통계적 데이타 분석방법을 위한 컴퓨터의 활용 I : 붓스트랩 이론과 응용+

전 명 식*

요 약

컴퓨터의 발전에 따른 통계방법 중에서 붓스트랩(bootstrap)에 대하여 연구하였다. 특히 추축통계량의 표본분포를 붓스트랩분포로 추정하는데 있어서 계산문제와 이론적인 정당성을 고려하였으며, 모분포의 성격을 나타내는 모수의 붓스트랩 신뢰영역을 몇 가지 사례들에 대해 살펴보았고 사례별로 붓스트랩 방법의 의미를 고찰하였다.

1. 소 개

현재 기존하는 대부분의 통계이론은 계산능력이 미미했던 1930년대에 이르러 대부분 이루어졌으며, 많은 부분이 정규성을 포함하는 강한 가정들을 전제로 수리적인 면에 치중됐었다. 한편, 컴퓨터의 빠른 발전과 사용가능으로 말미암아, 1980년대에 이르러는 많은 통계학자들이 계산처리 속도가 빠른 컴퓨터를 사용한 통계방법들의 중요성과 필연성에 커다란 관심을 가지게 만들었다.

통계방법에서 고속 컴퓨터를 사용할 때의 잇점으로는 다음 두 가지를 생각할 수 있다. 첫째, 기존 통계방법들의 대부분이 증명하기 힘든 기본 가정들을, 예를 들면 정규성등, 전제로 시작하는데 반하여 본 연구에서 다루고자하는 방법들은 그러한 가정들 없이도 가능하다. 둘째는 수리적인 해석이 가능한 통계방법만이 아니라, 컴퓨터의 계산능력을 사용한 몬테칼로(Monte Carlo) 방법 등으로 좋은 근사값을 구함으로써 훨씬 복잡한 통계문제도 효과적으로 처리할 수 있다. 이와 같은 분명한 이점들 대신 우리가 지불해야 하는 것은 고속 컴퓨터를 사용한 복잡한 계산의 반복이다. 그러나 지속적인 컴퓨터의 발전은 이미 이러한 계산의 벽을 넘었다고 할 수 있다. (Efron과 Diaconis, 1983)

본 연구에서는 고속 컴퓨터를 사용하는 통계방법들 중 가장 주목을 받고 있는 붓스트랩(bootstrap) 방법에 대해 연구하여, 그의 중심된 생각과 올바른 사용방법을 대표적인 예들

+ 이 논문은 아산사회복지사업재단의 1988년도 연구비 지원에 의하여 연구되었음.

* 고려대학교 통계학과, 서울 성북구 안암동

을 들어 설명하고 응용가능성을 알아보고자 한다. 여기서는, 적절한 추축통계량(pivotal statistics)을 사용하여, 모분포의 성격을 나타내는 모수(parameter)들의 신뢰영역(confidence region)을 구하는 데에 사용되는 함수적 추정방법으로서의 붓스트랩 기법을 중점적으로 다루기로 한다. 2장에서는 붓스트랩을 소개하고 붓스트랩신뢰영역의 일관성 등에 관해 논하겠으며, 3장에서는 대표적인 사례들과 붓스트랩의 변형을 설명하고, 4장에서는 이중붓스트랩(double bootstrap)에 대해 예를 들어 설명하기로 한다. 마지막으로 5장에서는 그 외의 문제들을 포함한 간단한 맺음을 하고자 한다. 물론 붓스트랩의 모든 영역을 다룰 수는 없었으며 이에 저자 개인의 주관이 개입되어 있음을 미리 밝혀둔다.

2. 붓스트랩 방법

2. 1. 붓스트랩의 소개

어떤 알려져 있지 않은 확률분포 F로부터 크기가 n인 무작위표본 $X = (X_1, X_2, \cdots, X_n)$를 얻었다고 하자. 이 때 관심있는 통계량으로 확률분포 F와 표본 $X = (X_1, X_2, \cdots, X_n)$에 근거한 추축통계량(pivotal statistic) $R_n(X, F)$가 있다고 하자. 여기서 통계량 $R_n(X, F)$의 확률분포를

$$J_n(x, F) = P_F[R_n(X, F) \leqq x] \qquad (2.1)$$

로 나타내기로 하자. 많은 경우에 $R_n(X, F)$의 표본분포 $J_n(x, F)$를 추정함으로써 확률분포 F의 모수 (또는 모수벡터) θ에 관한 통계적 추정이 가능하다. (예를 들면, μ와 $\overline{X}_n$와 S_n를 각각 모분포 F의 평균, 표본평균, 표본분산이라고 할 때, $R_n(X, F) = \sqrt{n}(\overline{X}_n - \mu)/S_n$의 분포를 정규분포로 근사추정하여 모평균 μ에 대한 통계적 추정을 한다.)

이제 표본 $X = (X_1, X_2, \cdots, X_n)$의 경험적 분포를 F_n으로 표기하기로 하자. 붓스트랩방법은 $J_n(x, F)$의 추정값으로, 알려져 있지 않은 F를 경험적 분포 F_n으로 대치한, 함수적 추정치 $J_n(x, F_n)$를 사용하여 $R_n(X, F)$의 표본분포를 추정하는 것으로서 그 사용가능성에 대한 연구가 활발히 진행되어 왔다. (Efron ; 1978, Bickel과 Freedman ; 1981, Beran ; 1984 등). 한편, Efron(1981, 82, 85, 87)에 의한 일련의 논문에서는 붓스트랩방법을 사용하여 편의(bias)와 표준오차(standard error)를 추정하였으며, 그에 근거한 모수의 신뢰구간(confidence interval)에 대한 연구가 주로 다루어졌다. 물론 위 두 방법에 의한 신뢰영역 사이에는 밀접한 관계가 있으며(Beran ; 1987) 많은 경우에 근사적으로 동등(asymptotically equivalent)하다.

이와 같은 붓스트랩방법을 실제로 사용하는 데는 다음의 두 가지 문제가 따른다. 첫째는 붓스트랩 추정값 $J_n(x, F_n)$의 실제 계산이며, 둘째는 이론적인 정당화라고 할 수 있다. 이론

적인 정당화 문제는 다음 2장에서 설명하기로 하고 실제 계산문제부터 다루기로 한다. 계산문제는 붓스트랩추정값 $J_n(x, F_n)$이 너무 복잡해서 폐쇄형(closed form)으로 나타내기가 거의 불가능하기 때문에 일어난다. 따라서 실제 계산에는, Taylor 전개에 의한 선형근사방법 등이 있으나, 다음과 같은 몬테칼로 근사방법이 일반적으로 받아들여지고 있다.

단계 1 : 주어진 n개의 관찰값 $X = (X_1, X_2, \cdots, X_n)$로부터 경험적 분포 F_n을 만든다.

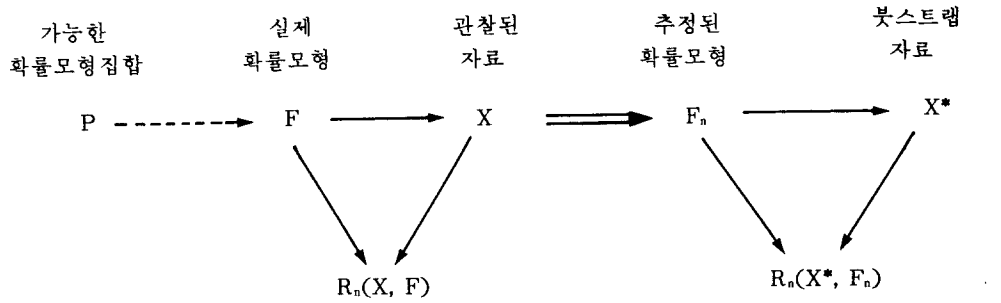
단계 2 : 무작위추출표본(붓스트랩표본이라고 통칭함) $X^* = (X_1^*, X_2^*, \cdots, X_n^*)$을 F_n으로부터 얻고 상응하는 $R_n(X^*, F_n)$을 계산한다.

단계 3 : 단계 2를 독립적으로 반복하여 $R_n(X^*, F_n)$의 값들을 구하여 '붓스트랩분포' $J_n(x, F_n)$의 몬테칼로 근사값을 구한다.

계산능력이 좋은 컴퓨터의 이용은 단계 3의 몬테칼로 근사값을 구하는 데 필수적이라 할 수 있다. (타당한 몬테칼로 근사값을 구하기 위해 필요한 반복회수에는 대한 연구는, 신뢰구간의 경우, Hall(1986)을 참조.)

