The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.6.811
31(6), 811–823

Identification of the out-of-control variable based on Hotelling's T^2 statistic

Sungim Lee[a,1]

[a]Department of Applied Statistics, Dankook University

(Received October 30, 2018; Revised November 1, 2018; Accepted November 1, 2018)

Abstract

Multivariate control chart based on Hotelling's T^2 statistic is a powerful tool in statistical process control for identifying an out-of-control process. It is used to monitor multiple process characteristics simultaneously. Detection of the out-of-control signal with the T^2 chart indicates mean vector shifts. However, these multivariate signals make it difficult to interpret the cause of the out-of-control signal. In this paper, we review methods of signal interpretation based on the Mason, Young, and Tracy (MYT) decomposition of the T^2 statistic. We also provide an example on how to implement it using R software and demonstrate simulation studies for comparing the performance of these methods.

Keywords: multivariate SPC, Hotelling's T^2 statistic, MYT decomposition, identification for out-of-control signal

1. 서론

호텔링의 T^2 통계량 (Hotelling, 1947)을 사용한 이단계 관리도(phase II control chart)는 연속형 다변량 품질 특성치의 평균벡터를 온라인 모니터링할 때 자주 사용된다 (Lim과 Lee, 2017). 이 관리도는 시점에 따라 관측된 통계량의 값으로부터 평균벡터의 변화유무를 검정할 수 있도록 하는데, 관리한계선 내에서 통계량의 값이 관측되면 관리상태 즉 평균벡터의 변화가 없다고 판단하고, 관리한계선 밖에서 관측될 때는 그 시점에서 평균벡터 변화의 신호 즉 이상신호가 발생했다고 해석한다. 그런데, 평균변화를 모니터링하는 일변량 관리도와 비교하여 평균벡터의 변화를 모니터링하는 다변량 관리도의 경우, 이상신호의 해석이 간단하지 않아 그 원인을 이해하기 위한 많은 연구가 있었다 (Alt, 1985; Murphy, 1987; Doganaksoy 등, 1991; Hayter와 Tsui, 1994; Mason 등, 1995, 1997; Timm, 1996; Lee, 2018). 가장 잘 알려진 방법 중 하나는 Mason 등 (1995)이 제안한 Mason-Young-Tracy (MYT) 분해인데, 호텔링의 T^2 통계량을 서로 독립적인 p개의 항으로 분할하여, 이상신호의 원인을 찾도록 한다. 그런데, Kourti와 MacGregor (1996)와 Mason과 Young (1999)이 언급했듯이 다변량 품질 특성치의 수가 10이 넘으면 MYT 분해를 위해 계산해야 할 항이 기하급수적으로 커지게 된다. 따라서, Mason 등 (1997)은 실제적인 접근방법으로 합리적인 계산량으로 이상신호의 원인을 찾는 절차를 제공하였다. Murphy (1987) 또한 다변량 데이터에 대한 이단계 관리도에서 이상신호에 대한 원인이 되는 변수를 선택하는 절차를 선택

[1]Department of Applied Statistics, Dankook University, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea. E-mail: silee@dankook.ac.kr

하였는데, 이 또한 MYT 분해의 특별한 경우가 된다. 본 연구에서는 이들 방법을 자세히 살펴보고, 실제 문제에서 어떻게 활용할 수 있는지 R을 활용한 사례분석을 통해 소개하고자 한다. 또한, 모의실험을 통해 각 절차의 특징을 비교해보고, 이상신호의 원인에 대한 추가 연구방향에 대해 고찰해 보기로 한다.

논문의 구성은 다음과 같다. 2절에서는 이단계 관리도에서의 호텔링 T^2 통계량의 이상신호의 원인을 식별하기 위한 절차들을 소개하기로 한다. 3절에서는 사례분석을 통해 2절에서 소개한 절차들을 어떻게 활용할 수 있는지 소개하기로 한다. 4절에서는 각 절차에 대해 모의실험을 통해 비교 분석해 보고자 한다. 5절에서는 연구결과를 요약하고 앞으로의 연구방향에 대해 고찰해 보기로 한다.

2. 이상신호의 원인 식별 방법

p변량 데이터 $X^T = (x_1, x_2, \ldots, x_p)$는 관리상태하에서 평균벡터가 $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^T$이고, 공분산 행렬이 $\Sigma = (\sigma_{ij})_{p\times p}$인 다변량 정규분포를 따른다고 가정한다. 일반적으로 μ와 Σ는 미지의 값이므로 관리상태하의 m개의 데이터 $X_i^T = (x_{1i}, x_{2i}, \ldots, x_{pi})$, $i = 1, \ldots, m$로부터 $\hat{\mu} = \bar{X} = (1/m)\sum_{i=1}^m X_i$와 $\hat{\Sigma} = S = (1/(m-1))\sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})^T$로 추정된다, 이를 바탕으로 새로운 관측치 $Y^T = (y_1, y_2, \ldots, y_p)$로부터 평균벡터의 변화를 온라인 모니터링하기 위한 호텔링의 T^2 관리통계량은 다음과 같이 정의되고 관리상태하에서 이 통계량은 다음의 F 분포를 따르게 된다.

$$T^2 = \big(Y - \bar{X}\big)^T S^{-1}\big(Y - \bar{X}\big) \sim \frac{p(m+1)(m-1)}{m(m-p)} F_{p,m-p}. \tag{2.1}$$

따라서, 위 통계량으로부터 다변량 관측치를 모니터링 하는 경우 관리 상한선(upper control limit; UCL)을 $p(m+1)(m-1)/\{m(m-p)\}F_{p,m-p}(1-\alpha)$로 두고 평균벡터의 변화여부를 판단하게 된다. Jackson (1991)이 언급했듯이 다변량 관리도에서 이상신호가 발생한 경우 그 문제가 무엇인지에 대해서도 답해야 한다. 그런데, T^2 통계량에 대한 이상신호의 경우 그 해석이 쉽지 않다 (Woodal과 Montgomery, 1999). 일반적으로 관측치 Y에 대한 이상신호의 원인이 무엇인지 그 원인을 찾아 제거하는 등의 조처를 해야 하는데, T^2 통계량에 기반한 다변량 관리도의 경우 이상신호의 원인을 식별하기 위한 또 다른 절차가 요구된다.

2.1. Murphy 방법

이 방법은 Murphy (1987)가 제안한 것으로 관리한계선을 넘는 관측치로부터 이상신호의 원인을 식별하기 위해 다음과 같은 절차를 수행한다.

