The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.6.1019
25(6), 1019–1026

Diagnosis of Observations after Fit of Multivariate Skew *t*-Distribution: Identification of Outliers and Edge Observations from Asymmetric Data

Seung-Gu Kim[1]

[1]Department of Data and Information, Sangji University

(Received October 22, 2012; Revised November 5, 2012; Accepted November 8, 2012)

Abstract

This paper presents a method for the identification of "edge observations" located on a boundary area constructed by a truncation variable as well as for the identification of outliers and the after fit of multivariate skew *t*-distribution(MST) to asymmetric data. The detection of edge observation is important in data analysis because it provides information on a certain critical area in observation space. The proposed method is applied to an Australian Institute of Sport(AIS) dataset that is well known for asymmetry in data space.

Keywords: Multivariate skew *t*-distribution, edge observation, outlier, ECM algorithm.

1. 서론

Azzalini (1985) 및 Azzalini와 Dalla-Valle (1996)에 의해 MSN(multivariate skew normal distribution)이 소개된 이후 비대칭 자료를 적합하기 위한 모형개발 노력이 계속되고 있다. 최근에는 MSN을 특수한 경우로서 포함하는 MST(multivariate skew *t*-distribution) 분포의 개발에 집중되고 있다. 정규분포모형을 적합한 후 이상치들(outliers)을 식별하는 것은 중요한 작업으로 여겨져왔다. 물론 비대칭 자료를 위한 MST 모형의 적합 후에도 이상치의 식별은 여전히 중요하지만, 추가로 관심을 가져야할 관측치들이 있다. 그것들은 MST의 어떤 임계영역에 위치한 관측치들이다. 이들은 MST의 밀도가 큰 부분에 위치하고 있음에도 이 관측치들이 이루는 경계 밖으로는 관측치들이 거의 발생할 수 없음을 알려주는 중요한 역할을 한다. 이러한 이유로 본 논문에서는 이들을 "에지 관측치(edge observations)"라 부를 것이다. 본 연구에서는 이상치 검출과 아울러 에지 관측치 식별을 위한 방법을 제공할 것이다.

Sahu 등 (2003)은 특별한 모수계의 한 구성원으로서 MST를 소개하였다. 이들의 MST는 Lin (2010)에 의해 혼합모형으로 확장되었다. 그리고 Pyne 등 (2009)은 Sahu 등 (2003)의 MST의 제약 버전을 소개하였다. 최근 Lo 등 (2008)과 Lo와 Gottardo (2012)는 Bickel과 Docksum (1981)의 확장된 Box-Cox 변환함수를 바탕으로하는 MST이 소개되기도 하였다. 그러나 본 연구에서는 Lachos 등 (2010) 및 Cabral 등 (2012)에서 사용한 SNI 족(skew-normal/independent family)의 구성원으로서의 MST를 사용할 것이다.

[1]Professor, Department of Data and Information, Sangji University, 660 Woosan-Dong, Wonju, KangWon-Do 220-702, Korea. E-mail: sgukim@sangji.ac.kr

다음 절에서는 SNI-MST를 소개하고 EM 알고리즘에 의한 적합방법을 간략히 제공한다. 3절에서는 이상치와 에지 관측치 검출을 위한 이론적 원리를 제공한다. 4절에서는 AIS 자료를 이용하여 제안된 검출법의 실효성을 보인다. 5절에서는 결론을 정리하고 몇가지 사안에 대한 토론을 제시하였다.