실지로 붓스트랩은 그리 복잡한 개념이 아니며 잭나이프(jackknnife)등을 포함한 비슷한 방법들은 이미 오래전부터 있었으나 계산능력의 부재로 실용성이 없어서 발전을 못했다고 하겠다. 이에 Efron(1979)의 통찰력은 컴퓨터의 발전에 따라 붓스트랩방법이 사용가능하며 여러가지 잇점이 있다는 사실을 밝혔다. 다른 유사한 표본재추출방법(resampling technique) 등에 대해서는 Efron(1982)에 붓스트랩과의 관계가 비교적 자세히 설명되고 있다. 이와 같은 붓스트랩과정을 도표를 통해 설명하면 다음 〈표 2. 1〉과 같다(Efron과 Tibshirani ; 1986).

〈표 2. 1〉 붓스트랩과정



여기서 근본적인 생각은 원래의 표본 $X = (X_1, X_2, \cdots, X_n)$에 근거한 추축통계량 $R_n(X, F)$와 조건부확률변수인 붓스트랩표본 $X^* = (X_1^*, X_2^*, \cdots, X_n^*)$에 의한 추축통계량 $R_n(X^*, F_n)$의 조건부확률분포가 적절한 조건하에서 유사할 것이라는 점이다. 따라서 관찰된 자료에 근거한 붓스트랩분포로서 표본분포를 추정하는 것이다.

한편 실제 표본분포 $J_n(x, F)$의 붓스트랩 추정값 $J_n(x, F_n)$는 모르는 F를 표본 $X_1, X_2, \cdots,$

X_n에 근거한 경험적 분포 F_n으로 대치한 함수적 추정값이다. 따라서 F를 꼭 F_n으로만 추정하여 대치하여야 할 필요는 없다. 가령 어떤 모수모형(parametric model)을 고려할 경우, 예를 들면 Poisson(θ) 모형, 표본분포 $J_n(x, \theta)$의 붓스트랩 추정치는 F를 비모수적으로 추정한 F_n보다는 θ를 그의 좋은 추정값, 예를 들면 최우도추정값, θ_n으로 대치한 $J_n(x, \theta_n)$이 더 타당해 보인다. 이와 같은 방법을 "모수적 붓스트랩(parametric bootstrap)"이라 부른다 (Efron ; 1979, Woodroofe와 Jhun ; 1989). 이외에도 기저분포 F를 확률밀도함수의커넬(kernel) 추정값 f_n을 사용하여 $\tilde{F}_n(x) = \int_{-\infty}^{x} f_n(t)dt$로 추정하여 사용하기도 하며 이를 평활 붓스트랩 (smoothed bootstrap)이라 칭한다(Efron ; 1979, Silverman ; 1986, Faraway와 Jhun ; 1989). 이와 같이 붓스트랩 방법의 여러가지 변화된 형태들을 고려할 수 있으며 그 중 어떤 것을 사용하느냐의 결정에는 이론적인 정당성과 동시에 계산 가능성이 관건이 된다고 하겠다.

2. 2. 이론적인 정당성

다음으로 이론적인 정당성은 '일치성(consistency)', '최적성(optimality)' 등이 있으나, 여기서는 일치성에 관한 경우를 주로 다루기로 한다. 다음 정리는 붓스트랩방법의 일치성이 성립하는 일반적인 가정을 제공하고 있다.

정리(Beran, 1984)

가능한 확률분포의 집합 F에 속하는 모든 F에 대하여 실제 확률분포 $J_n(x, F)$와 붓스트랩분포 $J_n(x, F_n)$가 같은 비퇴화분포(non-degenerate distribution) $J(x, F)$로 약수렴한다고 가정하자. 그러면 붓스트랩방법은 일치성이 있다고 말하며, 붓스트랩분포에 근거한 모수의 $(1-\alpha)\times 100\%$ 신뢰영역은 표본의 크기 n이 커짐에 따라 올바른 유의수준에 수렴한다. ■

이제 붓스트랩방법이 일치성이 있을 때, 붓스트랩분포 $J_n(x, F_n)$의 $\alpha\times 100\%$ 백분위수 $c_n(\alpha, F_n)$에 대해서는 다음 (2. 2)가 성립하며

$$\lim_{n\to\infty} P_F[R_n(X, F) \leqq c_n(\alpha, F_n)] = 1 - \alpha \qquad (2.\ 2)$$

위 정리에 근거한 벡터일 수도 있는 모수 θ에 대한 $(1-\alpha)$ 수준의 붓스트랩 신뢰영역은

$$\{\theta \mid R_n(X, F) \leqq c_n(\alpha, F_n)\} \qquad (2.\ 3)$$

으로 나타내진다.

위 정리의 중요성은 많은 통계문제에 응용될 수 있다는 점이다. 물론 붓스트랩분포 $J_n(x, F_n)$에 사용되는 경험적분포 F_n은 주어진 문제에 따라 모분포 F의 적절한 함수적 추정값으로 대치되어야 하며 다음 3장에서 위 정리의 응용을 예를 들어가며 설명하기로 한다.

3. 붓스트랩 방법의 사용사례

앞 절의 정리를 실제 통계문제에 응용하는데는 문제의 성격에 따른 적절한 붓스트랩 방법

의 적용이 결정되어야 하며 이에 따라 때로는 효용성을 증가시킬 수도 있으며 반면 잘못 사용되었을 경우는 틀린 결과를 가져오기도 한다. 여기서는 붓스트랩 방법의 적절한 사용을 사례에 따라 설명하기로 한다. 또한 사례마다 붓스트랩 방법의 통계적 의미를 설명하기로 한다.

3. 1. 모평균의 신뢰구간

평균이 μ이고 분산이 $\sigma^2 < \infty$인 모분포 F로부터 무작위추출표본 $X_1, X_2, \cdots, X_n$을 얻었다고 하자. 이 때 우리의 관심이 모평균 μ일 때 그에 대한 신뢰구간은 통계적 추론에 유용하게 사용된다. 사용되는 추축통계량으로

$$R_n(X, F) = \sqrt{n}(\overline{X}_n - \mu)/S_n \qquad (3.1)$$

단 $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$과 $S_n^2 = 1/(n-1)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, 을 고려할 수 있다. 여기서 전통적인 통계방법중의 하나는 중심극한정리를 사용하여 추축통계량의 분포를 표준정규분포로서 근사추정하는 것이다. 이에 따른 모평균 μ에 대한 $(1-\alpha)\times 100\%$ 신뢰구간은, z_α를 표준정규분포의 $(1-\alpha)\times 100\%$ 백분위수라 할 때,

$$\overline{X}_n - z_{\alpha/2}\cdot S_n/\sqrt{n} \leqq \mu \leqq \overline{X}_n + z_{\alpha/2}\cdot S_n/\sqrt{n} \qquad (3.2)$$

로 주어진다.

이제 〈자료 3. 1〉을 가지고 붓스트랩 방법을 통한 모평균 μ의 90% 신뢰구간을 구해보고 전통적인 통계방법에 의한 (3. 2)와 비교하여 본다.

〈자료 3. 1〉 Darwin의 Zea mays 자료 (단위 1/8 inches)

짝	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
차이	49	−67	8	16	6	23	28	41	14	29	56	24	75	61	−48

(이 자료는 15쌍의 mays의 키에 대해 cross-fertilized plants와 self-fertilized plants의 차이를 관찰한 것이며 차이의 모평균에 대한 통계적 추정을 하고자 한다.)