- 1단계: 이상신호가 발생한 p변량 데이터 $Y^T = (y_1, y_2, \ldots, y_p)$에 대해 개별적인 통계량 $T_i^2 = T^2(y_i)$ $(i = 1, 2, \ldots, p)$을 계산하고, 전체 T^2 통계량에서 개별 변수로 계산한 $T^2(y_i)$와의 차이를 계산한다. $D_{p-1}(i) = [T^2 - T^2(y_i)]$에 대해 $\min(D_{p-1}(i)) = D_{p-1}(r)$을 정의하고, 이 최솟값에 대해 다음과 같이 검정한다.
  $D_{p-1}(r) < \chi^2(p-1, 1-\alpha)$이면 r번째 변수를 이상신호의 원인으로 고려하고, $D_{p-1}(r) > \chi^2(p-1, 1-\alpha)$을 만족하면 다음 2단계를 진행한다.
- 2단계: 이제 $(p-1)$개의 통계량 $D_{p-2}(r, j) = [T^2 - T^2(y_r, y_j)]$ $(1 \le r, j \le p,\ r \ne j)$을 계산하고, $\min(D_{p-2}(r, j)) = D_{p-2}(r, s)$을 정의한 후 이 최솟값에 대해 다음과 같이 검정한다.
  $D_{p-2}(r, s) < \chi^2(p-2, 1-\alpha)$이면 r번째와 s번째 변수를 이상신호의 원인으로 고려하고, $D_{p-2}(r, s) > \chi^2(p-2, 1-\alpha)$을 만족하면 다음 3단계를 진행한다.
- 3단계: 2단계와 비슷하게 진행한다.

- $p-1$단계: 마지막으로 $D_{p-(p-1)} > \chi^2(1, 1-\alpha)$ 검정이 통계적으로 유의하면 모든 p개의 변수가 이상신호의 원인으로 고려될 수 있다.

Murphy (1987)는 관리상태하에서 $D_{p-i} \sim \chi^2(p-i)$임을 보였다. 이러한 원인 식별 절차는 회귀모형에서의 변수선택에 있어 전진선택법과 비슷한 절차를 나타내고 있음을 알 수 있다. 즉, 각 단계에서 D 통계량이 가장 작은 값을 기준으로 선택하는 것은 T^2 통계량의 값에서 가장 많은 부분을 차지하고 있는 변수를 선택하는 것과 같으며, 각 단계를 거쳐 더 고려할만한 변수가 없다고 판단될 때 중단하게 되는 절차로 이미 선택된 변수들만을 이상신호의 원인으로 고려하게 되는 방법이다. 이 절차는 원인 식별을 위해 계산해야 할 D의 가지 수가 가장 많은 경우라도 $p(p+1)/2$번이 된다. 주요 변수선택 시 사용되는 전진선택법의 문제점에서 알 수 있듯이 이러한 선택방법이 실제로 이상신호에 영향을 주는 변수들의 조합을 가장 잘 선택한다는 보장은 없지만, 우리가 고려해야 할 변수의 수가 많은 경우 합리적인 계산량으로 주요 변수들을 선택하게 된다는 장점이 있다.

2.2. Mason-Young-Tracy 분해

이상신호를 식별하기 위한 T^2에 대한 분할로 가장 잘 알려진 것은 Mason 등 (1995)이 처음 제안한 MYT 분할이다. 식 (2.1)에 있는 $(Y-\bar{X})$에 대해 y_p 변수를 제외한 관측벡터 $Y^{(p-1)} = (y_1, y_2, \ldots, y_{p-1})$와 y_p로 분할하면

$$\left(Y-\bar{X}\right)^T = \left[\left(Y^{(p-1)} - \bar{X}^{(p-1)}\right), \left(y_p - \bar{x}_p\right)\right]^T \tag{2.2}$$

이 되고, 이 때 $\bar{X}^{(p-1)}$는 그에 대응하는 평균벡터를 나타낸다. 평균벡터와 마찬가지로 분산행렬 S를 다음과 같이 분할하면,

$$S = \begin{bmatrix} S_{XX} & s_{xX} \\ s_{xX}^T & s_p^2 \end{bmatrix}.$$

식 (2.1)에 있는 T^2 통계량은

$$\begin{aligned} T^2 &= \left[\left(Y^{(p-1)} - \bar{X}^{(p-1)}\right), \left(y_p - \bar{x}_p\right)\right]^T \begin{bmatrix} S_{XX} & s_{xX} \\ s_{xX}^T & s_p^2 \end{bmatrix}^{-1} \begin{bmatrix} \left(Y^{(p-1)} - \bar{X}^{(p-1)}\right) \\ \left(y_p - \bar{x}_p\right) \end{bmatrix} \\ &= T_{p-1}^2 + s_{p\cdot 1,2,\ldots,p-1}^{-2}\left(y_p - \bar{x}_{p\cdot 1,2,\ldots,p-1}\right)^2 \end{aligned}$$

으로 분해된다. 단,

$$T_{p-1}^2 = \left(Y^{(p-1)} - \bar{X}^{(p-1)}\right)^T S_{XX}^{-1}\left(Y^{(p-1)} - \bar{X}^{(p-1)}\right), \tag{2.3}$$

$$T_{p\cdot 1,2,\ldots,p-1}^2 = \frac{\left(y_p - \bar{x}_{p\cdot 1,2,\ldots,p-1}\right)^2}{s_{p\cdot 1,2,\ldots,p-1}^2}, \tag{2.4}$$

$$s_{p\cdot 1,2,\ldots,p-1}^2 = s_p^2 - s_{xX}' S_{XX}^{-1} s_{xX}, \tag{2.5}$$

$$\bar{x}_{p\cdot 1,2,\ldots,p-1} = \bar{x}_p + B_p^T\left(Y^{(p-1)} - \bar{X}^{(p-1)}\right), \tag{2.6}$$

$$B_p^T = S_{XX}^{-1} s_{xX} \tag{2.7}$$

을 의미한다. 여기서 $B_p^T = S_{XX}^{-1} s_{xX}$는 $(p-1)$ 벡터로, 변수 $x_1, x_2, \ldots, x_{p-1}$를 변수 x_p에 적합시켜 구한 회귀계수로, $s_{p\cdot 1,2,\ldots,p-1}^2$은 조건부 분산으로 해석된다. 따라서, 식 (2.1)에 있는 T^2은 다음과 같이

분할된다.

$$T^2 = T^2_{p-1} + T^2_{p\cdot 1,2,\ldots,p-1}. \tag{2.8}$$

이 때, 식 (2.8)에 있는 첫 번째 항도 $(p-1)$개의 변수로 이루어진 호텔링의 T^2 통계량이므로, 다시 다음과 같은 두 개의 직교항으로 나눌 수 있다.

$$T^2_{p-1} = T^2_{p-2} + T^2_{p-1\cdot 1,2,\ldots,p-2}.$$

T^2_{p-2}은 Y벡터의 첫 $(p-2)$개의 성분으로 계산한 호텔링의 T^2 통계량이고, 두 번째 항 $T^2_{p-1\cdot 1,2,\ldots,p-2}$은 변수 $x_1, x_2, \ldots, x_{p-2}$가 주어졌을 때, x_{p-1}의 조건부 평균과 표준편차로 조정한 y_{p-1}의 제곱합을 가리킨다. 이런 방법으로 반복해서 분할을 한다면 T^2에 대한 분할은 다음과 같은 식으로 분해될 수 있다.

$$T^2 = T^2_1 + T^2_{2.1} + T^2_{3.1,2} + \cdots + T^2_{p\cdot 1,2,\ldots,p-1}. \tag{2.9}$$