2. MST에 대한 검토

2.1. MST 모형

j번째 임의표본인 p-변량 벡터 $\boldsymbol{Y}_j = (Y_{1j}, \ldots, Y_{pj})^T$가

$$f\left(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu\right) = 2t_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) T\left(r_j\sqrt{\frac{\nu+p}{\nu+D_j}}\,; \nu+p\right), \quad j = 1, \ldots, n \tag{2.1}$$

와 같은 확률밀도를 가질 때, MST 분포를 따른다고 정의한다. 그리고 $\boldsymbol{Y}_j \sim \mathrm{ST}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu)$으로 표시하자. 여기서 $t_p(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$는 위치모수와 척도모수가 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$이며, 자유도가 ν인 p-변량 t-밀도이며, $T(\cdot\,; df)$는 자유도가 df인 표준 단변량 t-분포함수를 나타낸다. 그리고 $D_j = D(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{y}_j - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})$으로서 Mahalanobis 거리제곱을 나타내며, $r_j = r(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}) = \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_j - \boldsymbol{\mu})$를 나타내며, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^T$은 밀도의 치우침(skewness) 모수벡터를 나타낸다. 그리고 $\boldsymbol{\Theta}$는 모든 (중복되지 않는) 모수를 포함하는 벡터를 나타내기로하자.

식 (2.1)의 MST는 만약 $\nu \to \infty$이면 MSN이 되며, $\boldsymbol{\delta} \to \boldsymbol{0}$이면 다변량 t-분포 밀도가 되고, 두 조건이 동시에 만족하면 다변량 정규분포 밀도로 축소되는 매우 신축적인 모형이다.

한편, 우도함수를 알기 쉽게 나타내기 위해 식 (2.1)에서 $\boldsymbol{\delta}^* = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}/\sqrt{1+\boldsymbol{\delta}^T\boldsymbol{\delta}}$ 및 $\boldsymbol{\Omega} = \boldsymbol{\Sigma} - \boldsymbol{\delta}^*\boldsymbol{\delta}^{*T}$와 같은 재모수화가 필요한데, 이것은 언제든지

$$\boldsymbol{\delta} = \frac{\left(\boldsymbol{\Omega} + \boldsymbol{\delta}^*\boldsymbol{\delta}^{*T}\right)^{-\frac{1}{2}}\boldsymbol{\delta}^*}{\left[1 - \boldsymbol{\delta}^{*T}\left(\boldsymbol{\Omega} + \boldsymbol{\delta}^*\boldsymbol{\delta}^{*T}\right)^{-\frac{1}{2}}\boldsymbol{\delta}^*\right]^{\frac{1}{2}}}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{\delta}^*\boldsymbol{\delta}^{*T} \tag{2.2}$$

와 같이 되돌릴 수 있다.

2.2. MST의 재표현 및 해석

$U_j \sim \mathrm{gamma}\,(\nu/2, \nu/2)$를 따른다고 할 때,

$$\begin{pmatrix} Z_j \\ \boldsymbol{Y}_j \end{pmatrix} \Big| U_j = u_j \sim N_{1+p}\left(\begin{pmatrix} 0 \\ \boldsymbol{\mu} \end{pmatrix}, \frac{1}{u_j}\begin{pmatrix} 1 & \boldsymbol{\delta}^* \\ \boldsymbol{\delta}^{*^T} & \boldsymbol{\Sigma} \end{pmatrix}\right) \tag{2.3}$$

을 따른다고 하자. 이때 조건부 변량 $\boldsymbol{Y}_j|(Z_j > 0, U_j = u_j)$은 식 (2.1)의 밀도를 가지는 것으로 알려져 있다. 여기서 $\mathrm{gamma}(\alpha, \beta)$은 평균이 α/β인 감마분포를 나타낸다.

식 (2.3)을 잠시 살펴보면 다음과 같은 해석이 가능하다. 즉, p-변량 $\boldsymbol{Y}_j$와 공분산이 $\mathrm{cov}(X_j, \boldsymbol{Y}_j) = \boldsymbol{\delta}^*$인 (즉 상관계수가 $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}^*$인) 어떤 단변량 확률변수 Z_j를 고려했을 때, Z_j가 0보다 큰 쪽으로 절단된 영역에서 $\boldsymbol{Y}_j$의 밀도는 상관계수에 비례하여 치우침이 발생하게 된다. 여기서 미관측 변수 Z_j를 "절단변수"라 부르자. 절단변수의 절단점은 0으로서 분명하다. 만약 관측변수 $\boldsymbol{Y}_j$의 모든 변량이 Z_j와 상관계수 1을 가지면 p개의 각 절단점들 역시 어떤 한 점으로 분명하게 되며 대응하여 매우 큰 비대칭적 치우침이 발생하게 된다. 그러나 각각의 상관계수가 1보다 작아지면 치우침이 약화되다가 Z_j가 $\boldsymbol{Y}_j$와 무상관이라면 $\boldsymbol{Y}_j$의 밀도는 치우침이 발생하지 않는다.