여기서 붓스트랩분포

$$J_n(x, F_n) = P_{F_n}[R_n(X^*, F_n) \leqq x] \text{ (여기서 } P_{F_n}\text{는 조건부확률)} \qquad (3.3)$$

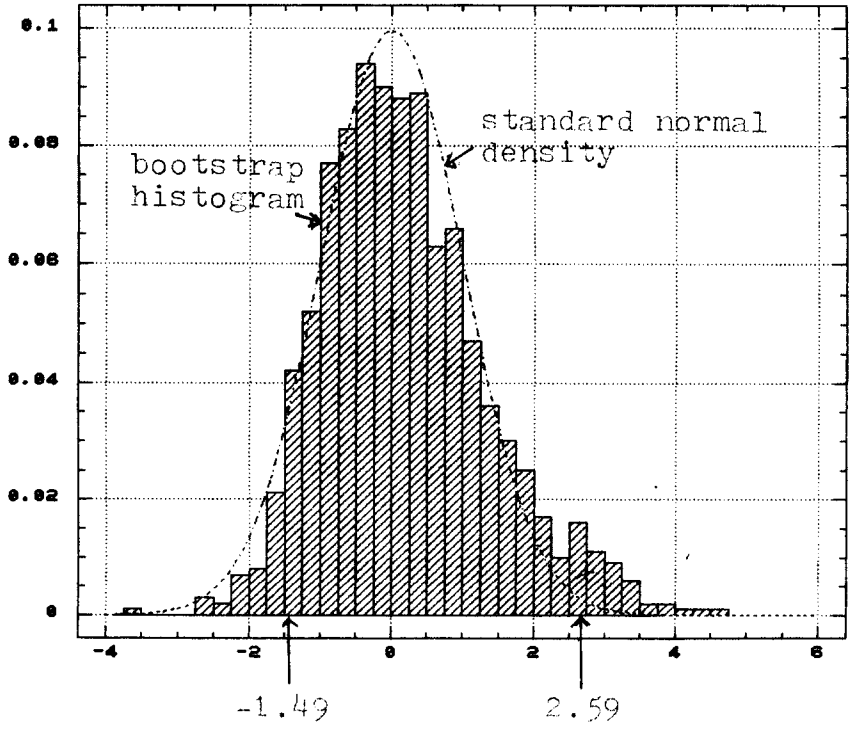
는 앞 장에서 설명된 세 단계의 몬테칼로방법에 의해 다음과 같이 근사추정된다.

우선 붓스트랩표본 $X^* = (X_1^*, X_2^*, \cdots, X_n^*)$을 경험적 분포 F_n으로부터 얻고 상응하는 $R_n(X^*, F_n) = \sqrt{n}(\overline{X}_n^* - \overline{X}_n)/S_n^*$을 계산한다. (단 $\overline{X}_n^* = (1/n)\sum_{i=1}^{n} X_i^*$, $S_n^{*2} = 1/(n-1)\sum_{i=1}^{n}(X_i^* - \overline{X}_n^*)^2$). 조건부확률변수 $R_n(X^*, F_n)$는 B=1000회의 독립반복시행에 의한 몬테칼로 근사방법을 통해 구했으며, 그에 따른 붓스트랩분포의 히스토그램은 〈그림 3. 1〉과 같다. 한편 추축

통계량의 붓스트랩분포 (3. 3)도 표준정규분포로 수렴한다는 점(Bickel과 Freedman ; 1981, Singh ; 1981 등)에 근거하여, 앞 2. 2절의 정리를 사용하여 모평균 μ에 대한 90% 붓스트랩 신뢰구간은 다음과 같이 구해진다. 식(2. 2)에 의한 붓스트랩분포의 5% 백분위값과 95% 백분위 값을 각각 −1.49, 2.59이며 식(2. 3)을 사용하면 $\{\mu \mid -1.49 \leqq R_n(X, F) = \sqrt{n}(X_n - \mu) \leqq 2.59\}$로부터 모평균 μ에 대한 90% 붓스트랩 신뢰구간은 [−4.28, 35.55]로 주어진다. 이와 같이 구해진 붓스트랩 신뢰구간은 표본평균 $\overline{X}_n = 21$에 대하여 대칭이 아니나, 통계량 $|R_n(X, F)|$를 사용하였을 경우 그의 붓스트랩분포의 90% 백분위수(=1.93)를 이용하여 표본평균에 대칭인 신뢰구간 (2.16, 39.94)를 구할 수 있다. 반면, 중심극한정리에 근거한 90%신뢰구간 (3. 2)는 [4.94, 37.06]이다. 위의 신뢰구간 들중에 어느 것이 더 좋은가는 기저분포를 모르기 때문에 말할 수 없다. 그러나 다른 몇 가지 기저분포에 대하여 여러번의 독립반복시행을 통한 모평균 μ에 대한 붓스트랩신뢰구간의 포함확률을 추정한 모의실험결과 (Ducharme과 Jhun : 1986 등)들은 붓스트랩방법의 가능성과 우월성을 보이고 있다.

실지 이론적으로도 기저분포 F가 연속일 때 중심극한정리에 의한 정규분포와 실지표본분포와의 차이가 $O(n^{-1/2})$인 반면에 붓스트랩분포와 실지표본분포와의 차이는 $O(n^{-1})$밖에 되지 않음을 Edgeworth연장 등을 통해 확인할 수 있다 (즉, $\sup |J_n(x, F) - J_n(x, F_n)| = O(n^{-1})$, 참조 : Bickel과 Freedman ; 1981, Singh ; 1981, Ducharme과 Jhun ; 1986). 이는 붓스트랩

〈그림 3. 1〉 Zea mays 자료에 대한 붓스트랩분포 (3. 3)의 히스토그램



방법의 실질적인 이론적 우월성이 확인되는 부분이다. 한편 기저분포가 격자점(lattice)에만 확률을 가지고 있을 때는, 예를 들면 이항분포 또는 포아송분포, 붓스트랩분포와 실지표본분포의 차이가 $O(n^{-1/2})$이나 평균신뢰확률(average coverage probability)을 고려하면 실용적인 면에 있어서 $o(n^{-1/2})$까지 그 차이를 줄일 수 있다는 이론이 가능하다 (Woodroofe와 Jhun ; 1989). 위의 결과중 '일치성'은 이표본의 경우로도 자연스럽게 연장되며 또한 다변량의 경우에도 평균벡터에 대한 붓스트랩 신뢰영역의 일치성(Beran ; 1984)이 성립된다.

3. 2. 회귀모형

주어진 자료 $Y' = (Y_1, Y_2, \cdots, Y_n)$와 크기가 $(n \times p)$인 계획행렬(design matrix)X에 대하여 다음과 같은 선형회귀모형을 고려해보자.

$$Y = X\beta + \epsilon \tag{3. 4}$$

단, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$는 회귀모수이고 $\epsilon' = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$은 각 성분 ϵ_i 들이 서로 독립이며 평균이 0이고 분산이 $\sigma^2 < \infty$인 공통분포 F를 가진 확률벡터이다.
이 경우 β의 최소제곱추정값은

$$\beta_n = (X'X)^{-1}X'Y \tag{3. 5}$$

으로 주어지며 이에 근거한 β의 신뢰영역 ; 은, 통계량 $R_n(X, F) = (X'X)^{1/2}(\beta_n - \beta)/\sigma$의 분포는 표본의 크기가 커짐에 따라 평균벡터가 0이고 공분산행렬은 $(p \times p)$ 단위행렬 I_p인 다변량정규분포로 수렴한다는, 극한분포이론에 근거를 둘 수 있다. 물론 모수 β의 신뢰영역을 구하기 위해서는 장애모수 σ를 추정해야 하는 문제도 남아 있으며, 유한(finite) 표본의 경우, 분포이론에도 어느 정도의 변화가 따른다.

이 경우에 통계량 $\sqrt{n}(\beta_n - \beta)$에 근거한 붓스트랩방법의 사용은 다음과 같다(Freedman ; 1981). 우선 잔차의 관찰값 $\epsilon = Y - X\beta_n = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$ 들을 구한 다음 그들의 평균이 0이 되도록 $\tilde{\epsilon}_i = \epsilon_i - (1/n)\sum_{i=1}^{n}\epsilon_i$으로 중심화(centering)한다. 이제 중심화된 잔차 $\tilde{\epsilon}_i (i = 1, 2 \cdots, n)$에 확률질량 1/n씩 놓은 분포를 F_n이라 할 때, 조건부 확률변수인 붓스트랩잔차 ϵ_i^*들은 $(i = 1, 2, \cdots, n)$ 서로 독립이며 공통분포 F_n를 갖는다. 즉, 붓스트랩표본은 중심화된 잔차의 관찰값들로부터 재추출(resampling)하여 얻는다. 다음으로 β_n와 ϵ^*를 사용하여 $Y^*(= X\beta_n + \epsilon^*)$를 생성하고 $\beta_n^* = (X'X)^{-1}X'Y^*$를 구한다.

