman.

[65] Tibshirani, R.(1989).Noninformative priors for one parameter of many. *Biometrika*, 76, 604-608.

[66] Van Dijk, H.K. and Kloek, T. (1984), Experiments with some alternatives for simple importance sampling in Monte Carlo integration, *Bayesian Statistics 2,* J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith eds. Amsterdam: North-Holland.

[67] Welch, B. and Peers, H.W.(1963). On formulae for confidence points based on integrals of weighted likelihoods, *Journal of Royal Statistical Society, Ser. B,* 25, 318-329.

[68] Zellner, A. and Siow, A. (1980). *Posterior Odds Ratios for Selected Regression Hypotheses*, In *Bayesian Statistics*, J. M. Bernardo, M. H. DeGroot, D. V. Lindeley and A. F. M. Smith (Eds), University Press, Valencia.