식 (2.3)의 관계는 즉시 다음과 같은 위계적 구조로도 나타낼 수 있다. 즉, $X_j \overset{\text{def}}{=} Z_j | Z_j > 0$이라 할 때

$$\boldsymbol{Y}_j | \left(X_j = x_j, U_j = u_j\right) \sim N_p\left(\boldsymbol{\mu} + \boldsymbol{\delta}^* x_j,\, \frac{\boldsymbol{\Omega}}{u_j}\right) \tag{2.4}$$

$$X_j | U_j = u_j \sim \text{HN}\left(0, u_j^{-1}\right), \tag{2.5}$$

$$U_j \sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{2.6}$$

여기서 $\text{HN}(\mu, \sigma^2)$는 평균과 분산이 (μ, σ^2)인 단변량 절반정규분포(half normal distribution)를 나타낸다.

식 (2.4)를 보면 알 수 있듯이 u_j와 x_j는 각각 관측치 $\boldsymbol{y}_j$의 산포와 평균적 위치를 결정한다. 만약 $\nu \to \infty$이면 $U_j = 1$로 퇴화되어 $\boldsymbol{Y}_j$가 정규분포를 따르도록 하지만, ν가 작을때 u_j는 0에 가까운 값을 취할 수 있는데, 이 경우 관측치 $\boldsymbol{y}_j$가 정규성에서 벗어난 관측치 즉 이상치가 된다.

한편, 식 (2.4)의 $\boldsymbol{\mu} + \boldsymbol{\delta}^* x_j$는 X_j에 대한 $\boldsymbol{Y}_j$의 회귀평면이라 할 수 있다. 이때 우리는 절단값 0에 가까운 x_j에 대응하는 관측치들이 $\boldsymbol{y}_j$ 공간상에서 어떤 경계면을 이루게 될 것으로 기대할 수 있다. 만약 x_j의 예측치 $\hat{x}_j$를 구할 수 있다면, 이 값이 0에 가까운 개체 j를 찾고 이에 대응하는 관측치 $\boldsymbol{y}_j$를 식별할 수 있다.

이상으로부터 우리는 미관측 자료인 u_j와 x_j의 예측치를

$$\hat{u}_j = E\left(U_j | \boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}\right) \quad \text{및} \quad \hat{x}_j = E\left(X_j | \boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}\right) \tag{2.7}$$

와 같은 조건부 기대값으로 구하여 이에 대응하는 관측치가 이상치인지 그리고 에지 관측치인지를 판별할 것이다. 그런데 이에 대한 구체적 방법은 먼저 모형 적합을 위한 EM 알고리즘을 소개한 후에 다시 논의하도록 하겠다.

2.3. EM 알고리즘

식 (2.4)–(2.6)의 조건부 관계로부터 완비자료 $\boldsymbol{u} = (u_1, \ldots, u_n)^T$, $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ 및 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$의 로그-우도는

$$\begin{aligned} L_c\left(\boldsymbol{\Theta} | \boldsymbol{u}, \boldsymbol{t}, \boldsymbol{y}\right) \propto \frac{1}{2} \sum_{j=1}^{n} \Big\{ & \log |\boldsymbol{\Omega}| + u_j \left(\boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{\delta}^* x_j\right)^T \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{\delta}^* x_j\right) \Big\} \\ & - \sum_{j=1}^{n} \left\{ \log \Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2} \log\left(\frac{\nu}{2}\right) - \frac{\nu}{2} \left(\log u_j - u_j\right) \right\} \end{aligned} \tag{2.8}$$