이 경우 붓스트랩확률변수 $\sqrt{n}(\beta_n^* - \beta_n)$의 조건부분포는 $\sqrt{n}(\beta_n - \beta)$의 표본분포와 같은 극한분포를 가지며 이는 β에 대한 붓스트랩신뢰구간의 일치성을 뒷받침하고 있다. 이 예에서 주목할 점은 잔차의 중심화가 필수적이라는 점이다. 물론 원래 가정에서 확률변수 $\epsilon_i (i = 1, \cdots, n)$의 기대치가 0이라고 가정한 사실과는 유관하다. 만일 중심화를 하지 않을 경우 그 차이는 0으로 수렴하지 않고 일종의 확률변수로 수렴하기 때문에 붓스트랩은 일치성을 잃게

되고 이론적인 정당성이 없어진다. 그러나 계획행렬 X의 행(column)들 중 어느 하나가 상수이면 잔차의 관찰값들은 평균이 0이 되므로 중심화가 필요없다. 또 한가지 주목할 점은 붓스트랩방법을 사용하는 경우 장애모수 σ의 추정을 피할 수 있다는 점이다. 이는 붓스트랩 방법의 중요한 잇점 중의 하나로 장애모수의 추정이 문제가 되는 많은 통계문제에 잠재적인 해결책이 되고 있다 (Beran ; 1984, Jhun ; 1985). 물론 통계량$(\beta_n - \beta)/\sigma(\beta_n)$사용도 마찬가지 과정을 통해 가능하며 기저분포에 덜 의존한다는 장점도 있다.

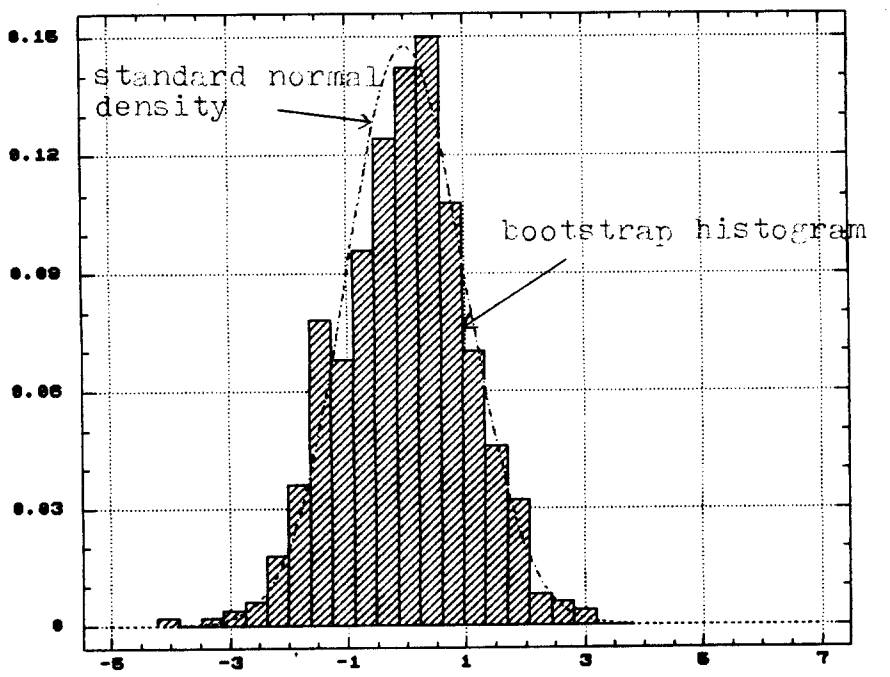
이제 〈자료 3. 2〉를 가지고 단순선형회귀모형 $Y = \beta_0 + \beta_1 X + \epsilon$을 고려하여 붓스트랩 방법을 통한 선형회귀모수 β_1의 90% 신뢰구간을 구해보고 전통적인 (t-분포)에 의한 그것과 비교하여 본다.

〈자료 3. 2〉 Monthly sales & monthly expenses for a certain firm.

Month	Sales(thousands)	Sales Expenses(thousands)
78/4	$ 181.7	$ 25.4
78/5	$ 179.5	$ 22.4
78/6	$ 157.0	$ 20.6
78/7	$ 197.0	$ 21.8
78/8	$ 239.4	$ 32.4
78/9	$ 217.8	$ 24.4
78/10	$ 227.1	$ 29.3
78/11	$ 233.4	$ 27.9
78/12	$ 242.0	$ 27.8
79/1	$ 251.9	$ 34.2
79/2	$ 190.0	$ 29.2
79/3	$ 295.8	$ 30.0

〈그림 3. 2〉에는 통계량 $(\beta_n - \beta)/\sigma(\beta_n)$의 붓스트랩 분포를 B = 500회의 독립 반복시행에 의해 얻어진 히스토그램으로 표현하였으며 회귀모수 β_1의 90% 신뢰구간은 (3. 23, 10. 27)로 구해졌다. 같은 자료에 대해 t-분포이론을 사용한 β_1의 90% 신뢰구간은 (3. 12, 10. 09)로 구해져 두 방법 사이에 큰 차이를 보이고 있지 않다.

〈그림 3. 2〉 자료 3. 2에 대한 $(\beta_n - \beta)/\sigma(\beta_n)$의 붓스트랩분포



3. 3 공분산행렬의 함수

p-차원 분포 F로부터 얻은 다변량관찰값 $X_1, X_2, \cdots, X_n$에 근거하여 모공분산행렬

$$\Sigma_F = \mathrm{Cov}(X_1) = E_F[(X_1 - \mu_F)(X_1 - \mu_F)'] \qquad (3.\ 6)$$

의 함수(예를 들면 상관계수, 고유값 등)에 대한 신뢰영역을 고려하기로 한다. 그러나 이러한 문제에 접근함에 있어서 많은 부분이 관찰값들의 모분포에 대해 다변량정규성을 가정하고 있으며 정규성하에서 조차도 해가 복잡한 경우가 많은 실정이다. Beran과 Srivastava(1985)는 다변량의 경우에 있어서 붓스트랩방법의 사용 가능성을 보여주고 있다.

이제 모공분산행렬의 추정값으로 $S_n = 1/(n-1)\sum_{i=1}^{n}(X_i - \overline{X_n})(X_i - \overline{X_n})'$을 사용하기로 한다. 적절한 조건하에서 $\sqrt{n}(S_n - \Sigma_F)$의 붓스트랩분포는 실제 표본분포와 같은 극한분포를 가지게 되는 일관성이 있다. 이제 모공분산행렬에 대해 미분가능한 함수 $g(\Sigma_F)$를 고려하면 'δ-방법'에 의해

$$R_n(X, F) = \sqrt{n}[g(S_n) - g(\Sigma_F)] \qquad (3.\ 7)$$

의 표본분포도 붓스트랩분포와 같은 극한분포를 가지며 따라서 붓스트랩 신뢰영역은 일치성을 가지게 된다.

가능한 $g(\Sigma_F)$의 예로써 모공분산행렬 Σ_F의 고유값 $\lambda_1, \lambda_2, \cdots, \lambda_p$들을 고려해 보기로 한다. 표본고유값들을 $l_1, l_2, \cdots, l_p$라고 할 때 모집단 고유값 $\lambda_1, \lambda_2, \cdots, \lambda_p$들의 동시신뢰영역을 구하는 데 적절한 추축통계량으로 $\max_{1\leq i\leq p} | \log(l_i) - \log(\lambda_i) |$을 사용할 수 있다.(log변환을 취한 것은 기저분포가 다변량정규분포인 경우 추축통계량의 극한분포가 다변량정규분포가 되기 때문이다.) 이때 단순(simple) 고유값은 공분산행렬의 미분가능한 함수이므로 붓스트랩 신뢰영역은 일치성이 있다. 반면 다중(multiple)한 경우는 미분가능하지 않으므로 붓스트랩 신뢰영역은 일치성이 없다.

다음과 같은 모의실험은 붓스트랩방법의 일치성을 보여주고 있다. 모의실험은 평균벡터가 (0, 0)이고 분산이 1이며 상관계수가 ρ인 이변량 정규분포로부터 생성한 크기가 $n = 20$인 독립확률변수에 근거하였다. 이제 모공분산행렬 Σ_F의 두 고유값 $\lambda_1 = 1 - \rho$, $\lambda_2 = 1 + \rho$의 90% 붓스트랩 신뢰영역을 추축통계량 $\max_{i=1, 2} | \log(l_i) - \log(\lambda_i) |$를 사용하여 $\rho = 0.0, 0.1, 0.5, 0.9$에 대하여 구하고 포함확률(coverage probability)을 1000번의 반복독립시행을 통해 추정한 결과가 아래 〈표 3. 1〉에 주어져 있다. (붓스트랩분포는 B=300개의 붓스트랩표본에 근거하여 근사추정되었다.)