이처럼 T^2 통계량은 p개의 항으로 분할될 수 있으며, 첫 번째 항은 y_1 변수에 관한 정보만 있는 통계량이지만, 두 번째 항부터는 y_1이 주어졌을 때, y_2 변수의 조건부 기대값과 표준편차로 표준화한 조건부 항으로 정의된다. 그러나 위 분할은 변수들의 순서를 고려할 때 $p!$ 가지수 중의 하나일 뿐이다. 다만 순서에 상관없이 모든 분할에 대해 T^2 통계량은 같은 값을 나타낸다. 이러한 분할을 MYT 분할이라고 한다. MYT 분할을 위해 필요한 서로 다른 항은 모두 $p \times 2^{p-1}$개가 되고, 각 항에 대한 확률분포는 모평균이 변화하지 않았다는 가정하에 다음의 분포를 따르게 된다.

$$T^2_j = \frac{(y_j - \bar{x}_j)^2}{s^2_j} \sim \frac{m+1}{m} F_{\alpha,1,m-1}, \quad j = 1, 2, \ldots, p, \tag{2.10}$$

$$T^2_{j\cdot 1,2,\ldots,j-1} = \frac{(y_j - \bar{x}_{j\cdot 1,2,\ldots,j-1})^2}{s^2_{j\cdot 1,2,\ldots,j-1}} \sim \frac{(m+1)(m-1)}{m(m-k-1)} F_{\alpha,1,m-k-1}. \tag{2.11}$$

이 때, 식 (2.11)에서 k는 조건부로 주어지는 변수의 갯수를 의미하고, 조건부 항에 대한 계산은 식 (2.8)에 의해 다음과 같이 계산할 수 있다.

$$T^2_{j\cdot 1,2,\ldots,j-1} = T^2_{(1,2,\ldots,j)} - T^2_{(1,2,\ldots,j-1)} \quad (2 \le j \le p). \tag{2.12}$$

이것은 T^2의 MYT 분해를 위한 조건부 항을 계산할 때, 모든 변수들의 부분집합에 대한 호텔링 T^2값만 있으면 계산이 가능하다는 것을 의미한다. 이처럼 MYT 분해를 할 수 있다면, 식 (2.10)으로부터 $\mathrm{UCL} = (m+1)/m\, F_{(\alpha,1,m-1)}$을 작성하여 이상신호의 원인으로 특정 변수를 고려하거나, 혹은 식 (2.11)로부터 작성된 $\mathrm{UCL} = (m+1)(m-1)/\{m(m-k-1)\}\, F_{(\alpha,1,m-k-1)}$은 변수들간의 선형관계가 과거 데이터에서의 패턴과 달라져서 이상신호가 발생했는지에 대한 정보를 알려준다.

예를 들어 $p = 3$인 경우에 T^2에 대한 모든 가능한 MYT 분해는 다음과 같다.

$$T^2 = T^2_1 + T^2_{(2\cdot 1)} + T^2_{(3\cdot 1,2)} \tag{2.13}$$

$$= T^2_1 + T^2_{(3\cdot 1)} + T^2_{(2\cdot 1,3)} \tag{2.14}$$

$$= T^2_2 + T^2_{(3\cdot 2)} + T^2_{(1\cdot 2,3)} \tag{2.15}$$

$$= T^2_2 + T^2_{(1\cdot 2)} + T^2_{(3\cdot 1,2)} \tag{2.16}$$

$$= T^2_3 + T^2_{(1\cdot 3)} + T^2_{(2\cdot 1,3)} \tag{2.17}$$

$$= T^2_3 + T^2_{(2\cdot 3)} + T^2_{(1\cdot 2,3)} \tag{2.18}$$

위에서 살펴본 것처럼, T^2 통계량에 대한 MYT 분해는 3!개로 6가지가 있으며, 이들로부터 얻을 수 있는 서로 다른 항은 모두 $3 \times 2^{3-1} = 12$개가 된다. 먼저 주변 통계량 T_1^2, T_2^2, T_3^2은 식 (2.10)으로부터 쉽게 구할 수 있다. 또한, $T_{(2\cdot1)}^2, \ldots, T_{(2\cdot3)}^2$은 식 (2.12)로부터 쉽게 계산된다. 즉, 두 변수로 이루어진 호텔링 $T_{(i,j)}^2$ $(i \neq j = 1, 2, 3)$ 통계량과 변수 하나로 이루어진 T_k^2 $(k = 1, 2, 3)$ 통계량이 주어지면 쉽게 구할 수 있다. 마찬가지로, 통계량 $T_{(3\cdot1,2)}^2, T_{(2\cdot1,3)}^2, T_{(1\cdot2,3)}^2$도 식 (2.12)의 관계로부터 쉽게 구할 수 있게 된다. 즉, p개의 변수로부터 $k\,(\leq p)$개의 변수를 선택하여 호텔링 T^2의 값을 주면 모든 조건부 항에 대한 계산이 가능해진다. 실제로 R 소프트웨어의 MSQC 패키지의 multi.chart()함수는 이들 조합에 대한 호텔링 T^2 값을 제공해 준다. 이와 관련한 좀 더 자세한 논의는 3절에서 하기로 한다.

2.3. Mason-Young-Tracy 방법을 이용한 이상원인 선택절차

$p = 3$인 경우에 살펴보았듯이 이상신호의 원인을 식별하기 위해 MYT 분해를 할 때는 p가 커지면서 계산해야 하는 항의 수가 기하급수적으로 커지게 된다. 따라서, Mason 등 (1997)은 MYT 분해로 생겨나는 모든 항을 계산하는 것 대신에 다음과 같은 계산 절차를 제안하였다.

- 1단계: 이상신호를 나타낸 관측벡터 Y의 개별적인 변수에 대해 T_i^2 $(i = 1, 2, \ldots, p)$을 계산한다. 그리고 통계적으로 유의미한 개별 T_i^2을 갖는 변수 y_i들을 제거한다. 이들 변수의 경우에는 다른 변수들과의 연관성을 검토할 필요가 없다. 이들 변수를 제거한 후 나머지 변수들로부터 T^2 통계량을 계산한 후 이상신호를 여부를 결정한다. 이상신호가 없다면, 1단계에서 제거된 개별변수를 이상신호의 원인으로 한다.
- 2단계: 1단계에서 제거된 변수들을 제외한 나머지 k개의 변수들로 구한 T^2 통계량에서 통계적으로 유의미한 관측값이 계산된다면, 이번에는 나머지 변수들을 2개씩 골라 모든 $T_{i\cdot j}^2$ 항을 계산해야 한다. 1단계에서와 마찬가지로 통계적으로 유의한 $T_{i\cdot j}^2$을 갖는 모든 가능한 쌍 (y_i, y_j)을 제거한다. 이것은 두 변수의 관계가 과거 데이터와 다르다는 것을 나타낸다. 이 때, 제거된 모든 변수를 이상신호의 원인으로 검사하고, 제거되지 않은 나머지 변수들로부터 다시 T^2 통계량을 계산한다. 더 이상 이상신호가 없다면, 2단계에서 제거된 두 변수의 관계와 1단계에서 제거된 변수를 이상신호의 원인으로 한다.
- 3단계: 2단계에서 제거된 변수들을 제외하고 나머지 변수들로 구한 T^2 통계량으로부터 이상신호가 탐지된다면, 남은 변수들로부터 모든 $T_{i\cdot j,k}^2$ 항을 계산한다. 이들로부터 통계적으로 유의미한 결과를 보여주는 (y_i, y_j, y_k)변수들의 조합을 제거하고, 나머지 변수들이 이상신호를 나타내는지 검토한다.
- 4단계: 지금까지의 방식으로 남은 변수들이 없어질 때까지 고차항에 대해서 이상신호의 원인을 계속 찾아볼 수 있다. 이러한 간편 계산식이 의미가 없는 경우는 이상신호가 발생했지만, 1단계, 2단계 등으로 거치면서 유의한 항이 나오지 않아, 최고차항으로 이루어진 항을 계산하게 되어 실제로 모든 경우를 다 계산해야 하는 경우가 된다.