와 같이 나타낼 수 있다. 이제 EM 알고리즘의 $(k+1)$번째 단계의 E-스텝에서 완비자료의 조건부 기대값 $Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(k)}) = E_{\boldsymbol{\Theta}^{(k)}}\left[L_c(\boldsymbol{\Theta}) | \boldsymbol{y}\right]$를 계산해야 한다. 이것은 식 (2.8)를 통해 알 수 있듯이 다음과 같은 몇 가지의 충분통계량에 대한 조건부 기대값을 계산해야 하는 문제이다. 즉,

$$\begin{aligned} u_j^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left[U_j | \boldsymbol{y}_j\right], \\ (ux)_j^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left[U_j X_j | \boldsymbol{y}_j\right], \\ \left(ux^2\right)_j^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left[U_j X_j^2 | \boldsymbol{y}_j\right] \end{aligned} \tag{2.9}$$

인데, 이 세 조건부 기대값은 Cabral 등 (2012)에서 명시적 형태로 주어져 있다. 이 기대값들을 바탕으로 M-스텝에서 모수 $Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(k)})$를 최대화함으로써 추정치 $\boldsymbol{\Theta}^{(k+1)}$을 얻는다. 그리고 이 과정을 모수들

이 수렴할 때까지 충분히 반복하여 최종적으로 추정치 $\widehat{\boldsymbol{\Theta}}$를 얻게된다 (M-스텝에 대한 구체적인 과정은 Cabral 등 (2012)나 Kim (2012)를 참조하기 바란다).

EM 알고리즘을 수행하면 식 (2.9)으로부터 $\hat{u}_j$들을 추가적인 계산없이 얻을 수 있다. 그러나 $\hat{x}_j = E_{\widehat{\boldsymbol{\Theta}}}[X_j|\boldsymbol{y}_j]$은 EM 알고리즘 수행 후에 계산해야 하는데 이것에 대해서는 다음 소절에서 자세히 설명하겠다.

2.4. $E_{\widehat{\boldsymbol{\Theta}}}\left[X_j|\boldsymbol{y}_j\right]$의 계산

$\boldsymbol{Y}_j \sim \mathrm{ST}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu)$일 때, 조건부 확률변수 $U_j|\boldsymbol{y}_j$는

$$h(u_j|\boldsymbol{y}_j) = g\left(u; \frac{\nu+p}{2}, \frac{\nu+D_j}{2}\right) \Phi\left(u_j^{\frac{1}{2}} r_j\right) T^{-1}\left(r_j \sqrt{\frac{\nu+p}{\nu+D_j}}; \nu+p\right) \tag{2.10}$$

의 확률밀도를 가진다 (Kim, 2012). 여기서 $g(u;\alpha,\beta)$는 모수가 (α,β)인 감마분포 밀도이며 $\Phi(\cdot)$는 $N(0,1)$의 분포함수를 나타내고, $r_j = m_j/V$인데 $m_j = V^2 \boldsymbol{\delta}^{*T} \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})$, $V^2 = (1+\boldsymbol{\delta}^{*T} \boldsymbol{\Omega}^{-1} \boldsymbol{\delta}^*)^{-1}$를 각각 나타낸다. 그리고 Cabral 등 (2012)에 따르면 조건부 확률변수 $X_j|(\boldsymbol{y}_j, U_j = u_j)$는

$$f_x\left(x_j|\boldsymbol{y}_j, u_j\right) = \phi\left(x_j; m_j, u_j^{-1} V_j^2\right) \Phi^{-1}\left(u_j^{\frac{1}{2}} r_j\right), \quad 0 < x_j < \infty \tag{2.11}$$

와 같은 확률밀도를 가진다. 여기서 $\phi(\cdot;\mu,\sigma^2)$은 $N(\mu,\sigma^2)$의 확률밀도이다. 식 (2.11)은 결국 $N(m_j, u_j^{-1}V_j^2)$ 분포의 양으로의 절단분포를 의미한다.