〈표 3. 1〉 붓스트랩 신뢰영역에 대한 포함확률의 추정값

ρ	0.0	0.1	0.5	0.9
명목신뢰확률(=90%)	80.4%	86.6%	90.8%	90.1%

(주의 : ρ=0.0인 경우는 고유값이 다중이므로 공분산행렬의 미분가능한 함수가 아니며 붓스트랩신뢰영역의 정당성이 없는 사실이 모의실험에서도 증명되고 있다.)

위의 예에서 붓스트랩은 다변량자료의 경우에도, 새삼 강조하지만, 정규성이 필요없는 훨씬 약한 가정하에서도 성립되는 이론을 제공한다는 점에 많은 의미가 있다고 볼 수 있다. 이러한 결과는 이표본의 경우로도 연장가능하며 연관된 몇몇에 의해 연구가 진행중이다. 이외에도 공분산행렬 Σ_F의 미분가능한 함수로는 상관계수, 선형모형의 계수 등이 있다.

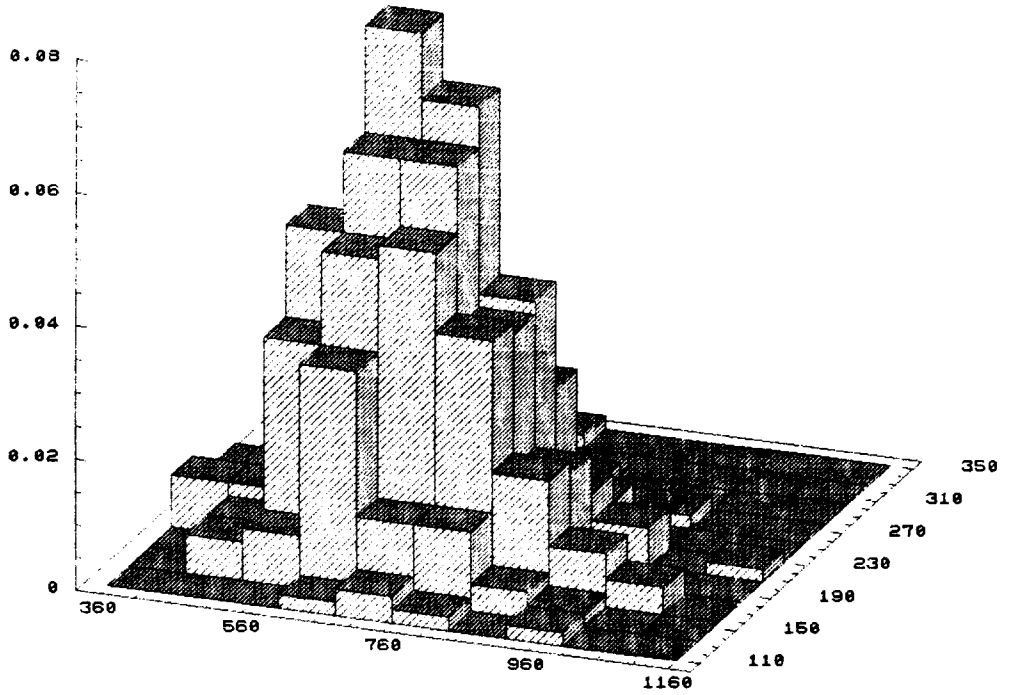
이제 〈자료 3. 3〉에 주어진 다변량확률변수의 표본고유값과 표본고유벡터에 대해 붓스트랩방법(B=500)을 적용하여 그 결과를 살펴보기로 한다.

〈자료 3. 3 ; 출처 Mardia et al(p3−4)〉

다섯 과목에 대한 시험성적(O와 C는 각각 Open-book과 Closed-book 시험임)

Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)	Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
77	82	67	67	81	46	61	46	38	41
63	78	80	70	81	40	57	51	52	31
75	73	71	66	81	49	49	45	48	39
55	72	63	70	68	22	58	53	56	41
63	63	65	70	63	35	60	47	54	33
53	61	72	64	73	48	56	49	42	32
51	67	65	65	68	31	57	50	54	34
59	70	68	62	56	17	53	57	43	51
62	60	58	62	70	49	57	47	39	26
64	72	60	62	45	59	50	47	15	46
52	64	60	63	54	37	56	49	28	45
55	67	59	62	44	40	43	48	21	61
50	50	64	55	63	35	35	41	51	50
65	63	58	56	37	38	44	54	47	24
31	55	60	57	73	43	43	38	34	49
60	64	56	54	40	39	46	46	32	43
44	69	53	53	53	62	44	36	22	42
42	69	61	55	45	48	38	41	44	33
62	46	61	57	45	34	42	50	47	29
31	49	62	63	62	18	51	40	56	30
44	61	52	62	46	35	36	46	48	29
49	41	61	49	64	59	53	37	22	19
12	58	61	63	67	41	41	43	30	33
49	53	49	62	47	31	52	37	27	40
54	49	56	47	53	17	51	52	35	31
54	53	46	59	44	34	30	50	47	36
44	56	55	61	36	46	40	47	29	17
18	44	50	57	81	10	46	36	47	39
46	52	65	50	35	46	37	45	15	30
32	45	49	57	64	30	34	43	46	18
30	69	50	52	45	13	51	50	25	31
46	49	53	59	37	49	50	38	23	9
40	27	54	61	61	18	32	31	45	40
31	42	48	54	68	8	42	48	26	40
36	59	51	45	51	23	38	36	48	15
56	40	56	54	35	30	24	43	33	25
46	56	57	49	32	3	9	51	47	40
45	42	55	56	40	7	51	43	17	22
42	60	54	49	33	15	40	43	23	18
40	63	53	54	25	15	38	39	28	17
23	55	59	53	44	5	30	44	36	18
48	48	49	51	37	12	30	32	35	21
41	63	49	46	34	5	26	15	20	20
46	52	53	41	40	0	40	21	9	14