이처럼 실제로 MYT 분해로 나타나는 모든 항을 다 계산해야 하는 경우 이것은 마치 변수선택법의 전역적 탐색법과 비슷하다고 할 수 있을 것이다. 그러나 위에서 제안한 절차는 통계적으로 유의한 변수들을 하나씩 이상신호의 원인을 고려하고, 나머지 변수들의 T^2 통계량이 유의하지 않으면 중단하게 되어 계산의 부담을 줄여주게 된다. 만약 모든 단계를 다 계산해야 하는 경우라면 이것은 MYT 분해를 통해 검정하는 것과 마찬가지의 계산 횟수를 갖게 된다.

3. 사례분석

이 절에서는 R 소프트웨어 MSQC 패키지를 사용해서 2절에서 소개한 이상신호의 원인을 탐지하는 절차를 소개하기로 한다. 예제 데이터로 MSQC 패키지에 있는 $p = 5$인 수질오염 데이터를 사용하기로

```
library(MSQC)
data("water1", "water2")
phase1=mult.chart(water1,type="t2",alpha=0.01)
Xmv=phase1$Xmv        # 식 (2.1)의 평균벡터 추정
S=phase1$covariance   # 식 (2.1)의 공분산행렬 추정
colm=nrow(water1)     # 식 (2.1)의 m
phase2=mult.chart(water2,type="t2",Xmv=Xmv, S=S, colm=colm,alpha=0.01)
```

Figure 3.1. R code for a phase II control chart based on the T^2 statistic.

한다. 이것은 수질검사에서 측정된 5개의 변수(수소이온농도(pH), 인산염(mg/l), 질산염(mg/l), 용존산소 및 증발잔류물(mg/l))로 구성되어 있으며, *water1*과 *water2*로 두 개의 데이터집합이 주어졌다. 수질검사에서의 5가지 특성치를 이단계 모니터링하기 위해 MSQC 패키지의 mult.chart()함수를 사용하여 Figure 3.1과 같이 프로그래밍할 수 있다.

이를 실행하면 $m = 30$인 *water1* 데이터로부터 관리상태일때의 표본 평균벡터($\bar{X}$)와 표본 공분산행렬(S)이 다음과 같이 추정된다.

$$\bar{X} = \begin{pmatrix} 6.89 \\ 0.12 \\ 0.54 \\ 99.03 \\ 173.29 \end{pmatrix}; \qquad S = \begin{pmatrix} 0.700 & 0.0330 & 0.070 & 1.40 & 2.70 \\ 0.033 & 0.0026 & 0.005 & 0.08 & 0.23 \\ 0.070 & 0.0050 & 0.024 & 0.22 & 0.61 \\ 1.400 & 0.0800 & 0.220 & 6.40 & 10.00 \\ 2.700 & 0.2300 & 0.610 & 10.00 & 40.00 \end{pmatrix}. \tag{3.1}$$

이 추정치를 식 (2.1)에 적용하면, *water2* 데이터의 관측치 Y에 대하여 온라인 모니터링이 가능하다. 그 결과로써 Figure 3.2를 살펴보면 *water2*의 18번째 데이터 $y_{18} = (6.07, 0.18, 0.55, 102.64, 169.97)$이 이상상태로 확인되는데, mult.chart()함수는 이상상태가 발생한 관측치에 대하여 Figure 3.2처럼 5개 품질 특성치에 대한 부분집합으로 31개의 조합에 대한 호텔링 T^2 통계량을 제공해 준다. 이로부터 2절에서 소개한 각 절차를 적용할 수 있다.

3.1. Murphy 방법의 이상원인 선택 절차

이상신호가 발생한 *water2* 데이터의 18번째 관측치에 대하여 그 원인을 알아보기 위해 2절에서 소개한 절차를 실시한다. 이때 유의수준은 각각 $\alpha = 0.01$로 하였다.

- 1단계: $\min(D_4(i)) = T^2 - T^2(y_4) = 26.1105 - 2.0363 = 24.0742$로 자유도 4인 카이제곱 분포를 고려할 때, 이러한 차이는 통계적으로 유의한 것으로 나타나 다음 단계를 고려한다.
- 2단계: $\min(D_3(4, j)) = T^2 - T^2(y_4, y_1) = 26.1105 - 8.6168 = 17.4937$로 자유도 3인 카이제곱 분포를 고려할 때, 이러한 차이는 유의한 것으로 나타나 다음 단계를 고려한다.
- 3단계: $\min(D_2(4, 1, k)) = T^2 - T^2(y_4, y_1, y_2) = 26.1105 - 14.8012 = 11.3093$로 자유도 2인 카이제곱 분포를 고려할 때, 이러한 차이는 유의한 것으로 나타나 다음 단계를 고려한다.
- 4단계: $\min(D_1(4, 1, 2, l)) = T^2 - T^2(y_4, y_1, y_2, y_5) = 26.1105 - 26.0037 = 0.1068$로 자유도 1인 카이제곱 분포를 고려할 때, 이러한 차이는 유의하지 않은 것으로 나타나 이 단계에서 멈추게 되며, 이 경우 이상신호의 원인으로 y_4, y_1, y_2, y_5의 변수를 고려한다.

```
The following(s) point(s) fall outside of the control limits[1] 1
8
$`Decomposition of`
[1] 18

      t2 decomp     ucl p-value 1 2 3 4 5
 [1,]    0.9606  7.8509  0.3351 1 0 0 0 0
 [2,]    1.3846  7.8509  0.2489 2 0 0 0 0
 [3,]    0.0042  7.8509  0.9490 3 0 0 0 0
 [4,]    2.0363  7.8509  0.1643 4 0 0 0 0
 [5,]    0.2756  7.8509  0.6036 5 0 0 0 0
 [6,]   10.2810 11.6719  0.0004 1 2 0 0 0
 [7,]    1.4585 11.6719  0.2492 1 3 0 0 0
 [8,]    8.6168 11.6719  0.0012 1 4 0 0 0
 [9,]    0.9614 11.6719  0.3942 1 5 0 0 0
[10,]    2.1567 11.6719  0.1339 2 3 0 0 0
[11,]    2.1746 11.6719  0.1318 2 4 0 0 0
[12,]    5.1720 11.6719  0.0120 2 5 0 0 0
[13,]    2.8282 11.6719  0.0755 3 4 0 0 0
[14,]    0.5257 11.6719  0.5967 3 5 0 0 0
[15,]    5.3303 11.6719  0.0107 4 5 0 0 0
[16,]   10.6117 15.3193  0.0001 1 2 3 0 0
[17,]   14.8012 15.3193  0.0000 1 2 4 0 0
[18,]   15.2719 15.3193  0.0000 1 2 5 0 0
[19,]    8.6568 15.3193  0.0003 1 3 4 0 0
[20,]    1.6335 15.3193  0.2032 1 3 5 0 0
[21,]   10.6050 15.3193  0.0001 1 4 5 0 0
[22,]    3.5478 15.3193  0.0265 2 3 4 0 0
[23,]    5.2499 15.3193  0.0051 2 3 5 0 0
[24,]    7.8219 15.3193  0.0006 2 4 5 0 0
[25,]    5.3498 15.3193  0.0047 3 4 5 0 0
[26,]   16.1631 19.0863  0.0000 1 2 3 4 0
[27,]   15.2999 19.0863  0.0000 1 2 3 5 0
[28,]   26.0037 19.0863  0.0000 1 2 4 5 0
[29,]   10.7571 19.0863  0.0000 1 3 4 5 0
[30,]    8.1885 19.0863  0.0002 2 3 4 5 0
[31,]   26.1105 23.1040  0.0000 1 2 3 4 5
```