이때

$$X_j|\boldsymbol{y}_j \sim tt_{\nu+p}\left(m_j, V^2\left[\frac{\nu+D_j}{\nu+p}\right]; (0,\infty)\right) \tag{2.12}$$

이다. 즉, 조건부 확률변수 $X_j|\boldsymbol{y}_j$은 위치모수와 척도모수가 $(m_j, V^2[(\nu+D_j)/(\nu+p)])$이며 자유도 $\nu+p$인 단변량 t-분포의 양으로의 절단 분포를 따른다. 이에 대한 증명은 부록에 수록하였다.

이제 Kim (2008)의 Theorem 3.1 혹은 Ho 등 (2012)의 Theorem 1에 의해

$$\hat{x}_j = E\left(X_j|\boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}\right) = \hat{m}_j + \hat{V}^2 \sqrt{\frac{\hat{\nu}+\hat{D}_j}{\hat{\nu}+p}} \left(\frac{\hat{\nu}+p}{\hat{\nu}+p-1}\right) \times \hat{\kappa}_j, \quad j = 1, \ldots, n \tag{2.13}$$

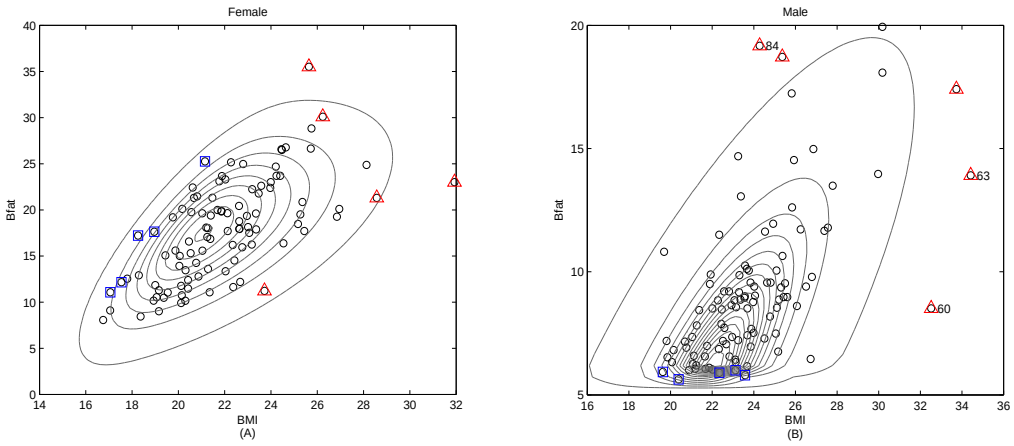
단,

$$\hat{\kappa}_j = \frac{1}{\sqrt{\pi(\hat{\nu}+p)}} \frac{\Gamma\left(\frac{\hat{\nu}+p+1}{2}\right)}{\Gamma\left(\frac{\hat{\nu}+p}{2}\right)} T^{-1}\left(\frac{\hat{m}_j}{\hat{V}} \sqrt{\frac{\hat{\nu}+p}{\hat{\nu}+\hat{D}_j}}; \hat{\nu}+p\right)$$

과 같이 얻을 수 있다.

3. 실자료 예제

AIS(Australian Institute of Sport) 자료 (Cook과 Weisberg, 1994)는 자료집단의 비대칭성 때문에 치우친 분포 모형의 적합능력을 평가하는 벤치마킹 사례로 자주 이용된다. 이 자료는 100명의 여성과 102명의 남성 운동선수들의 11가지의 특성을 조사한 자료로서 우리는 이 중에 시각적 판정이 용이하도



(a) Female (b) Male

Figure 3.1. Results of Identification of Outliers and Edge Observations (<Symbols> black ◯: observations, red △: outlying observations, blue □: edge observations, contour: fitted density).