〈그림 3. 3〉 첫 두 표본고유값의 붓스트랩 결합분포

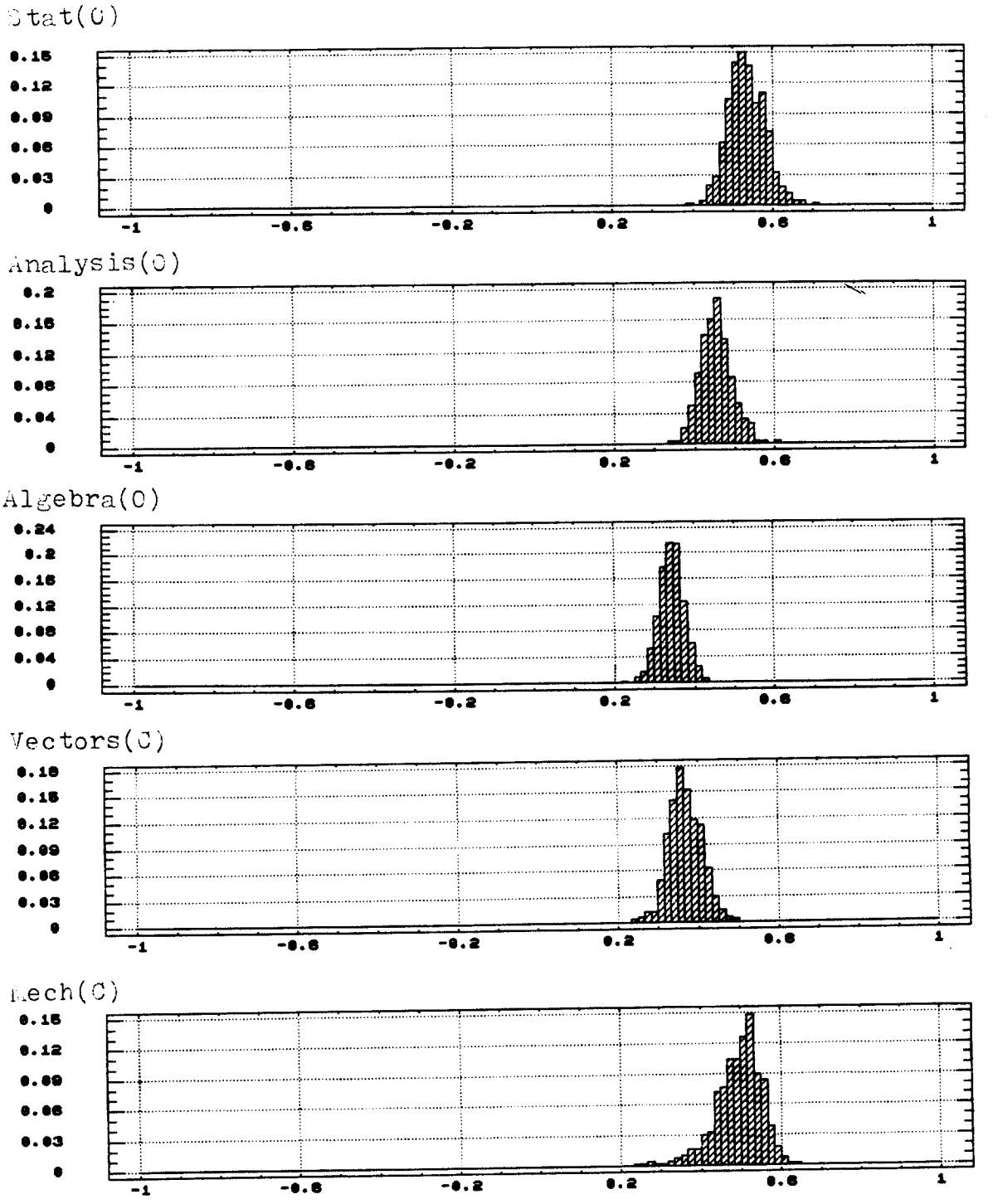


〈자료 3. 3〉에 주어진 관찰값들의 표본고유값은(679.2, 199.8, 102.6, 83.76, 31.8)으로 그에 상응하는 표본고유벡터는 아래와 같다.

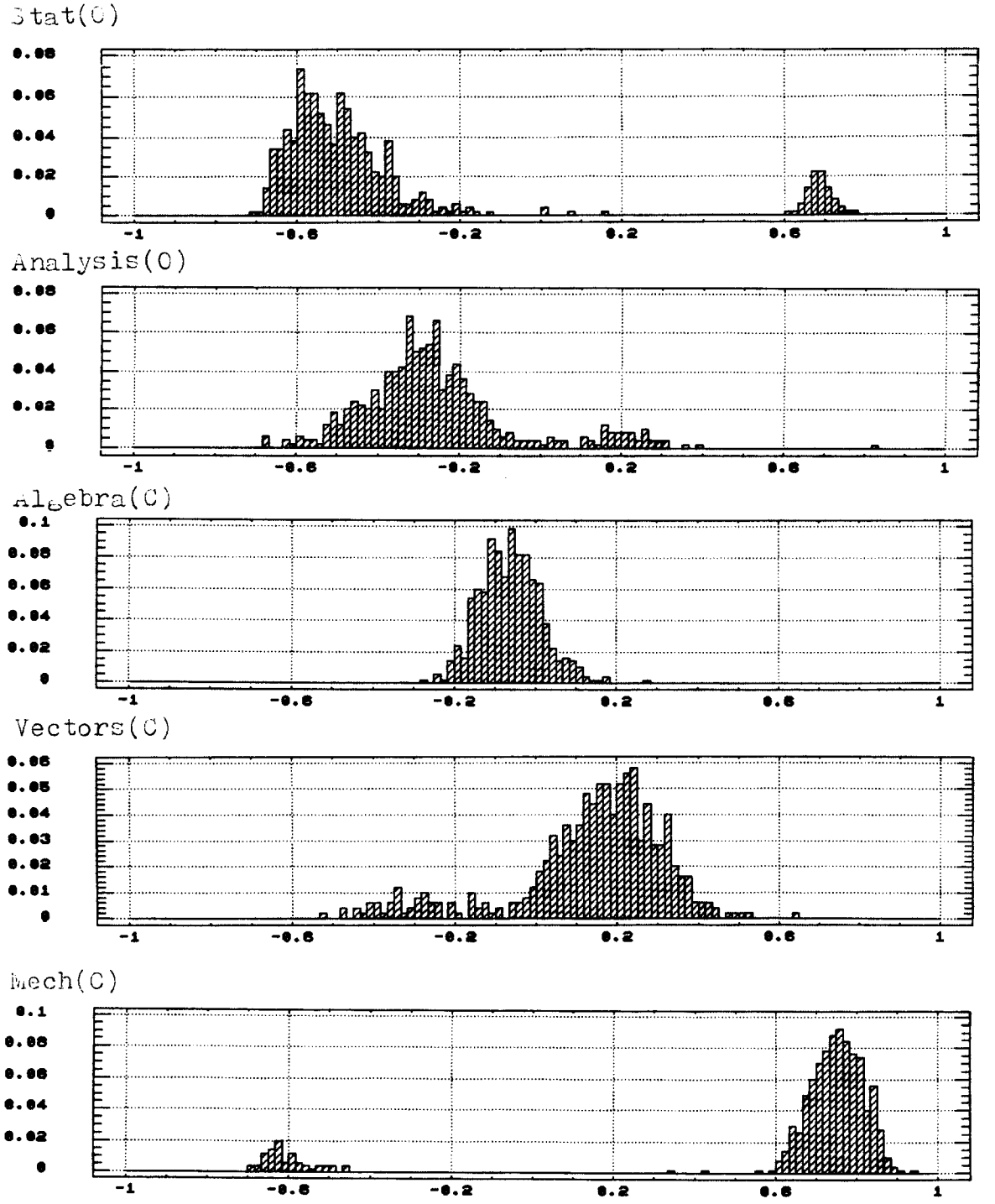
변수	Mech(C)	Vectors(C)	Algebra(C)	Analysis(O)	Stat(O)
제1고유벡터	0.51	0.37	0.35	0.45	0.53
제2고유벡터	0.75	0.21	−0.08	−0.30	−0.55
제3고유벡터	−0.30	0.42	0.15	0.60	−0.60
제4고유벡터	0.30	−0.78	−0.00	0.52	−0.18
제5고유벡터	0.08	0.19	−0.92	0.29	0.15

이러한 고유값과 고유벡터는 주성분분석 등에서 필수로 하는 통계량이나 그에 대한 분포이론은 정규성하에서 조차도 매우 복잡한 실정이다. 여기서 주어진 자료에 근거한 붓스트랩 방법에 의한 첫 두 고유값과 고유벡터의 붓스트랩 표본분포는 편의상 〈그림 3. 3〉, 〈그림 3. 4〉와 〈그림 3. 5〉에 각각 요약되어 있다. 〈그림 3. 4〉와 〈그림 3. 5〉는 맨 아래 그림으로부터 첫 변수(Mech)에 대한 가중계수의 분포가 시작된다. 여기서 분명히 알 수 있는 결과는 첫번째 표본고유벡터가 두번째의 그것보다는 훨씬 안정되어 있으며 따라서 통계적 정도(statistical accuracy)가 좋다는 사실이다. 또한 표본고유값의 분포를 구함으로써 그것의 편의나 분산이 부산물로 얻어진다.

〈그림 3. 4〉 제1고유벡터 가중계수의 붓스트랩분포



〈그림 3. 5〉 제2고유벡터 가중계수의 붓스트랩 분포



3. 4 확률밀도함수

주어진 관찰치 X_1, X_2, …, X_n로부터 알려져 있지 않은 기저확률밀도함수 f(x)를 다음과 같이 적절한 평활커넬(smooth kernel) K를 사용하여

$$f_n(x ; h) = 1/(nh) \sum_{i=1}^{n} K\{(x - X_i)/h)\} \qquad (3.\ 8)$$

으로 추정하는 방법은 컴퓨터의 발전과 더불어 실용성이 커짐에 따라 많은 흥미를 끌고 있다. 이제 확률밀도함수 f(x)가 smooth하다는 가정이 있을 때, f(x)의 신뢰영역(confidence band)을 구하는 문제를 고려해 보기로 한다. Bickel과 Rosenblatt(1973)은

$$R_n(X, f) = \sup_{0 \le x \le 1} | (nh)^{1/2}\{f_n(x ; h) - Ef_n(x ; h)\} | \qquad (3.\ 9)$$

라는 추축통계량의 극한분포를 구하고 나아가 편의(bias), $Ef_n(x ; h) - f(x)$를 0으로 수렴시킴으로써 수리적으로 매우 복잡하나마 f(x)의 신뢰영역을 얻어냈다. Lo(1987)와 Jhun(1988)은, 서로 독립적으로 경험적분포로부터 붓스트랩표본을 재추출하여 얻은 붓스트랩신뢰영역의 일치성을 보였으며 그에 따른 잇점에 대해 설명하고 있다.