Figure 3.2. Output for the mult.chart function: the MYT decomposition of the T^2 statistic for an out-of-control observation.

위 결과를 요약하면 18번째 데이터의 경우 수소, 이온농도, 질산염과 용존산소 및 증발잔류물 간의 관계가 *water1* 데이터집합에서 관측된 관계와 차이가 난다는 것을 의미한다. 즉, 각 품질특성치의 값이 변화했다기 보다는 y_3 측정치를 제외하고 다른 변수들 간의 선형관계가 *water1* 데이터에서의 관계와 달라졌다는 것을 의미한다.

3.2. Mason-Young-Tracy 분해

위 에제 18번째 데이터에 대한 MYT 분해를 시행하면, T^2 통계량에 대한 MYT 분해는 모두 $5! = 120$가지가 있으며, 우리가 구해야 할 주변 통계량과 조건부 통계량의 갯수는 모두 $5 \times 2^{5-1} = 80$개가 된다. 2.2절에서 $p = 3$인 경우와 마찬가지로, 식 (2.12)로부터 MYT 분해로 생성되는 모든 항에 대해 계산할 수 있고, 이것은 Table 3.1과 같다. 특히, 굵은 글씨는 식 (2.11)을 통해 $\alpha = 0.01$로 하여 $T_{i\cdot j}$의 경우 UCL $= 8.1719$, $T_{i\cdot j,k}$의 경우 UCL $= 8.5202$, $T_{i\cdot j,k.l}$의 경우 UCL $= 8.8992$, 그리고 $T_{i\cdot j,k,l,m}$의 경우 UCL $= 9.3134$이 넘어 T^2 통계량이 큰 값으로 관측된 원인에 대한 정보를 주게 된다. 예를 들어,

18번째 데이터에서는 y_1과 y_2의 관계가 달라졌음을 알 수 있다.

Table 3.1로부터 T^2 통계량의 MYT 분해에 대한 예 2가지를 살펴보면 다음과 같다.

$$\begin{aligned} T^2 &= T_1^2 + T_{2\cdot 1}^2 + T_{3\cdot 1,2}^2 + T_{4\cdot 1,2,3}^2 + T_{5\cdot 1,2,3,4}^2 \qquad (3.2)\\ &= 0.9606 + 9.3204 + 0.3307 + 5.5514 + 9.9474 = 26.1105,\\ T^2 &= T_4^2 + T_{2\cdot 4}^2 + T_{1\cdot 2,4}^2 + T_{5\cdot 1,2,4}^2 + T_{3\cdot 1,2,4,5}^2 \qquad (3.3)\\ &= 2.0363 + 0.1383 + 12.6266 + 11.2025 + 0.1068 = 26.1105. \end{aligned}$$

변수의 순서를 어떻게 고려하는 가에 따라 필요한 주변 통계량과 조건부 통계량은 달라지지만, 결국 T^2 통계량은 같다는 것을 확인할 수 있다.

3.3. Mason-Young-Tracy 분해를 이용한 이상원인 선택절차

Table 3.1과 같이 모든 MYT 항을 계산하는 것을 대신해서 원인 식별을 위해 Mason 등 (1997)이 제시한 좀 더 간편한 절차를 대입해 보면 다음과 같다.

- 1단계: *water2*의 18번째 관측치에 대하여, 개별적인 변수에 대해 주변 통계량 T_i^2을 계산한 결과를 Table 3.1에서 보듯이 $\alpha = 0.01$에서 관리한계선(식 (2.11))을 구하면 UCL $= 7.8059$로 $T_i^2 <$ UCL $(i = 1, 2, \ldots, 5)$이 되어 통계적으로 의미있는 변수들이 없어 다음의 2단계를 실시한다.
- 2단계: 5개의 변수들에 대해 모든 $T_{i\cdot j}^2$항을 계산해야 한다. $\alpha = 0.01$일 때 식 (2.11)로부터 관리한계선은 UCL $= 8.1719$로 주어지고, $T_{2\cdot 1}$과 $T_{1\cdot 2}$이 통계적으로 유의미하게 되어 (y_1, y_2)를 제거한다. 이것은 두 변수의 관계가 과거 데이터와 다르다는 것을 의미한다. 이 때, 제거 된 모든 변수를 신호의 원인으로 검사하고, 제거되지 않은 나머지 변수들로부터 다시 $T_{3,4,5}^2$통계량을 살펴보는데, 통계적으로 유의하지 않아 첫 번째 변수와 두 번째 변수를 이상신호의 원인으로 한다.

이 절차는 MYT 분해를 위해 80개 항을 계산하는 것을 대신하여, 1단계에서 5개의 항 (Table 3.2), 2단계에서 20개의 항 (Table 3.3) 등 25개 항에 대한 계산만으로 주된 이상신호의 원인을 좀 더 빠르게 탐지할 수 있음을 알 수 있었다. 또한 이상신호의 원인이 개별적인 변수의 변화에 기인한 것인지 아니면 변수들 간의 직선관계가 변한 것인지에 대한 정보를 알려준다. 이것은 T^2 통계량이 변수들의 순서에 따라 개별적인 변수항과 조건부 항으로 모형선택의 변수선택과 비슷한 개념으로 주효과와 교호작용효과를 찾는 것과 비슷하다고 할 것이다. 다만, 이 방법은 앞절에서 살펴본 것처럼 모든 항을 살펴보는 것이 아니기 때문에, 개별적인 효과의 영향이 발생한 경우 다른 변수와의 연관성에 대한 정보는 생략된 정보를 주게 된다. 그러나 p가 큰 경우 효과적으로 계산량을 줄이며 이상신호의 원인을 알아 볼 수 있는 효과적인 방법임을 이해할 수 있다.