Table 3.1. Estimates of parameters and $\hat{u}_j$ and $\hat{x}_j$ for a few observations

	Parameters	μ_1	μ_2	ω_1^2	ω_{12}	ω_2^2	ν	δ_1	δ_2
Female	Estimates	19.56	16.98	3.11	7.94	28.39	75.30	2.92	−0.11
	Outlying Obs.	75	56	70	53	63			
	$\hat{u}_j$	0.87	0.89	0.94	0.95	0.95			
	Edge Obs.	50	94	5	86	37			
	$\hat{x}_j$	0.16	0.18	0.19	0.19	0.20			
Male	Estimates	21.82	5.78	2.22	0.08	0.02	3.63	8.80	23.00
	Outlying Obs.	60	63	84	78	82			
	$\hat{u}_j$	0.13	0.14	0.17	0.20	0.22			
	Edge Obs.	42	56	72	69	55			
	$\hat{x}_j$	0.03	0.03	0.05	0.06	0.08			

록 두 가지 특성변량 BMI(body mass-index: Y_1)과 Bfat(percentage of body fat: Y_2)만을 사용할 것이다.

Figure 3.1에 여성 집단 (a)과 남성 집단 (b) 자료들을 산점도(심볼 ◯)로 나타내었다. 두 집단 모두 치우침을 가지는 비대칭성을 보여주고 있는데, 등고선으로 표현된 2-변량 MST 분포로 잘 적합되고 있다 (참고로 여성집단과 남성집단에 대한 적합모형의 로그-우도는 각각 −514.1394 및 −441.4471였다). 그리고 Table 3.1에 두 집단에 대한 MST 모형의 모수 추정치들과 함께 가장 작은 5개의 $\hat{u}_j$와 $\hat{x}_j$ 그리고 그에 대응하는 관측치의 번호를 제공하였다.

3.1. 이상치 진단

먼저 여성집단의 경우 자유도 추정치 $\hat{\nu} = 75.30$로서 추정된 분포가 t-분포라기보다는 정규분포에 더 가깝다. 대다수 $\hat{u}_j$들이 거의 1에 가까운 값을 나타내었는데, 그 중 가장 작은 5개의 관측치의 $\hat{u}_j$들은 각각 0.87, 0.89, 0.94, 0.95, 0.95였다. 이는 0보다는 1에 가까운 값으로서 정규성을 넘어서는 이상치들이

라고 말할 수 없으며, 다만 대부분의 자료보다 상대적으로 중심에서 조금 멀리 떨어져 있는 자료임을 말해 주고 있는 것이다. 이것들은 Figure 3.1(a)에서 빨간 △로 표시되어 있다.

반면 남성집단의 경우 가장 작은 5개의 $\hat{u}_j$는 각각 0.13, 0.14, 0.17, 0.20, 0.22을 나타내었다 (Figure 3.1(b)의 빨간 △). 특히 처음 60, 63, 84번 관측치는 t-분포가 아닌 정규분포 하에서는 발생할 가능성이 거의 없는 것들로서 분명하게 이상치라고 판정할 수 있을 것이다. 실제로 남성집단의 자유도 추정치는 $\hat{\nu} = 3.63$로서 추정된 밀도는 매우 두꺼운 꼬리를 가지고 있는 t-분포임을 말하고 있다.

3.2. 에지 관측치 진단

여성집단과 남성집단에 대해 가장 작은 5개의 $\hat{x}_j$를 Table 3.1에 수록하였다. 그리고 이에 대응하는 관측치들을 Figure 3.1(a)–(b)에 파랑 □로 표시하였다. 이 관측치들은 일정한 경계를 형성하는데 적합 등고선이 압축되어 나타난 임계영역과 아주 잘 일치하고 있음을 보여주고 있다. 특히 여성집단보다 남성집단의 $\hat{x}_j$가 훨씬 작은데, 이것은 남성집단의 비대칭성이 훨씬 더 심하다는 것을 의미한다. 실제로 Table 3.1에서 치우침 모수 추정치 $(\hat{\delta}_1, \hat{\delta}_2)$가 여성집단의 경우 $(2.92, -0.11)$인 반면 남성집단은 $(8.80, 23.00)$로서 매우 크다는 것이 이를 확인해 주고 있다.