다음의 모의실험결과 〈표 3. 2〉는 붓스트랩방법의 가능성을 보여주고 있다. 모의실험은 표본의 크기가 n = 100인 서로 독립인 N(0, 1)의 확률변수에 근거하여 구간 [−2, 2] 사이의 정규확률밀도함수의 90, 95, 99%의 붓스트랩신뢰영역(bootstrap confidence band)을 구하고 그의 포함확률을 B=100번의 반복독립시행을 통해 추정하였다. 이 때의 추축통계량은

$$\sup\{ | f_n(x ; h) - f(x) | /f(x)^{1/2} : -2 \leqq x \leqq 2 \} \qquad (3.\ 10)$$

이며 커넬은 정규확률밀도함수를 사용하였으며 평활계수 h는 근사적인 수리조건을 만족시키는 5가지를 시도하였다. (붓스트랩분포는 B=200개의 붓스트랩표본에 의해 근사추정되었다.)

〈표 3. 2〉 붓스트랩 신뢰영역(confidence band) 포함확률의 추정값

명목포함확률 / 평활계수	90%	95%	99%
$h = n^{-.25}$	90%	95%	99%
$h = n^{-.30}$	91%	94%	99%
$h = n^{-.35}$	90%	94%	98%
$h = n^{-.40}$	90%	94%	98%
$h = n^{-.45}$	89%	94%	98%

그러나 위의 방법들은 $Ef_n(x ; h) - f(x)$를 0으로 수렴시키도록 평활계수 h를 선택해야 하며 이는 최적수렴율을 갖지 못한다는 단점이 있다. 실지로 h의 최적수렴율은 $n^{-.20}$이다. Faraway

와 Jhun(1989)은 최적수렴율을 사용한 $f_n(x ; h)$에 대해서도 평활붓스트랩을 사용하여 신뢰영역을 구하는 방법을 제시하였으며 평활붓스트랩은 적분평균제곱오차(integrated mean squared error)를 올바르게 추정함으로써 평활계수 h의 선택에도 사용될 수 있다는 것이 보여졌다. (실제로 붓스트랩표본을 평상적인 경험적분포로부터 재추출하는 경우 적분평균제곱오차 중에서 적분분산오차만 옳게 추정하며 적분편의제곱오차는 올바로 추정하지 못한다. 여기서 적분평균제곱오차 = 적분분산오차+적분편의제곱오차이다.) 이러한 결과들은 붓스트랩방법을 적절히 변환하여 사용함으로써 그 성능을 이론적으로도 발전시킬 수 있다는 사실을 말하고 있으며 또한 수렴율이 $n^{1/2}$이 아닌 경우에도 붓스트랩의 일관성이 가능함을 보이고 있다.

3. 5 방향형 자료(directional data)

p-차원 구체(sphere) S_p상에 분포되어 있는 확률변수를 고려하는 문제는 지질학이나 생물학 등을 포함한 많은 분야에서 제기되고 있다. 이제 모분포 F로부터 주어진 관찰값 X_1, X_2, …, X_n에 근거하여 평균방향벡터(mean directional vector), $\theta(F) = \mu(F)/\|\mu(F)\|$의 신뢰영역을 구하는 문제를 고려해 본다 (단 $\|\cdot\|$은 p-차원 실수공간에서의 유클리드 norm). 이제 주어진 관찰값에 대한 경험적분포를 F_n으로 표기하기로 하자. 이 때 $\theta(F)$의 추정값으로는 $\theta(F_n)$가 사용되며 신뢰영역은 $\theta(F_n)$이 축이며 사이각이 $\varphi(F_n)$인 콘(cone)모양으로

$$C = \{\nu \in S_p \mid \langle \nu, \theta(F_n)\rangle \geqq \varphi(F_n)\} \tag{3. 11}$$

으로 나타내진다. (단 $\langle \cdot, \cdot \rangle$은 보통 사용되는 스칼라곱(scalar product)을 나타냄.)

이 때 사용되는 추축통계량은

$$R_n(X, F) = n\{1 - \langle \theta(F), \theta(F_n) \rangle\} \tag{3. 12}$$

이며, 기존 방법들은 기저분포 F에 대해 von Mises-Fisher 분포 등의 모수분포를 가정한 $R_n(X, F)$의 극한분포에 근거를 두고 있다(Watson ; 1983). (이때의 극한분포는 가중카이제곱분포의 합(weighted sum of chi-squared)의 꼴로 나타내진다.) Ducharme외 3인(1985)은 붓스트랩의 일치성을 보였으며, 다음은 여러가지 종류의 변환된 붓스트랩에 대한 모의실험 결과의 일부이다. 우선 붓스트랩 방법의 변환된 꼴들을 포함한 몇 가지 가능한 신뢰콘(confidence cone)에 관한 설명은 다음과 같다.

C_B : 경험적 분포 F_n으로부터 붓스트랩표본을 추출함.

C_{RS} : 기저분포 F가 회전대칭(rotationally symmetric)이므로 경험적 분포를 회전대칭꼴로 변환시킨 분포로부터 붓스트랩표본을 추출함.

C_{PB} : 주어진 표본으로부터 기저분포의 모수를 추정한 모수적 붓스트랩 방법을 사용.

C_L : 기저분포를 안다는 가정하에서 구한 모수적 극한분포에 의한 방법

모의실험은 모평균방향벡터가 북극 (0, 0, 1)이고 밀집정도를 나타내는 모수 k = 3.0인 von

Mises-Fisher 분포로부터 표본을 생성하여 명목신뢰확률이 0.9인 신뢰콘들을 만든 다음 모평균벡터 (0, 0, 1)이 포함되는지의 여부를 1000번 반복독립시행에 대하여 살폈으며 그 결과가 다음 〈표 3. 3〉에 주어져 있다. (붓스트랩분포는 B=200개의 붓스트랩표본에 근거하여 근사추정하였다.)

〈표 3. 3〉 신뢰콘(confidence cone)의 포함확률 추정값

신뢰콘 / 표본의 크기(n)	C_B	C_{RS}	C_{PB}	C_L
10	0.827	0.824	0.822	0.825
20	0.873	0.871	0.877	0.876
50	0.897	0.888	0.890	0.887

물론 모의실험자료를 von Mises-Fisher 분포로부터 생성했으므로 C_L과 다른 붓스트랩 방법과의 비교가 결코 붓스트랩에는 유리한 바가 전혀 없다. 위의 모의실험결과에서 몇 가지 변형된 붓스트랩 신뢰콘들 사이의 큰 차이는 없음을 볼 수 있으며, 비모수적 방법으로서 붓스트랩 방법에 만족할 만하다. 단 추정값들이 명목신뢰확률 0.9에 대체로 가까우면서도 표본의 크기가 커짐에 따라 아래로부터 0.9에 가까와짐을 발견할 수 있다. 실제로 이러한 문제는 다음장에서 설명될 사전추축(prepivoting)을 통해 개선할 수 있으며 붓스트랩의 성능을 더 발전시킬 수 있다.

4. 이중 붓스트랩(double bootstrap)

여기서는 붓스트랩 방법을 중복해서 사용하는 이중 붓스트랩에 대하여 다음 두가지 예를 통해서 간단히 설명하고자 한다.

4. 1 사전추축화(prepivoting)

앞 2. 2절의 정리를 사용하여 붓스트랩 신뢰영역을 만드는 데는 추축통계량 $R_n(X, F)$가 중요한 역할을 하며 $R_n(X, F)$를 다음 (4. 1)과 같이 변환함으로써 붓스트랩 신뢰영역의 신뢰오차(error in confidence level)를 줄일 수 있다. 이제 R_n을 R_n 자신의 추정된 분포함수에 $J_n(x, F_n)$에 의해 변환하여

$$R_{n,1}(X, F) = J_n(R_n(X, F), F_n) \qquad (4. 1)$$

이라 하였을 때 변환된 $R_{n,1}(X, F)$의 극한분포는 (0, 1)상의 일양(uniform)분포를 가지며 이러한 과정을 'prepivoting'이라 한다. 이제 $R_{n,1}(X, F)$의 분포 $J_{n,1}(x, F)$를 $J_{n,1}(x, F_n)$으로 추정한다. 이에 따른 붓스트랩 신뢰영역은

$$\{\theta \in H \mid R_n(X, F) \leqq J_n^{-1}[\{J_{n,1}^{-1}(1-\alpha, F_n)\}, F_n]\} \qquad (4. 2)$$

으로 주어진다. 이와 같은 작업을 하는데는 R_n의 분포 $J_n(x, F)$를 $J_n(x, F_n)$으로 추정하는데 처음 붓스트랩 방법이 사용되고 각각의 붓스트랩 추정량에 대하여 R_{n-1}의 분포 $J_{n-1}(x, F)$를 $J_{n-1}(x, F_n)$으로 추정하는데 두번째 붓스트랩이 중복 사용되며 이를 '중복(double) 붓스트랩' 이라 한다. 이러한 사전추축화에 대한 이론적 내용은 Beran(1987)에 비교적 자세히 설명되어 있다.