4. 이상신호의 원인에 대한 성능 비교

이번 절에서는 모의실험을 통해 이상신호의 원인을 찾기 위한 3가지 방법의 성능을 비교 평가해 보고자 한다. 이를 위한 모의실험 절차는 다음과 같다. $p = 3$인 T^2 관리도를 통해 평균벡터에 대한 이단계 모니터링을 가정하였다. 관리상태 하에서 품질 특성치는 다음을 가정한다.

$$X \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}\right) \qquad (4.1)$$

Table 3.1. All possible MYT decompositions for an 18th observation which falls outside the upper control limit

	T^2	계산		T^2	계산
1	T_1^2	0.9606	41	$T_{3\cdot 2,4}^2$	$3.5478 - 2.1746 = 1.3732$
2	T_2^2	1.3846	42	$T_{1\cdot 2,4}^2$	$14.8012 - 2.1746 = \mathbf{12.6266}$
3	T_3^2	0.0042	43	$T_{5\cdot 2,4}^2$	$7.8219 - 2.1746 = 5.6473$
4	T_4^2	2.0363	44	$T_{1\cdot 2,5}^2$	$15.2719 - 5.1720 = \mathbf{10.0999}$
5	T_5^2	0.9606	45	$T_{3\cdot 2,5}^2$	$5.2499 - 5.1720 = 0.0779$
6	$T_{2\cdot 1}^2$	$10.2810 - 0.9606 = \mathbf{9.3204}$	46	$T_{4\cdot 2,5}^2$	$7.8219 - 5.1720 = 2.6499$
7	$T_{3\cdot 1}^2$	$1.4585 - 0.9606 = 0.4979$	47	$T_{1\cdot 3,4}^2$	$8.6568 - 2.8282 = 5.8286$
8	$T_{4\cdot 1}^2$	$8.6168 - 0.9606 = 7.6562$	48	$T_{2\cdot 3,4}^2$	$3.5478 - 2.8282 = 0.7196$
9	$T_{5\cdot 1}^2$	$0.9614 - 0.9606 = 0.0008$	49	$T_{5\cdot 3,4}^2$	$5.3498 - 2.8282 = 2.5216$
10	$T_{1\cdot 2}^2$	$10.2810 - 1.3846 = \mathbf{8.8964}$	50	$T_{1\cdot 3,5}^2$	$1.6335 - 0.5257 = 1.1078$
11	$T_{3\cdot 2}^2$	$2.1567 - 1.3846 = 0.7721$	51	$T_{2\cdot 3,5}^2$	$5.2499 - 0.5257 = 4.7242$
12	$T_{4\cdot 2}^2$	$2.1746 - 1.3846 = 0.7900$	52	$T_{4\cdot 3,5}^2$	$5.3498 - 0.5257 = 4.8241$
13	$T_{5\cdot 2}^2$	$5.1720 - 1.3846 = 3.7874$	53	$T_{1\cdot 4,5}^2$	$10.6050 - 5.3303 = 5.2747$
14	$T_{1\cdot 3}^2$	$1.4585 - 0.0042 = 1.4543$	54	$T_{2\cdot 4,5}^2$	$7.8219 - 5.3303 = 2.4916$
15	$T_{2\cdot 3}^2$	$2.1567 - 0.0042 = 2.1525$	55	$T_{3\cdot 4,5}^2$	$5.3498 - 5.3303 = 0.0195$
16	$T_{4\cdot 3}^2$	$2.8282 - 0.0042 = 2.8240$	56	$T_{4\cdot 1,2,3}^2$	$16.1631 - 10.6117 = 5.5514$
17	$T_{5\cdot 3}^2$	$0.5257 - 0.0042 = 0.5215$	57	$T_{5\cdot 1,2,3}^2$	$15.2999 - 10.6117 = 4.6882$
18	$T_{1\cdot 4}^2$	$8.6168 - 2.0363 = 6.5805$	58	$T_{3\cdot 1,2,4}^2$	$16.1631 - 14.8012 = 1.3619$
19	$T_{2\cdot 4}^2$	$2.1746 - 2.0363 = 0.1383$	59	$T_{5\cdot 1,2,4}^2$	$26.0037 - 14.80128012 = \mathbf{11.2025}$
20	$T_{3\cdot 4}^2$	$2.8282 - 2.0363 = 0.7919$	60	$T_{3\cdot 1,2,5}^2$	$15.2999 - 15.2719 = 0.0280$
21	$T_{5\cdot 4}^2$	$5.3303 - 0.0042 = 5.3261$	61	$T_{4\cdot 1,2,5}^2$	$26.0037 - 15.2719 = \mathbf{10.7318}$
22	$T_{1\cdot 5}^2$	$0.9614 - 0.6036 = 0.3578$	62	$T_{2\cdot 1,3,4}^2$	$16.1631 - 8.6568 = 7.5063$
23	$T_{2\cdot 5}^2$	$5.1720 - 0.2756 = 4.8964$	63	$T_{5\cdot 1,3,4}^2$	$10.7571 - 8.6568 = 2.1003$
24	$T_{3\cdot 5}^2$	$0.5257 - 0.2756 = 0.2501$	64	$T_{2\cdot 1,3,5}^2$	$15.2999 - 1.6335 = \mathbf{13.6664}$
25	$T_{4\cdot 5}^2$	$5.3303 - 0.6036 = 4.7267$	65	$T_{4\cdot 1,3,5}^2$	$10.7571 - 1.6335 = \mathbf{9.1236}$
26	$T_{3\cdot 1,2}^2$	$10.6117 - 10.2810 = 0.3307$	66	$T_{2\cdot 1,4,5}^2$	$26.0037 - 10.6050 = \mathbf{15.3987}$
27	$T_{4\cdot 1,2}^2$	$14.8012 - 10.2810 = 4.5202$	67	$T_{3\cdot 1,4,5}^2$	$10.7571 - 10.6050 = 0.1521$
28	$T_{5\cdot 1,2}^2$	$15.2719 - 10.2810 = 4.9909$	68	$T_{1\cdot 2,3,4}^2$	$16.1631 - 3.5478 = \mathbf{12.6153}$
29	$T_{2\cdot 1,3}^2$	$10.6117 - 1.4585 = \mathbf{9.1532}$	69	$T_{5\cdot 2,3,4}^2$	$8.1885 - 3.5478 = 4.6407$
30	$T_{4\cdot 1,3}^2$	$8.6568 - 1.4585 = 7.1983$	70	$T_{1\cdot 2,3,5}^2$	$15.2999 - 5.2499 = \mathbf{10.0500}$
31	$T_{5\cdot 1,3}^2$	$1.6335 - 1.4585 = 0.1750$	71	$T_{4\cdot 2,3,5}^2$	$8.1885 - 5.2499 = 2.9386$
32	$T_{2\cdot 1,4}^2$	$14.8012 - 8.6168 = 6.1844$	72	$T_{1\cdot 2,4,5}^2$	$26.0037 - 7.8219 = \mathbf{18.1818}$
33	$T_{3\cdot 1,4}^2$	$8.6568 - 8.6168 = 0.0400$	73	$T_{3\cdot 2,4,5}^2$	$8.1885 - 7.8219 = 0.3666$
34	$T_{5\cdot 1,4}^2$	$10.6050 - 8.6168 = 1.9882$	74	$T_{1\cdot 3,4,5}^2$	$10.7571 - 5.3498 = 5.4073$
35	$T_{2\cdot 1,5}^2$	$15.2719 - 0.9614 = \mathbf{14.3105}$	75	$T_{2\cdot 3,4,5}^2$	$8.1885 - 5.3498 = 2.8387$
36	$T_{3\cdot 1,5}^2$	$1.6335 - 0.9614 = 0.6721$	76	$T_{1\cdot 2,3,4,5}^2$	$26.1105 - 8.1885 = 17.9220$
37	$T_{4\cdot 1,5}^2$	$10.6050 - 0.9614 = \mathbf{9.6436}$	77	$T_{2\cdot 1,3,4,5}^2$	$26.1105 - 10.7571 = \mathbf{15.3534}$
38	$T_{1\cdot 2,3}^2$	$10.6117 - 2.1567 = \mathbf{8.4550}$	78	$T_{3\cdot 1,2,4,5}^2$	$26.1105 - 26.0037 = 0.1068$
39	$T_{4\cdot 2,3}^2$	$3.5478 - 2.1567 = 1.3911$	79	$T_{4\cdot 1,2,3,5}^2$	$26.1105 - 15.2999 = \mathbf{10.8106}$
40	$T_{5\cdot 2,3}^2$	$5.2499 - 2.1567 = 3.0932$	80	$T_{5\cdot 1,2,3,4}^2$	$26.1105 - 16.1631 = \mathbf{9.9474}$