마지막으로 본 논문의 주제와는 다소 괴리가 있지만 흥미로운 관심거리 하나를 추가로 제공하면서 논문을 마치도록 하겠다. 미지의 절단 전 확률변수 Z와 관측변수 (Y_1, Y_2)의 상관계수 추정치는 $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\delta}}^*$로서 얻을 수 있는데, 그 결과는 다음과 같았다. 즉, 미지의 절단변수 Z는 여성 BMI와 거

Group		BMI (Y_1)	Bfat (Y_2)
Female	Z	0.95	−0.03
Male		0.36	0.93

의 절대적인 상관성이 있고 여성 Bfat와는 무관한 변수이며, 반대로 남성 Bfat과는 매우 큰 상관성이 있으며 남성 BMI와는 다소 무관한 변수로 나타났다. 이러한 변수 Z가 절단이 전제됨으로써 (BMI, Bfat) 자료의 비대칭성(즉 skewness)을 유발하고 있는 것이다. 스포츠 생리학자라면 아마 변수 Z가 무엇인지 금방 알지도 모른다.

4. 결론 및 토의

본 논문에서는 비대칭 자료에 대한 MST 모형적합 후 이상치 검출과 "에지 관측치"를 식별하는 방법을 제안하였다. 이들은 관측자료 $\boldsymbol{y}_j$에 대한 미관측 자료의 조건부 기대값 $E(U_j|\boldsymbol{y}_j)$ 및 $E(X_j|\boldsymbol{y}_j)$으로 각각 예측한 후 이들이 0에 가까운지 검토하는 것이다.

본 논문에서 제공한 이상치 검출방법은 McLachlan과 Peel (2000)에서 t-분포 적합에 관해 소개되었던 방법이다. 다만 본 논문에서는 skew t-분포 적합에도 이 방법이 유효함을 보인 것이다. 그러나 저자가 아는 한 에지 관측치 개념은 본 논문에서 처음 언급한 것으로서 방법 또한 처음일 것이라 사료된다. 비대칭 자료의 적합에 대한 관심이 증가할수록 에지 관측치 식별에 대한 관심도 함께 커질 것으로 기대된다.

이울러 제안된 에지 관측치 식별 기법은 MST 혼합모형으로 확장하여 시도해 볼 수 있을 것이다. 이때 군집화된 그룹에서 식별된 에지 관측치들은 각 군집의 특성을 밝히는데 중요한 역할을 할 것으로 예상된다. 이 문제는 본 연구의 향후 과제이다.

부록

식 (2.12)의 증명: 식 (2.12)를 증명하기 전에 먼저 다음 성질을 알아두자.

Lemma A.1 $U \sim \text{gamma}(\alpha, \beta)$를 따를 때, 임의의 p-차원 벡터 $\boldsymbol{x}$에 대해

$$E\left[\phi\left(\boldsymbol{x}\sqrt{U}; \mathbf{0}, \boldsymbol{\Sigma}\right) U^{\frac{p}{2}}\right] = t_{2\alpha}\left(\boldsymbol{x}\sqrt{\frac{\alpha}{\beta}}; \mathbf{0}, \boldsymbol{\Sigma}\right)\left(\frac{\alpha}{\beta}\right)^{\frac{p}{2}} \tag{A.1}$$

이다.