4. 2 검정함수(power function)의 추정

주어진 자료 $X = (X_1, X_2, \cdots, X_n)$을 확률모형 $P_{\theta, \xi}$로부터 얻었을 때 다음과 같은 가설검정을 고려해 보자.

$$H_0 : \xi = \xi_0 \quad \text{vs} \quad H_1 : \xi \neq \xi_0 \tag{4. 3}$$

모수벡터 (θ, ξ)은 적절한 모수공간 Ω에 속하며 θ는 알려져 있지 않은 장애모수이다. 이제 (θ, ξ)의 consistent 추정량을 (θ_n, ξ_n)으로 표기한다. 이 때 주어진 추축검정 통계량을 $R_n(X, \xi)$라 할 때 그의 분포

$$J_n(x, (\theta, \xi)) = P_{\theta, \xi}[R(X, \xi) \leqq x]\text{는,} \tag{4. 4}$$

(θ, ξ)를 (θ_n, ξ_n)으로 추정하여 붓스트랩분포 $J_n(x, (\theta_n, \xi_n))$로 추정될 수 있다.
이제 붓스트랩신뢰영역이

$$\{\xi : R_n(X, \xi) \leqq c_n(\alpha : \theta_n, \xi_n)\} \tag{4. 5}$$

으로 주어졌을 때 검정방법(test rule)은

$$\psi_n(X) = \begin{cases} 1 & (R_n(X, \xi) > c_n(\alpha : \theta_n, \xi_n)\text{인 경우}) \\ 0 & (\text{그외의 경우}) \end{cases}$$

이 되며 그의 검정함수는

$$\beta_n(\alpha : \xi, \theta) = P_{\xi, \theta}[R_n(X, \xi) > c_n(\alpha : \theta_n, \xi_n)] \tag{4. 6}$$

이 된다. 이제 알고자 하는 검정함수 $\beta_n(\alpha : \xi, \theta)$의 추정량으로는

$$\beta_n(\alpha : \xi, \theta_n) = P_{\xi, \theta_n}[R_n(X, \xi) > c_n(\alpha : \theta_n, \xi_n)] \tag{4. 7}$$

을 생각할 수 있다. 여기서 식 (4. 5)에서 $c_n(\alpha : \theta_n, \xi_n)$을 구하는 데 첫 단계 붓스트랩이 사용되며, 확률분포 P_{ξ, θ_n}으로부터 재추출한 표본에 근거한 $R_n(X, \xi)$의 두번째 붓스트랩 시뮬레이션 결과에 의해 검정함수의 추정값 (4. 7)을 구할 수 있다. (이러한 추정값 (4. 7)의 이론적 성격에 대한 자세한 설명은 Beran(1986)에 있다.)

5. 결 언

이상과 같이 부족하나마 붓스트랩 방법의 사용에 대해 살펴보았다. Efron의 연구가 붓스트랩 방법에 의한 편의와 표준편차의 추정을 통해 모수의 신뢰구간을 구하는 데 중점을 둔 반면 본고에서는 적절한 추축통계량의 표본분포를 붓스트랩 분포로 추정하는 과정과 그의 통계적 의미에 촛점을 두었다. 물론 표본분포를 추정함에 있어 편의와 표준편차는 부산물로 얻어진다. 본고에서 언급하지 못한 관련된 주요 문제들로는 다음과 같은 것들을 들 수 있다.

(1) 붓스트랩 방법의 최적성 (optimalilty)
(2) 붓스트랩 표본의 크기에 관한 문제 (Hall ; 1986b)
(3) 다변량 자료 분석에의 적용(예를 들면 단계적 판별 분석)
(4) non-iid의 경우 붓스트랩
(5) empirical process의 붓스트랩
(6) 기저분포의 F의 추정에서 특이값(outlier)의 처리
(7) 오차의 측정 및 분석
(8) 그외 저자가 생각못한 주요 통계 문제들

컴퓨터의 놀라운 계산능력은 우리에게 붓스트랩방법을 가능케했으며 앞으로도 그에 대응하는 통계방법의 이론과 응용의 발전이 새삼 기대된다.

참 고 문 헌

(1) Beran R. (1984) "Bootstrap methods in statistics" Jber. d. Dt. Math.-Verein. 86 14−30
(2) Beran R. & Srivastava M. (1985) "Bootstrap tests and confidence regions for functions of a covariance matrix" Ann. Statist. 13 95−115
(3) Beran R. (1986) "Simulated power functions" Ann. Statist. 14 151−173
(4) Beran R. (1987) "Prepivoting to reduce level error of confidence sets" Biometrika 74, 3, 457−68
(5) Bickel P. & Rosenblatt M. (1973) "On some gloval measures of the deviations of density function estimates" Ann. Statist. 6, 1071−95
(6) Bickel P. & Freedman D. (1981) "Some asymptotic theory for the bootstrap" Ann. Statist. 9 1196−1217
(7) Ducharme G., Jhun M., Romano J. & Truong K. (1985) "Bootstrap confidence cones for directional data" Biometrika 72 637−645
(8) Ducharme G. & Jhun M. (1986) "A note on the bootstrap procedure for testing linear hypotheses" Stat. 17 527−531
(9) Efron B. (1979) "Bootstrap methods ; another look at the jackknife" Ann. Statist. 7 1−26

(10) Efron B. (1981) "Nonparametric standard errors and confidence interval" Can. J. Statist. 9 139-172

(11) Efron B. (1982) "The jackknife, the bootstrap, and other resampling plans" SIAM monograph #38

(12) Efron B. (1985) "Bootstrap confidence intervals for a class of parametric problems" Biometrika 72 45-58

(13) Efron B. & Diaconis P. (1983) "Computer-intensive methods in Statistics" Scientific America 116-130

(14) Efron B. (1987) "Better bootstrap confidence intervals" J. Amer. Statist. Ass. 82 171-185

(15) Efron B. & Tibshirani R. (1986) "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy" Statistical Science 1 1 54-77

(16) Faraway J. &. Jhun M. (1988) "Bootstrap choice of bandwidth for density estimation" Tech. Report. Stat. Dept. Univ. Michigan

(17) Freedman D. (1981) "Bootstrapping regression models" Ann. Statist. 9 1218-28

(18) Hall P. (1986a) "On the bootstrap and confidence intervals" Ann. Statist. 14 1431-52

(19) Hall P. (1986b) "On the number of bootstrap simulations required to construct a confidence interval" Ann. Statist. 4 1453-62

(20) Jhun M. (1985) "Bootstrapping k-means clustering"

(21) Jhun M. (1988) "Bootstrapping density estimates" CommStatA 17 61-78

(22) Lo A. (1987) "A large sample study of the Bayesian bootstrap" Ann. Statist. 15, 1, 360-375

(23) Mardia, K, Kent, J., Bibby, J. (1979) "Multivariate Analysis" Academic Press.

(24) Silverman B. (1986) "Density estimation for Statistics and data analysis" Chapman and Hall London

(25) Singh K. (1981) "On the asymptotic accuracy of Efron's bootstrap" Ann. Statist. 9 1218-28

(26) Watson J. (1983) "Statistics on spheres" Wiley New York

(27) Woodroofe M. & Jhun M. (1989) "Singh's theorem in the lattice case" StatProb Letters 7 201-205

A Computer Intensive Method for Modern Statistical Data Analysis I ; Bootstrap Method and Its Applications

Myoungshic Jhun*

Abstract

Computer intensive bootstrap methods are studied as a tool of statistics. Practical calculation and theoretical justification problem of the methods in estimating the sampling distribution and construction confidence region of parameters are discussed through several examples. Statistical meaning of the methods are also considered.

* Dept. of Statistics, Korea University, Seongbook-gu, Anam-dong, Seoul.