단, 굵은 글씨는 식 (2.11)을 통해 $\alpha = 0.01$로 하여 $T_{i\cdot j}$의 경우 UCL = 8.1719, $T_{i\cdot j,k}$의 경우 UCL = 8.5202, $T_{i\cdot j,k,l}$의 경우 UCL = 8.8992, 그리고 $T_{i\cdot j,k,l,m}$의 경우 UCL = 9.3134이 넘어 T^2 통계량이 큰 값으로 관측된 원인에 대한 정보를 제공한다.

MYT = Mason-Young-Tracy; UCL = upper control limit.

Table 3.2. Summary of the individual T^2 statistics

	T^2	UCL
1	$T_1^2 = 0.9606$	7.8509
2	$T_2^2 = 1.3846$	7.8509
3	$T_3^2 = 0.0042$	7.8509
4	$T_4^2 = 2.0363$	7.8509
5	$T_5^2 = 0.9606$	7.8509

MYT = Mason-Young-Tracy; UCL = upper control limit.

Table 3.3. Summary of all T_{ij}^2

	T^2		T^2
1	$T_{2\cdot1}^2 = \mathbf{9.3204}$	11	$T_{4\cdot3}^2 = 2.8240$
2	$T_{3\cdot1}^2 = 0.4979$	12	$T_{5\cdot3}^2 = 0.5215$
3	$T_{4\cdot1}^2 = 7.6562$	13	$T_{1\cdot4}^2 = 6.5805$
4	$T_{5\cdot1}^2 = 0.00089$	14	$T_{2\cdot4}^2 = 0.1383$
5	$T_{1\cdot2}^2 = \mathbf{8.8964}$	15	$T_{3\cdot4}^2 = 0.7919$
6	$T_{3\cdot2}^2 = 0.7721$	16	$T_{5\cdot4}^2 = 5.3261$
7	$T_{4\cdot2}^2 = 0.7900$	17	$T_{1\cdot5}^2 = 0.3578$
8	$T_{5\cdot2}^2 = 3.7874$	18	$T_{2\cdot5}^2 = 4.8964$
9	$T_{1\cdot3}^2 = 1.4543$	19	$T_{3\cdot5}^2 = 0.2501$
10	$T_{2\cdot3}^2 = 2.1525$	20	$T_{4\cdot5}^2 = 4.7267$

단, 굵은 글씨는 $\alpha = 0.01$일 때 식 (2.11)로부터 UCL = 8.1719를 넘는 경우를 나타낸다.
MYT = Mason-Young-Tracy; UCL = upper control limit.

새로운 데이터 Y는 $N(\mu, \Sigma)$를 가정하는데, 평균벡터는 $\mu = (\delta, 0, 0)$이고 공분산행렬 Σ는 식 (4.1)의 분산과 동일하다. 즉, 새로운 데이터는 분산의 변화가 없고 첫 번째 변수 x_1의 평균만 δ만큼 이동시켜, 호텔링 T^2 통계량으로부터 이상신호가 발생하는 경우 이상신호의 원인은 첫 번째 변수의 평균이 변화한 것임을 가정한다. 각 절차의 성능을 정의하기 위해 1,000개의 Y 관측치에 대해 첫 번째 변수의 이동을 탐지한 건수 즉, 각 절차의 원인 파악 비율을 계산한다. 이를 위해 각 방법에 대해 이상신호를 탐지한 경우는 다음과 같이 정의한다.

- MYT 분해 (MYT-D): x_1 변수의 평균 이동을 탐지하기 위해, 이상신호가 발생한 데이터에 MYT 분해들 중 x_1과 관련이 있는 T_1^2, $T_{1\cdot2}^2$, $T_{1\cdot3}^2$, $T_{1\cdot2,3}^2$에 대해 유의수준 α의 관리한계선을 적용하여 이들 중 관리한계선을 벗어나는 경우가 있으면 이상신호의 원인을 찾은 것으로 간주한다.
- MYT 방법 (MYT): T_1^2 통계량이 관리한계선을 벗어나면 이상신호의 원인을 찾은 것으로 간주한다.
- Murphy 방법 (Murphy): $D = T^2 - T_1^2$ 통계량의 값이 작다면 이상신호의 원인을 찾은 것으로 간주한다.

위 방법으로 비교하기 위해 x_1 변수를 $\delta = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$으로 변화 시켰으며, 변수들 간 상관관계는 $\rho = 0.2$, $\rho = 0.5$, $\rho = 0.8$을 사용하였다. 먼저, 변수들 간 상관정도에 관계없이 δ가 작은 경우 세 가지 절차로부터 변수의 원인을 확인하는 검정력은 낮은 편이었다. MYT 분해는 주변 통계량 뿐 아니라 조건부 통계량을 통해 검정을 함으로써 다른 두 가지 간단한 방법들에 비해 x_1의 평균변화를 탐지한 비율이 높았지만 δ가 작은 경우에는 검정비율의 변화가 그리 크지 않았다. Murphy 방법의 경우 변수들 간 상관성이 큰 경우에는 $\delta = 3$으로 평균변화가 큰 경우에도 검정비율이 작게 나타나는 것을 알 수 있다. 그러나 MYT 분해에 의한 방법은 검정비율이 99.8%로 거의 1에 가깝게 나타나 MYT 분해가 평

Table 4.1. Comparison of the percentages of correct identification of the cause for out-of-control signals

δ	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
	MYT-D	MYT	Murphy	MYT-D	MYT	Murphy	MYT-D	MYT	Murphy
0.5	3.8	3.4	3.3	5.5	3.9	3.6	7.0	2.7	2.6
1.0	9.8	8.1	7.7	13.4	8.9	8.4	31.9	13.4	9.6
1.5	20.8	18.3	17.6	28.8	21.3	19.3	65.3	30.2	14.6
2.0	35.6	34.0	31.4	52.4	41.5	30.7	88.1	51.5	11.4
2.5	54.1	52.7	47.0	71.5	62.5	41.9	98.7	71.2	3.1
3.0	73.9	72.8	64.9	87.6	80.6	41.9	99.8	85.7	0.8

MYT = Mason-Young-Tracy; MYT-D = MYT decomposition.