위 결과는 Lin (2010)의 Proposition 1에 주어진 등식의 양변을 $\boldsymbol{x}$에 관해 미분하면 즉시 얻을 수 있다. 이제 관측의 첨자 j를 떼고 $r = m/V$라 쓰면서,

$$\begin{aligned} f_x(x|\boldsymbol{y}) &= \int_0^\infty f(x, u|\boldsymbol{y})du = \int_0^\infty f_x(x|\boldsymbol{y}, u)h(u|\boldsymbol{y})du \\ &= T^{-1}\left(r\sqrt{\frac{\nu+p}{\nu+D}}; \nu+p\right)\int_0^\infty \phi\left(x; m, u^{-1}V^2\right) g\left(u; \frac{\nu+p}{2}, \frac{\nu+D}{2}\right) du \\ &= T^{-1}\left(r\sqrt{\frac{\nu+p}{\nu+D}}; \nu+p\right)\int_0^\infty u^{\frac{1}{2}2}\phi\left(u^{\frac{1}{2}}(x-m); 0, V^2\right) g\left(u; \frac{\nu+p}{2}, \frac{\nu+D}{2}\right) du \\ &= T^{-1}\left(r\sqrt{\frac{\nu+p}{\nu+D}}; \nu+p\right) E\left[\phi\left(U^{\frac{1}{2}}(x-m); 0, V^2\right) U^{\frac{1}{2}}\right] \end{aligned} \tag{A.2}$$

을 얻을 수 있다. 여기서 $p = 1$인 경우에 대하여 Lemma A.1을 이용하면,

$$\begin{aligned} f_x(x|\boldsymbol{y}) &= T^{-1}\left(r\sqrt{\frac{\nu+p}{\nu+D}}; \nu+p\right) t_{\nu+p}\left((x-m)\sqrt{\frac{\nu+p}{\nu+D}}; 0, V^2\right)\left(\frac{\nu+p}{\nu+D}\right)^{\frac{1}{2}} \\ &= T^{-1}\left(r\sqrt{\frac{\nu+p}{\nu+D}}; \nu+p\right) t_{\nu+p}\left(x; m, V^2\left[\frac{\nu+D}{\nu+p}\right]\right), \quad 0 < x < \infty \end{aligned} \tag{A.3}$$

을 얻는다. 그런데

$$\int_0^\infty t_{\nu+p}\left(x; m, V^2\left[\frac{\nu+D}{\nu+p}\right]\right) dx = T\left(r\sqrt{\frac{\nu+p}{\nu+D}}; \nu+p\right)$$

이므로, $f_x(x|\boldsymbol{y})$은 $tt_{\nu+p}(m_j, V^2[(\nu+D)/(\nu+p)]; (0, \infty))$의 확률밀도임을 알 수 있다.

References

Azzalini, A. (1985). A class of distribution which includes the normal ones, *Scandinavian Journal of Statistics*, **33**, 561–574.

Azzalini, A. and Dalla-Valle, A. (1996). The multivariate skew normal distribution, *Biometrika*, **83**, 715–726.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited, *Journal of American Statistical Association*, **76**(374), 296–311.

Cabral, C. S., Lachos, V. H. and Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distribution, *Computational Statistics and Data Analysis*, **56**, 126–142.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, **56**, Wiley, New York.

Ho, H. J., Lin, T. I., Chen, H.-Y. and Wang, W.-L. (2012). Some results on the truncated multivariate t distribution, *Journal of Statistical Planning & Inference*, **142**, 25–40.

Kim, H. J. (2008). Moments of truncated Student-t distribution, *Journal of Korean Statistical Society*, **37**, 81–87.

Kim, S.-G. (2012). ECM Algorithm for fitting of mixtures of multivariate Skew t-Distribution, *Communications of the Korean Statistical Society*, **19**, 673–684.

Lachos, V. H., Ghosh, P. and Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed model, *Statistica Sinica*, **20**, 303–322.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions, *Statistics and Computing*, **20**, 343–356.

Lo, K., Brinkman, R. R. and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, **73**, 321–332.

Lo, K. and Gottardo, R. (2012). Flexible mixture modeling via the multivariate t distribution with the Box-Cox transformation: An alternative to the skew-t distribution, *Statistics and Computing*, **22**, 33–52.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L. and Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis, *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8519–8524.

Sahu, S. K., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distribution with application to Bayesian regression model, *The Canadian Journal of Statistics*, **31**, 129–150.