균벡터의 변화를 민감하게 탐지할 것으로 기대된다. 다만 p가 큰 경우에는 이러한 MYT 분해를 계산하는 것이 복잡하기 때문에 MYT 방법처럼 간단한 절차를 쓰는 것도 꽤 유용함을 알 수 있다 (Table 4.1).

5. 앞으로의 연구 방향

호텔링 T^2 통계량은 다변량 품질 특성치에 대한 관리도로 가장 일반적인 통계량이다. 이 통계량의 값으로부터 평균벡터의 이상신호가 감지되면, 적절한 조치를 하기 위해 그 원인을 검토하는 것은 매우 중요한 문제이다. 이 경우 호텔링 T^2 통계량의 값을 변수들의 순서를 고려하여 1개의 주변 항과 $(p-1)$개의 조건부 항으로 분해하고, 각 통계량의 크기를 통해 이상신호의 원인을 설명한 MYT 분해는 실제 문제에서 이상신호를 이해하는데 실질적인 도움을 줄 수 있으리라 짐작할 수 있다. 그러나, p의 크기가 커짐에 따라 계산해야 할 MYT 성분은 기하급수적으로 늘어나기 때문에 이상신호의 원인을 식별하기가 쉽지 않다. 이에 이상신호의 원인을 찾기 위한 변수 선택의 알고리즘이 제안되었는데, 모의실험을 통해 Murphy (1987)의 절차보다는 MYT 분해에 대한 간편절차가 원인 식별을 더 잘한다는 것을 알 수 있었다. 특히 Murphy의 절차가 변수들의 상관이 높은 경우 원인 식별 능력이 떨어진다는 것을 알 수 있었다. 그러나 이것은 제한된 모의실험의 결과로써 이에 대한 추가 연구가 필요할 것이다. 예를 들어, T^2 통계량에 대한 p개의 MYT 분해를 통해 살펴보면 각각의 성분 또한 F 분포를 따르게 되는데 , 모의실험에서 원인 식별을 위해 MYT 각 항에 대해 모두 유의수준 α 검정법을 사용하였다. 이상신호의 원인 식별을 위해서는 전체 관측치에 대한 T^2 통계량이 통계적으로 유의하게 큰 관측치에 대해서만 살펴보기 때문에, MYT 항에 대한 조건부 통계량의 확률분포를 알아보는 등 이에 대한 좀 더 정확한 근사가 필요하다는 것을 알 수 있다. 이에 앞으로의 다변량 관리도 연구에서는 이상신호 발생 후 그 원인이 되는 변수를 식별하기 위한 검정절차를 동시에 모니터링할 수 있는 방안에 대해 추가로 연구할 필요가 있을 것이다.

References

Alt, F. B. (1985). Multivariate quality control, In Kotz, S., Johnson, N. L., Read, C.R. (eds.) *Encyclopedia of Statistical Sciences*, **6**, 111–122.

Doganaksoy, N., Faltin, F. W., and Tucker, W. T. (1991). Identification of out-of-control multivariate characteristic in a multivariable manufacturing environment, *Communications in Statistics - Theory and Methods*, **20**, 2775–2790.

Hayter, A. J. and Tsui, K. L. (1994). Identification and quantification in multivariate quality control problems, *Journal of Quality Technology*, **26**, 197–208.

Hotelling, H. (1947). *Multivariate Quality Control, Techniques of Statistical Analysis*, Eisenhart, Hastay, and Wallis (eds), McGraw-Hill, New York.

Jackson, J. E. (1991). *A User's Guide to Principal Components*, John Wiley & Sons, New York.

Kenett, R. S. and Halevy, A. (1984). Some statistical aspects of quality conformance inspection in military specifications documents, *Proceedings of the 5th International Conference of the Israeli Society for Quality Assurance*, Tel Aviv, 23–35.

Kourti, T. and MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring, *Journal of Quality Technology*, **28**, 409–428.

Lee, S. (2015). Effects of parameter estimation in phase I on phase II control limits for monitoring autocorrelated data, *The Korean Journal of Applied Statistics*, **28**, 1025–1034.

Lee, S. (2018). Notes on identifying source of out-of-control signals in phase II multivariate process monitoring, *The Korean Journal of Applied Statistics*, **31**, 1–12.

Lim, J. and Lee, S. (2017). Phase II monitoring of changes in mean from high-dimensional data, *Applied Stochastic Models in Business and Industry*, **33**, 626–639.

Mason, R. L., Tracy, N. D., and Young, J. C. (1995). Decomposition of T^2 for multivariate control chart interpretation, *Journal of Quality Technology*, **27**, 99–108.

Mason, R. L., Tracy, N. D., and Young, J. C. (1997). A practical approach for interpreting multivariate T^2 control chart, *Journal of Quality Technology*, **29**, 396–406.

Mason, R. L. and Young, J. C. (1999). Improving the sensitivity of the 62 statistic in multivariate process control, *Journal of Quality Technology*, **31**, 155–165.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control* (5th ed), John Wiley & Sons, New York.

Murphy, B. J. (1987). Selecting out-of-control variables with 62 multivariate quality procedures, *The Statistician*, **36**, 571–583.

Shewhart, W. A. (1926). Quality control charts, *Bell System Technical Journal*, **2**, 593–603.

Timm, N. H. (1996). Multivariate quality control using finite intersection tests, *Journal of Quality Technology*, **28**, 233–243.

Woodall, W. H. and Montgomery, D. C. (1999). Research issues and ideas in statistical process control, *Journal of Quality Technology*, **31**, 376–385.

호텔링 T^2의 이상신호 원인 식별

이성임[a,1]

[a]단국대학교 응용통계학과

(2018년 10월 30일 접수, 2018년 11월 1일 수정, 2018년 11월 1일 채택)

요 약

호텔링 T^2 통계량에 근거한 다변량 관리도는 공정의 이상상태를 식별하는 통계적 공정관리의 강력한 도구 중 하나이다. 다수의 품질 특성치를 동시에 모니터링하는데 사용된다. T^2 관리도를 통해 이상신호가 탐지된다는 것은 평균벡터의 변화가 있다는 것을 의미하게 된다. 그러나, 이러한 다변량 통계량의 신호는 이상신호에 대한 원인을 식별하기 어렵게 한다. 이 논문에서는 T^2 통계량을 서로 독립인 항으로 분해한 Mason, Young, Tracy (MYT) 분해에 기반한 원인 식별 방법들을 살펴본다. 또한, R 소프트웨어를 사용하여 사례분석을 하고, 모의실험을 통해 각 절차의 성능을 비교 평가해보고자 한다.

주요용어: 다변량 SPC, 호텔링 T^2, MYT 분해, 이상신호 원인 식별

[1](16890) 경기도 용인시 수지구 죽전로 152, 단국대학교 응용통계학과. E-mail: silee@dankook.ac.kr